Estimating the Proportion of Cases of Lung Cancer Legally Attributable to Smoking: A Novel Approach for Class Actions Against the Tobacco Industry

Jack Siemiatycki, PhD, Igor Karp, MD, PhD, Marie-Pierre Sylvestre, PhD, and Javier Pintos, MD, PhD

Objectives. The plaintiffs' lawyers for a class action suit, which was launched in Quebec on behalf of all patients with lung cancer whose disease was caused by cigarette smoking, asked us to estimate what proportion of lung cancer cases in Quebec, if they hypothetically could be individually evaluated, would satisfy the criterion that it is "more likely than not" that smoking caused the disease.

Methods. The novel methodology we developed is based on the dose–response relationship between smoking and lung cancer, for which we use the pack-years as a measure of smoking, and the distribution of pack-years of smoking among cases.

Results. We estimated that the amount of smoking required to satisfy the "more likely than not" criterion is between 3 and 11 pack-years. More than 90% of the Quebec cases satisfied even the most conservative of these thresholds.

Conclusions. More than 90% of cases of lung cancer in Quebec are legally attributable to smoking. The methodology enhances the ability to conduct class action suits against the tobacco industry. (*Am J Public Health.* 2014;104:e60–e66. doi:10.2105/AJPH.2014.302040)

Legal actions have been launched in different jurisdictions seeking compensation from the tobacco industry for harms done to alleged victims of tobacco-related diseases. If an identified individual launches such an action, that individual has to demonstrate that the tobacco industry shares responsibility for the fact that he or she smoked and that it is more likely than not that smoking caused the disease.

As an alternative, a class action may be launched on behalf of an unnamed group of individuals who, collectively, may be claimed to be victims of tobacco-related disease. In this situation, the same components of proof are required. In jurisdictions where there may be thousands or millions of cases of lung cancer, the feasibility of a class action suit is challenged by the requirement to demonstrate that each claimant meets the "more likely than not" criterion.

This project was initiated in response to a request to one of the authors to participate as an expert witness in a class action undertaken in the Province of Quebec, Canada, on behalf of people with tobacco-related diseases. In Quebec, the law allows for plaintiffs to request a "collective recovery" against a defendant. The law states:

> The court orders collective recovery if the evidence produced enables the establishment with sufficient accuracy of the total amount of the claims of the members; it then determines the amount owed by the debtor even if the identity of each of the members or the exact amount of their claims is not established.[1]

In such a regime, if the court finds for the plaintiffs and orders a collective recovery, it could authorize individuals seeking redress to make their claims against the collective recovery fund, to a court-appointed administrator, through a procedure that would be simpler and less contentious than a court case. The court would stipulate the conditions that a claimant would have to satisfy to secure a share of the collective recovery fund.

We examined the issue of quantifying the number of people on behalf of whom "collective recovery" could be claimed in a class action suit brought by diseased smokers against the tobacco industry.

In this article, we do not deal with the issue of tobacco industry responsibility for smoking behaviors, though a suit would clearly need to address this. Nor do we deal with the algorithm that might be used for disbursing money from the collective recovery fund, though a finding for the plaintiffs would clearly need to address this. Finally, we do not deal with the nature of evidence in a case where the state seeks to recover health care costs from the industry.

METHODS

When an epidemiologist is asked to estimate the number of cases of a given disease that were caused by a given factor in a given population, a typical practice is to compute the attributable fraction in the population, AF_p, and apply this fraction to the total number of cases of the illness in the population. We define AF_p as excess fraction of the lung cancer rate in the entire target population attributable to the risk factor at issue. We can represent AF_p as

$$(1)\quad (R_p - R_u)/R_p,$$

where R_p = rate in the entire target population, and R_u = rate in the hypothetical unexposed subpopulation with the same covariate structure as in the entire target population.

Computational formulae have been developed to estimate AF_p based on the causal rate ratio for the association between the exposure and the disease and the prevalence of exposure either in the entire population or among cases of the disease.[2–4]

But the AF_p, so pertinent in attempts to quantify the public health and financial impacts of a risk factor, is less so in the context of a suit to redress the alleged harms done to individuals. In US jurisprudence, the latter situation is generally associated with the standard of proof that the probability that the risk factor at issue was causal to the individual's disease must be greater than 50% for the court to affirm the

plaintiff's claim.[5–10] The probability that the risk factor was causal to the individual's disease is generally referred to as "probability of causation" (PC).

The concept of probability of causation has been used in determining compensation for workers who developed radiation-related cancers[5,11,12] as well as in other workers' compensation investigations.[13] For a given individual with a disease who had exposure to a given risk factor, and under certain conditions,

$$(2)\quad PC = (RR - 1)/RR,$$

where *RR* denotes the causal rate ratio associated with the risk factor the individual was exposed to, taking into account as far as possible the relevant aspects of the exposure and modifiers of the RR.[13–15] If formula 2 holds, it can be seen that probability of causation is greater than 0.50 when RR is greater than 2.0. Typically, the RR would be estimated from epidemiological studies that quantified the causal association between the risk factor and the disease at issue.

