Is Residual Confounding a Reasonable Explanation for the Apparent Protective Effects of Beta-carotene Found in Epidemiologic Studies of Lung Cancer in Smokers?

Daniel O. Stram, Mark Huberman,† and Anna H. Wu

The results of three randomized trials of beta-carotene supplementation for the prevention of lung cancer among smokers are in contradiction to a large body of epidemiologic evidence for the reduction of risk of lung cancer among smokers with higher intake and/or higher serum levels of beta-carotene. Complicating this issue are widely noted negative associations between tobacco use and intake or serum levels of beta-carotene. Although observational studies attempt to control for reported smoking histories, the accuracy of self-reported smoking is uncertain; correlations as low as 0.5 between reported and true smoking exposure are not inconsistent with studies of biomarkers of cigarette exposure. The authors developed a simple statistical model for random errors in reported smoking (relative to true tobacco exposure) and assumed a modest (inverse) relation between true tobacco exposure and serum beta-carotene. Calculations from this model, combined with a model for lung cancer contemplated by Doll and Peto (J Epidemiol Community Health 1978;78:303–13), suggest that biases in assessment of smoking exposure between smokers with low versus high beta-carotene intake may plausibly explain much or all of the observed protective effect of high beta-carotene levels. Appropriate cohort studies of lung cancer in smokers, utilizing biomarkers of smoking, are needed and are presently ongoing. Am J Epidemiol 2002;155:622–8.

beta carotene; lung neoplasms; measurement error; smoking

The failure of three prospective randomized trials (1–3) to find beneficial effects of beta-carotene supplementation on the risk of lung cancer among smokers was unexpected, given the consistent association in observational studies of low beta-carotene levels with elevated risk of lung cancer among smokers.

In their report on the results of the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study, Albanes et al. (2) discussed three possible explanations of this apparent paradox: 1) that serum levels and “usual” diet represent long-term exposure to beta-carotene and that effects on risk may be different from that of short-term exposure; 2) that beta-carotene measured in blood or estimated from dietary intake may be acting as a marker for other correlated intakes that may, in fact, be protective; 3) that high beta-carotene intake or serum concentrations are general indicators of other beneficial dietary or protective lifestyle practices or are inversely associated with exposures to “harmful or carcinogenic factors.” In this paper we discuss the possibility that the harmful factor in explanation 3 was, in fact, tobacco itself. Specifically we consider here the possibility that low beta-carotene levels may simply serve as a marker for higher than reported exposure to tobacco.

Several studies have reported that beta-carotene intakes (4–8) and/or serum levels of beta-carotene (5) are negatively associated with reports of current smoking levels. Thus, at least to some extent, reductions in beta-carotene appear to act as a biomarker for smoking. All observational studies that found protective effects of beta-carotene on risk of lung cancer in smokers controlled for reported smoking history. However, because the metric of smoking exposure that is biologically relevant to lung cancer risk is undoubtedly less than perfectly correlated with reported numbers of cigarettes used per day, distortions in the beta-carotene effect due to residual confounding remain a distinct possibility. In this paper we outline the issues involved in evaluating the magnitude of potential distortions using relevant statistical models for residual confounding.

Several authors have suggested that residual confounding of beta-carotene intake with true smoking exposure may have induced the apparent protection afforded to smokers with high intake of beta-carotene. Henderson et al. (9) discussed the difficulty of interpreting beta-carotene effects given their findings that beta-carotene intake was lower in heavy smokers than in light smokers. Marshall and Hastrup (10) performed an extensive analysis of the extent of residual confounding expected in epidemiologic settings when a
very strong confounder exists and specifically discussed the
issues involved in joint analyses of beta-carotene, smoking,
and lung cancer. This paper is an expansion of these previ-
ous discussions; in particular, we both elaborate and refine
the approach of Marshall and Hastrup to deal with the fol-
lowing: evidence from studies of smoking biomarkers
that may limit the possible magnitude of random errors in
self-reports of current smoking; relative risk models for
lung cancer as a function of observed smoking that are
preferred based upon epidemiologic data; and an expansion
of the models used for the relation between self-reports of
smoking and true exposure to allow for the quite different
effects of “classical” versus “Berkson” error in smoking
self-reports. More discussion of this last issue is given
below.

