ELSEVIER

Journal of Econometrics 112 (2003) 135–151

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey

Elizabeth Johnson*, Francesca Dominici, Michael Griswold, Scott L. Zeger

Department of Biostatistics, School of Public Health, The Johns Hopkins University Bloomberg, 615N. Wolfe Street, Baltimore, MD 21205-3179, USA

Abstract

We estimate the fraction of disease cases, and the fraction of their total medical expenditures, attributable to smoking for two disease groups: (LC) lung and laryngeal cancer and chronic obstructive pulmonary disease, (CHD) cardiovascular disease, stroke and other smoking-caused cancers. We use a generalized additive model to predict the probability of disease; and a semi-parametric, two-part cost model to estimate the average difference in medical expenditures for persons with and without disease. We estimate that 53% and 13% of the medical expenditures for persons with LC or CHD are attributable to smoking.

JEL classification: C1; C13; C14; I1; I10

Keywords: Smoking-attributable fraction; Smoking-attributable expenditure; Generalized additive model; Two-part model

1. Introduction

Since seminal papers by Doll and Hill (1956) and Wynder et al. (1956) identified a possible association between smoking and lung cancer, public health and medical researchers have investigated the effects of smoking on health for half a century. U.S. Surgeon General Reports have determined that smoking causes lung and laryngeal cancer, chronic obstructive pulmonary disease (COPD), coronary heart disease, stroke, and premature death, as well as other major diseases and health conditions (Department

* Corresponding author. Tel.: +1-410-955-0958; fax: +1-410-955-0958.
E-mail address: ejohnson@jhsph.edu (E. Johnson).

0304-4076/02/$ - see front matter
PII: S0304-4076(02)00157-4

of Health and Human Services, 1984, 1989, 1990). More recently, researchers have broadened their concerns to include the health policy and services issues related to smoking, one component of which is to quantify the costs of treating diseases and conditions caused by smoking (e.g. Strassels et al., 2001).

This body of research has focused on two distinct quantities for estimation. The first is the fraction of actual medical expenditures for a particular population in a fixed interval that is attributable to smoking (Luce and Schweitzer, 1978; Rice et al., 1986; Miller et al., 1998, 1999). The second target is the difference in health expenditures for a population with a particular smoking pattern, and the expected expenditures for that same population absent smoking (Manning et al., 1989; Hodgson and Kopstein, 1984). These two targets are referred to as gross and net smoking-attributable expenditures (Warner et al., 1999). The main difference between them is that the net value includes the savings from smoking causing premature death in addition to increased prevalence of major diseases, referred to as the *death benefit* (Zeger et al., 2000; Rubin, 2000). Warner et al. (1999) discuss the relevance of these two different targets of inference in more detail.

In this paper, we focus on statistical methods for estimating the gross effect that we call the smoking-attributable fraction of expenditures (SAFE). We directly estimate the SAFE using the 1987 National Medical Expenditure Survey or NMES (National Center For Health Services Research, 1987), a population-based survey of non-institutionalized persons that includes information on smoking dose, disease occurrence including all the major diseases caused by smoking, and medical expenditures.

Early investigations of smoking costs relied upon indirect estimates of expenditures, and fractions of disease cases attributable to smoking. For example, Luce and Schweitzer (1978) relied upon expert judgments about the risk of diseases attributable to smoking. Bartlett et al. (1994) made the first direct estimate of the SAFE using NMES data. This work was updated a few years later by Miller et al. (1998), who estimated a SAFE for Medicare expenditures for each state. Their statistical analysis derived from a system of latent variable models including a joint bivariate probit model for the occurrence of disease and smoking, and then a set of expenditure models based upon a log-normal distribution of expenditures in which both the mean and variance of the distribution could vary by smoking level. Their estimates of the SAFE included smoking effects on the occurrence of major diseases, as well as more subtle effects mediated through general poor health among persons without diseases. Independently, Miller et al. (1999) estimated a SAFE from NMES using a logistic regression of the chance of disease given a categorical smoking level, and a log-normal model for expenditures. Warner et al. (1999) present a thorough summary of these earlier studies of smoking-attributable expenditures.

Previous studies have relied upon parametric models for the risk of disease and for average expenditures for a given disease. In particular, most analyses have used the working assumption that, when positive, the logarithm of expenditures can be well approximated by a normal distribution. Estimates of mean differences in expenditures for persons with and without disease are central to the SAFE. When a log-normal model is used, these differences depend upon model assumptions for both the mean and variance of the log expenditures. For example, Miller et al. (1998) allow both the

mean and the variance to change with disease and/or smoking status. An important question is whether the estimated SAFE values depend critically on this methodology and its inherent assumptions.

In this paper, we also model the probability of disease as a function of smoking dose and the mean expenditures as a function of disease status. We use semi-parametric models (Hastie and Tibshirani, 1990) in which we assume that the risk of disease is a smooth but arbitrary function of smoking dose, age, and other confounders. More importantly, we avoid a particular parametric model for estimating the mean difference in expenditures by using a case—control matching algorithm. Our approaches are distinct and complementary to others used to estimate smoking attributable expenditures.

Section 2 briefly reviews the National Medical Expenditure Survey. Section 3 presents the estimation of the smoking-attributable fractions. Miller et al. (1999) provide further details. Results and sensitivity analyses to model assumptions are summarized in Section 4, followed by a discussion in Section 5.

