American Journal of Epidemiology
Copyright © 2002 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved

Vol. 156, No. 9
Printed in U.S.A.
DOI: 10.1093/aje/kwf122

Modeling Smoking History: A Comparison of Different Approaches

Karen Leffondré[1,2], Michal Abrahamowicz[1,2], Jack Siemiatycki[3,4], and Bernard Rachet[3]

[1] Department of Epidemiology and Biostatistics, Faculty of Medicine, McGill University, Montreal, Quebec, Canada.
[2] Division of Clinical Epidemiology, The Montreal General Hospital, Montreal, Quebec, Canada.
[3] Department of Epidemiology and Biostatistics, INRS–Institut Armand-Frappier, Université du Québec, Laval-des-Rapides, Quebec, Canada.
[4] Canada Research Chair in Environmental Epidemiology, Université de Montréal, Montreal, Quebec, Canada.

Received for publication June 11, 2001; accepted for publication June 21, 2002.

The impact of cigarette smoking on various diseases is studied frequently in epidemiology. However, there is no consensus on how to model different aspects of smoking history. The aim of this investigation was to elucidate the impact of several decisions that must be made when modeling smoking variables. The authors used data on lung cancer from a case-control study undertaken in Montreal, Quebec, Canada, in 1979–1985. The roles of smoking status, intensity, duration, cigarette-years, age at initiation, and time since cessation were investigated using time-dependent variables in an adaptation of Cox's model to case-control data. The authors reached four conclusions. 1) The estimated hazard ratios for current and ex-smokers depend strongly on how long subjects are required to not have smoked to be considered "ex-smokers." 2) When the aim is to estimate the effect of continuous smoking variables, a simple approach can be used (and is proposed) to separate the qualitative difference between never and ever smokers from the quantitative effect of smoking. 3) Using intensity and duration as separate variables may lead to a better model fit than using their product (cigarette-years). 4) When estimating the effects of time since cessation or age at initiation, it is still useful to use cigarette-years, because it reduces multicollinearity.

case-control studies; epidemiologic methods; multicollinearity; multivariate analysis; neoplasms; proportional hazard; smoking; time-dependent covariate

Abbreviation: AIC, Akaike's Information Criterion.

Smoking is one of the most investigated risk factors in epidemiologic studies. However, there is little consensus regarding how to represent this multidimensional phenomenon. We screened 40 articles published in 2000 in epidemiologic and clinical journals that assessed the impact of smoking on various outcomes. We found considerable variation in both the nature of the data collected on smoking history and the way the data were used in the analyses. While some studies reported smoking status (never/current/ex-smoker) only (1–5), others reported detailed information on the number of cigarettes or packs smoked per day (intensity), duration of smoking (6–11), age at initiation (12–16), and/or time since cessation (14–17). However, even studies with similar objectives and the same data on smoking used this information differently. For example, among cancer studies, the amount of time for which subjects were required to have refrained from smoking to be considered ex-smokers varied from 1 day (12) to 5 years (18). When both intensity and duration were known, the most commonly used variable was cigarette-years or pack-years (7, 19), calculated as the product of the two variables—which implies that intensity and duration have the same impact. However, some cancer studies suggested that duration was more important than intensity (20), while in others, duration was not significant after adjustment for intensity (12). In addition, little is known about the effects of age at smoking initiation and time since smoking cessation, variables that are often not adjusted for duration of smoking (15, 16). Among studies with detailed information, few considered more than one smoking variable at a time (12, 14)—partly because of the difficulty

Reprint requests to Dr. Michal Abrahamowicz, Division of Clinical Epidemiology, The Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec H3G 1A4, Canada (e-mail: michal@michal.ri.mgh.mcgill.ca).

TABLE 1. Demographic characteristics of subjects in a case-control study of environmental exposures and cancer at the time of diagnosis/interview, Montreal, Quebec, Canada, 1979–1985

Characteristic	Lung cancer cases (*n* = 640)		Cancer controls (*n* = 485)		Population controls (*n* = 430)	
	%	Mean	%	Mean	%	Mean
Age (years)*		59.4 (7.0)†		58.8 (8.0)		59.7 (7.7)
Respondent type						
Self	77.7		85.2		87.0	
Proxy	22.3		14.8		13.0	
Ethnic group						
French	68.5		59.2		64.2	
Anglophone	13.1		16.7		13.5	
Italian	6.1		8.9		7.4	
Other European	7.8		5.8		7.9	
Jewish	1.4		4.7		2.6	
Other	3.1		4.7		4.4	
Occupational index (years)‡		20.5 (16.6)		16.4 (15.7)		18.5 (16.9)
Annual income (Canadian dollars)§		22,547 (7,791)		24,557 (8,630)		26,334 (8,363)

* Controls were age-stratified to match the age distribution of cases.

† Numbers in parentheses, standard deviation.

‡ Total duration of exposure to substances or jobs considered to be risk factors for lung cancer.

§ Approximately 54% of subjects responded directly to the question on income; for the rest of the subjects, we used the median income of the census tract of residence as a proxy measure, with the latter estimate being normalized to the distribution of the self-responses.

in simultaneously modeling several aspects of the same exposure (21, 22), especially those directly related to time. It remains unclear how ignoring some aspects of smoking history may affect the estimated effects of other smoking-related variables. Thus, there is a need to understand how different approaches may lead to different conclusions regarding the role of various aspects of smoking history.

The overall objective of this investigation was to examine the limitations and advantages of the different approaches used in the literature for modeling smoking-related variables, using data on lung cancer. Specifically, we focused on 1) the operational definition of the distinction between current smokers and ex-smokers; 2) the impact, on the estimated effects of continuous smoking-related variables, of including never smokers in the analysis; 3) the implications of replacing intensity and duration, as separate variables, by cigarette-years; and 4) the simultaneous modeling of several smoking-related variables.

MATERIALS AND METHODS

Data

We used data from a case-control study undertaken in Montreal, Quebec, Canada, during 1979–1985 to investigate associations between environmental and occupational exposures and several types of cancer in males. The study design has been described in detail elsewhere (23–25). A total of 857 primary lung cancer cases, 533 population controls, and approximately 3,000 patients with cancer at other sites ("cancer controls") were interviewed. For the present analysis, we excluded subjects who had not completed the detailed questionnaire and controls with cancers thought to be related to smoking (26), which left 640 lung cancer cases, 430 population controls, and 485 cancer controls. Data on smoking included the ages at which regular cigarette smoking began and ceased and the average amount smoked daily.