The use of formula 2 for estimating probability of causation has been contested.[14,16–21] Greenland and Robins[18–20] have argued that this formulation of probability of causation assumes that the risk factor acts only to induce the disease among persons in whom it would not otherwise occur, ignoring other possible processes by which a risk factor could contribute to disease occurrence, such as hastening the onset of disease. They further showed that "under commonly used assumptions," formula 2 provides a lower bound for the probability of causation.[16,19] Dawid et al.[21] have similarly asserted that the equality in formula 2 should be replaced by a greater than or equal to symbol.

Notwithstanding its imperfections, we will use formula 2 to estimate probability of causation for a few reasons. First of all, as illustrated by the temporal and geographical variations in lung cancer occurrence, the disease barely exists in populations in which smoking is not practiced, thus providing strong support to the presumption that smoking (predominantly, if not exclusively) serves to induce all-or-none occurrence rather than hastened occurrence of the disease. Second, the RR associated with smoking is so high that most smoking cases are likely to have probability of causation much greater than 0.50, and any modest underestimation of the probability of causation will have little impact on estimating the numbers of individuals who might be eligible for compensation. Third, it has been argued that, notwithstanding the limitations of formula 2, when properly interpreted, it should have a role in adjudicating compensation claims.[15,22,23] Fourth, unless and until more valid approaches are developed and adopted by courts, the RR greater than 2 criterion derived from formula 2 continues to be widely used in tort cases.[9,10]

Whereas AF_p is a measure that pertains to a population, probability of causation pertains to an individual. For each case of lung cancer in the population, there is a corresponding, particular value of probability of causation that depends on the individual's smoking history. If every case could be presented in court, and if for each one it were possible to estimate the probability of causation, then it would be possible to establish for how many of them the probability of causation exceeds 0.50. But in a class action where there may be thousands or millions of diseased people, there may be no practical means to compute the probability of causation for each individual.

Our mandate was to address the following question: What fraction of lung cancer cases diagnosed in Quebec in 1995 through 2006 would satisfy the criterion of probability of causation greater than 0.50? We refer to this quantity as the "legally attributable fraction in the population," and we designate it LAF_p. To better explain the meaning of LAF_p we demonstrate how LAF_p differs from AF_p.

Let N equal the total number of lung cancer cases in the population, and let us consider the division of N into 5 proper subsets. We denote the number of nonsmoking cases as N_A; N_B, N_C, N_D, and N_E represent numbers of smoking cases in subsets distinguished on the basis of 2 characteristics: whether they would have occurred irrespective of smoking, and whether the probability of causation is less than or equal to or greater than 0.50. Table 1 defines the 5 subsets.

The excess cases in the population are represented by $N_D + N_E$ and

$$(3)\quad AF_p = (N_D + N_E)/N.$$

The cases who have PC greater than 0.50 are represented by $N_C + N_E$ and

$$(4)\quad LAF_p = (N_C + N_E)/N.$$

A graphical illustration of this notion is shown in Appendix A (available as a supplement to this article at http://www.ajph.org).

Estimating the Legally Attributable Fraction in the Population

There are 2 distinct steps in estimating LAF_p: (1) based on the dose–response relationship between smoking and lung cancer, estimate the critical amount of smoking, *X*, that corresponds to the probability of causation of 0.50 or, equivalently, to the RR of 2.0; (2) estimate what fraction of cases of lung cancer in the target population have a history of smoking in the amount that exceeds *X*. This fraction estimates LAF_p.

One overarching constraint on the choice of data sources for these 2 steps is that whatever smoking metric is used for step 1 has to be available for step 2, and vice versa. We chose the pack-years variable, a metric that satisfies this constraint, and that is strongly associated with lung cancer. (See Appendix B for further justification for this choice, available as a supplement to this article at http://www.ajph.org.).

Step I—estimating the critical amount X. There are several conceivable sources of information

TABLE 1—Defining Characteristics of 5 Proper Subsets of Lung Cancer Cases

Subgroup	Smoker	Cases Would Have Occurred Irrespective of Smoking	Probability of Causation	No. in Subgroup
A	No	Yes	0.00	N_A
B	Yes	Yes	≤ 0.50	N_B
C	Yes	Yes	> 0.50	N_C
D	Yes	No	≤ 0.50	N_D
E	Yes	No	> 0.50	N_E

on the dose–response relationship between pack-years and lung cancer. Dose–response estimates might be available from a local study in the target area or from a published meta-analysis of studies conducted in multiple populations. For reasons given in Appendix B, we recommend using results from an appropriate meta-analysis rather than from any single study. When we were conducting our analysis, we found no meta-analysis of the pack-year–lung cancer association, so we conducted our own.