2. Data

The 1987 National Medical Expenditure Survey (NMES, US Department of Health and Human Services, Public Health Service, 1987) provides data on annual medical expenditures and disease status for a representative sample of the U.S. civilian, non-institutionalized population. We analyze data on 13,505 persons whose ages range from 40 to 94 years from among the 38,446 participants in the survey. For each person, the survey collects information on socioeconomic factors, medical conditions, medical-care expenditures, insurance coverage, and personal characteristics including gender, age, race, income level, marital status, and education level.

To provide additional information on smoking and health risk behaviors, NMES is supplemented by the Adult Self-Administered Questionnaire Household Survey (ASAQHS). Of the 13,505 persons used in our analysis, 16% did not return their ASAQHS, and thus have missing smoking information. Another 6% returned their ASAQHS, but did not complete one or more of the questions regarding smoking characteristics. Table 1 presents summary statistics of the NMES variables used in this analysis.

In the analyses described below, we focus on two disease groups, which we refer to as LC and CHD. LC includes lung and laryngeal cancer and chronic obstructive pulmonary disease for which smoking is the predominant cause. CHD includes coronary heart disease, stroke and several cancers for which smoking is a substantial contributing factor. See Table 2 for the complete list of diseases with their corresponding ICD-9 codes included in each group. A case of disease is defined to be a person who reported that in 1987 he or she either saw a doctor or was hospitalized for one of the diseases in either LC or CHD; or experienced some disability days because of a disease in either LC or CHD.

One advantage of NMES is that expenditure data is ascertained in up to five quarterly interviews conducted by trained personnel. Medical expenses are often verified by supporting data obtained from treating clinicians and hospitals. A second advantage is

Table 1
Summary statistics of the NMES variables used in the analysis

Variable		Mean(count)	% Missing(count)
LC	Case	0.02(273)	—
	Non-case	0.98(13,230)	
CHD	Case	0.11(1480)	—
	Non-case	0.89(11,750)	
Annual medical expenditure ($)		3324.5(8833.8)[a]	—
Ever smoker	Yes	0.48(6507)	0.14(1946)
	No	0.37(5050)	
Current smoker|ever smoker	Yes	0.35(2941)	0.02(186)
	No	0.40(5050)	
Age started to smoke		19.3(6.2)[a]	0.17(2328)
Cigarettes per day		20.2(12.4)[a]	0.18(2476)
Length of smoking		31.7(14.5)[a]	0.17(2328)
Age quit		45.2(14.6)[a]	0.18(2437)
For former smokers			
Years since quitting		16.7(13)[a]	0.18(2437)
For former smokers			
Age		60.7(13.2)[a]	—
Gender	Male	0.43(5799)	—
	Female	0.57(7704)	
Race	African American	0.17(2285)	—
	Other	0.83(11,218)	
Marital status	Never married	0.05(657)	0.02(315)
	Widowed/divorced/separated	0.29(3973)	
	Married	0.63(8558)	
Census region	Northeast	0.20(2722)	—
	Midwest	0.25(3300)	
	South	0.37(5027)	
	West	0.18(2454)	
Poverty status	Poor	0.11(1533)	—
	Near poor	0.05(718)	
	Low income	0.15(2014)	
	Middle income	0.31(4173)	
	High income	0.38(5065)	
Education	4+ years of college	0.13(1809)	—
	1–3 years of college	0.14(1880)	
	Some/all high school	0.49(6639)	
	Less than high school	0.24(3175)	
Seat belt use	Sometimes/seldom/never	0.39(5271)	
	Nearly always/always	0.52(7005)	0.09(1227)

[a]Values in parentheses are standard deviations.

that the smoking questions are detailed and include: whether the person ever smoked more than a total of 100 cigarettes; whether he or she is a current smoker; the duration and frequency of smoking; and finally the date a person stopped among those that quit.

Table 2
Diseases that have been determined in the U.S. Surgeon General Report (1989) to be caused by smoking

Variable name	Diseases	ICD-9
LC	Laryngeal cancer	161
	Lung cancer	162
	COPD	491, 492, 496
CHD	Oral cancer	140, 141, 143, 144, 145, 146, 148, 149
	Esophageal cancer	150
	Stomach cancer	151
	Pancreatic cancer	157
	Bladder cancer	188
	Kidney cancer	189
	Cerebrovascular disease	342, 430, 431, 432, 433. 434, 435, 436, 437, 438
	Arteriosclerosis	440, 441, 444
	Coronary heart disease	410, 411, 412, 413, 414, 425, 427, 428
	Other arterial disease	443.1, 443.9
	Other respiratory disease	515, 516.3
	Peptic ulcer disease	531, 532, 533

NMES data derive from 1987. Recent updates (e.g. Medical Expenditure Panel Survey, 1997) have insufficient sample size and smoking information to allow a similar analysis, although an update to our analysis may be possible with future data releases. Hence, NMES is still the best source of information for this kind of study (Strassels et al., 2001). Further details about NMES are available from U.S. Department of Health and Human Services (National Center For Health Services Research, 1987).

3. The smoking attributable fractions

The *population attributable fraction* is commonly used in epidemiology to describe the proportion of disease that is due to a particular causal factor (Levin, 1953). It is defined as the proportional difference in average disease risk between an exposed and otherwise similar unexposed group (e.g. Gordis, 1996).

