

## ON THE DISCREPANCY BETWEEN EPIDEMIOLOGIC STUDIES IN INDIVIDUALS OF LUNG CANCER AND RESIDENTIAL RADON AND COHEN'S ECOLOGIC REGRESSION

Jay H. Lubin\*

**Abstract**—There is still substantial confusion in the radiation effects community about the inherent limitations of ecologic analysis. As a result, inordinate attention has been given to the discrepant results of Cohen, in which a negative estimate is observed for the regression of county mortality rates for lung cancer on estimated county radon levels. This paper demonstrates that Cohen's ecologic analysis cannot produce valid inference on the exposure-response relationship for individuals unless lung cancer risk factors (smoking, age, occupation, etc.) for individuals are statistically uncorrelated with indoor radon level within counties or unless risk effects for radon and other factors are additive. Both of these assumptions are contradicted in the literature. Thus, contrary to common assumption, when a linear no-threshold model is the true model for radon risk for individuals, higher average radon concentration for a county does not necessarily imply a higher lung cancer rate for the county. In addition, valid inference from county-level ecologic analysis and the elimination of the ecologic bias cannot be achieved with the addition of county-wide summary variables (including "stratification" variables) to the regression equation. Using hypothetical data for smoking and radon and assuming a true positive association for radon and lung cancer for individuals, the analysis demonstrates that a negative county-level ecologic regression can be induced when correlation coefficients for smoking and radon within county are in the range  $-0.05$  to  $0.05$ . Since adverse effects for radon at low exposures are supported by analysis of miner data (all data and data restricted only to low cumulative exposures), a meta-analysis of indoor radon studies, and molecular and cellular studies, and since ecologic regressions are burdened by severe limitations, the negative results from Cohen's analysis are most likely due to bias and should be rejected.

Health Phys. 75(1):4-10; 1998

**Key words:**  $^{222}\text{Rn}$ ; epidemiology; cancer; lungs, human

### INTRODUCTION

RECENT publications and various Internet discussion groups suggest that there is still substantial confusion in

\* Division of Cancer Epidemiology and Genetics, National Cancer Institute, Executive Plaza North, Rm 403, Bethesda, MD 20892.  
For correspondence or reprints contact Jay H. Lubin at the above address.

(Manuscript received 12 August 1997; revised manuscript received 30 September 1997, accepted 6 April 1998)

0017-9078/98/\$3.00/0

Copyright © 1998 Health Physics Society

the radiation effects community about the inherent limitations of ecologic analyses of indoor radon exposure and risk of lung cancer (Cohen 1997; see the radsafe listserv on the Internet, address [gopher://romulus.ehs.uiuc.edu:70/11/radiation](mailto:gopher://romulus.ehs.uiuc.edu:70/11/radiation)). In spite of the positive association between radon progeny exposure and lung cancer in all epidemiologic studies of underground miners (NRC 1988, 1998; Lubin et al. 1995), the consistency of the miner results with studies in the general population (Lubin and Boice 1997), the molecular and cellular basis for low dose effects from alpha particles (NRC 1998), and the known limitations of ecologic studies (Piantadosi et al. 1988), an inordinate amount of attention has been given to the discrepant results of the ongoing ecologic study of lung cancer mortality in U.S. counties by Cohen (1995, 1997). In his most recent paper, Cohen seeks a "not implausible potential explanation" for the difference between his results and most epidemiologic studies. This paper demonstrates the probable basis for this discrepancy.

Cohen regresses age-adjusted county lung cancer mortality rates on an estimate of the average county-level radon concentration and obtains a negative relationship (Cohen 1995). He interprets the results as directly relevant to inference about the functional form of the relationship between radon progeny exposure and lung cancer risk for the individual. Based on this presumed relevance, he concludes that the negative regression coefficient implies that the linear no-threshold model for risk estimation is not valid and thus the exposure-response relationship for individuals is not positive at the low exposures commonly experienced in the general population. His conclusions imply that there is a fundamental misspecification of the miner-based risk model as applied to low indoor exposures. As shown below, the absolute link that Cohen presupposes between the county-level regression model and the risk model for individuals is fallacious.

The limitations of ecologic studies in general have been widely discussed in the epidemiologic literature (Brenner et al. 1992; Greenland 1992; Morgenstern 1995; Piantadosi et al. 1988) and will not be considered in detail here. However, results from ecologic studies of indoor radon and lung cancer should be viewed with particular skepticism, due to substantial misclassification

of exposure risk, 20 comparing in

The logic a national efficient sion co to the r estimat downw (1991) rates a With d positive for gam level w tive co gamma Finally, thetical ecologi individ smokin to nons county

The regressi demons respons related even w for risk higher a necessa county. Unbiase ship for regressi certain r with in regressi through assured for the e radon le is positi

### RISK

The determin underlyi simplify and  $c =$  individu

of exposure and the small expected level of radon-related risk, 20–30% excess in perhaps 5% of the population, compared with a 1,000–2,000% excess risk from smoking in 30–40% of the population.