RESIDUAL CONFOUNDING AS AN EFFECT OF
MEASUREMENT ERROR

Distortion of the joint relation between a disease outcome
and covariates is one of the effects of random measurement
error (10–12). Suppose that beta-carotene intake actually
has no effect upon the risk of lung cancer, but that it is
negatively correlated with the biologically relevant true
exposure that we imagine to be only imperfectly correlated
with reported smoking level. If true exposure were known
for subjects in an epidemiologic study, then joint analyses
would give an unbiased estimate of no effect for the influ-
ence of beta-carotene intake or serum levels on lung cancer
risk conditional upon lung dose. If, however, lung dose is
imperfectly measured (using reported cigarettes per day),
then beta-carotene would remain somewhat negatively cor-
related with true lung dose conditional on reported smoking
level. Beta-carotene would appear protective because
groups of smokers with the same reported levels of smoking
but different true lung doses would tend also to have dif-
ferent levels of beta-carotene.

A model for residual confounding

We adopt for illustrative purposes a simple multivariate
model incorporating the following:

1. a bivariate lognormal distribution between true lung
dose and reported smoking. True lung dose is stan-
dardized to units equivalent to cigarettes/day, and the
marginal mean of true lung dose is taken as equal to
the mean of reported smoking. One “cigarette” of true
lung dose is the value that is on average dispensed by
consumption of one cigarette, where the averaging is
over the population of smokers.

2. a linear relation between the log of serum beta-caro-
tene and (the arithmetic value of) true lung dose. This
is similar to that used for the data analysis by Stryker
et al. (5).

3. a model for the relative risk of lung cancer in smokers.
We focus upon the relation described by Doll and
Peto (13).

Consider the following analysis. Smokers are stratified by
level of reported smoking, and beta-carotene levels are
measured for each smoker. Within smoking category smok-
ers are subcategorized by quartiles of beta-carotene level.
We ask then, for each smoking category, what is the differ-
ence in the distribution of true lung dose, and hence in lung
cancer risk, between smokers in the highest and lowest
quartile of beta-carotene level?

Bivariate distribution of true lung dose and reported
smoking. Reports of the cigarettes/day consumed by
smokers range from 20 to 30 cigarettes with lower estimates
coming from self-reports by individual smokers than from
dividing the total sales of cigarettes by the number of
smokers in the US population (14). We parameterize our
model for reported smoking so that the median of reported
smoking is equal to 20 cigarettes per day and so that 95
percent of smokers report between 5 and 60 cigarettes.

The following relation between the log of true lung dose,
X, and the log of reported smoking, Z, is assumed:

\[ X = (1 - b)\mu_Z + bZ + e. \]  

(1)

Here \( \mu_Z \) is the mean of the log of reported smoking, and \( e \)
is a random variable with mean zero and variance \( \sigma_e^2 \). The
choices of \( b \) and \( \sigma_e^2 \) determine the measurement error model.

We note that if \( b = 1 \) then this model corresponds to a
so-called Berkson error model in which the log true dose is
distributed symmetrically around log reported smoking as

\[ X = Z + e. \]

Choosing \( b \) to be equal to the reliability coefficient

\[ b = \frac{\sigma_X^2}{\sigma_Z^2} \]

leads to the classical error model in which (log) reported
smoking is distributed symmetrically around log true dose as

\[ Z = X + \text{error}. \]

(See Appendix.)

A Berkson error model is appropriate if it is believed that
smokers reported cigarette intake accurately but that true
lung dose was influenced by other factors, such as the
number of puffs, depth of inhalation, and so on, that varies
from smoker to smoker. A classical model is appropriate if
random errors in the reported intakes are believed to be of
overriding importance. By varying \( b \) we produce three mod-
els, a Berkson, a classical, and an intermediate variant,
which we term the “mixed” model. In the mixed error
model, a Berkson relation between true lung dose and true
cigarette consumption and a classical error relation between
reported smoking and true cigarette intake are both assumed
and are assumed to have equal error variances. The error
model (equation 1) describes a multiplicative relation be-
tween errors in reported smoking and true lung dose. We
regard a multiplicative model as intuitively reasonable, be-
cause large true exposures are likely to have larger reporting
efforts than small values. Our purpose in exploring the
general form (equation 1) of the measurement error model is
to ensure, by varying \( b \), that our results are not highly
dependent upon subtleties of our modeling assumptions.