In this paper, we estimate two quantities from the NMES survey: (1) the fraction of cases of a particular group of diseases that is attributable to smoking (SAF); and (2) the fraction of total medical expenditures for persons with these diseases that is attributable to smoking (SAFE). By *attributable*, we imply a comparison of smokers to *otherwise similar* non-smokers. That is, we estimate for the population of people for which the NMES sample is representative, the difference in rates of disease for smokers and similar non-smokers. The target of our inference is this population difference, expressed as a fraction of the rate in the smoking group and averaged over the covariate

distribution of the smokers. By *similar*, we mean that individuals have similar values of the covariates including age, gender, race, income level and others available in NMES as detailed below.

As the survey sample size goes to the population size, the probability limit of our two statistics is the corresponding value for the entire population of which NMES is representative. Other investigators (e.g. Rubin, 2001) have discussed estimation of the *causal effects* of smoking, namely the difference in disease rates or expenditures for a population of smokers compared to what would have occurred had they never smoked. These counterfactual quantities are not directly observable. Their estimation or extrapolation, is beyond the scope of this paper.

More specifically, we estimate:

$$\mathrm{SAF} = \left(\sum_{i=1}^{n} D_i \times w_i \times AF_i\right) \Big/ \left(\sum_{i=1}^{n} D_i \times w_i\right),$$

$$\mathrm{SAFE} = \left(\sum_{i=1}^{n} D_i \times C_i \times w_i \times \mathrm{AFE}_i \times \mathrm{AF}_i\right) \Big/ \left(\sum_{i=1}^{n} D_i \times C_i \times w_i\right), \tag{1}$$

where n is the sample size; i indexes the subject; D_i is the binary disease indicator; C_i denotes his or her reported medical expenditure for the year; w_i is the sampling weight for subject i; AF_i represents the smoking-attributable fraction of disease for subject i with covariate profile X_i; and AFE_i represent the disease-attributable fraction of expenditures for subject i with covariate profile X_i (Woodard, 1999; Miller et al., 1998).

In the expression for the SAF, the numerator is the number of disease cases attributable to smoking and the denominator is the total number of cases. In the SAFE, the numerator is the total expenditures attributable to diseases caused by smoking, and the denominator is the total expenditures for all people with the diseases of interest regardless of their cause. Standard errors for the estimates of the SAF and SAFE are obtained using a bootstrap with $m=100$ replications (Efron and Tibshirani, 1991, 1993).

The smoking-attributable fraction of disease (AF) is defined by

$$\mathrm{AF}_i = \begin{cases} \dfrac{\mathrm{P}(D_i|\mathrm{dose}_i, X_i) - \mathrm{P}(D_i|\mathrm{dose}_i=0, X_i)}{\mathrm{P}(D_i|\mathrm{dose}_i, X_i)} & \text{if } i \text{ is a current or former smoker,} \\ 0 & \text{if } i \text{ is a never smoker,} \end{cases} \tag{2}$$

where $\mathrm{P}(D_i|\mathrm{dose}_i, X_i)$ is the probability of disease for smokers with covariate profile X_i, and $\mathrm{P}(D_i|\mathrm{dose}_i = 0, X_i)$ is the probability of disease for non-smokers ($\mathrm{dose}_i = 0$) with the same covariate profile X_i.

The disease-attributable fraction of medical expenditures (AFE) is defined by

$$\mathrm{AFE}_i = \begin{cases} \dfrac{\mathrm{E}(C_i|D_i, X_i) - \mathrm{E}(C_i|D_i=0, X_i)}{\mathrm{E}(C_i|D_i, X_i)} & \text{if } D_i=1, \\ 0 & \text{if } D_i = 0, \end{cases} \tag{3}$$

where $E(C_i|D_i,X_i)$ is the expected expenditure for subjects with the disease and covariate profile X_i, and $E(C_i|D_i=0,X_i)$ is the expected expenditure of subjects with the same covariate profile X_i, but without disease ($D_i=0$). The next section details our models for the probability of disease and for the mean expenditures for a given disease.

3.1. The disease model

NMES provides data on potential predictors of disease prevalence including: (1) duration and degree of smoking; (2) key demographic characteristics including age, gender, and race; and (3) additional covariates including socio-economic status, education, use of seat belt as a surrogate for risk taking behavior, marital status, and geographic region (Doll et al., 1994; McBride, 1992; Sherman, 1992; Krieger et al., 1999; Osann, 1998; Tockman et al., 1987).

We model the probability of disease given the level of smoking exposure (dose_i) and other predictors (X_i) using a generalized additive model with logit link (Hastie and Tibshirani, 1990). Here, the log odds of disease is assumed to be an additive smooth function of dose, age, age × gender, a linear function of an indicator of having recently quit (quit within 1 year), and linear functions of a subject's personal characteristics. These variables are described in Table 1.

More specifically, we model the main effect of age on the probability of disease as a smooth function with three degrees of freedom. We also include an interaction term between age and gender, by allowing a different smooth curve for each gender.

The main effect of dose for current smokers on the probability of disease is modeled as a smooth function with three degrees of freedom; while the effect of dose for former smokers is modeled as a bivariate smooth function with 12 degrees of freedom.

To adjust for a subject's personal characteristics, we include indicator variables for gender (male = 1, female = 0) and race (African American = 1, 0 otherwise). The additional confounders included in the model are: poverty status, marital status, education level, seat belt use, and census region. All these variables enter in the model as linear terms.