Individual studies typically reported results in categories of pack-years with never smokers as the reference. Estimating the dose–response relationship in the form of a smooth parametric curve requires decisions to be made about several analytic features, including the treatment of results that are reported in categories of pack-years and the functional form of the relationship. For the present purpose, we made the following choices, with justifications shown in Appendix B. In computing the slope for each study, we used the midpoint of the category as the independent variable in the regression analysis. But the top category in each study was open-ended (e.g., > 50 pack-years), which posed a problem of how to represent the smoking variable for that category. We dealt with it by excluding the top open-ended pack-year category from each study. Among possible functional forms, we chose both a linear model and a log-linear (sublinear) model.

The models can be expressed as follows:

$$(5)\quad \text{Linear model}: \mathrm{RR} = 1.0 + \beta_{\mathrm{RR}} * \mathrm{PY};$$

$$(6)\quad \text{Log-linear model}: \ln\mathrm{RR} = \beta_{\ln\mathrm{RR}} * \mathrm{PY};$$

where PY = lifetime cumulative pack-years of smoking, β_{RR} = slope of the linear relationship between *PY* and *RR*, and β_{lnRR} = slope of the linear relationship between *PY* and *lnRR*.

We used both the classic general linear modeling and its modified version, suggested by Greenland and Longnecker,[24] to take account of the correlated estimates of RR from categorical analyses. See Appendix B for detailed explanations of these various choices, as well as for the details of the meta-analyses to estimate slopes.

Once a meta-slope, $\hat{\beta}_{RR}$ or $\hat{\beta}_{lnRR}$, has been determined, the critical point on the pack-year scale at which the RR estimate equals 2.0 can be derived as follows.

For the linear model, by setting RR to 2.0,

$$(7)\quad \mathrm{X} = 1.0 \Big/ \hat{\beta}_{RR}.$$

For the log-linear model, by setting RR to 2.0,

$$(8)\quad \mathrm{X} = \ln(2.0) \Big/ \hat{\beta}_{lnRR}.$$

Step II—Estimating the fraction of cases whose cumulative pack-years exceeds X. Unlike the slope quantifying the dose–response relationship between smoking and cancer, which can generally be presumed to be a fairly invariant in time and place, the distribution of smoking pack-years cannot be presumed to be so. Thus it would generally be necessary to base this step on data collected in the target population during the relevant time period. The method for accomplishing this step depends on the nature of smoking history data available locally.

If the investigator has access to smoking history data from a representative sample of lung cancer cases in the target population, and if these data can be portrayed in pack-years, then it is quite straightforward to estimate the fraction of cases whose smoking history puts them above *X*, and thus to estimate $\mathrm{LAF_p}$. We had access to such a database.

Our team had conducted a case–control study of lung cancer etiology in the metropolitan area of Montreal, which comprises about 35% of the population of the Province of Quebec. The fieldwork was conducted in 1996 through 2000 and involved data collection regarding 1165 patients with lung cancer (735 men and 430 women).[25] The study, aiming to explore associations between lung cancer and a host of occupational and lifestyle factors, included collection of lifetime smoking histories. The main use of this study for the present purpose was to provide data for estimating the distribution of pack-years of smoking among lung cancer cases. We assumed that the case series in this study provided a reasonably valid estimate of smoking distribution for all cases in the target population, Quebec Province. We refer to this study as Montreal Study II to distinguish it from Montreal Study I,[26] a case–control study conducted earlier and included in our meta-analysis, because it satisfied the inclusion criteria outlined in Appendix B.

A more challenging problem arises if there is no reliable local set of data on the smoking profile of lung cancer cases, and available data pertain only to the smoking profile of the target population.

A national smoking survey was conducted by Statistics Canada.[27] Each participant was asked a series of questions that allow the estimation of pack-years. We obtained descriptive tables showing smoking distributions for people resident in Quebec, by age and gender, based on the survey conducted in 2002 and 2003.

Two indirect methods for step II. We will outline here 2 indirect methods that can be used, one approximate and very simple (indirect method I), and a second, more exact but more complex (indirect method II).

Indirect method I. Indirect method I requires only estimates of the rate ratio of lung cancer among smokers versus never smokers ($\mathrm{RR_{ever}}$) and of the distribution of pack-years in the general population. Let $\mathrm{P_s}$ = proportion of the target population who were ever smokers; and $\mathrm{Q_s}$ = proportion of lung cancer cases in the target population who were ever smokers.

As RR_{ever} may be presented as follows,

$$(9)\quad \mathrm{RR_{ever}} = \mathrm{Q_s}(1 - \mathrm{P_s})/\mathrm{P_s}(1 - \mathrm{Q_s}),$$

then Q_s may be derived from P_s and RR_{ever} as follows:

$$(10)\quad \mathrm{Q_s} = \mathrm{P_s}\,\mathrm{RR_{ever}}/(1 + \mathrm{P_s}\,\mathrm{RR_{ever}} - \mathrm{P_s}).$$

Let $\mathrm{P_x}$ = proportion of ever smokers in the target population whose smoking history puts them above the critical value *X*; and $\mathrm{Q_x}$ = proportion of ever smoking patients with lung cancer whose smoking history puts them above the critical value *X*.