Demographic and smoking-related characteristics of the subjects at diagnosis/interview are shown in tables 1 and 2, respectively. The distributions of the smoking variables were very similar among persons in the two control groups, who were less exposed to smoking than cases. These differences were not due to age at diagnosis/interview, since controls were age-stratified to match the age distribution of the cases (24).

Statistical methods

Logistic regression, the standard method for analyzing case-control data, cannot directly account for variation in lifetime exposures, such as smoking habits. Therefore, to incorporate time-dependent covariates, we used an adaptation of Cox's model (27) for case-control studies (28, 29). The time axis was represented by age, from birth to diagnosis. This implied that the effects of all variables, expressed by the hazard ratio, were adjusted for current age. Controls were censored at the age of diagnosis/interview. To account for retrospective identification of cases, we included in the risk set at each age of diagnosis only the cases diagnosed at

TABLE 2. Smoking-related characteristics of subjects in a case-control study of environmental exposures and cancer at the time of diagnosis/interview, Montreal, Quebec, Canada, 1979–1985

Characteristic	Lung cancer cases (*n* = 640)		Cancer controls (*n* = 485)		Population controls (*n* = 430)	
	%	Mean	%	Mean	%	Mean
Smoking status						
Never smoker	1.3		19.8		18.6	
Current smoker	70.9		47.8		47.2	
Ex-smoker*	27.8		32.4		34.2	
Time since cessation of smoking						
1 day–2 years	42.1		17.8		10.9	
2–5 years	20.8		19.1		23.1	
5–10 years	23.1		17.8		21.1	
10–15 years	8.4		13.4		12.9	
>15 years	5.6		31.9		32.0	
Intensity of smoking (cigarettes/ day)†		37.1 (18.4)‡		27.5 (15.8)		28.1 (15.3)
Duration of smoking (years)†		41.3 (8.7)		36.7 (12.7)		36.8 (11.9)
Cigarette-years (cigarettes/ day × years)†		1,532 (811)		1,017 (697)		1,027 (641)
Age at initiation of smoking (years)†		16.7 (4.3)		17.9 (5.6)		18.1 (4.7)

* Subject had stopped smoking at least 1 day before the interview.

† Mean values and standard deviations among current and ex-smokers.

‡ Numbers in parentheses, standard deviation.

that age and all controls not censored before (30). Thus, each case was included in only one risk set. To account for 1) interdependence of subsequent observations, corresponding to different values of time-dependent covariates, for the same subject and 2) manipulation of the risk sets (30), we relied on a robust variance estimator (31–33).

In each model, smoking-related variables were represented by time-dependent covariates, updated at each age. For example, a 60-year-old subject who started smoking at age 20 years and stopped at age 50 years was assigned 0, 10, 30, and 30 years of duration at ages 15, 30, 50, and 60 years, respectively. In analyses restricted to current smokers, this subject would be included only for age intervals from 20 to 50 years. Age at initiation, intensity, duration, and cigarette-years were represented by continuous variables. In one set of analyses, they were assumed to have linear effects on risk, while in another set, they underwent logarithmic transformation. Since the two sets of findings were similar, we present only those obtained with untransformed variables. Since its effect on the hazard was not monotonic (figure 1), time since cessation was categorized (table 2). All models included all of the potential confounders listed in table 1. Occupational index was a time-dependent variable representing the duration of exposure to substances or jobs considered to be risk factors for lung cancer (26). Ethnicity was represented by dummy variables, and annual income was log-transformed. For analyses involving the same data set, the goodness of fit of different models was compared using Akaike's Information Criterion (AIC), computed as –2(log-likelihood) + 2(number of estimated parameters); a lower AIC indicates a better fit (34).

Analyses were initially conducted separately for population controls and cancer controls, but because results were similar, only results obtained with all controls pooled are shown.

RESULTS

Distinction between current smokers and ex-smokers

The most common smoking variable used in epidemiologic studies is smoking status, defined as never smoker/ever smoker or as never smoker/current smoker/ex-smoker. While there is general agreement regarding the definition of ever smokers (7, 12, 14, 17, 20, 35), there is much greater variability regarding the discrimination between current smokers and former smokers, even among studies of lung cancer (11, 18). However, as is shown below, the estimated hazard ratio for current and ex-smokers depends substantially on this definition.

In table 3, model 1 includes only an indicator of ever smoking, while models 2–5 distinguish current smokers from ex-smokers. All five models used never smokers as the reference category, but the cutoff for the identification of ex-smokers varied. For example, model 3 defined ex-smokers as those who had stopped smoking at least 1 year previously. Table 3 shows that, as expected, the risk of developing lung cancer was systematically lower for ex-smokers than for current smokers. In models 2–5, increasing the cutoff implied that the category "ex-smokers" excluded more and

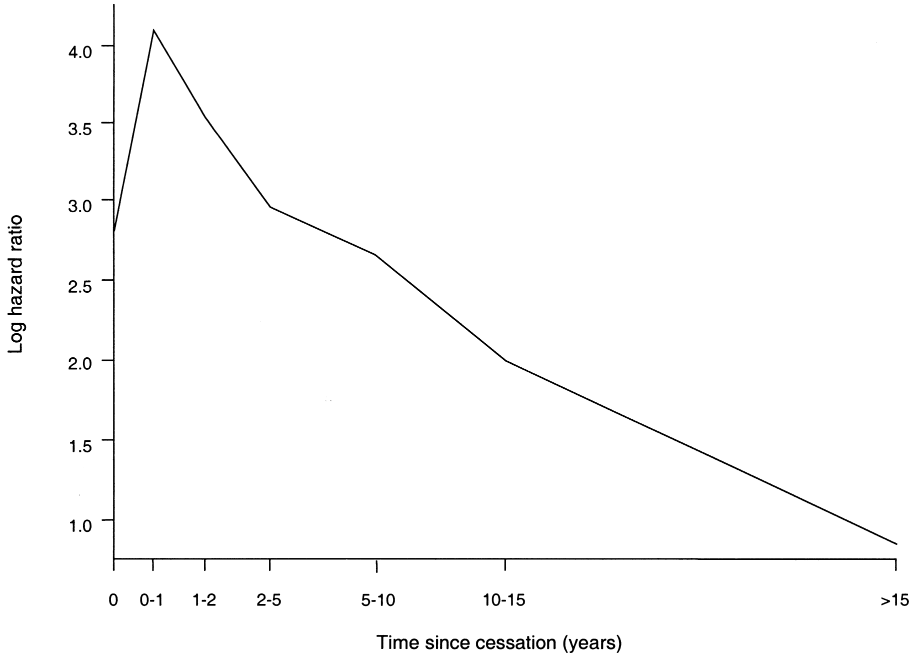

FIGURE 1. Log hazard ratio for lung cancer by time since smoking cessation, adjusted for respondent type, ethnic group, occupational index, and annual income, with never smokers used as the reference group, Montreal, Quebec, Canada, 1979–1985.