After preliminary exploratory analyses, we have defined the dose variable to encompass the duration and degree of the smoking exposure as follows:

$$\text{dose} = \begin{cases} \text{pack year} & \text{if current smoker,} \\ \text{pack year} \times \text{years since quit} & \text{if former smoker,} \\ 0 & \text{if never smoker,} \end{cases} \tag{4}$$

where, for current or former smokers,

$$\text{pack year} = \frac{\text{reported number of cigarettes/day}}{20} \times \text{reported number of years smoked.} \tag{5}$$

The variable pack year is a measure of cumulative exposure which combines self-reported information on the degree and duration of smoking.

3.2. The expenditure model

To estimate the average expenditures in the disease and control groups for a given covariate profile, we use a two-part model (Duan et al., 1983; Lipscomb et al., 1998). In the first part, we model the probability of incurring any cost, $P(C_i > 0 \mid D_i, X_i)$. In the second part, we estimate the average medical expenditures, given a positive expenditure, $E(C_i | C_i > 0, D_i, X_i)$.

We assume that the probability of incurring any cost follows a generalized additive model with logit link (Hastie and Tibshirani, 1990). The key predictors are two indicators: whether the person has a disease in the LC category, and whether the person has a disease in the CHD category, but does not have any disease in the LC category. The model also includes all the explanatory variables used in the disease model and listed in Table 1, with the exception of the smoking variables. We examine only those smoking-related costs that are associated with the selected diseases listed in Table 2. Although these are the major diseases caused by smoking (e.g. Surgeon General Report, 1989), smoking has been implicated as a factor in other diseases and conditions, for example cataracts and macular degeneration (Munoz et al., 2000; Smith et al., 2001), respiratory morbidity (Surgeon General Report, 1984) and low birthweight (Lightwood et al., 1999; Adams and Young, 1999). In our study, we address neither costs that stem from such conditions, nor increased costs associated with poorer general health associated with smoking (e.g. Miller et al., 1998). In addition, we ignore costs that might be associated with exposure to environmental tobacco smoke (Environmental Protection Agency (EPA), 1993).

In the second part of the model, we estimate $E(C_i | C_i > 0,\ D_i{=}1, X_i)$ and $E(C_i | C_i > 0,\ D_i = 0, X_i)$ using only those subjects with positive expenditures. We use a $K \times (1:10)$ matching algorithm, where for each case in the sample ($D_i{=}1$), we estimate $E(C_i | C_i > 0,\ D_i = 1, X_i)$ by taking the sample mean of the positive expenditures for the K *closest* cases to the selected case in terms of their propensity scores (Cochran and Rubin, 1973; Rosenbaum and Rubin, 1984; Rubin and Thomas, 2000; D'Agostino Jr. and Rubin, 2000). The propensity score is the probability of having disease adjusted by confounding factors, and it is estimated by using the disease model described in Section 3.1. We then estimate $E(C_i | C_i > 0,\ D_i = 0, X_i)$ in a similar fashion, by taking the sample mean of the positive expenditures for the $K \times 10$ *closest* controls to the same selected case in terms of their propensity scores.

For our analysis we choose $K = 5$. To ensure that the SAFE is not sensitive to this choice, we conduct sensitivity analyses as detailed below. In addition, sensitivity analyses are conducted to compare our non-parametric approach to estimating $E(C_i | C_i > 0,\ D_i,\ X_i)$ with more traditional approaches.

4. Results

The disease model describes the log odds of disease as a function of a subject's smoking information and personal characteristics. Fig. 1 displays the association between the log odds of LC and the smooth function of age, and dose. For both genders,