Three examples highlight the problems with ecologic analysis. Piantadosi et al. (1988) used data from a national nutrition survey and compared regression coefficients obtained from data on individuals with regression coefficients obtained from aggregated data. Relative to the regression coefficients for individuals, coefficients estimated from aggregated data were biased upward, downward, and even reversed sign. Muirhead et al. (1991) carried out an ecologic regression of leukemia rates and levels of indoor radon and gamma radiation. With data aggregated at the county level, they found a positive coefficient for radon and a negative coefficient for gamma radiation. With data aggregated at the district level within county, coefficients were reversed, a negative coefficient for radon and a positive coefficient for gamma radiation, suggesting district-level confounding. Finally, Greenland and Robins (1994) presented a hypothetical example that mimicked the results of Cohen's ecologic analysis. Within counties, lung cancer risk for individuals increased with increasing radon for fixed smoking level, and risk was higher in smokers compared to nonsmokers for fixed radon level. Analysis based on county rates resulted in a negative association between county risk and average radon levels.

This paper illustrates analytically the reversal of a regression coefficient when data are aggregated and demonstrates the fallacy that a negative exposure-response relationship at the county level necessarily is related to risks for individuals. Thus, it is shown that even when a linear no-threshold model is the true model for risk in individuals from residential radon exposure, higher average radon concentration for a county does not necessarily imply a higher lung cancer rate for the county. We also make the following observations: (1) Unbiased inference on the exposure-response relationship for an individual cannot be achieved with ecologic regression using only county-level data, except under certain restrictive conditions, which are unlikely to occur with indoor radon; (2) No adjustment in an ecologic regression using standard county-wide data, either through added covariates or "stratification," can be assured to correct the bias; and (3) A negative estimate for the ecologic regression of lung cancer rate on county radon level, when the true association for individual risk is positive, can occur in practical situations.

### RISK MODELS FOR INDIVIDUALS AND FOR COUNTIES

The (true) county-level risk model for lung cancer is determined by averaging over all residents the true underlying risk model for lung cancer for individuals. To simplify, assume that there are only two counties,  $c = 1$  and  $c = 2$ , a single age group, say ages 65–69 y, and all individuals are exposed at a constant exposure rate.

Further, assume that there is only one additional risk factor, smoking, and that all data are measured without misclassification. While the development of the effects of confounding variables is presented in terms of smoking, the implications extend to other lung cancer risk factors, such as age and occupational exposures.

Cumulative radon progeny exposure is expressed in terms of radon concentration in  $\text{Bq m}^{-3}$ , denoted  $w$ , assuming 70% home occupancy, 0.5 equilibrium factor, and 30-y exposure. For concentration  $w$  and smoking status  $s$ , where  $s = 1$  denotes smoker and  $s = 0$  denotes nonsmoker, the lung cancer mortality rate for an individual,  $r(s, w)$ , is defined as

$$r(s, w) = r_0 \theta^s (1 + \beta_1 w), \quad (1)$$

where  $r_0 = r(0, 0)$  is the background lung cancer rate in nonsmokers who are not exposed to radon, i.e., exposed only at ambient levels. To a first order approximation, eqn (1) is consistent with epidemiologic data. The parameter  $\theta$  defines the relative risk in smokers compared to nonsmokers, and  $\beta_1$  defines the true excess relative risk for 30 y residence at  $1 \text{ Bq m}^{-3}$ . The values for  $r_0$ ,  $\theta$ , and  $\beta_1$  are assumed the same for both counties. Thus, by design, county of residence does not affect risk. With data on  $s$  and  $w$  for individuals and risk described by eqn (1), case-control and cohort studies allow, at least conceptually, unbiased estimation of  $\theta$  and  $\beta_1$ . Eqn (1) is a linear no-threshold model in  $w$ , meaning that given smoking status halving exposure  $w$ , halves the added excess risk.