The choice of \( b \) and \( \sigma_e^2 \) determines the correlation \( R_{x,z} \)
between reported smoking and true lung dose, a value crucial to all that follows. Various studies describe associations of self-reported smoking with biomarkers including blood, urine, salivary or hair levels of nicotine or cotinine, blood thiocyanate or carboxyhemoglobin, and carbon monoxide expiration, with cotinine generally found to be the most sensitive (15–18). Correlations for cotinine ranged from 0.40 to 0.70 in various studies (17, 19–23). We consider two models: first, one in which the correlation between true lung dose and reported smoking, \( R_{x,z} \), equals 0.55 (the median of the correlations with cotinine measurements in the studies cited above); the second sets this correlation to 0.85, which we consider to be a very high value for the validity of a self-reported exposure. We denote those models with \( R_{x,z} = 0.55 \) as large error models and \( R_{x,z} = 0.85 \) as small error models.

**Correlation between log beta-carotene and true lung dose.** Although few studies directly examine the correlations between beta-carotene and reported smoking, reduced beta-carotene intake or serum levels in smokers versus nonsmokers and in heavy smokers versus light smokers are very evident (4–8). We assume a relatively weak negative correlation, \( R_{x,z} = -0.25 \), between log beta-carotene and true lung dose. Note that the correlations between reported smoking and (log) serum beta-carotene are attenuated by measurement errors compared with the true association. We have the convenient relation that for any of our error models

\[
\text{Corr}(\text{log beta-carotene, reported smoking}) = \text{Corr}(\text{log beta-carotene, true lung dose}) \\
\times \text{Corr}(\text{reported smoking, true lung dose}).
\]

Under the large error model, a true correlation of \(-0.25\) yields an observed correlation of \(-0.14\), which is not implausible compared with the reports cited above. Reasons for negative correlations include a general tendency for smokers to have poor diets (7–9), perhaps as a direct effect of tobacco intake upon the perceived tastefulness of fruits and vegetables, and (for serum levels) a direct action of cigarette smoke on beta-carotene levels (24).

**Models for risk of lung cancer**

Doll and Peto (13) assess the dose-response relation between smoking and lung cancer using careful age-specific comparisons of subjects who began smoking at approximately the same age and continued smoking throughout follow-up. Their model for the relative risk (RR) of lung cancer mortality may be written as

\[
RR = \left( 1 + \frac{1}{6} \times \text{cigarettes/day} \right)^2
\]

We use this model as well as a linear model

\[
RR = 1 + \frac{13}{9} \times \text{cigarettes/day},
\]

agreeing with the Doll and Peto model at the values of 0 and 40 cigarettes/day but linear in dose response. Both models 2 and 3 involve reported smoking. If \( b < 1 \) in the error model for smoking, then the dose response will be even more marked after adjustment for random errors in reported smoking. In particular (Appendix), if a model for risk is linear in reported smoking, \( z \), then the model for risk on the true dose scale, \( x \), will involve a higher power of \( x \), namely, \( x^{1/b} \), where \( b \) is from model 1. Similarly a quadratic term \( z^2 \) implies a greater than quadratic term \( x^{2/b} \) in the true relation. There are important differences between the Berkson model and the classical model in the attenuation of risk implied by errors in smoking reports. The classical model for error produces a profound attenuation of the strength of the dose response. For example (appendix equation A2), after correction for attenuation using the classical large-error model combined with the quadratic Doll-Peto model, the relative risk of lung cancer for a smoker whose “true lung dose” equals 60 cigarettes/day will be equal to 1,649, compared with a nonsmoker, compared with 121 on the observed dose scale. However, a purely Berkson multiplicative error model has quite the opposite effect. The relative risk on the scale of true dose, when \( x = 60 \) cigarettes/day, is actually lower (RR = 50) than on the self-report scale (RR = 121). This somewhat unusual feature of the Berkson model is due to the fact that the average value, \( E(x|z) \), of true dose, \( x \), given observed dose, \( z \), is higher than \( z \) on the arithmetic scale. The mixed model is intermediate in the degree of resulting attenuation (relative risk of 144 at true \( x = 60 \) cigarettes/day compared with 121 on the observed dose scale).