more subjects who had only recently stopped smoking. Consequently, the hazard ratio for ex-smokers gradually decreased from model 2 to model 5. However, while "current smokers" included an increasing number of smokers who had already stopped smoking, the hazard ratio for current smokers increased from model 2 to model 5. This occurred because smokers who had stopped smoking less than 2 years previously had a higher risk than actual current smokers (time since cessation = 0), as shown in figure 1. One reason may be that some case patients with early symptoms of lung cancer stopped smoking just before diagnosis, inducing a reverse causality bias, in that recent smoking

TABLE 3. Impact of the definition of an "ex-smoker" on estimated hazard ratios for lung cancer, Montreal, Quebec, Canada, 1979–1985

Model	Cutoff*	Smoking status†	Cases (*n* = 640)		Controls‡ (*n* = 915)		HR§,¶	95% CI§	AIC§,#
			No.**	%**	No.**	%**			
1	N/A§	Ever smoker	632	98.7	739	80.8	15.8	7.6, 32.8	7,340
2	1 day	Ex-smoker	178	27.8	304	33.2	14.2	6.7, 30.3	7,339
		Current smoker	454	70.9	435	47.5	16.6	7.9, 34.8	
3	1 year	Ex-smoker	129	20.1	285	31.2	11.0	5.1, 23.7	7,316
		Current smoker	503	78.6	454	49.6	18.3	8.7, 38.4	
4	2 years	Ex-smoker	103	16.1	260	28.4	9.4	4.3, 20.4	7,297
		Current smoker	529	82.6	479	52.4	18.9	9.0, 39.8	
5	5 years	Ex-smoker	66	10.3	196	21.4	7.3	3.3, 16.0	7,277
		Current smoker	566	88.4	543	59.3	19.0	9.0, 39.9	

* Cutoff corresponding to the minimum time interval for which the subjects were required to have stopped smoking to be considered ex-smokers.

† Smoking status was ascertained at each age.

‡ Pooled cancer and population control groups.

§ HR, hazard ratio; CI, confidence interval; AIC, Akaike's Information Criterion; N/A, not applicable.

¶ Hazard ratio relative to never smokers, adjusted for respondent type, ethnic group, occupational index, and annual income.

# A lower AIC indicates the best fit to the data for the same data set.

** Number or percentage of ever, current, or ex-smokers at the time of diagnosis/interview. Hazard ratios were estimated from the time-varying smoking status variables.

TABLE 4. Impact of the inclusion of never smokers on the estimated effect of cigarette-years of smoking on lung cancer risk, Montreal, Quebec, Canada, 1979–1985

Model	Data set*	Smoking variable(s)	Unit or category	HR†,‡	95% CI†	AIC†,§
6	Smokers only	Cigarette-years	800	1.91	1.60, 2.27	6,900
7	All subjects	Cigarette-years¶	800	2.03	1.73, 2.40	7,093
8	Never smokers plus half of smokers#	Cigarette-years¶	800	2.30	1.79, 2.95	3,127
9	All subjects	Ever smoking	Yes/no	14.83	6.96, 31.61	7,046
		Cigarette-years**	800	1.91	1.60, 2.27	

* Data set used in the analysis.

† HR, hazard ratio; CI, confidence interval; AIC, Akaike's Information Criterion.

‡ Hazard ratio for a one-unit increase in the explanatory variable, as indicated in the fourth column, adjusted for respondent type, ethnic group, occupational index, and annual income.

§ A lower AIC indicates the best fit to the data for the same data set.

¶ For never smokers, cigarette-years = 0.

# Fifty percent of smokers were randomly eliminated from the data set to increase the proportion of never smokers.

** To allow direct interpretation of the hazard ratio for ever smoking, cigarette-years was centered. For never smokers, centered cigarette-years = 0; for ever smokers, centered cigarette-years = cigarette-years minus mean(cigarette-years).

cessation may be a marker for early symptoms of lung cancer (36). If so, it may be preferable to discount some recent exposure (37) and use, for example, a cutoff of approximately 2 years for defining ex-smokers. In any case, these results illustrate the importance of both carefully defining ex-smokers and reporting this definition.

Including never smokers

In some studies, the effect of continuous smoking-related variables is estimated while including in the analysis never smokers, who are assigned a value of 0 for the relevant variable(s) (6, 14, 20). This assumes that the difference between ever and never smokers is quantitative rather than qualitative. Table 4 explores this issue by focusing on the cigarette-years variable. In model 6, which used only smokers, the estimated effect of cigarette-years is not distorted by the qualitative difference between ever and never smokers. In models 7–10, never smokers were included and assigned a value of 0 for cigarette-years. Model 7, which did not include an indicator of ever smoking, slightly overestimated the impact of cigarette-years compared with model 6. Artificially increasing the proportion of never smokers yielded an even higher estimate (model 8), suggesting that the overestimation might be due to the inclusion of never smokers.

Figure 2 provides an explanation for this overestimation. The solid curve, representing a flexible 3-df smoothing spline estimate from the generalized additive model (38), shows the nonlinearity of the cigarette-years effect. The dashed line shows the linear estimate of the cigarette-years effect, obtained by pooling data from never and ever smokers and imposing the (incorrect) linearity assumption. Because the slope of the dashed line is the weighted average of the local slopes of the solid curve (39), it underestimates the initial difference between never and light smokers and overestimates the continuous effect of increasing cigarette-years among smokers.