The lung cancer rate for a county, denoted  $r^c$ , is the average risk, eqn (1), for  $N_0^c$  nonsmokers and  $N_1^c$  smokers, with total population  $N^c = N_0^c + N_1^c$ . Suppose  $P_0^c = N_0^c/N^c$  and  $P_1^c = N_1^c/N^c$  are the proportions of nonsmokers and smokers, respectively. The lung cancer rate for the county  $r^c$  is

$$\begin{aligned} r^c &= \sum_{s, w} r_0 \theta^s (1 + \beta_1 w) / N^c \\ &= r_0 [P_0^c \sum_{s=0, w} (1 + \beta_1 w) / N_0^c \\ &\quad + \theta P_1^c \sum_{s=1, w} (1 + \beta_1 w) / N_1^c] \\ &= r_0 (P_0^c + \theta P_1^c) [1 + \beta_1 (\lambda_0^c \bar{W}_0^c + \lambda_1^c \bar{W}_1^c)], \end{aligned}$$

where  $\lambda_0^c = (P_0^c) / (P_0^c + \theta P_1^c)$  and  $\lambda_1^c = (\theta P_1^c) / (P_0^c + \theta P_1^c)$ , and where  $\bar{W}_0^c$  and  $\bar{W}_1^c$  are the average radon concentrations in nonsmokers and smokers, respectively. Setting  $\bar{W}^c = \lambda_0^c \bar{W}_0^c + \lambda_1^c \bar{W}_1^c$ , the lung cancer rate for county  $c$  is

$$r^c = r_0 (P_0^c + \theta P_1^c) (1 + \beta_1 \bar{W}^c). \quad (2)$$

Note that  $\bar{W}^c$  is a weighted average of  $\bar{W}_0^c$  and  $\bar{W}_1^c$  with  $\lambda_0^c$  and  $\lambda_1^c$  as weights.

Eqn (2) represents the true relationship between the county lung cancer rate and the two risk factors, radon and smoking. The product  $r_0 (P_0^c + \theta P_1^c)$  is the county lung cancer rate in the mixed population of smokers and nonsmokers who are not exposed to radon. The factor  $1 + \beta_1 \bar{W}^c$  describes the county-level relative risk of radon. Since  $\bar{W}^c$  is not the simple average radon concen-

tration for the county and since neither  $\bar{W}^c$  nor mean radon levels for smokers and nonsmokers,  $\bar{W}_0^c$  and  $\bar{W}_1^c$ , are typically known, unbiased estimates of  $\beta_t$  are problematic.

In an ecologic analysis, data at the county level are typically available on the proportion of smokers  $P_1^c$  and the overall average radon level, denoted  $\bar{W}^c$ , which can be expressed as the weighted average

$$\begin{aligned}\bar{W}^c &= \sum_{s,w} w/N^c \\ &= P_0^c \sum_{s=0,w} w/N_0^c + P_1^c \sum_{s=1,w} w/N_1^c \\ &= P_0^c \bar{W}_0^c + P_1^c \bar{W}_1^c.\end{aligned}$$

Both  $\bar{W}^c$  and  $\bar{W}^c$  are computed from mean radon levels for smokers and nonsmokers, but use different weights. Ecologic analysis regresses disease rates on  $\bar{W}^c$  and other factors based on models of the form

$$r^c = r_0(P_0^c + \theta P_1^c)(1 + \beta_e \bar{W}^c). \quad (3)$$

Although eqn (3) is similar in form to eqn (2) and indeed is an example of a linear no-threshold model, differences between  $\bar{W}^c$  and  $\bar{W}^c$  can result in a biased estimate of the true exposure-response relationship (i.e.,  $\beta_e \neq \beta_t$ ).

With the simple average,  $\bar{W}^c$ , each individual contributes equally to the mean radon level for the county. However, eqn (2) defines the county lung cancer rate as a function of both smoking status and radon level, in which each individual does not contribute equally to the overall rate. For example, the risk of lung cancer for a nonsmoker is  $r_0(1 + \beta_t w)$ , while the risk of lung cancer for a smoker is  $r_0 \theta(1 + \beta_t w)$ . Thus, use of eqn (3) with  $\bar{W}^c$  can result in an estimate  $\beta_e$ , which is biased for  $\beta_t$ . The differential contribution of nonsmokers and smokers to county risk is subsumed within the weights used for  $\bar{W}^c$ .

Cohen uses  $\bar{W}^c$  as a regressor variable in his analysis [see eqn (1) in Cohen (1997) or eqn (2) in Cohen (1995)], rather than  $\bar{W}^c$ . Since mean radon levels for smokers and nonsmokers are unknown, the inclusion of  $P_1^c$ , or any transformation of  $P_1^c$ , or the product  $P_1^c \times \bar{W}^c$  in the regression cannot eliminate the difference between  $\bar{W}^c$  and  $\bar{W}^c$ . As discussed below, use of  $\bar{W}^c$  introduces bias, except under certain restrictive conditions.