The proportion of cases with history of smoking such that the number of pack-years of smoking surpassed *X* can be computed as

$$(11)\quad \mathrm{LAF_p} = \mathrm{Q_s} \times \mathrm{Q_x}.$$

P_x is readily available by examining the frequency distribution of pack-years in the population survey. A very conservative estimate of Q_x can be derived by assuming that the distribution of pack-years is similar among

smoking cases as it is among smokers in the general population. Namely,

$$(12)\quad Q_x = P_x.$$

This is a conservative assumption because it is quite certain that the diseased smokers were, on average, heavier smokers than the nondiseased smokers. Thus, by assuming $Q_x = P_x$, we certainly underestimate LAF_p.

Indirect method II. Indirect method II is conceptually similar to indirect method I, but instead of using the simple dichotomy of never or ever smoker and the simplifying assumption of formula 12, it uses multicategory smoking distribution and the RR_i pertaining to each of those smoking categories. That is, we first take the i smoking categories that are provided by the general population smoking survey. For each category we derive an estimate of the RR_i for smokers in that category by using the RR function derived from the meta-analysis of dose–response relationships. Once we have the RR_i and the proportion of the general population that falls into each smoking category, we can then configure the problem in such a way as to use formula 10 to deduce the proportion of lung cancer cases that fall into each smoking category. Namely, we can define i subsets of the population, where the i^{th} subset includes never smokers plus all individuals in the i^{th} smoking category. The never smokers are common to all subsets.

For instance, if one of the smoking categories is 20 to 25 pack-years, then we can define a subset of the population that includes all never smokers plus all participants with 20 to 25 pack-years. We can then proceed to do an analysis in that subset that is analogous to the binary analysis we did under method I, whereby we deduce the proportion of smokers among cases from the proportion of smokers in the population via formula 10. Such an analysis would lead to an estimate of the proportion of cases who are smokers in the 20 to 25 pack-year category, but with the denominator being the number of cases in this i^{th} subset of the population. If we then want to estimate the proportion of all cases in the population who are in the 20 to 25 pack-year category, we need to calibrate it to the denominator of the cases in the population. Once we have the distribution of smokers among cases in this way, and knowing the RR function, we can estimate what proportion of cases fall into smoking categories that confer RR greater than 2.0. We have described the principles here. The operational algebraic derivation can be found in Appendix D (available as a supplement to this article at http://www.ajph.org).

RESULTS

Table 2 shows the slope estimates from the meta-analyses described in Appendix B, based on linear and log-linear models, as well as the corresponding critical values, X, as derived from formulas 7 and 8. Using the linear model, we estimated the meta-slope to be 0.270 (95% confidence interval [CI] = 0.225, 0.314) among men and 0.363 (95% CI = 0.238, 0.487) among women. For men, the 3 point estimates of X were 3.7 PY, 9.0 PY, and 11.2 PY, respectively. For women, the 3 point estimates of X were 2.8 PY, 8.4 PY, and 9.0 PY, respectively.

Many of the meta-estimates involved statistically significant heterogeneity, but the resulting X value under a linear model was below 12 PY even in the study with the lowest slope.

Table 3 shows estimates of the prevalence of smoking and the pack-years distribution among lung cancer cases in Montreal Study II and among Quebec residents from the Statistics Canada survey. Because the Statistics Canada survey was based on a representative sample of adults in Quebec, we weighted the age-specific patterns to the age distribution of patients with lung cancer.

The direct information based on the Montreal case series shows that the proportion of cases with history of regular smoking, Q_s, was 0.976 for men and 0.931 for women.

Had Q_s been unavailable from a case series, then Q_s could have been derived from RR_{ever} and P_s. As shown in Appendix C (available as a supplement to this article at http://www.ajph.org), for men, the meta-RR_{ever} was 8.8 (95% CI = 7.5, 10.4), whereas for women the meta-RR_{ever} was 7.8 (95% CI = 5.9, 10.3). Based on the Statistics Canada survey, the age-adjusted proportion of ever-smokers in the population, P_s, was 0.738 among men and 0.425 among

TABLE 2—Estimates of the Slope Parameter Quantifying the Dose–Response Association Between Pack-Years and Lung Cancer, Based on 3 Sets of Meta-analyses, and the Corresponding Estimates of the Critical Pack-Year Amount *X*, by Gender

Design		Meta-slope		Critical Value X[a]
Model[b]	Slope Estimation Method[c]	No. Data Sets[d]	$\hat{\beta}$ (95% CI)	X (Range Based on 95% CI of $\hat{\beta}$)
Men				
Linear	Conventional	15	0.270 (0.225, 0.314)	3.7 (3.2, 4.4)
Log-linear	Conventional	15	0.077 (0.051, 0.104)	9.0 (6.7, 13.6)
Log-linear	G-L	9	0.062 (0.044, 0.079)	11.2 (8.7, 15.8)
Women				
Linear	Conventional	14	0.363 (0.238, 0.487)	2.8 (2.1, 4.2)
Log-linear	Conventional	14	0.082 (0.059, 0.105)	8.4 (6.6, 11.7)
Log-linear	G-L	11	0.077 (0.061, 0.093)	9.0 (7.4, 11.4)