**MATERIALS AND METHODS**

For each choice of \( b \) and \( \sigma_x^2 \) in model 1, we can find relative risk models, \( RR(x) \), on the true dose scale \( x \) that correspond to models 2 and 3 on the scale of \( z \) and that are functions of one or both of \( x^{1/b} \) and \( x^{2/b} \), respectively. We integrate \( RR(x) \) over the conditional distribution, \( f(x|z, B) \), of \( x \) given \( z \) and \( B \). To compare the lung cancer risk of two smokers with the same \( z \) but different beta-carotene levels \( B_1 \) and \( B_2 \), we form the new relative risk ratio

\[
RR = \frac{\int_0^\infty \text{RR}(x) f(x|z, B_2) dx}{\int_0^\infty \text{RR}(x) f(x|z, B_1) dx}.
\]

Studies in smokers typically compare the risk of lung cancer of subjects with “low” values of beta-carotene with those of smokers with “high” values of beta-carotene. In our calculations of \( RR \) we do the same thing, picking low values of \( B \) to be the average for the lowest quartile of beta-carotene and high values to be the average in the highest quartile. Details of the relevant calculations are given in the Appendix.
High values of RR indicate important residual confounding between beta-carotene and true lung dose in observational studies. Many observational studies have found the risks for smokers with low levels of beta-carotene to be about twice the risks for those with high levels of beta-carotene intake or serum measurements. If RR in equation 4 approaches 2, then we argue that these observational findings may in fact be purely an artifact of errors in the measurement of smoking.

RESULTS

Table 1 gives the risk ratio RR expected between smokers with low and high beta-carotene intake, conditional on the same level of reported smoking, according to the following factors:

1. the choice of \( b \) in the measurement error model 1.
2. the degree of correlation, \( R_{z,x} \), between true lung dose and reported smoking, which itself is dependent on \( b \) and \( \sigma^2 \) in model 1.
3. the choice of risk model for lung cancer as being either linear-quadratic (model 2) or linear (model 3).
4. whether reported smoking \( z \) is taken to be a continuous or a grouped variable (with three categories) in the conditioning.

In all cases the calculations are performed under the assumption that the correlation, \( R_{z,x} \), between the true lung dose and the log beta-carotene level is equal to \(-0.25\).

In table 1 a combination of the linear-quadratic model for risk with the high-error models for reported smoking leads to risk ratios, RR, which range from approximately 1.8 to 2.3 depending upon whether a Berkson, classical, or mixed measurement error model is chosen for the relation between true and reported smoking; the risk ratio is slightly higher for the classical and mixed models than it is for the Berkson model. Using the linear model lowers the risk ratio, with RR now between 1.3 and 1.6; again RR is slightly higher for the classical and mixed models than for the Berkson model. As expected, the results are quite dependent upon whether the high or low error measurement model is chosen: The maximum risk ratio for the low error model is 1.2 for the quadratic risk model using mixed or classical errors.

DISCUSSION

Cigarette smoking is by far the most important cause of lung cancer, and the dose-response relation between true smoking level and lung cancer risk may be even stronger than that which can be observed using self-reports of smoking. Because of the extraordinary strength of the dose-response relation between true exposure and risk, any variable such as beta-carotene that may be providing a modest amount of additional information about true exposure needs to be very carefully examined as a confounder before making conclusions about its own likely contribution to risk.

This general point (that residual confounding is important) has been made repeatedly by statisticians (11, 12) and specifically by Marshall and H构造 (10) in the case of beta-carotene’s apparent protective effects. Our expanded discussion of this issue here is intended to reinforce the need to secure better data on the accuracy of the measurement of smoking and on the strength of the relation between beta-carotene intake and smoking exposure.