## INTERPRETATION OF MODELS

There are instances in which county-level regression under eqn (3) can lead to unbiased estimates. Suppose there were no smoking effect, i.e.,  $\theta = 1$ . Then  $\lambda_0^c = P_0^c$  and  $\lambda_1^c = P_1^c$ , which implies that  $\bar{W}^c = \bar{W}^c$  and therefore that  $\beta_t$  is estimable from county data. Suppose smoking status and radon exposure were independent (or uncorrelated) within county. Then, the mean radon concentration is the same for nonsmokers and smokers, i.e.,  $\bar{W}_0^c = \bar{W}_1^c = \bar{W}^c$ , which implies  $\bar{W}^c = \bar{W}^c$ , and  $\beta_t$  is again estimable using standard ecologic data on counties. Note that if  $P_1^1 \neq P_1^2$  then the ecologic regression will be biased due to the usual between-county confounding,

although this confounding can be removed by covariate adjustment.

If neither of the above situations apply, then  $\bar{W}^c \neq \bar{W}^c$  and the estimate of  $\beta_t$  will be biased. Consider a situation similar to Cohen's results in which a positive exposure-response relationship for individuals reverses to a negative relationship at the county level. Suppose the lung cancer rate for county 1 is less than the lung cancer rate in county 2, while the mean (observed) radon concentration in county 1 is greater than the mean in county 2. This is expressed symbolically as (1)  $r^1 < r^2$  and (2)  $\bar{W}^1 > \bar{W}^2$ .

Assume  $r_0$ ,  $\theta$ , and the probability of being a smoker,  $P_1^1 = P_1^2 = P_1$ , are the same in both counties. Under these assumptions, eqn (2) implies that condition (1)  $r^1 < r^2$  is equivalent to the condition  $\bar{W}^1 < \bar{W}^2$ , which can be rewritten as  $\lambda_1(\bar{W}_1^1 - \bar{W}_1^2) > \lambda_0(\bar{W}_0^1 - \bar{W}_0^2)$ . Similarly, condition (2) can be expressed as  $P_1(\bar{W}_1^2 - \bar{W}_1^1) < P_0(\bar{W}_0^2 - \bar{W}_0^1)$ . Under these inequalities, a positive trend in individuals will reverse and become a negative trend at the county-level if eqn (3) with  $\bar{W}^c$  is fit. When  $\theta > 1$ ,  $\lambda_1 > P_1$ . After a little manipulation, conditions (1) and (2) imply  $\bar{W}_1^2 > \bar{W}_1^1$  and  $\bar{W}_0^2 < \bar{W}_0^1$ . For two counties, the bias,  $B$ , ( $=\beta_e/\beta_t$ ) can be expressed as

$$\begin{aligned}B &= \frac{\bar{W}^2 - \bar{W}^1}{\bar{W}^2 - \bar{W}^1} \\ &= \frac{\lambda_0(\bar{W}_0^2 - \bar{W}_0^1) + \lambda_1(\bar{W}_1^2 - \bar{W}_1^1)}{P_0(\bar{W}_0^2 - \bar{W}_0^1) + P_1(\bar{W}_1^2 - \bar{W}_1^1)} \\ &= \frac{\lambda_0 + \lambda_1 \times \Delta}{P_0 + P_1 \times \Delta},\end{aligned}$$

where  $\Delta = (\bar{W}_1^2 - \bar{W}_1^1)/(\bar{W}_0^2 - \bar{W}_0^1)$ . While eqn (2) is linear in  $\bar{W}$  with parameter  $\beta_t$ , this expression shows that for the ecologic regression in  $\bar{W}$  the exposure-response parameter  $\beta_e = B \times \beta_t$  is a complex non-linear function that depends through  $\Delta$  on the levels of radon. Fig. 1 plots the bias for  $P_1 = 0.2, 0.8$  and  $\theta = 2, 15$ , with the dotted line representing no bias  $B = 1$ . The figure shows that the ecologic bias can be of any magnitude while reversal of trend occurs in a narrow range of values which depend on the proportion exposed to the confounding factor. For two counties, numerical examples of reversals in trend are given in the next section.

Interpretation simplifies if we assume that smoking and radon concentration are independent in county 2, which implies  $\bar{W}^2 = \bar{W}^2$ . The two conditions for reversal of trend can be expressed as  $\bar{W}^1 < \bar{W}^2 < \bar{W}^1$ . The inequality  $\bar{W}^1 < \bar{W}^1$  is equivalent to

$$\lambda_0 \bar{W}_0^1 + \lambda_1 \bar{W}_1^1 < P_0 \bar{W}_0^1 + P_1 \bar{W}_1^1.$$

Since  $\lambda_1 > P_1$ , the expression implies  $\bar{W}_1^1 < \bar{W}_0^1$ . Thus, when smoking and radon are independent in county 2, a reversal of the regression trend occurs if mean radon in