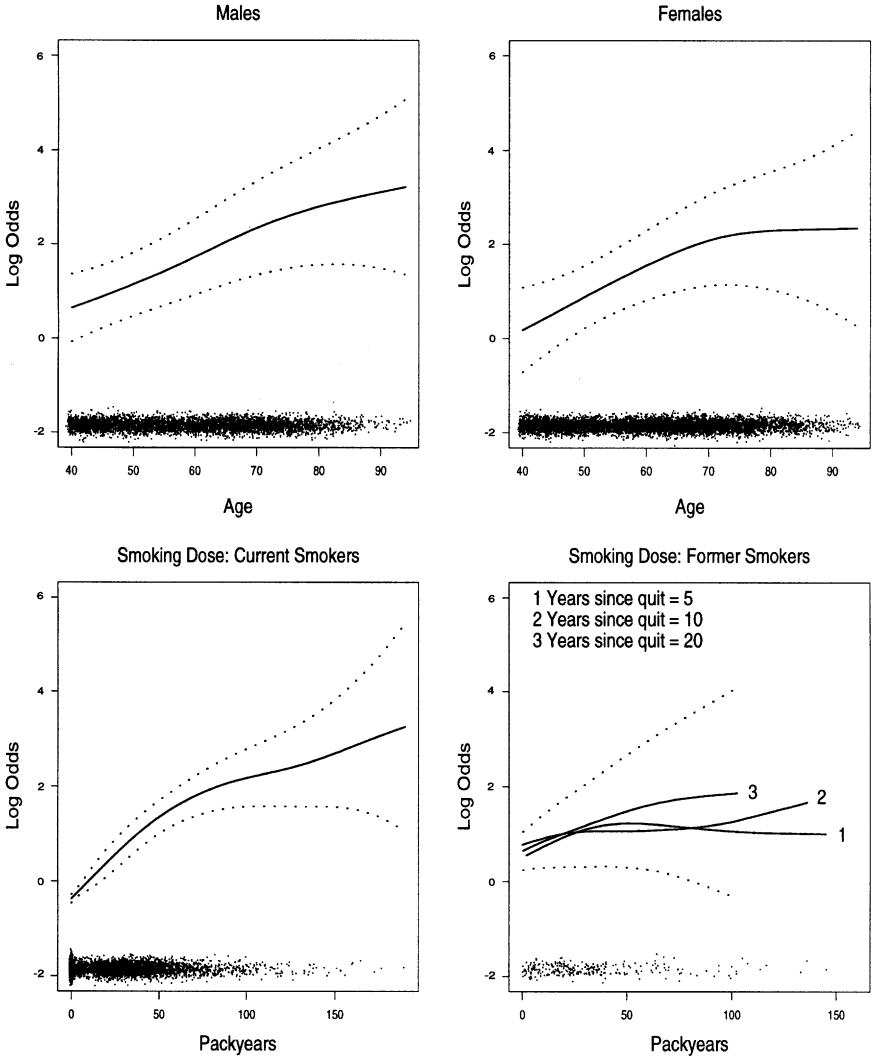

Fig. 1. Log odds of lung cancer, laryngeal cancer or COPD (LC) modeled as a smoothing spline with 3 degrees of freedom for the variables: age by gender, and dose by smoking status. The points along the x-axis of the display represents the observed values for each subject in NMES. The dashed lines represent approximately 95% confidence bounds for the estimated functions.

the odds of LC increase as a function of age. For men, the log odds continue to increase throughout the age range, whereas for women, the risk plateaus after about age 70. The estimated log odds of LC increase by roughly 3 log units, more rapidly at lower doses, for current smokers. For former smokers, the estimated log odds of LC increase at low doses and then tend to plateau at higher doses; this trend appears consistent for all values of years since quit. The graphical display of the estimated

Table 3
Point estimates and standard errors of the coefficients in the disease models

Variable		LC		CHD	
		Coefficient	Std. error	Coefficient	Std. error
Intercept		−8.99	1.05	−5.53	0.36
Race	Other	—	—	—	—
	African American	−1.21[a]	0.49	−0.30	0.15
Gender	Female	—	—	—	—
	Male	0.94	1.00	−1.35[a]	0.40
Recent quitter	No	—	—	—	—
	Yes	0.83[a]	0.32	0.77[a]	0.19
Poverty status	Poor	—	—	—	—
	Near poor	0.16	0.33	0.07	0.18
	Low income	−0.33	0.28	−0.10	0.14
	Middle income	−0.81[a]	0.26	−0.09	0.13
	High income	−0.96[a]	0.29	−0.28	0.14
Census region	Northeast	—	—	—	—
	Midwest	−0.01	0.23	0.02	0.10
	South	−0.09	0.22	0.03	0.10
	West	−0.08	0.25	−0.06	0.11
Marital status	Never married	—	—	—	—
	Widowed/divorced/separated	0.98	0.59	−0.16	0.17
	Married	0.90	0.59	−0.25	0.16
Education	4+ years of college	—	—	—	—
	1–3 years of college	−0.06	0.37	0.29	0.14
	Some/all high school	0.22	0.29	0.17	0.12
	Less than high school	0.39	0.32	−0.08	0.14
Seat belt use	Sometimes/seldom/never	—	—	—	—
	Always	−0.001	0.16	0.08	0.07

[a]Indicate coefficients significant at the 5% level.

associations between the log odds of CHD and the smooth function of age by gender, and dose by smoking status have not been included, as the association of CHD with these variables is qualitatively similar as was seen for LC.

Table 3 presents the estimated coefficients and standard errors of all other variables in the disease models. Based on the model fit, the odds of LC and CHD for Non-African Americans are 3.35 (95% confidence interval: 1.26–8.94) and 1.35 (1.00–1.82) times the odds of LC and CHD for African Americans, respectively. Comparing subjects who are recent quitters to those who are not, we estimate that the odds of LC and CHD are, respectively, 2.29 (1.21–4.35) and 2.16 (1.48–3.16) times higher for the recent

Table 4
Estimated SAFs and SAFEs for 1987 NMES

		All persons	Male		Female	
			< 65	65+	< 65	65+
SAF						
	LC	70.2	73.0	74.6	64.0	63.2
		(5.1)	(7.2)	(5.6)	(9.4)	(6.1)
	CHD	19.6	28.3	25.4	17.2	10.7
		(3.2)	(4.8)	(3.8)	(4.2)	(2.1)
SAFE						
	LC	53.4	56.9	55.3	51.2	48.8
		(5.4)	(4.7)	(6.8)	(8.0)	(9.7)
	CHD	13.4	14.3	16.6	14.5	10.1
		(2.3)	(3.8)	(3.2)	(4.6)	(2.3)

All estimates are expressed in percentages. ( ) standard errors are given in parentheses.

quitters. The additional potential confounders, including poverty status, marital status and education, showed little association with the log odds in either disease group. This is consistent with previous investigations by Thun et al. (2000).

Using our matching algorithm, we can compare the estimated average positive expenditures for the cases to the estimated average positive expenditures for the matched controls. In the LC and CHD groups, the average positive expenditures for the cases are \$8810 and \$7258, respectively. For the confounder score matched controls, the corresponding values are \$2770 and \$2759. When we restrict our attention to just those cases and controls who smoke, we obtain an average positive expenditure of \$8659 (LC) and \$7221 (CHD) for the cases and \$2821 (LC) and \$2854 (CHD) for the matched controls. Therefore, the effect of having a major smoking attributable disease on positive medical expenditures is similar for smokers and non-smokers.