In a recent account, Albanes (25) discounts the possibility that residual confounding could be at the root of observations of a protective effect of beta-carotene, indicating that most studies involved “adequate control” of smoking history and other known confounders. We disagree and believe that residual confounding must in fact be considered as an important candidate among potential explanations. Some epidemiologists dealing with cohort analyses evidently now also take this view. A recent report of a cohort study analysis (26) noted that careful control for self-reported cigarette exposure led to reductions in the apparent protective effect of beta-carotene intake and noted that the remaining protective effect seen may be indicative of confounding by unmeasured smoking characteristics.

In our modest correlation \( R_{z,x} = -0.25 \), high measurement error \( R_{z,x} = 0.55 \) model, the relative risk induced purely by residual confounding lies between 1.3 and 2.3 depending upon whether it is the linear or quadratic risk model and whether it is the Berkson model or classical error

---

**TABLE 1.** Model predictions of the ratio of risks of lung cancer for lowest versus highest quartile of beta-carotene conditional on observed smoking treated as either a continuous or a grouped (by tertile) variable

<table>
<thead>
<tr>
<th>Risk model</th>
<th>High measurement error model*</th>
<th>Low measurement error model†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Berkson Classical Mixed</td>
<td>Berkson Classical Mixed</td>
</tr>
<tr>
<td>Linear-quadratic‡</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grouped</td>
<td>1.8</td>
<td>1.2</td>
</tr>
<tr>
<td>Continuous</td>
<td>1.8</td>
<td>1.1</td>
</tr>
<tr>
<td>Linear§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grouped</td>
<td>1.4</td>
<td>1.1</td>
</tr>
<tr>
<td>Continuous</td>
<td>1.3</td>
<td>1.1</td>
</tr>
</tbody>
</table>

* Correlation between reported smoking and true lung dose = 0.55.
† Correlation between reported smoking and true lung dose = 0.85.
‡ Risk model of Doll and Peto: ratio of risks of lung cancer (RR) = \( \frac{1}{\beta_0} (1 + \frac{1}{\beta} \times \text{cigarettes/day})^2 \).
§ Linear risk model: RR = \( 1 + \frac{1}{\beta} \times \text{cigarettes/day} \).
model that is used. Relative risks of this magnitude approach the levels seen for the apparent protective effect of beta-carotene in the observational studies of interest, so that most or all of the protective effects seen among smokers are plausibly due to residual confounding with tobacco given what we know (or don’t know) today about the magnitude of random errors in smoking self-reports. The fact that our findings are only partly dependent upon whether the Berkson, classical, or mixed model is assumed for errors in self-reports satisfies us that our results are robust to the details of the measurement model used; this is true despite the profound differences in the degree of attenuation of the true smoking model induced by these three different error models. This points out a truism about exposure measurement error: that the loss of control of confounding variables, that is, residual confounding, is a more characteristic effect of random measurement error than is the well-known “bias toward the null” because the latter is primarily a feature of the classical measurement model.

Our high error ($R_{z,c} = 0.55$) model is consistent with the literature regarding correlations between cotinine and self-reports only if we assume that cotinine is a nearly perfect reflection of true dose, an issue that is clearly open to dispute. Several studies indicate that cotinine, in particular, is associated with individual smoking behaviors (puffing, cigarette time) (27), as well as total cigarettes (16), and may be more predictive of adverse health outcomes than is reported smoking (28, 29). Nevertheless we recognize that the correlation between biomarkers and true lung dose is itself attenuated to some unknown degree by measurement error and that the high measurement error model would allow little play for this factor.