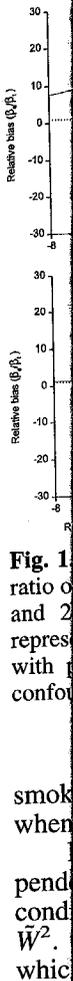
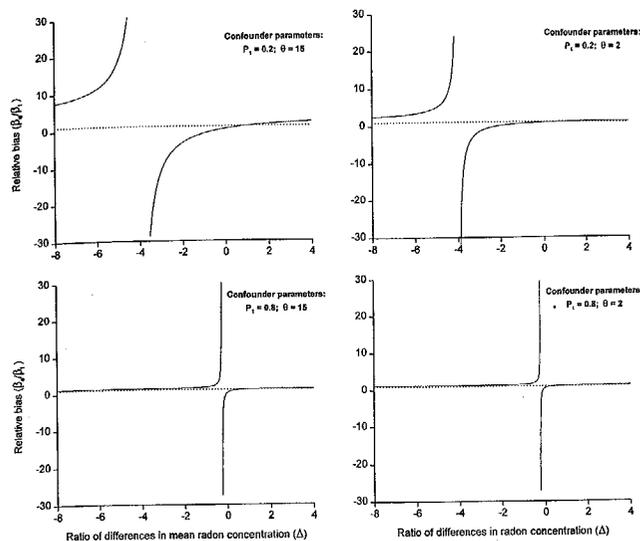


Fig. 1. ratio of  $\beta_e$  and  $\beta_t$  represent with confounding factor.

smoker when  $\bar{W}^2 < \bar{W}^1$  and  $\bar{W}_0^2 < \bar{W}_0^1$  which



**Fig. 1.** Relative bias ( $B = \beta_e/\beta_r$ ) in ecologic regression and the ratio of the differences in mean radon concentration for counties 1 and 2 for smokers and nonsmokers (see text). Dotted lines represent no bias,  $B = 1$ . Confounding factor is assumed binary, with panels showing bias for different proportions having the confounder ( $P_1$ ) and relative risks for the confounder ( $\theta$ ).

smokers in county 1 is less than in nonsmokers, i.e., when smoking and radon are negatively correlated.

In contrast, suppose smoking and radon are independent in county 1, implying  $\bar{W}^1 = \bar{W}_0^1$ . The two conditions for reversal can be expressed as  $\bar{W}^2 < \bar{W}^1 < \bar{W}_0^2$ . The inequality  $\bar{W}^2 < \bar{W}_0^2$  reduces to  $\bar{W}_0^2 < \bar{W}_1^2$ , which indicates that reversal in trend occurs if mean

radon level in county 2 is greater in smokers than in nonsmokers, i.e., smoking and radon are positively correlated in county 2.

These two illustrations indicate that the sign of the overall correlation between smoking and radon within county is not the determinant of a reversal in the regression. The Appendix shows that conditions for reversal are determined by the magnitude of the difference in the correlation coefficients between smoking and radon within the counties.

## EXAMPLES

The following numerical examples show that reversals can occur in practical situations. We assign parameters the following values,  $r_o = 0.0005$ ,  $P_1^1 = P_1^2 = P_1 = 0.4$ ,  $\theta = 15$ , and  $\beta_r = 0.00146$  per Bq m<sup>-3</sup> (for 30 y at a constant exposure rate, or equivalently, 0.010 per Working Level Month). Radon concentration in U.S. homes is approximately log-normally distributed with geometric mean (GM) 25 Bq m<sup>-3</sup> and geometric standard deviation (GSD) 3.0, with concentrations lowest in the Pacific Northwest region (GM = 13 Bq m<sup>-3</sup>) and highest in the Upper Plains region (GM = 60 Bq m<sup>-3</sup>) (Marcinowski et al 1994). These values serve as the basis for the values in Table 1.

In Table 1 the lung cancer rate,  $r$ , is lower in county 1 than in county 2 (since the value for  $\bar{W}$  is lower in county 1), but the average radon level for county 1 is greater than county 2 (column labeled  $\bar{W}$ ). Consider set A. In county 1, radon and smoking are negatively correlated, with mean radon level greater in nonsmokers ( $\bar{W}_0^1 = 55$  Bq m<sup>-3</sup>) than in smokers ( $\bar{W}_1^1 = 42$  Bq m<sup>-3</sup>).