To avoid such difficulties, model 9 included an indicator for ever smoking (table 4). Accordingly, the effect of cigarette-years was estimated by comparing only subjects who had the same value as the indicator, that is, only smokers. In addition, we centered cigarette-years by subtracting the mean cigarette-years value from the original value for all smokers, while keeping 0 for never smokers. Such a linear transformation of cigarette-years does not change its estimated effect (40), but it allows the effect of ever smoking to compare average smokers with never smokers, since both groups are assigned a value of 0 for centered cigarette-years. Accordingly, the hazard ratios for ever smoking in models 1 and 9 are very similar. Without this transformation of cigarette-years, the estimated hazard ratio for ever smoking would be more difficult to interpret, as it would compare never smokers and hypothetical smokers with 0 cigarette-years. Thus, model 9 provides interpretable estimates of both the qualitative effect of smoking status and the quantitative effects of smoking exposure. Moreover, the very significant effect of cigarette-years in model 9 shows how much information is lost if only smoking status is included in the model.

Replacing intensity and duration by cigarette-years

Among studies that have information on both intensity and duration, most use the product of the two (7, 19, 41)—that is, pack-years or cigarette-years—rather than two separate variables (10, 20). We assessed the implications of such an approach.

First, to see whether the effects of intensity, duration, and cigarette-years differed between current smokers and ex-smokers, we tested their interactions with current smoker/ex-smoker status (data not shown). The effect of intensity was significantly stronger among current smokers than among ex-smokers (p = 0.05 for interaction), whereas the effect of duration was significantly lower (p = 0.03). In contrast, the interaction between cigarette-years and status was not signif-

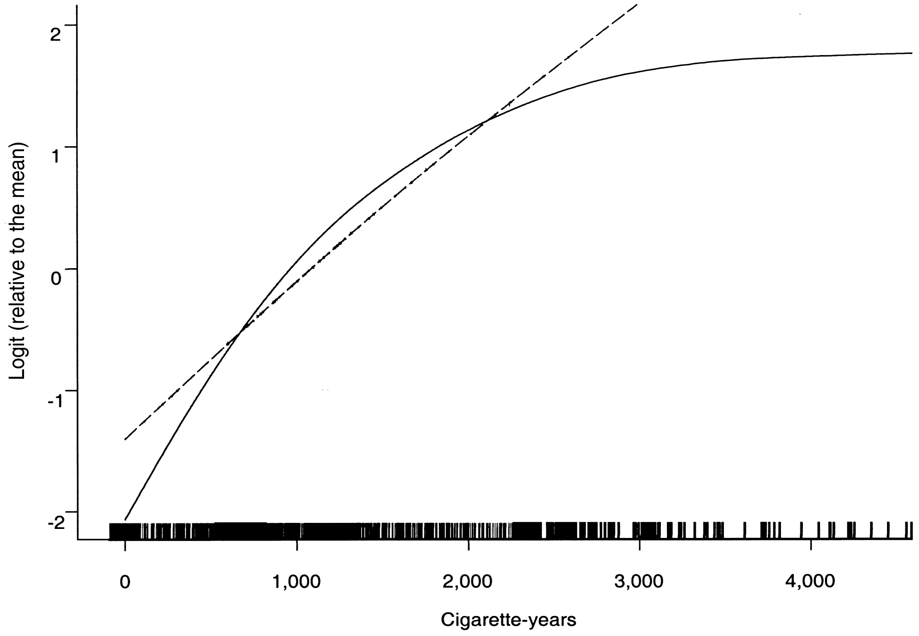

FIGURE 2. Effect of cigarette-years of smoking on the risk of lung cancer, adjusted for respondent type, ethnic group, occupational index, and annual income, Montreal, Quebec, Canada, 1979–1985. The solid curve represents the flexible estimate from the generalized additive model (a 3-df smoothing spline), with the difference in risks between light smokers (500 cigarette-years) and never smokers (0 cigarette-years) being much larger than the slope for actual smokers. The dashed line represents the linear estimate of the effect of cigarette-years over its entire range (0–4,000). The tick marks on the horizontal axis represent the distribution of cigarette-years.

icant (p = 0.48), probably because the differences in the effects of intensity and duration compensated for each other.

Because of these differences and because some studies compared the effects of intensity and duration using both current smokers and ex-smokers (20), we created four models using either current smokers or all smokers (table 5). For current smokers, intensity only (model 10) fitted the data better than duration only (model 11). Model 12 shows that duration was no longer significant after adjustment for intensity. This may result from insufficient variation in duration among current smokers who were compared at the same age and started smoking at similar ages (interquartile range: 14.5–18.6 years). Finally, the difference in AIC of 21.0 between model 12 and model 13 indicates that model 12, which included intensity and duration as separate variables, fitted the data for current smokers substantially better than model 13, which used cigarette-years (42). This result further confirms that the two ways of modeling these variables are not equivalent. For all smokers, intensity only (model 10) also fitted the data better than duration only

TABLE 5. Comparison of the effects of intensity of smoking, duration of smoking, and cigarette-years of smoking on lung cancer risk among current smokers and all smokers, Montreal, Quebec, Canada, 1979–1985

Model	Smoking variable	Unit	Current smokers*			All smokers†		
			HR‡,§	95% CI‡	AIC‡,¶	HR	95% CI	AIC
10	Intensity	15 cigarettes/day	1.76	1.49, 2.08	4,675	1.50	1.31, 1.71	7,021
11	Duration	10 years	1.64	1.15, 2.35	4,857	1.84	1.54, 2.20	7,064
12	Intensity	15 cigarettes/day	1.73	1.47, 2.05	4,674	1.57	1.37, 1.79	6,866
	Duration	10 years	1.20	0.84, 1.70		1.83	1.50, 2.24	
13	Cigarette-years	800	1.84	1.49, 2.27	4,695	1.91	1.60, 2.27	6,900

* Models were fitted using current smokers only. Smoking status was ascertained at each age, and at a given age, only subjects who were smoking at that age were considered current smokers.

† Models were fitted using all smokers (never smokers were not included).

‡ HR, hazard ratio; CI, confidence interval; AIC, Akaike's Information Criterion.

§ Hazard ratio for a one-unit increase in the explanatory variable, as indicated in the third column, adjusted for respondent type, ethnic group, occupational index, and annual income.

