Commentary

Synthesizing the Lifetime History of Smoking

Jack Siemiatycki

Centre hospitalier de l’Université de Montréal, University of Montreal, Montreal, Quebec, Canada

Over 50 years have elapsed since a series of remarkable epidemiologic studies highlighted the critical role that cigarette smoking plays in the etiology of lung cancer (1-5). The recent passing of Sir Richard Doll was a poignant reminder of the importance of that research, both for our understanding of cancer etiology and as a portent of the development of modern epidemiology. It can be argued that in terms of its potential public health effect, this cluster of studies has been the most important research yet carried out on cancer, including clinical, basic, and epidemiologic research.

Notwithstanding the obvious fact that the most important priority in this area is to reduce or eliminate smoking, there remain many outstanding scientific questions about the smoking-lung cancer relationship. One question that has come to the fore (6, 7) and is addressed in the article in this issue by Thurston et al. (8) is how to synthesize the various dimensions of smoking information (age started, average amount, duration, gaps, time since quitting, etc.) to better encapsulate the smoking history for statistical analyses.

The most common approach has been to compute a simple unweighed index of cumulative exposure, known as the pack-years (or cigarette-years) variable. This simple and widely used index has intuitive appeal. But increasingly, questions have been raised as to whether this parameterization is optimal.

Whereas smoking represents an important context in which to discuss this methodologic issue, the problem applies to most epidemiologic variables. Apart from our genetic endowment, there are not many human characteristics that do not embody a time- and dose-related aspect. For many variables, the time-related aspect is often ignored. For example, whereas breastfeeding is of different duration and different intensity for different babies, and social class can vary through life, these variables are usually treated as time-invariant variables. How to create variables for analysis of time-varying exposures is an important and pervasive problem. Addressing the issue for smoking and lung cancer is important not only because of its face value, but also because the approaches that are used to deal with this issue in this context can provide a template for how to approach it for other exposure-disease relationships.

Let us first consider why a composite variable of exposure is needed. Table 1 lists some potential uses and provides brief comments on their relevance to the smoking-lung cancer area. Because of the strength of the association between smoking and lung cancer, almost any index of exposure would have sufficed to show the existence of an association. In any case, we no longer need to establish that smoking causes lung cancer. However, we do need indices that can be used to assess other putative risk factors for lung cancer and to assess interactions between smoking and other factors. So long as we are interested in the potential role of radon, air pollution, diet, alcohol, occupational chemicals, and many other factors, we need to be attentive to the optimal way of parameterizing a smoking history. Fitting different smoking indices to lung cancer data sets may provide insights into the mechanisms of carcinogenesis. For example, the multistage model of carcinogenesis was in part built on results from studies of the smoking-lung cancer relationship (9). Although the mechanisms thereby inferred have not been substantiated in detail, they have given rise to fruitful avenues of research (10). Another example is from analyses of the interaction between smoking and radon that have led to some hypotheses about possible mechanisms of action (11, 12). An innovative use of models for smoking and lung cancer comes from a clinical research application. Bach et al. (13) have shown how such a model can be used in evaluating lung cancer screening or chemoprevention. That is, such a model can be used to predict risk for study subjects and thereby complement or even substitute for contemporaneous randomized controls.

Although several purposes can be served by the creation of synthetic smoking indices, it is unlikely that a single parameterization is optimal for all purposes. For instance, to evaluate interaction with another variable, there may be an advantage to synthesizing the smoking history into a single variable. The interpretation of interactions with multiple smoking variables could be complicated. On the other hand, if the smoking history is treated as a confounder, or if it is used as a marker of risk for clinical studies, then there is little cost in using more than one variable to describe the smoking history. Investigators who are exploring methods of modeling smoking should explicitly indicate the uses to which the models might apply.

The statistical methods used are obviously constrained first of all by the data that are collected. Different studies have collected different aspects of lifetime smoking history, ranging from the simplest binary smoker/non-smoker variable to information on age started, duration, gaps in smoking, amounts smoked at different ages, quitting, etc. In addition, information might be collected on inhalation and on types of cigarettes used. In the future, we may see collection of more detailed information on smoking amounts at different ages and we may see increasing collection of biomarkers of smoking. Any advice on which smoking indices to use is therefore conditional on the type of data collected, and it will need to be reviewed when new types of data are collected. However, this is a two-way street. Development of robust indices of exposure may guide epidemiologists in designing their data collection protocols.