Table 4 displays the estimated SAF and SAFE for both disease groups. Bootstrap standard errors are presented in parentheses. For LC and CHD, we estimated that 70.2% (95% confidence interval: 60.0–80.4) and 19.6% (13.2–26.0) of the cases are attributable to smoking, respectively. When we stratify by age and sex, we found that the attributable fraction of disease cases is larger for males than females, and tends to be larger for people younger than 65 than for people older than 65 years old.

Table 4 also presents results for the smoking attributable expenditures. We estimated that 53.4% (95% confidence interval: 42.6–64.2) of the 9.57 billion dollars that was expended in 1987 on people in the LC disease group is attributable to smoking. For CHD, we estimated that 13.4% (9.6–18.2) of the 47.3 billion dollars expended for persons in the CHD disease group is attributable to smoking.

4.1. Sensitivity analysis

In our analysis, we have used statistical methods that are less dependent upon specific parametric assumptions when estimating the fraction of disease cases and expenditures

Table 5
Estimated SAFEs using the $K \times (1:10)$ matching algorithm with $K = 1, 5, 10$, or 20

	Matching scheme	All persons	Male		Female	
			< 65	65+	< 65	65+
LC						
	1:10	51.5	55.9	53.1	47.8	47.2
		(6.1)	(5.7)	(8.8)	(9.4)	(10.0)
	5:50	53.4	56.9	55.3	51.2	48.8
		(5.4)	(4.7)	(6.8)	(8.0)	(9.7)
	10:100	52.4	55.6	54.3	50.6	47.9
		(5.5)	(4.8)	(7.1)	(8.4)	(9.0)
	20:200	52.2	54.7	54.3	50.9	47.9
		(5.2)	(4.9)	(6.0)	(8.2)	(8.8)
CHD						
	1:10	13.6	14.6	16.9	14.9	10.2
		(2.2)	(3.7)	(3.1)	(4.5)	(2.2)
	5:50	13.4	14.3	16.6	14.5	10.1
		(2.3)	(3.8)	(3.2)	(4.6)	(2.3)
	10:100	13.4	14.3	16.6	14.4	10.2
		(2.4)	(3.8)	(3.3)	(4.6)	(2.3)
	20:200	13.4	14.3	16.6	14.4	10.1
		(2.4)	(3.8)	(3.3)	(4.6)	(2.3)

All estimates are expressed in percentages. ( ) standard errors are given in parentheses.

that are attributable to smoking. In doing so, however, we have made methodologic choices. In this section, we assess the sensitivity of our findings to two key choices: the size of the matched case—control groups in our expenditure analysis, and the use of the matching algorithm compared to more traditional parametric approaches.

4.1.1. Choice of K in the matching algorithm

An advantage of our method is that we do not depend upon a particular parametric model for medical expenditures. Instead, we have introduced a matching algorithm in which a number of cases (K) and controls ($K \times 10$) must be specified. Here, we examine how sensitive our estimates are to the choice of K.

Table 5 presents the estimated SAFEs for $K = 1, 5, 10$ and 20. The bootstrap standard errors are given in parentheses. Note that there are very small differences in the SAFEs across the various choices of K. Also note that the standard errors for the estimates tend to decrease as the choice of K increases.

4.1.2. Comparing the matching algorithm to log-normal models

In this section, we will compare our estimates of the SAFEs using $K = 5$ to more traditional parametric approaches. Specifically, we consider using a log-normal model to estimate $\mathrm{E}(C_i | C_i > 0,\ D_i,\ X_i)$ for those subjects with and without disease. The log-normal model includes all the explanatory variables which were included in the

Table 6
Estimated SAFEs using our matching algorithm with $K = 5$ vs. the log-normal models with one or two smearing estimates

Model	All persons	Male		Female	
		< 65	65+	< 65	65+
LC					
5:50 matching algorithm	53.4	56.9	55.3	51.2	48.8
	(5.4)	(4.7)	(6.8)	(8.0)	(9.7)
Log-normal model with one smearing estimate	57.3	60.8	59.4	55.4	52.4
	(4.3)	(4.3)	(4.9)	(8.4)	(7.1)
Log-normal model with two smearing estimates	55.6	58.9	57.6	53.8	50.8
	(4.7)	(4.6)	(5.0)	(8.4)	(7.7)
CHD					
5:50 matching algorithm	13.4	14.3	16.6	14.5	10.1
	(2.3)	(3.8)	(3.2)	(4.6)	(2.3)
Log-normal model with one smearing estimate	15.1	16.2	18.8	16.2	11.5
	(2.5)	(4.0)	(3.2)	(5.0)	(2.4)
Log-normal model with two smearing estimates	14.2	15.2	17.6	15.3	10.8
	(2.4)	(3.7)	(3.0)	(4.8)	(2.3)

All estimates are expressed in percentages. ( ) standard errors are given in parentheses.

model for $P(C_i > 0|D_i, X_i)$ as described in Section 3.2. When obtaining our predicted values, we use a smearing estimate as proposed by Duan (1983). We first use a single smearing estimate and then allow the smearing estimate to differ for the two disease groups.

Table 6 presents the estimated SAFEs for our matching algorithm using $K = 5$ and the log-normal model. Note that the estimated SAFEs using our matching algorithm are conservative relative to the estimates using the log-normal model. Also note that estimated standard errors are similar in magnitude for the two approaches, although slightly smaller for the parametric model as would be expected. See Dominici and Zeger (2001) for a detailed comparison of non-parametric and parametric two-part estimates.