¶ A lower AIC indicates the best fit to the data for the same data set.

TABLE 6. Hazard ratios for lung cancer from two different combinations of the variables “intensity of smoking” and “duration of smoking” that entail the same number of cigarette-years, based on model 12,* Montreal, Quebec, Canada, 1979–1985

Cigarette-years (cigarettes/day × years)	Combination 1		Combination 2		Hazard ratio† for combination 1 vs. combination 2	
	Intensity (cigarettes/day)	Duration (years)	Intensity (cigarettes/day)	Duration (years)	Current smokers	All smokers
900	30	30	30	30	1.00	1.00
900	25	36	36	25	0.81‡	1.40§
900	20	45	45	20	0.62	2.15
900	15	60	60	15	0.43	3.96
900	12	75	75	12	0.31	6.87

* Model 12, identified in table 5, included intensity and duration as separate variables.

† Hazard ratio adjusted for respondent type, ethnic group, occupational index, and annual income.

‡ Among current smokers, the estimated coefficients from model 12 for intensity and duration were 0.0367 and 0.0179, respectively. Accordingly, the estimated hazard ratio for smoking 25 cigarettes/day over a period of 36 years, relative to smoking 36 cigarettes/day over 25 years, was exp(0.0367 × 25 + 0.0179 × 36)/exp(0.0367 × 36 + 0.0179 × 25) = 0.81.

§ Among all smokers, the estimated coefficients from model 12 for intensity and duration were 0.0301 and 0.0607, respectively. Accordingly, the estimated hazard ratio for smoking 25 cigarettes/day over a period of 36 years, relative to smoking 36 cigarettes/day over 25 years, was exp(0.0301 × 25 + 0.0607 × 36)/exp(0.0301 × 36 + 0.0607 × 25) = 1.40.

(model 11), but the effect of duration remained statistically significant after adjustment for intensity (model 12). The estimated effect of duration appeared stronger among ex-smokers, probably because of confounding by time since cessation. Indeed, given the small variation in age at initiation, longer duration was systematically associated with more recent cessation among ex-smokers of the same age. Thus, failure to account for a strong impact of time since cessation (figure 1) inflated the effect of duration among ex-smokers.

Table 6 further explores the implications of replacing intensity and duration (model 12) by the product of the two (model 13). Each row in table 6 compares model 12 estimates for two combinations of intensity and duration corresponding to the same amount of cigarette-years (900), separately for current smokers and all smokers. For example, among current smokers, the hazard ratio for smoking 25 cigarettes per day over a period of 36 years, relative to smoking 36 cigarettes per day over 25 years, was 0.81. In contrast, according to model 13, both groups have the same risk (hazard ratio = 1) by definition, as both smoked 900 cigarette-years. Subsequent rows of table 6 indicate that the discrepancy between the predictions of model 12 (hazard ratio = 0.62, 0.43, or 0.31) and model 13 (hazard ratio = 1) becomes even bigger as the contrast between intensity and duration increases. Overall, table 6 shows that, among current smokers, using cigarette-years underestimates the impact of intensity and overestimates the impact of duration. However, the last column of table 6 shows an opposite pattern among all smokers, for whom duration has a stronger effect than intensity, which reflects the interactions of the indicator of current/ex-smoking with intensity and with duration.

In summary, whereas replacing intensity of smoking and duration of smoking by the product of the two variables may induce a loss of information, the implications are different for current smokers versus all smokers. Our results also suggest the need to adjust for time since quitting smoking in analyses including ex-smokers. However, such an adjustment might introduce some problems, which are evaluated below.

Simultaneous modeling of several time-related smoking variables

To investigate the implications of simultaneous modeling of age at smoking initiation, duration of smoking, and/or time since smoking cessation, we focused on the impact of age at initiation.

All five models shown in table 7 included an indicator of ever smoking and selected centered continuous smoking variables. Older age at initiation appears to be associated with significantly lower risk if it is not adjusted for other smoking variables (model 14) or is adjusted only for intensity (model 15). After additional adjustment for duration, the effect of age at initiation is inverted (model 16), because among subjects of similar age and duration, those who started smoking earlier automatically have a greater time since cessation. Therefore, the “protective” effect of younger age at initiation in model 16 probably reflects confounding by time since cessation. Additional adjustment for time since cessation in model 17 resulted in higher estimated effects of both age at initiation and duration, in comparison with model 16. However, the confidence intervals for both estimates were much wider in model 17, indicating their numerical instability. This occurred because the four time-related variables (age at initiation, duration, time since cessation, and

TABLE 7. Results from simultaneous estimation of the effects of several smoking-related variables on lung cancer risk among all subjects, Montreal, Quebec, Canada, 1979–1985

Model	Smoking variable(s)*	Unit or category	HR†,‡	95% CI†	AIC†,§
14	Ever smoking	Yes/no	15.74	7.53, 32.88	7,289
	Age at initiation	5 years	0.73	0.62, 0.84	
15	Ever smoking	Yes/no	15.05	7.08, 32.00	7,130
	Age at initiation	5 years	0.77	0.65, 0.91	
	Intensity	15 cigarettes/day	1.48	1.30, 1.68	
16	Ever smoking	Yes/no	13.32	6.23, 28.47	6,993
	Age at initiation	5 years	1.39	1.11, 1.75	
	Intensity	15 cigarettes/day	1.63	1.42, 1.88	
	Duration	10 years	2.56	1.88, 3.49	
17	Ever smoking	Yes/no	11.45	4.93, 26.56	6,996
	Age at initiation	5 years	1.64	0.91, 2.97	
	Intensity	15 cigarettes/day	1.63	1.42, 1.87	
	Duration	10 years	3.54	1.17, 10.74	
	Time since cessation¶	>15 years	1.63	0.15, 18.02	
		10–15 years	1.43	0.30, 6.90	
		5–10 years	1.66	0.65, 4.68	
		2–5 years	1.48	0.81, 2.70	
18	Ever smoking	Yes/no	16.85	7.84, 36.21	7,009
	Age at initiation	5 years	1.00	0.85, 1.17	
	Cigarette-years	800	1.83	1.51, 2.21	
	Time since cessation¶	>15 years	0.22	0.10, 0.46	
		10–15 years	0.44	0.22, 0.86	
		5–10 years	0.80	0.52, 1.25	
		2–5 years	1.06	0.70, 1.61	

* Since all models included an indicator of ever smoking, all of the continuous smoking variables were centered.