**Table 1.** Examples of a true positive trend for radon and lung cancer rates for individuals ( $\beta_r$ ) reversing to a negative trend for lung cancer rates<sup>a</sup> at the county level ( $\beta_e$ ), resulting in a negative estimate of the excess relative risk.<sup>b</sup>

Set	County	Nonsmokers			Smokers			Corr[W,S]	$\bar{W}$	$\bar{W}$	$r \times 1,000$	$\beta_e \times 1,000$
		GM	GSD	$\bar{W}_0$	GM	GSD	$\bar{W}_1$					
A	1	29	3.1	55	25	2.8	42	-0.08	44	50	3.51	
	2	25	3	46	25	3	46	0.00	46	46	3.52	-0.70
B	1	17	2.3	24	13	2.2	18	-0.14	18	22	3.39	
	2	13	2.6	21	13	2.6	21	0.00	21	21	3.40	-3.21
C	1	63	2.7	103	55	2.7	90	-0.05	91	98	3.74	
	2	60	2.6	95	60	2.6	95	0.00	95	95	3.76	-1.56
D	1	29	2.7	47	25	2.7	41	-0.05	42	45	3.50	
	2	25	2.8	42	25	3	46	0.03	45	44	3.52	-5.14
E	1	29	3	53	25	2.4	37	-0.12	38	46	3.48	
	2	25	3	46	25	2.6	39	-0.05	40	43	3.49	-0.83
F	1	25	2.9	44	25	3	46	0.01	46	45	3.52	
	2	25	2.6	39	30	2.8	51	0.10	50	44	3.54	-9.68

<sup>a</sup> The lung cancer rates for counties are determined from eqn (2), with a probability of smoking of 0.4, a smoking relative risk of 15, a true excess relative risk of 0.00146 Bq m<sup>-3</sup> (corresponding to 0.010/Working Level Month), and a lung cancer rate among nonsmokers with ambient radon exposure of 0.0005. Exposures are based on 30-y residence at a constant exposure rate, under standard occupancy and equilibrium assumptions. Radon is assumed log-normally distributed and specified by the geometric mean (GM) and geometric standard deviation (GSD). Radon values are based on a national survey (Marcinowski et al. 1994). Correlation coefficients for radon and smoking, corr[W,S], are shown.

<sup>b</sup> Mean radon for a county is derived from the weighted average of the mean radon concentration for nonsmokers  $\bar{W}_0$  and smokers  $\bar{W}_1$ . True county rates ( $r$ ) are based on the mean  $\bar{W}$  using weights that account for differential risks of lung cancer by smoking status. The biased regression estimate  $\beta_e$  is based on the simple overall mean  $\bar{W}$  and computed as  $(r^2 - r^1)/[0.0005 \times (0.6 + 15 \times 0.4) \times (\bar{W}^2 - \bar{W}^1)]$  where superscripts refer to county 1 and 2.

The "smoking-risk-adjusted" mean radon for the county is  $\bar{W}^1 = 44 \text{ Bq m}^{-3}$ , resulting in a lung cancer rate for county 1 of  $3.51/1,000 [=0.0005 \times (0.6 + 15 \times 0.4)(1 + 0.00146 \times 44)]$ . In county 2 smoking and radon are independent with  $\bar{W}_0^2 = \bar{W}_1^2 = \bar{W}^2 = 46 \text{ Bq m}^{-3}$  and the lung cancer rate is  $3.52/1,000$ . The average radon levels are  $\bar{W}^1 = 50 \text{ Bq m}^{-3}$  and  $\bar{W}^2 = 46 \text{ Bq m}^{-3}$  in counties 1 and 2, respectively, resulting in a negative slope,  $\beta_e = -0.0007 \text{ Bq m}^{-3}$ .

Condition (1)  $\bar{W}^1 < \bar{W}^2$  can be rewritten as  $\lambda_0 \bar{W}_0^1 + \lambda_1 \bar{W}_1^1 < \bar{W}^2$ , while condition (2)  $\bar{W}^2 < \bar{W}^1$  can be rewritten as  $\bar{W}^2 < P_0 \bar{W}_0^1 + P_1 \bar{W}_1^1$ . Given values for  $\bar{W}_0^2$  and  $\bar{W}_1^2$  for county 2, the two conditions define a 2-dimensional region of values for  $\bar{W}_0^1$  and  $\bar{W}_1^1$  in which a true positive regression for individuals reverses to a negative regression in  $\bar{W}$  at the county level. The shaded area in Fig. 2 shows the reversal region for  $\bar{W}_0^1$  and  $\bar{W}_1^1$  for the radon values for county 2 in Table 1 set A. The asterisk denotes the specific set A values for county 1 in Table 1. The boundary of the shaded area identifies values where the county-level estimate of  $\beta_e$  is zero.

Table 1 illustrates that a reversal of the regression can occur when the correlation of radon and smoking within county 1 is negative and the correlation within county 2 is zero (sets A–C) or positive (set D), both correlations are negative (set E) or both positive (set F).