Note. CI = confidence interval; G-L = Greenland–Longnecker; RR = rate ratio.
[a]Number of pack-years at which RR = 2.0, based on the slope estimate. For linear model, $X = 1/\hat{\beta}$. For log-linear model, $X = (\ln 2)/\hat{\beta}$.
[b]Linear: $RR = 1.0 + \beta * PY$; log-linear: $\ln RR = \beta * PY$, where PY = pack-years and RR = rate ratio.
[c]This refers to the method for estimating the slope of each study. The "conventional" method simply takes the midpoint of each category as the value of the independent variable representing the "dose," and the RR for each category as the dependant variable. The G-L approach[24] adjusts for the correlated nature of subcategory estimates. This method assumes a log-linear model. It requires information that was not available in all studies, and thus could not include results from all the studies.
[d]The number of data sets exceeds the number of studies because some studies published results on multiple subpopulations, which we treated as separate data sets for the meta-analysis.

TABLE 3—Estimates of Prevalence of Ever-Smoking and Pack-Year Distribution Among Lung Cancer Cases in Montreal and Among the General Population of Quebec Province, by Gender

Smoking Index	Montreal Study Cases,[a] %		Estimates for General Population,[b] %	
	Men	Women	Men	Women
Ever smoker (among entire population)	97.6	93.1	73.8	42.5
Pack-years (among ever smokers)				
≤ 5	0.4	1.1	6.0	13.2
> 5 to ≤ 10	1.0	1.3	8.3	11.2
> 10 to ≤ 15	0.7	1.1	8.6	9.1
> 15 to ≤ 20	1.1	2.6	7.4	8.1
> 20 to ≤ 25	1.4	2.4	9.3	10.3
> 25 to ≤ 37.5	6.8	15.2	18.1	15.5
> 37.5 to ≤ 50	13.0	22.2	13.7	14.9
> 50	75.1	47.4	28.8	17.6

[a]The first 2 columns come directly from the case–control Montreal Study II.
[b]The last 2 columns describe the smoking profiles in the general population of Quebec as ascertained in a Statistics Canada survey and age-standardized according to the age distribution of lung cancer cases.

women. Using formula 10, we derived estimates of Q_s equal to 0.961 among men and 0.852 among women. The directly and indirectly derived estimates of Q_s were almost identical for men, and only slightly discordant for women.

National smoking surveys are usually based on large samples, which entail little statistical variability. In our case, the estimate of P_s from the Statistics Canada survey for the population of Quebec older than 45 years was based on a total sample of about 6000 men and 8000 women. With such large sample sizes, the values of Q_s corresponding to the lower and upper 95% CIs of the RR_{ever} estimate are very tight indeed: 0.955 to 0.967 for men and 0.813 to 0.884 for women. Such small variability of estimated Q_s values will have very little impact on the estimates of LAF_p. As a consequence, we will treat the values of Q_s derived from the population survey as quasi-constants.

As mentioned previously, indirect method I requires assuming that the pack-year distribution among smoking cases is the same as that among smokers in the general population. Table 3 shows that this assumption is incorrect and that it will indeed lead to underestimation of the LAF_p. The extent of underestimation may be deduced from some of our empirical findings, which follow.

Table 4 shows estimates of LAF_p under 9 scenarios, combining 3 meta-analysis methods for estimating slopes, and 3 methods for estimating the smoking distribution among lung cancer cases. Furthermore, we show possible ranges of these scenario estimates by using the 95% confidence limits of the meta-slope estimates. For men, 7 of the 9 point estimates of LAF_p are above 0.90, and 24 of the 27 estimates in the table are above 0.80. The lowest values, as expected, are obtained when we used a log-linear model and when we used the very conservative indirect method I to estimate smoking distribution among lung cancer cases. The values of LAF_p are over 0.90 in all the scenarios when we use the smoking information from the Montreal case series because there are very few lung cancer cases with low pack-year amounts (Table 3). Indirect method II produces higher values of LAF_p than indirect method I. Within each scenario, the range of values embodied in the 95% CI for X estimates has very little impact on the range of the LAF_p estimates.

For women, the pattern is similar to that for men, though the values of LAF_p estimates are about 0.05 to 0.10 lower among women than among men for the same scenarios. Seven of the 9 point estimates of LAF_p are above 0.79 and 22 of the 27 estimates in Table 4 are above 0.70.

Using Miettinen's formula[3] with RR_{ever} equal to 8.8 for men and 7.8 for women, and frequency of smoking derived from the Montreal case series, 0.976 for men and 0.931 for women, we estimate AF_p is 0.865 among men and 0.812 among women.