† HR, hazard ratio; CI, confidence interval; AIC, Akaike's Information Criterion.

‡ Hazard ratio for a one-unit increase in the explanatory variable, as indicated in the third column, adjusted for respondent type, ethnic group, occupational index, and annual income.

§ A lower AIC indicates the best fit to the data for the same data set.

¶ The reference category was defined as current smokers and ex-smokers who had quit smoking less than 2 years previously.

current age) were nearly multicollinear. As figure 3 illustrates, the value of each time-related variable can be deduced from the other three variables, so their effects cannot be separated. We were able to derive an estimate in model 17 only because time since cessation was categorized, which avoided strict multicollinearity. However, interpretation of the resulting estimates is impossible. For instance, the hazard ratio for more than 15 years of cessation in model 17 in fact compares that category of ex-smokers with those who had stopped less than 2 years previously, assuming that both groups had the same current age and duration and started smoking at the same age, which is logically impossible. Indeed, it is evident that model 17 yields misleading estimates: Subjects who had stopped smoking less than 2 years previously seemed to have about 40–70 percent lower risks than those who had quit many years earlier, which would contradict findings on the widely accepted benefits of smoking cessation.

Figure 3 also indicates that for current smokers (time since cessation = 0), multicollinearity occurs between age at initiation, duration, and current age. Because we used current age as the time axis, this multicollinearity makes it impossible, for current smokers, to simultaneously model age at initiation and duration. Moreover, the fit obtained with age at initiation and intensity is exactly the same as that obtained with duration and intensity, and age at initiation and duration have exactly opposite coefficients in their respective models (data not shown), making it impossible to separate their effects.

One way to avoid multicollinearity is to reduce the number of variables directly related to time. In model 18, where intensity and duration are replaced with cigarette-years, both

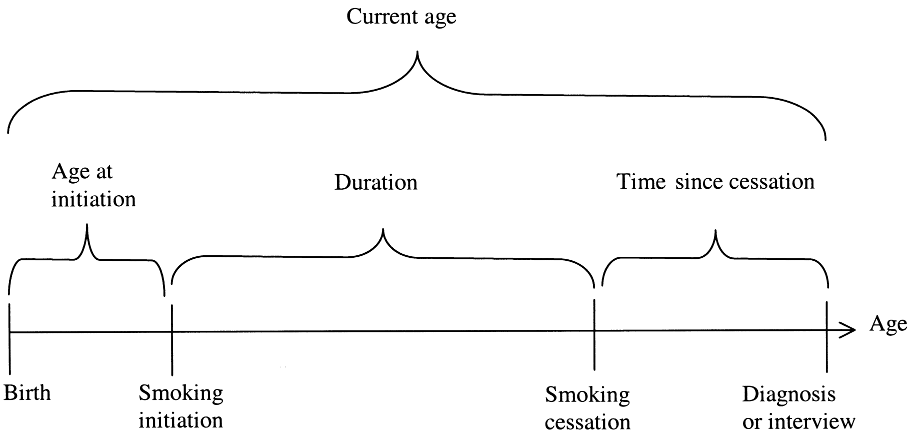

FIGURE 3. Theoretical relations between age at initiation of smoking, duration of smoking, time since cessation of smoking, and current age. The value of each variable can be deduced from that of the three other variables.

the estimated effect of cigarette-years and its confidence interval are very close to those obtained using only current smokers and including only cigarette-years (model 13 in table 5). Model 18 also shows that after adjustment for cigarette-years and time since cessation, age at initiation did not independently contribute to lung cancer risk, whereas longer time since cessation was still associated with significantly lower risk.

Thus, when investigating or adjusting for time since cessation and/or age at initiation, use of cigarette-years is preferable to the use of intensity and duration.

DISCUSSION

These analyses suggested four things. First, when estimating the effect of smoking status or adjusting for smoking status, the distinction between current smokers and ex-smokers should be carefully defined and reported. Second, when using never smokers in the analysis, including an indicator of ever smoking and centering the continuous smoking variables allows for easier interpretation of the estimates. Third, the effects of intensity and duration should be compared after excluding ex-smokers, to avoid confounding by time since cessation; and separating intensity and duration may lead to a better fit than cigarette-years. Fourth, when investigating the effects of time since cessation and/or age at initiation, adjusting for cigarette-years rather than for duration and intensity reduces multicollinearity between several time-related variables.

Overall, our results show that the estimated effects of various aspects of smoking history may depend substantially not only on their exact definitions but also on the inclusion of never smokers and/or ex-smokers in the analysis and on the other smoking variables taken into account. Considerable variation in the way different studies address these issues may limit the comparability of their findings.

Our study emphasizes the importance of information on smoking duration, intensity, and time since cessation. The importance of their independent effects suggests that some studies which considered smoking a potential confounder but adjusted only for smoking status (1–4) may have been subject to residual confounding (43). However, after current age is accounted for, multicollinearity makes it impossible to separate the effects of age at initiation, duration, and time since cessation among ex-smokers, as well as the effects of the first two variables among current smokers. Thus, a priori knowledge, possibly based on previous analyses, is necessary to decide which collinear variables should be included. For example, Barbone et al. (16) chose to include age at initiation and time since cessation, in addition to current age, in a multiple logistic regression analysis of the impact of smoking on lung cancer. Whereas both age at initiation and time since cessation had significant effects (16), these estimates could be confounded by duration, especially for age at initiation, which did not independently affect risks in our study.