5. Discussion

This paper has estimated the fraction of medical expenditures attributable to persons suffering major diseases caused by smoking. Using semi-parametric statistical methods applied to the 1987 NMES, we estimate that 53% and 13% of medical expenditures to treat the LC and CHD disease groups are attributable to smoking, respectively. For both groups, the percentage is larger for persons under 65 years of age because their relative risk of disease comparing smokers to non-smokers is greater.

We estimate that 6.6% of medical expenditures for the 1987 NMES population who are 40–94 years of age are attributable to smoking. This corresponds to 4.6% for

persons 19–94 years of age. It is not surprising that this value is slightly lower than the 6–8% range of estimates previously reported (Miller et al., 1998). Our approach is conservative because we have focused only upon the major diseases caused by smoking and have only considered the effects of smoking on costs that are mediated through disease. That is, in our expenditure model, we did not allow for a direct effect of smoking on expenditures over and above the effect that results from higher prevalence of disease among smokers. We have compared expenditures for persons with disease to others without disease whether or not they are smokers and hence have ignored the tendency for smokers to have poorer health and greater expenditures absent a major disease (Miller et al., 1998).

Unlike previous studies, our analysis has used a continuous measure of cumulative smoking dose, allowed to vary as a smooth function of cumulative pack years for current smokers and allowed to vary as a two-dimensional, smooth function of cumulative pack years and time since quit for former smokers.

A key difference between previous approaches and ours is the use of a matching algorithm instead of parametric regression models to estimate the mean difference in expenditures for persons with and without diseases caused by smoking. Our matching algorithm uses a semi-parametric model for the probability to have disease as a way to control for potential confounders. However, we avoid modeling of both the mean and the variance of the log-positive expenditures as a function of disease, which is necessary for estimating the difference in average expenditures between groups in a log-normal model.

We have used the difference between the mean expenditure for matched cases and matched controls in our analysis. Dominici and Zeger (2001) have developed an alternate, more efficient estimate of the difference in means between two skewed distributions. Their method, called SQUARE, compares the quantile functions for cases and controls within a stratum of the propensity score. Under the assumption that the ratio of quantile functions is a smooth function of percentile, they developed an alternate estimate of the mean difference that has reduced variance as compared to the difference in sample means; this is applicable in situations such as the NMES data.

This paper has not specifically addressed the fact that approximately 20% of the NMES sample has missing smoking information. However, we used a combination of Bayesian data augmentation (Gelman et al., 1995), empirical sampling, and multiple imputation (Rubin, 1987) to account for missing smoking information. Our estimates were not sensitive to the adjustment for missing data, therefore only the estimates based on the complete data are presented here.

The methods used in this research have relaxed some of the parametric assumptions commonly applied to estimate smoking attributable expenditures. We have allowed non-linear dependence of disease risk on smoking level, age and other covariates by using generalized additive models (Hastie and Tibshirani, 1990). We have developed an approach in which we need not specify that positive expenditures follow a particular distribution, nor model the dependence of higher moments of the expenditure distribution on disease or smoking. Our statistical methods are complementary to those used by previous investigators. Our estimate of the smoking attributable expenditures due

only to direct effects of smoking on a limited set of major diseases are consistent with the slightly larger estimates previously published that incorporate direct and indirect pathways.

Acknowledgements

We are grateful for the support from NIMH Grant R01 MH56639. Funding for Francesca Dominici was also provided by a grant from the Health Effect Institute (Walter A. Rosenblith New Investigator Award). Scott L. Zeger has served as an expert and consultant on behalf of Federal and State agencies and private health plans in lawsuits against the tobacco industry.

References

Adams, E.K., Young, T.L., 1999. Costs of smoking: a focus on maternal, childhood, and other short-run costs. Medical Care Research and Review 56, 3–29.

Bartlett, J., Miller, L.S., Rice, D.P., Max, W., 1994. Medical-care expenditures attributable to cigarette smoking—United States, 1993. Morbidity Mortality Weekly Report 43, 469–472.

Cochran, W.G., Rubin, D.B., 1973. Controlling bias in observational studies: a review. Sankhyā, Series A, Indian Journal of Statistics 35, 417–446.

D'Agostino Jr., R.B., Rubin, D.B., 2000. Estimating and using propensity scores with partially missing data. Journal of the American Statistical Association 95, 749–759.

Department of Health and Human Services, 1984. The health consequences of smoking—chronic obstructive lung disease. A Report of the Surgeon General, US Government Printing Office, Washington, DC.

Department of Health and Human Services, 1989. Reducing the health consequences of smoking. 25 years of progress. A Report of the Surgeon General. US Government Printing Office, Washington, DC.

Department of Health and Human Services, 1990. Health Benefit of Smoking Cessation. A Report of the Surgeon General. US Government Printing Office, Washington, DC.

Doll, R., Hill, A., 1956. Lung cancer and other causes of death in relation to smoking. British Medical Journal 2, 1071–1081.

Doll, R., Peto, R., Wheatly, K., Gray, R., Sutherland, I., 1994. Mortality in relation to smoking: 40 years' observations on male British doctors. British Medical Journal 309, 901–911.

Dominici, F., Zeger, S.L., 2001. Smooth quantile ratio estimation (SQUARE). Technical Report, Johns Hopkins University, Baltimore, MD.

Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. Journal of the American Statistical Association 78, 605–610.

Duan, N., Manning Jr., W.G., Morris, C.N., Newhouse, J.P., 1983. A comparison of alternative models for the demand for medical care. Journal of Business and Economic Statistics 1, 115–126.