The article by Thurston et al. in this issue (8) addressed the use of smoking-cancer models for the purpose of investigating the interactions between smoking and a number of genes (14, 15). In their comparison of four different parameterizations of smoking history, they found that a metric called logcig-years is more nearly linearly related to the log relative risk of lung cancer and less sensitive to the incorporation of other aspects of smoking history (smoking status or years since quitting) than three alternative metrics: pack-years, square root of pack-years, and including intensity (transformed) and duration as...
two separate variables. They also fit their data to the well-known Doll and Peto model (9) and found a rather poor fit. (It is worth noting that although the Doll and Peto model for absolute risk involves a dependence on the square of smoking intensity and the 4.5 power of duration, the excess relative risk is obtained by dividing this quantity by the baseline risk as a function of attained age, leading to a complex dependency on duration, age at starting, and time since quitting exposure.) Thurston et al. used their findings and a rationalization based on the supposed relationship among smoking, DNA adducts, and lung cancer as a basis for recommending their metric. Every mathematical model is a simplification that is more or less valid and more or less useful. Their metric does have some desirable qualities, including parsimony and approximate linearity, at least in their data set.

Their analysis raises several issues about how to find a way to model the smoking history. The first, alluded to above, is that the purpose of the model needs to be clear. In their case, it is for the purpose of testing interactions; thus, it is preferable to have a single smoking variable. Second is the issue of what should be the role of biological criteria for motivating and justifying a model. It seems self-evident that models should be built on our understanding of biological and physiologic processes. But when our understanding of those processes is minimal, it may be an inefficient and invalid diversion to place too much weight on those minimal scraps of knowledge. In that case, a data-driven approach is probably the best we can achieve. If we know how much we know about the basic biological phenomena, it would be possible to judge how much weight to place on that knowledge. But usually, we do not even know how extensive and valid our understanding is. In the example of the smoking, DNA adduct, and lung cancer picture, my impression is that our understanding is very fragmentary indeed, and that therefore, the biological justification for their model is no more than an interesting starting point for empirical exploration. Third, should models be developed among all subjects or among smokers only? This has both technical and conceptual implications. (Is the difference between nonsmokers and smokers qualitative as well as quantitative?). Fourth is the issue of which statistical criteria to use in comparing different models. Although linearity and parsimony are interesting criteria, should there be some other criteria such as analysis of residuals, likelihood ratio tests, and goodness-of-fit tests? What should be the hierarchy of criteria? A fifth issue concerns generalizability of their findings. There are many population-specific factors that could affect the performance of a given metric in a given population, including the pattern of smoking behavior; the correlation between smoking and other covariates; the prevalence and effect of effect modifiers, both environmental and genetic; and the nature of the tobacco products consumed. It would not be surprising if the optimal parameterization of the smoking history varied by place and time. Add to these such methodologic features as differing measurement error and differing study designs and it seems very hopeful to expect any one metric to be optimal for all populations.

There is an advantage to using metrics that have some intuitive meaning and that are widely used; thus, the results can be compared. Although this is less true for a confounder variable than for an exposure variable, there is still some merit in fostering standardization of methods, even for confounders. The conventional variables (duration, daily amount, and pack-years) have these qualities. On the other hand, there is an advantage to using the metric that seems to best embody the smoking history in that data set. Perhaps the investigator should be encouraged to examine both. If we cannot expect a uniformly optimal metric, then before any metric is used, its properties should be explored in the data set to which it will be applied.

Thurston et al. present a metric that is simple to implement, has some intuitive appeal, and that showed some desirable qualities in their data set. They have also illustrated an approach to judging the value of a metric. Other investigators have suggested other metrics [e.g., Doll and Peto (9), Leffondre et al. (6), and Hoffmann and Bergmann (7)]. Standard textbooks of epidemiology deal with variables that are binary, categorical, or continuous; however, they provide little guidance on how to create variables that reflect complex exposure histories. It would be a service to epidemiologists if methodologists could provide some practical guidance, possibly based on simulation studies and on theory, as to how to go about evaluating different metrics of smoking history for various purposes.

**References**

## Cancer Epidemiology, Biomarkers & Prevention

### Synthesizing the Lifetime History of Smoking

Jack Siemiatycki


<table>
<thead>
<tr>
<th>Updated version</th>
<th>Access the most recent version of this article at:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><a href="http://cebp.aacrjournals.org/content/14/10/2294">http://cebp.aacrjournals.org/content/14/10/2294</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cited articles</th>
<th>This article cites 12 articles, 6 of which you can access for free at:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><a href="http://cebp.aacrjournals.org/content/14/10/2294.full.html#ref-list-1">http://cebp.aacrjournals.org/content/14/10/2294.full.html#ref-list-1</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Citing articles</th>
<th>This article has been cited by 5 HighWire-hosted articles. Access the articles at:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>/content/14/10/2294.full.html#related-urls</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E-mail alerts</th>
<th>Sign up to receive free email-alerts related to this article or journal.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reprints and Subscriptions</td>
<td>To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at <a href="mailto:pubs@aacr.org">pubs@aacr.org</a>.</td>
</tr>
<tr>
<td>Permissions</td>
<td>To request permission to re-use all or part of this article, contact the AACR Publications Department at <a href="mailto:permissions@aacr.org">permissions@aacr.org</a>.</td>
</tr>
</tbody>
</table>