## DISCUSSION

Correlations of radon with other risk factors, such as smoking, within a county can bias ecologic regression

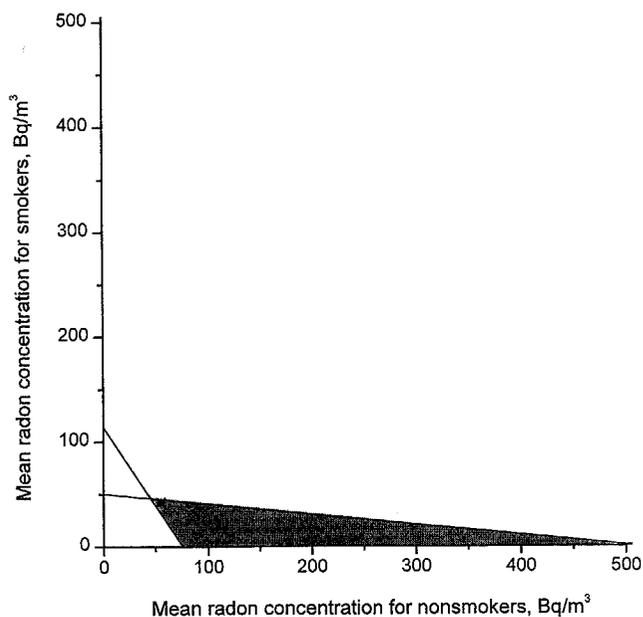


Fig. 2. Shaded area is the region of all values for mean radon level for smokers-response ( $\beta_s$ ) for radon and lung cancer rate for individual exposure results in a negative ecologic regression estimate ( $\beta_e$ ). Radon and smoking assumed independent in county 2. The asterisk denotes the mean values for county 1 from set A Table 1.

estimates of the exposure-response relationship for individuals. This paper demonstrates that bias arises because the simple average radon concentration for a county does not account for the different contributions to county-level risk from subgroups (smokers and nonsmokers) within the county. Moreover, county-level variables, including "stratification" or interaction variables, cannot eliminate the regression bias; for example, in our presentation inclusion of  $P_1 \times W$  does not eliminate the bias, i.e.,  $\bar{W}$  cannot be turned into  $\bar{W}$ . Unbiased ecologic studies can potentially be carried out, but in our example only if data on the joint distribution of radon and smoking within county are available, either from more detailed county data or from secondary sources such as a population survey, or if radon and other risk factors are independent. Methods for unbiased analysis of aggregated data are considered by Sheppard et al. (1996) and Prentice and Sheppard (1995).

In the examples, smoking was the confounder which was assumed correlated with radon level within county, but similar considerations apply to all lung cancer risk factors. For example, age is a potent risk factor for lung cancer. The U.S. lung cancer mortality rate for ages 70–74 y is over 20 times the rate for ages 40–44 y (NRC 1988). Thus, ecologic regression analysis could be biased if age and radon concentration were correlated within counties, which is quite likely since radon level is related to housing characteristics (Marcinowski et al. 1994; Cohen 1991), income status and other factors (Cohen 1991) that vary with age. An unbiased ecologic analysis (using  $\bar{W}$  as a regressor variable) requires data on the joint distribution within county of radon and age (and smoking status). As with smoking, using age-adjusted lung cancer rates or stratifying counties by the proportions of the population in particular age groups does not address within county confounding by age. Accounting for age is further complicated by the observation that the effect of radon exposure for ages 75 y and over is about one-fifth the effect for ages under 55 y (Lubin et al. 1995). Cohen uses 1970–1979 lung cancer mortality rates, and radon concentration data from the late 1980's, potentially 20 y or more between disease-relevant exposure and effect. Lung cancer mortality rates have increased over time, 40% between 1973–1992 (Kosary et al. 1995), and thus calendar time may also bias Cohen's results, since indoor radon levels have varied over time (Swedjemark et al. 1987); i.e., radon and calendar year are likely correlated within county.

Eqn (1) defines a multiplicative association between radon exposure and smoking, and is consistent with available miner data (NRC 1988; Lubin et al. 1995). Because excess lung cancer risk from radon depends on smoking status, a similar bias in ecologic regression arises if the joint risk for radon and smoking were intermediate between additive and multiplicative, an association which is consistent with available data (Lubin et al. 1995). If, however, the joint effects were additive, i.e.,  $r(s, w) = r_0(\theta^s + \beta w)$ , then, corresponding to eqn (2), the lung cancer disease rate averaged over

all  
 $\beta_e$   
with  
corr  
cons  
and

expo  
eqn  
the  
line  
does  
nam  
intri  
mod  
 $\beta_e$   
edly  
mag  
thre  
true  
the  
are  
are

situ  
addi  
and  
error  
betw  
quite  
C an  
of p  
corr  
with  
radon  
large  
F in  
0.4,  
occu

dealt  
only  
plaus  
betw  
For t  
the p  
syste  
1 and  
corre  
found  
count  
a rev  
corre  
corre  
with  
betw  
with  
have  
1991

all residents of a county is  $r^c = r_0[(P_0^c + \theta P_1^c) + \beta_r \bar{W}^c]$ . Thus, unbiased estimates of  $\theta$  and  $\beta_r$  are possible with ecologic data even if smoking and radon were correlated within county, although this model is not consistent with miner data (Lubin et al. 1995; Hornung and Meinhardt 1987).