Most of the phenomena discussed here are sufficiently general to apply to any study design and type of analysis. Indeed, all of our major findings were replicated using multiple logistic regression instead of Cox's model (data not shown). Still, the relative advantage of Cox's model versus logistic regression in analyses of case-control studies with time-varying exposures needs further systematic evaluation. Whereas some numerical results may depend on whether a linearity assumption is imposed on the effects of continuous smoking-related variables, the general phenomena illustrated in this study are probably quite robust. Indeed, in our sensitivity analyses, the logarithmic transformations of intensity, duration, and cigarette-years improved the goodness of fit and, for current smokers only, eliminated the difference in AIC between using intensity and duration (model 12) and using cigarette-years (model 13). However, all other conclusions were not affected by nonlinear transformations. Moreover, some numerical results may depend on study population, design, and/or outcome. For example, in contrast to the case with our findings, Karlson et al. (35) found that the effect of smoking duration on risk of rheumatoid arthritis was the same for both current smokers and ex-smokers. Our finding that ex-smokers were at higher risk than current smokers for several years after quitting smoking (figure 1) may also be outcome-specific, as it may reflect a substantial lag in the impact of smoking on lung cancer (44). This lag implied that the "effective" cumulative exposure for current smokers was much smaller than that for ex-smokers with the same duration. If so, the effect of time since cessation may be quite different for other outcomes such as cardiovascular disease, where the lag time seems to be much shorter (44). Accordingly, the patterns of confounding between time since cessation and duration may also differ substantially between different diseases. Moreover, while the small number of nonsmoking case patients ($n = 8$) might have affected the numerical stability of analyses based on the use of never smokers as the reference group, the findings illustrated in table 3 were replicated when we used ex-smokers as the reference group (data not shown). Finally, while the effects of particular variables may be different for women and men, the methodological problems illustrated by our analyses probably apply to women as well.

Another limitation was that we had no information on interruptions in a subject's smoking throughout his lifetime or on variation in smoking intensity over time. While the former information might be helpful, as it would reduce some of the multicollinearity problems, the latter would be useful to enhance the accuracy of the estimated effect of intensity. In our own study population, we doubt that interruptions represented a large enough fraction of total exposure time to make a difference. Most of our subjects were born between 1910 and 1935 and incurred their smoking exposure prior to the 1980s, before there were widespread antismoking programs.

As in most other studies of the impact of smoking, we had no data on the actual composition of cigarettes smoked by our subjects in different periods. As has been demonstrated in regard to esophageal cancer (15) and other outcomes, the kind of tobacco smoked can influence risk. This could have distorted some of our findings. If the higher tar levels of cigarettes in the past increased their carcinogenic potency, our inability to take this into account might have increased the estimated impact of a long duration of smoking. To account for temporal changes in cigarette composition, one could attempt to adjust for calendar time (45). However, this would further complicate the problem of multicollinearity between several time-related aspects of smoking history. This issue requires further investigation.

The relevance of some methodological issues addressed in this study is not limited to smoking. The reverse-causality bias discussed in relation to table 3 has also been discussed in the context of the association between antidepressant use and breast cancer (37, 46), and solutions proposed in that context were analogous to ours. Moreover, including never-exposed subjects in the analysis may affect the estimated effect of any quantitative exposure. Furthermore, the problem of multicollinearity concerns not only time-related smoking variables but also other lifestyle, occupational, and environmental risk factors, including (for example) different nutrients that add up to total energy intake (47).

In contrast to some previous sophisticated proposals for dealing with specific aspects of smoking history (36, 48–50), all solutions suggested here can be easily implemented with any standard statistical software. We believe that the use of these methods would result in more comprehensive, robust, and comparable assessments of the impact of different aspects of smoking history on various health outcomes.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada, awarded to Dr. Michal Abrahamowicz. The collection of original data was supported by grants from the Institut de Recherche en Santé et Sécurité du Travail du Québec, the National Health Research and Development Program, and the National Cancer Institute of Canada (NCIC) (Principal Investigator: Dr. Jack Siemiatycki).

Dr. Karen Leffondré was supported by an international research fellowship from La Ligue Nationale Contre le Cancer, Paris, France; a postdoctoral fellowship from the Division of Clinical Epidemiology of the Montreal General Hospital, Québec, Canada; and a grant from the NCIC, under the PREECAN [Programme of Research in Environmental Etiology of Cancer] program. Dr. Michal Abrahamowicz is a Health Scientist of the Canadian Institutes for Health Research. Dr. Jack Siemiatycki holds a Canada Research Chair in Environmental Epidemiology at the Université de Montréal, Montréal, Canada. Dr. Bernard Rachet was supported by an international research fellowship from La Ligue Nationale Contre le Cancer; a postdoctoral fellowship from the Institut National de la Recherche Scientifique, Québec, Canada; and a grant from the NCIC, under the PREECAN program.

REFERENCES

1. Hu FB, Wang B, Chen C, et al. Body mass index and cardiovascular risk factors in a rural Chinese population. Am J Epidemiol 2000;151:88–97.
2. Coello SD, De Léon AC, Ojeda FB, et al. High density lipoprotein cholesterol increases with living altitude. Int J Epidemiol 2000;29:65–70.
3. Roivainen M, Viik-Kajander M, Palosuo T, et al. Infections, inflammation, and the risk of coronary heart disease. Circulation 2000;101:252–7.
4. Meigs JB, Mittleman MA, Nathan DM, et al. Hyperinsulinemia, hyperglycemia, and impaired hemostasis: The Framingham Offspring Study. JAMA 2000;283:221–8.
5. Roger VL, Farkouh ME, Weston SA, et al. Sex differences in evaluation and outcome of unstable angina. JAMA 2000;283: 646–52.
6. Gauderman WJ, Morrison JL. Evidence for age-specific genetic relative risks in lung cancer. Am J Epidemiol 2000;151:41–9.
7. Mori M, Hara M, Wada I, et al. Prospective study of hepatitis B and C viral infections, cigarette smoking, alcohol consumption, and other factors associated with hepatocellular carcinoma risk in Japan. Am J Epidemiol 2000;151:131–9.
8. Luce D, Bugel I, Goldberg P, et al. Environmental exposure to tremolite and respiratory cancer in New Caledonia: a case-control study. Am J Epidemiol 2000;151:259–65.