Efron, B., Tibshirani, R.J., 1991. Statistical data analysis in the computer age. Science 253, 390–395.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Environmental Protection Agency (EPA), 1993. Respiratory health effects of passive smoking: lung cancer and other disorders. Technical Report EPA/600/6-90/006F, US Environmental Protection Agency.

Gelman, A., Carlin, J., Stern, H., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & Hall, New York.

Gordis, L., 1996. Epidemiology. W.B. Saunders Company, Philadelphia.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, New York.

Hodgson, T., Kopstein, A., 1984. Health care expenditures for major diseases in 1980. Health Care Financing Review 64, 489–547.

Krieger, N., Quesenberry, C., Peng, T., Horn-Ross, P., Stewart, S., Brown, S., Swallen, K., Guillermo, T., Suh, D., Alvarez-Martinez, L., Ward, F., 1999. Social class, race/ethnicity, and incidence of breast, cervix, colon, lung, and prostate cancer among Asian, black, Hispanic, and white residents of the San Francisco Bay Area, 1988–92. Cancer Causes and Control 10, 525–537.

Levin, M., 1953. The occurrence of lung cancer in man. Acta Union International Contra Cancrum 9, 531–541.

Lightwood, J.M., Phibbs, C.S., Glantz, S.A., 1999. Short-term health and economic benefits of smoking cessation: low birth weight. Pediatrics 104, 1312–1320.

Lipscomb, J., Ancukiewicz, M., Parmigiani, G., Hasselblad, V., Samsa, G., Matchar, D.B., 1998. Predicting the cost of illness: a comparison of alternative models applied to stroke. Medical Decision Making 18S, S39–S56.

Luce, B., Schweitzer, S., 1978. Smoking and alcohol abuse: a comparison of their economic consequences. New England Journal of Medicine 298, 569–571.

Manning, W., Keeler, E., Newhouse, J., 1989. The taxes of sin. Do smokers and drinkers pay their way? Journal of American Medical Association 261, 1604–1609.

McBride, P.E., 1992. The health consequences of smoking: cardiovascular diseases. Medical Clinics of North America 76, 333–349.

Medical Expenditure Panel Survey, 1997. Agency for healthcare research and quality. National Center for Health Statistics, http://www.meps.ahcpr.gov.

Miller, L.S., Zhang, X., Rice, D.P., Max, W., 1998. State estimates of total medical expenditures attributable to cigarette smoking, 1993. Public Health Reports 113, 447–458.

Miller, V.P., Ernst, C., Collin, F., 1999. Smoking-attributable medical care costs in the USA. Social Science and Medicine 48, 375–391.

Munoz, B., West, S.K., Rubin, G.S., Schein, O.D., Quigley, H.A., Bressler, S.B., Bandeen-Roche, K., 2000. Causes of blindness and visual impairment in a population of older Americans: the salisbury eye evaluation study. Archives of Ophthalmology 118, 819–825.

National Center For Health Services Research, 1987. National Medical Expenditure Survey. Methods II. Questionnaires and data collection methods for the household survey and the Survey of American Indians and Alaska Natives. National Center for Health Services Research and Health Technology Assessment.

Osann, K., 1998. Epidemiology of lung cancer. Current Opinion in Pulmonary Medicine 4, 198–204.

Rice, D.P., Hodgson, T.A., Sinsheimer, P., Browner, W., Kopstein, A.N., 1986. The economic costs of the health effects of smoking, 1984. The Milbank Quarterly 64, 489–547.

Rosenbaum, P.R., Rubin, D.B., 1984. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79, 516–524.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Rubin, D.B., 2000. Statistical issues. In: Gastwirth, J. (Ed.), Statistical Science in the Court-room. Springer, New York, pp. 19–55.

Rubin, D.B., 2001. Estimating the causal effects of smoking. Statistics in Medicine 20, 1395–1414.

Rubin, R.B., Thomas, N., 2000. Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association 95, 573–585.

Sherman, C.B., 1992. The health consequences of smoking: pulmonary diseases. Medical Clinics of North America 76, 355–372.

Smith, W., Assink, J., Klein, R., Mitchell, P., Klaver, C., Klein, B., Hofman, A., Jensen, S., Wang, J.J., de Jong, P.T.V.M., 2001. Risk factors for age-related macular degeneration: pooled findings from three countries. Ophthalmology 108, 697–704.

Strassels, S.A., Smith, D.H., Sullivan, S.D., Mahajan, P.S., 2001. The costs of treating copd in the United States. Chest 119, 344–352.

Thun, M.J., Apicella, L.F., Henley, S.J., 2000. Smoking vs. other risk factors as the cause of smoking-attributable deaths: confounding in the courtroom. Journal of American Medical Association 284, 706–712.

Tockman, M.S., Anthonisen, N.R., Wright, E.C., Donithan, M., 1987. Airways obstruction and the risk for lung cancer. Annals of Internal Medicine 106, 512–518.

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, 290–300.

Woodard, M., 1999. Epidemiology: Study Design and Data Analysis. Chapman & Hall/CRC, Boca Raton, FL.

Wynder, E., Bross, I., Cornfield, J., O'Donnell, W., 1956. Lung cancer in women: a study of environmental factors. New England Journal of Medicine 255, 1111–1121.

Zeger, S., Wyant, T., Miller, L., Samet, J., 2000. Statistical testimony on damages in Minnesota versus the Tobacco Industry. In: Gastwirth, J. (Ed.), Statistical Science in the Courtroom. Springer, New York, pp. 19–55.