The numbers of lung cases attributable and legally attributable to smoking may be estimated by multiplying the numbers of cases in the target population during the time period at issue by the values of AF_p and LAF_p, respectively. According to data obtained from the government-run Registre des Tumeurs du Québec (e-mail communication, Rabia Louchini, September 13, 2013), for the 12-year period from 1995 to 2006, the average annual number of incident cases of lung cancer was about 3800 in men and 2400 in women. When we combine data for men and women, we estimate that, each year, approximately 5200 cases of lung cancer were attributable to smoking, and approximately 5700 cases would have satisfied the "more likely than not" criterion.

DISCUSSION

In a class action against the tobacco industry, the defense might argue, even if the "general causation" between smoking and lung cancer is acknowledged, plaintiffs must demonstrate, one by one, that each member of the class has a demonstrable specific causal link to smoking. As there are several known causes of lung cancer, and as there is no biological basis for determining whether any given one of them actually was causal to any given individual's cancer, the defense might argue further that only by individually examining each individual's personal history can a court determine whether it is "more likely than not" that the individual's cancer was caused by smoking, and, after all the individual files are evaluated, how many individuals satisfy the criterion of probability of causation greater than 0.50. Although the method we describe in this article does not permit the certain identification of individuals whose cancer was caused by smoking, it allows the estimation of the number of cases that would satisfy the criterion of probability of causation greater than 0.50.

The LAF_p is a function of the distribution of probability of causation. Although the estimation of probability of causation by means of formula 2 has been contentious,[14,16–20]

TABLE 4—Estimates of the Legally Attributable Fraction of Cases in the Population for Smoking and Lung Cancer, Based on 3 Sets of Meta-analyses and 3 Methods of Estimating Smoking Distributions Among Lung Cancer Cases, by Gender

Meta-Analysis Design		Meta-Analysis Results		Smoking Distribution			
Model[a]	Slope Method[b]	X[c]	Range[d]	Source[e]	Method[f]	LAF_p[g]	Range
Men							
Linear	Conventional	3.7	3.2–4.4	Case series	Direct	0.971	0.971–0.971
Linear	Conventional	3.7	3.2–4.4	Population	Indirect I	0.915	0.903–0.915
Linear	Conventional	3.7	3.2–4.4	Population	Indirect II	0.978	0.975–0.978
Log-linear	Conventional	9.0	6.7–13.6	Case series	Direct	0.965	0.959–0.967
Log-linear	Conventional	9.0	6.7–13.6	Population	Indirect I	0.840	0.758–0.872
Log-linear	Conventional	9.0	6.7–13.6	Population	Indirect II	0.953	0.907–0.966
Log-linear	G-L	11.2	8.7–15.8	Case series	Direct	0.961	0.954–0.965
Log-linear	G-L	11.2	8.7–15.8	Population	Indirect I	0.791	0.727–0.840
Log-linear	G-L	11.2	8.7–15.8	Population	Indirect II	0.927	0.884–0.953
Women							
Linear	Conventional	2.8	2.1–4.2	Case series	Direct	0.920	0.920–0.920
Linear	Conventional	2.8	2.1–4.2	Population	Indirect I	0.794	0.739–0.794
Linear	Conventional	2.8	2.1–4.2	Population	Indirect II	0.923	0.909–0.923
Log-linear	Conventional	8.4	6.6–11.7	Case series	Direct	0.907	0.905–0.916
Log-linear	Conventional	8.4	6.6–11.7	Population	Indirect I	0.663	0.613–0.701
Log-linear	Conventional	8.4	6.6–11.7	Population	Indirect II	0.877	0.845–0.895
Log-linear	G-L	9.0	7.4–11.4	Case series	Direct	0.907	0.905–0.914
Log-linear	G-L	9.0	7.4–11.4	Population	Indirect I	0.663	0.613–0.682
Log-linear	G-L	9.0	7.4–11.4	Population	Indirect II	0.877	0.845–0.886

Notes. G-L = Greenland–Longnecker; RR = rate ratio.

[a]Linear: $RR = 1.0 + \beta * PY$; log-linear: $\ln RR = \beta * PY$.

[b]This refers to the method for estimating the slope of each study. The "conventional" approach simply takes the midpoint of each category as the value of the independent variable representing the "dose," and the RR for each category as the dependent variable. The G-L approach adjusts for the correlated nature of subcategory estimates. This method assumes a log-linear model. It requires information that was not available in all studies, and thus could not include results from all the studies.

[c]Number of pack-years at which RR = 2.0, based on the slope estimate. For linear model, $X = 1/\beta$. For log-linear model, $X = (\ln 2)/\beta$.

[d]The range of X is copied from Table 2.

[e]Case series is from the Montreal II study. Population data come from the Statistics Canada survey.

[f]Direct: When using case series, we simply compute the proportion of cases whose pack-year total is above *X*. When using population data, we implement either indirect method I or indirect method II.

[g]LAF_p = the fraction of cases whose pack-year values would put them above the PC = 0.50 threshold. Range of LAF_p is based on range of *X*.

this remains the standard technique and it provides a lower bound for the parameter of interest.[19–22] The real problem with the possible underestimation of probability of causation occurs with risk factors that are associated with much smaller RRs than smoking. In such a situation the underestimation of probability of causation for exposed cases by formula 2 might result in erroneous classification of a substantial fraction of these cases as having probability of causation below 0.50 (even if their true probability of causation may be above 0.50).