More careful examination of the relation between intake and serum levels of beta-carotene and reported levels of smoking in existing data sets is needed. One relevant reference is Stryker et al. (5) who reported a negative value for the slope of the bivariate regression of log beta-carotene consumption on cigarettes/day significant at the $p = 0.0004$ level among 137 males (25 percent of whom were current smokers). This significance level corresponds to $R = -0.296$ among the same number of subjects using Fisher’s $z$ transformation to do the hypothesis test. Although this correlation would be smaller in magnitude if restricted only to smokers, this aspect is counterbalanced by attenuation due to measurement errors in using reported smoking level rather than the biologically relevant true exposure. Given these considerations, we do not believe that a negative correlation between beta-carotene and true lung dose of the magnitude assumed in this paper (i.e., $-0.25$) is at all outside the range of plausibility. The risk ratios given in table 1 are quite dependent on the assumption of negative correlations this large: for example, for the classical model with high error, assuming this correlation to be equal to $-0.15$ instead of $-0.25$ reduces the risk ratio from 1.6 to 1.3 for the linear model. We note a very recent report (30) of a strongly significant inverse relation between systemic markers of inflammation and the serum beta-carotene level in both smokers and nonsmokers. Further examination of these sorts of data will also help in evaluating whether a correlation of $-0.25$ between the biologically relevant aspects of smoking exposure (such as those causing inflammation and presumably enhancing cancer induction and/or promotion) and serum beta-carotene is, in fact, reasonable.

Our calculations have been based upon the Doll and Peto (13) model in which it is the current smoking amount that affects the current risk of lung cancer, not total accumulated dose (puck-years) or duration of smoking. Duration of smoking must be dealt with carefully as a risk factor because it is confounded both with current age (a huge risk factor) and age at start (a potentially important risk factor). It is our own belief that studies that deal properly (by stratification) with these confounders, but still find that duration of smoking is much more predictive of risk than smoking amount is, may, in fact, have been especially subject to large errors in reported smoking.

To date there has been little or no work in observational studies of lung cancer controlling for biomarkers for smoking, and this would be very difficult to do in case-control studies. We are aware of a project at the Hawaii Cancer Research Center in Honolulu that will collect urine and serum samples at baseline, for prospective follow-up, for approximately 40,000 members of the Hawaii-Los Angeles multiethnic cohort study (31). One of the aims of the lung cancer component of this study (L. Marchand, Principal Investigator) is to investigate the relation between lung cancer and cotinine and various markers of nutrition intake including beta-carotene utilizing the stored samples. Clearly a finding from this study that the protective effect of beta-carotene is much weaker after conditioning on cotinine than when conditioning on reported levels of smoking would further implicate residual confounding as a cause of the beta-carotene effect seen in previous observational studies. On the other hand, if beta-carotene appeared protective after conditioning on cotinine, this could mean either that beta-carotene is actually protective or that serum cotinine is simply a poor biomarker. However, this latter conclusion would be an important indirect argument in favor of the protective effect of beta-carotene. Because of measurement error attenuation, the high error models being relied on here are inconsistent with reports in the literature of correlations of $\geq 0.4$ between cotinine and observed smoking unless cotinine is actually a very good biomarker for true exposure. Therefore, we regard the University of Hawaii work as likely to add considerably to the state of current knowledge about whether residual confounding is at the root of the apparent protective effects of beta-carotene.

ACKNOWLEDGMENTS

This work has been supported in part by Public Health Service grant R01 CA54281 from the National Cancer Institute.

The authors particularly acknowledge Dr. Malcolm C. Pike and Dr. Donald A. Pierce for their comments on earlier versions of this manuscript.
REFERENCES


APPENDIX

Let $x$, $y$, and $z$ be the true lung dose, true level of cigarette consumption, and reported consumption, respectively. Assume that $Z = \log z$ is normally distributed with mean $\mu_Z = \log 20$ and variance $\sigma_Z^2 = 0.38$, implying that 95 percent of the values of reported smoking fall between 5 and 60 cigarettes/day. Let $Y = \log y$. Assume that $Z$ conditional on $Y$ is normal with mean $\muZY$ and variance $\sigmaZY^2$ (classical errors). Let $X = \log x$. Assume that $Y$ conditional on $X$ is also normal with mean $\muYX$ and variance $\sigmaYX^2$ and that the Berkson and classical errors are independent. It follows that the distribution of $X$ conditional on $Z$ is normal with mean

$$\mu_{XZ} = \left(1 - \frac{\gamma_Y^2}{\sigma_Y^2}\right)Z + \frac{\gamma_Y^2}{\sigma_Y^2}\mu_Z$$

and variance

$$\sigma_{XZ}^2 = \gamma_Y^2 + \frac{\gamma_Y^2}{\sigma_Y^2} \left(1 - \frac{\gamma_Y^2}{\sigma_Y^2}\right).$$