9. Dikshit RP, Kanhere S. Tobacco habits and risk of lung, oropharyngeal and oral cavity cancer: a population-based case-control study in Bhopal, India. Int J Epidemiol 2000;29:609–14.
10. Jourenkova-Mironova N, Mitrunen K, Bouchardy C, et al. High-activity microsomal epoxide hydrolase genotypes and the risk of oral pharynx and larynx cancers. Cancer Res 2000;60: 534–6.
11. Husgafvel-Pursiainen K, Boffetta P, Kannio A, et al. p53 mutations and exposure to environmental tobacco smoke in a multicenter study on lung cancer. Cancer Res 2000;60:2906–11.
12. Yuan JM, Wang XL, Xiang YB, et al. Non-dietary risk factors for nasopharyngeal carcinoma in Shanghai, China. Int J Cancer 2000;85:364–9.
13. Holmen TL, Barrett-Connor E, Holmen J, et al. Health problems in teenage daily smokers versus nonsmokers, Norway, 1995–1997: The Nord-Trøndelag Health Study. Am J Epidemiol 2000;151:148–55.
14. Brüske-Hohlfeld I, Möhner M, Pohlabeln H, et al. Occupational lung cancer risk for men in Germany: results from a pooled case-control study. Am J Epidemiol 2000;151:384–95.
15. Launoy G, Milan C, Faivre J, et al. Tobacco type and risk of squamous cell cancer of the oesophagus in males: a French multicenter case-control study. Int J Epidemiol 2000;29:36–42.
16. Barbone F, Bovenzi M, Cavallieri F, et al. Cigarette smoking and histologic type of lung cancer in men. Chest 1997;112: 1474–9.
17. Nelson LM, McGuire V, Longstreth WT Jr, et al. Population-based case-control study of amyotrophic lateral sclerosis in western Washington State. I. Cigarette smoking and alcohol consumption. Am J Epidemiol 2000;151:156–63.
18. Kreuzer M, Kreienbrock L, Gerken M et al. Risk factors for lung cancer in young adults. Am J Epidemiol 1998;147:1028–37.
19. Crawford EL, Khuder SA, Durham SJ, et al. Normal bronchial epithelial cell expression of glutathione transferase P1, glutathione transferase M3, and glutathione peroxidase is low in subjects with bronchogenic carcinoma. Cancer Res 2000;60: 1609–18.
20. Lagergren J, Bergström R, Lindgren A, et al. The role of tobacco, snuff and alcohol use in the aetiology of cancer of the oesophagus and gastric cardia. Int J Cancer 2000;85:340–6.
21. McKnight B, Cook LS, Weiss NS. Logistic regression analysis for more than one characteristic of exposure. Am J Epidemiol 1999;149:984–92.
22. Dorfman A, Kimball AW, Friedman LA. Regression modeling of consumption or exposure variables classified by type. Am J Epidemiol 1985;122:1096–107.
23. Siemiatycki J, Krewski D, Franco E, et al. Association between cigarette smoking and each of 21 types of cancer: a multi-site case-control study. Int J Epidemiol 1995;24:504–14.
24. Siemiatycki J. Risk factors for cancer in the workplace. Boca Raton, FL: CRC Press, 1991.
25. Siemiatycki J, Day NE, Fabry J, et al. Discovering carcinogens in the occupational environment: a novel epidemiological approach. J Natl Cancer Inst 1981;66:217–25.
26. International Agency for Research on Cancer. Tobacco smoking. (IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans, vol. 38). Lyon, France: International Agency for Research on Cancer, 1986.
27. Cox DR. Regression models and life tables (with discussion). J R Stat Soc B 1972;34:187–220.
28. Prentice RL, Breslow NE. Retrospective studies and failure time models. Biometrika 1978;65:153–8.
29. Chen K, Lo SH. Case-cohort and case-control analysis with Cox's model. Biometrika 1999;86:755–64.
30. Barlow WE, Ichikawa L, Rosner D, et al. Analysis of case-cohort designs. J Clin Epidemiol 1999;52:1165–72.
31. MathSoft, Inc. S-Plus. Version 4.0. Seattle, WA: MathSoft, Inc, 1997.
32. Barlow WE. Robust variance estimation for the case-cohort design. Biometrics 1994;50:1064–72.
33. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. J Am Stat Assoc 1989;84:1074–8.
34. Akaike H. A new look at the statistical model identification. IEEE Trans Automatic Control 1974;19:716–23.
35. Karlson EW, Lee IM, Manson JE, et al. A retrospective cohort study of cigarette smoking and risk of rheumatoid arthritis in female health professionals. Arthritis Rheum 1999;42:910–17.
36. Hauptmann M, Lubin JH, Rosenberg P, et al. The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer. Stat Med 2000;19:2185–94.
37. Cotterchio M, Kreiger N, Darlington G, et al. Antidepressant medication use and breast cancer risk. Am J Epidemiol 2000; 151:951–7.
38. Hastie TJ, Tibshirani RJ. Generalized additive models. New York, NY: Chapman and Hall, Inc, 1990.
39. Abrahamowicz M, du Berger R, Grover SA. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. Am J Epidemiol 1997;145:714–29.
40. Kleinbaum DG, Kupper L, Muller KE. Applied regression analysis and other multivariable methods. 2nd ed. Boston, MA: PWS-Kent Publishing Company, 1988.
41. Cascorbi I, Henning S, Brockmöller J, et al. Substantially reduced risk of cancer of the aerodigestive tract in subjects with variant –463A of the myeloperoxidase gene. Cancer Res 2000; 60:644–9.
42. Quantin C, Abrahamowicz M, Moreau T, et al. Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. Am J Epidemiol 1999;150:1188–200.
43. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. Epidemiology 1997;8:429–34.
44. Hrubec Z, McLaughlin JK. Former cigarette smoking and mortality among U.S. veterans: a 26-year follow-up, 1954–1980. (Smoking and tobacco control monographs, no. 8). Bethesda, MD: National Cancer Institute, 1997:501–30. (NIH publication no. 97-4213).
45. Blizzard L, Dwyer T. Declining lung cancer mortality of young Australian women despite increased smoking is linked to reduced cigarette 'tar' yields. Br J Cancer 2001;84:392–6.
46. Kelly JP, Rosenberg L, Palmer JR, et al. Risk of breast cancer according to use of antidepressants, phenothiazines, and antihistamines. Am J Epidemiol 1999;150:861–8.
47. Willett WC, Howe GR, Kushi LH. Adjustment for total energy intake in epidemiological studies. Am J Nutr 1997;65(suppl): 1220S–8S.
48. Whittemore AS. Effect of cigarette smoking in epidemiological studies of lung cancer. Stat Med 1988;7:223–38.
49. Mark SD, Robins JM. Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. Stat Med 1993;12: 1605–28.
50. Koehler KJ, McGovern PG. An application of the LFP survival model to smoking cessation data. Stat Med 1990;9:409–21.