Furthermore, when the RR at the highest levels of exposure to some risk factor is less than 2.0, no exposed case would be able to demonstrate that the probability of causation was greater than 0.50, even though some of the exposed cases would not have occurred but for the exposure at issue. For these reasons, it cannot be automatically assumed that the methods outlined in this article would provide approximately valid estimates of LAF_p in the case of a risk factor with a much lower RR than that characterizing the smoking–lung cancer association.

To provide options to epidemiologists who might be called upon to present their expert opinion in such tobacco cases, we have shown a variety of methods that could be used under different circumstances. To ensure that the LAF_p estimates are not exaggerated, we have systematically adopted conservative assumptions and methods. For instance, in selecting models for the dose–response relationship, we chose linear and sublinear (log-linear) models, though we may just as well have chosen a supralinear model, which gives higher risk estimates than the linear model at low exposure levels, and thus leads to even higher LAF_p estimates than the linear model. Similarly, the presentation of indirect method I, despite providing an underestimate of LAF_p, is useful because it is simple to implement and can be explained to lay audience. It provides an unequivocal floor to the estimate of LAF_p.

It should be noted that it cannot be assumed that LAF_p will always be greater or always be smaller than AF_p. As shown in Appendix A, the relative magnitudes of the 2 parameters will depend on the location of *X*, the slope of the dose–response line, and the distribution of pack-years in the population.

Two types of input are needed to implement the estimation of LAF_p for smoking: estimates quantifying the association between smoking and lung cancer and data on the distribution of smoking among lung cancer cases in the target population.

Irrespective of the methods and models, and including the wide ranges of 95% confidence intervals, the estimate of the critical amount of smoking required to induce an RR equal to 2.0 was in the range of 2 to 16 pack-years. The seemingly wide variation in critical amount has very little impact on the estimate of LAF_p, as relatively few smokers accumulate low amounts of pack-years. Only 3% of lung cancer cases who were smokers in Montreal smoked less than 20 pack-years. Thus, whether we calculate the critical amount *X* to be 2 pack-years or 16 pack-years, it really has only a slight impact on the estimate of LAF_p. Irrespective of the methods and smoking data used, the estimated LAF_p was greater than 0.80 for men and greater than 0.66 for women. The best estimates of LAF_p, based on the smoking profiles of Montreal II case series,

are greater than 0.96 among men and greater than 0.91 among women.

It may appear that our estimates of critical amount of pack-years of exposure represent trivial exposure. This is not so; each pack-year represents about 7300 cigarettes.

In Quebec, where there were about 6200 newly diagnosed cases of lung cancer per year from 1995 to 2006, this translates to a total of about 5700 cases per year that would satisfy the criterion of probability of causation greater than 0.50. For other populations with similar age distributions and smoking profiles as Quebec, we would expect similar values of LAF_p. For instance, if we assume a similar smoking profile in the United States as in Quebec, then, of the approximately 225 000 incident cases per year (116 000 men and 109 000 women in 2012), the numbers of legally attributable cases in the United States would be approximately 210 000 per year. ■

About the Authors

Jack Siemiatycki, Igor Karp, and Marie-Pierre Sylvestre are with the Department of Social and Preventive Medicine, School of Public Health, University of Montreal, Montreal, Quebec, and Health Risks Division, Centre de Recherche du CHUM, University of Montreal. Javier Pintos is with Health Risks Division, Centre de Recherche du CHUM, University of Montreal.

Correspondence should be sent to Jack Siemiatycki, 850, Rue St-Denis, Room S02-422, Montreal (Quebec), H2X 0A9, Canada (e-mail: j.siemiatycki@umontreal.ca). *Reprints can be ordered at http://www.ajph.org by clicking the "Reprints/Eprints" link.*

This article was accepted April 12, 2014.

Contributors

J. Siemiatycki defined the problem, conceptualized the methodology, and supervised the implementation of the methodology. He led the interpretation of the results and the drafting of the article. I. Karp, M. P. Sylvestre, and J. Pintos contributed to the development of the methodology, to the data analysis, to the interpretation of results, and to the drafting of the article.

Acknowledgments

J. Siemiatycki holds the Guzzo–Cancer Research Society Chair in Environment and Cancer. I. Karp holds a salary award from the Canadian Institutes for Health Research. The Montreal II case–control study of lung cancer was funded by Health Canada and the Canadian Institutes for Health Research.

We gratefully acknowledge that access to the Statistics Canada Canadian Community Health Survey was provided by the Centre Interuniversitaire Québécois de Statistiques Sociales. We are grateful to the director of the Registre Québécois du Cancer for data on the annual numbers of incident cancer cases. We are grateful for the critical comments on an earlier draft of this article on the part of Ben Armstrong (London School of Hygiene and Tropical Medicine), Karen Leffondré (Université de Bordeaux), Jay Lubin (US National Cancer Institute), Jamie Robins (Harvard University), Kyle Steenland (Emory University), Aiden Talai (University of Toronto), and Duncan C. Thomas (University of Southern California). The following colleagues contributed to the meta-analyses reported in the appendices: Aihua Liu, Aude Lacourt, Sally Campbell, and Lesley Richardson. Jean-Francois Sauvé helped with the figures in the appendices.