In equation 1 in the text we identify $b = 1 - \gamma_Y^2/\sigma_Y^2$. For the fully Berkson model, $b$ equals 1, while for the fully classical error model $b = \gamma_Y^2/(\sigma_X^2 + \gamma_Y^2)$, which is the reliability coefficient. In the mixed error model, we set $\gamma_Y^2 = \gamma_Y^2$. We...
note that \( E(x'|z) = \exp[n(1 - b)\mu_z + n^2\sigma^2_{x'z}/2]z^b \). The correlation between \( z \) and \( x \) equals

\[
R_{z,x} = \frac{\exp(\sigma^2_z - \gamma^2_z) - 1}{\sqrt{\exp(\sigma^2_z + \gamma^2_z - 1) - 1}}.
\]

Let \( B \) be the log of reported beta-carotene intake or serum level with mean \( \alpha + \beta x \) and variance \( \sigma^2 \). The correlation of \( B \) and \( x \) is

\[
R_{B,x} = \frac{\beta \sigma_z}{\sqrt{(\beta \sigma_x)^2 + \sigma^2}}
\]

and under the assumption that \( B \) is independent of \( Z \) given \( X \), it can readily be shown that \( R_{B,z} = R_{B,x} \times R_{z,x} \). Let \( \mu(z) \) be the relative risk of lung cancer in terms of reported smoking level \( z \). Let \( \lambda(x) \) be the relative risk of lung cancer in terms of true lung dose \( x \). These two quantities are related by

\[
\mu(z) = E[\lambda(x)|z].
\]

(A1)

Our models for \( \mu(z) \) can be summarized by

\[
\mu(z) = 1 + Az + Bz^2
\]

where \( A = 1/3 \) and \( B = 1/36 \) for the Peto-Doll relation, and \( A = 13/9 \) and \( B = 0 \) for the linear relation. Substituting for \( E(x'|z) \) in equation A1,

\[
\lambda(x) = 1 + A \exp[-(1 - b)\mu_x/b - \sigma^2_{\mu}/2b^2]x^{1/b} + B \exp[-2(1 - b)\mu_x/b - 2\sigma^2_{\mu}/b^2]x^{2/b}.
\]

(A2)

Appendix table 1 gives the parameter values for equation A2 corresponding to the high and low error version of the classical Berkson and mixed models. Because \( 0 < b \leq 1 \), \( \lambda(x) \) may depend on higher powers of the exposure covariate than \( \mu(z) \). The numbers in table 1 were obtained as

\[
E[\lambda(x)|Z_-, Z < Z_+, B < B_-] = \int_{-\infty}^{\mu} \int_{-\infty}^{Z_-} \int_{-\infty}^{Z_+} \int_{-\infty}^{\infty} \lambda(x) f(X, Z, B) dX dZ dB
\]

These integrals were calculated by forming a three-dimensional grid and summing the values of the integrand at the points of the grid. The fineness of the grid and the dimensions of the grid were successively increased until no further changes in the results were noted (to the accuracy quoted in table 1). Further details of the calculations may be obtained from the authors.

### APPENDIX TABLE 1. Parameter values for the models used for errors in self-reports of smoking

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Classical</th>
<th>Mixed</th>
<th>Berkson</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma_0 )</td>
<td>0.2491</td>
<td>0.0918</td>
<td>0.1534</td>
</tr>
<tr>
<td>( \gamma_b )</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( 1/b )</td>
<td>2.9</td>
<td>1.3</td>
<td>1.7</td>
</tr>
<tr>
<td>( \mu_{z} )</td>
<td>Log 20</td>
<td>Log 20</td>
<td>Log 20</td>
</tr>
<tr>
<td>( \sigma^2_{x,z} )</td>
<td>0.0858</td>
<td>0.0696</td>
<td>0.2450</td>
</tr>
</tbody>
</table>