Note. J. Siemiatycki has been an expert witness in a class action suit on behalf of plaintiffs against the tobacco industry in the Province of Quebec, Canada.

Human Participant Protection

This project was conducted without original data collection from human participants. There was some theoretical development of methods and the methods developed were applied to data that were abstracted from epidemiological results previously published by many authors in many journals. Some of the data used came from our own studies in Montreal. For these studies, ethical approval was obtained from all participating institutions, and all participants provided informed consent.

References

1. Article 1031, Province of Quebec Code of Civil Procedure, RSQ, c-25, art 1031. Available at: http://www.canlii.org. Accessed May 19, 2014.

2. Levin ML. The occurrence of lung cancer in man. *Acta Unio Int Contra Cancrum.* 1953;9(3):531–541.

3. Miettinen OS. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am J Epidemiol.* 1974;99(5):325–332.

4. Steenland K, Armstrong B. An overview of methods for calculating the burden of disease due to specific risk factors. *Epidemiology.* 2006;17(5):512–519.

5. Lagakos SW, Mosteller F. Assigned shares in compensation for radiation-related cancers. *Risk Anal.* 1986;6(3):345–357.

6. Black B, Lilienfeld DE. Epidemiologic proof in toxic tort litigation. *Fordham Law Rev.* 1984;52(5):732–785.

7. Cox LA. Probability of causation and the attributable risk. *Risk Anal.* 1984;4(3):221–230.

8. Bailey LA, Gordis L, Green M. *Reference Guide on Epidemiology. Reference Manual on Scientific Evidence.* Washington, DC: Federal Judicial Center; 1994:168–169.

9. Miller C. Causation in personal injury: legal or epidemiological common sense? *Leg Stud.* 2006;26 (4):544–569.

10. Mengersen K, Moynihan SA, Tweedie RL. Causality and association: the statistical and legal approaches. *Stat Sci.* 2007;22(2):227–254.

11. Hatch OG. The Radiogenic Cancer Compensation Act. *Congr Rec.* 1983:128.

12. Report of the National Institutes of Health ad hoc working group to develop radioepidemiological tables. Washington, DC: National Institutes of Health; 1985.

13. Armstrong B, Theriault G. Compensating lung cancer patients occupationally exposed to coal tar pitch volatiles. *Occup Environ Med.* 1996;53(3):160–167.

14. Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol.* 1988;128(6):1185–1197.

15. Thomas DC. Probability of causation and compensation. In: *Statistical Methods in Environmental Epidemiology.* Oxford, England: Oxford University Press; 2009:337–350.

16. Robins J, Greenland S. The probability of causation under a stochastic model for individual risk. *Biometrics.* 1989;45(4):1125–1138.

17. Parascandola M. What is wrong with the probability of causation? *Jurimetrics.* 1998;39(Fall):29–44.

18. Greenland S. Relation of probability of causation to relative risk and doubling dose: a methodologic error that has become a social problem. *Am J Public Health.* 1999; 89(8):1166–1169.

19. Greenland S, Robins JM. Epidemiology, justice, and the probability of causation. *Jurimetrics.* 2000;40:321–340.

20. Robins J. Should compensation schemes be based on the probability of causation or expected years of life lost? *J Law Policy.* 2004;12:537–548.

21. Dawid AP, Faigman DL, Fienberg SE. Fitting science into legal contexts: assessing effects of causes or causes of effects. *Sociol Methods Res.* 2013;Epub ahead of print.

22. Thomas DC. Resolved: the probability of causation can be used in an equitable manner to resolve radiation tort claims and design compensation schemes. *Radiat Res.* 2000;154(6):717–718.

23. A review of the draft report of the NCI-CDC Working Group to Revise the 1985 Radioepidemiological Tables. Washington, DC: Committee on an Assessment of Centers for Disease Control and Prevention Radiation Studies from DOE Contractor Sites; 2000: 23–25.

24. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose–response data, with applications to meta-analysis. *Am J Epidemiol.* 1992;135 (11):1301–1309.

25. Pintos J, Parent ME, Richardson L, Siemiatycki J. Occupational exposure to diesel engine emissions and risk of lung cancer: evidence from two case–control studies in Montreal, Canada. *Occup Environ Med.* 2012;69(11):787–792.

26. Siemiatycki J, Krewski D, Franco E, Kaiserman M. Associations between cigarette smoking and each of 21 types of cancer: a multi-site case–control study. *Int J Epidemiol.* 1995;24(3):504–514.

27. Statistics Canada. Canadian Community Health Survey (CCHS) - Cycle 1.1. Available at: http://www.statcan.gc.ca/start-debut-eng.html. Accessed September 13, 2012.