A linear no-threshold model implies that reducing exposure reduces risk. Eqn (1), the induced county-level eqn (2), and the ecologic regression eqn (3) (similar to the eqn used in Cohen's analysis) are all examples of linear no-threshold models. While Cohen's approach does in fact result in an unbiased estimate of effect, namely  $\beta_e$ , the estimated quantity is not a parameter of intrinsic interest. His model is a linear no-threshold model in  $\bar{W}$  with parameter  $\beta_e$ , not in  $\tilde{W}$  with parameter  $\beta_r$ . Moreover, Fig. 1 indicates that  $\beta_e$  may differ markedly from the true  $\beta_r$  of eqn (1). Since the direction and magnitude of bias are unknown, Cohen's linear no-threshold analysis is uninformative with respect to the true exposure-response relationship in  $\tilde{W}$ , except under the unlikely conditions that other lung cancer risk factors are independent of radon or their joint effects with radon are additive.

The examples in this paper represent a simple situation, with a single age group, no mobility, only one additional risk factor, and all information on exposures and disease outcome characterized precisely without error. Reversals in trend occurred when the correlations between smoking status and radon within county were quite small, in the range  $-0.05$  to  $0.05$  (see Table 1, sets C and D). In an actual ecologic analysis, there is a myriad of potential risk factors for lung cancer that may be correlated with radon level and which may be measured with great uncertainty. Because the expected risk from radon is small, the effects of confounders need not be large to bias results. For example, using the values of set F in Table 1 for  $\bar{W}_0^1$ ,  $\bar{W}_1^1$ ,  $\bar{W}_0^2$  and  $\bar{W}_1^2$ , and with  $P_1 = 0.4$ , a reversal of trend when  $\bar{W}$  is used rather than  $\tilde{W}$  occurs with 40% excess risk, i.e.,  $\theta = 1.4$ .

The examples were further limited because they dealt with the exposure-response trend for two counties only. With two counties, a reversal in trend could plausibly result from random variation in the association between radon and confounding factors within counties. For the examples to apply to ecologic studies in general, the patterns of correlations within counties must apply systematically to large numbers of counties. From Table 1 and the Appendix, reversals in trend occurred when the correlation between radon concentration and a confounder within county was greater in the higher rate county. Thus, in an ecologic analysis, results suggest that a reversal of trend is possible when there is a positive correlation between county lung cancer rate and the correlation of radon and other lung cancer risk factors within county (or when there is a negative correlation between the lung cancer rate and the correlation of radon with protective factors). Examples of reversal of trend have been published for national data (Muirhead et al. 1991; Piantadosi et al. 1988), for simulate data (Stidley

and Samet 1994), and for a hypothetical example (Greenland and Robins 1994).

The negative regressions by Cohen (1995) are at odds with the overwhelming evidence from epidemiological and biological studies and are likely the result of confounding within county. Although the factors responsible for the bias are unknown, it seems likely that within county confounding from smoking and age may play a role, although many other factors could be involved. While ecologic studies have proved useful in epidemiology in identifying possible disease associations, they are widely recognized as only a first step in an etiologic investigation, which must be followed with analytic studies. Fundamental problems with the ecologic approach limit its usefulness to hypothesis generation and preclude its use for general hypothesis testing, including testing a linear no-threshold model. Ecologic studies that contradict both biologically plausible models of disease causation and epidemiologic studies should be rejected.

*Acknowledgements*—I wish to thank Drs. William Field, Ethel Gilbert and Charles Land, and Mr. Brian Smith for useful discussions of this and related topics and comments by Duncan Thomas and a reviewer which greatly improved the paper.

## REFERENCES

- Brenner, H.; Savitz, D. A.; Jockel, K.-H.; Greenland, S. Effects of nondifferential misclassification in ecologic studies. *Am. J. Epidemiol.* 135:85-95; 1992.
- Cohen, B. L. Test of the linear no-threshold theory of radiation carcinogenesis for inhaled radon decay products. *Health Phys.* 68:157-174; 1995.
- Cohen, B. L. Variation of radon levels in U.S. homes correlated with house characteristics, location, and socioeconomic factors. *Health Phys.* 60:631-642; 1991.
- Cohen, B. L. Problems in the radon vs. lung cancer test of the linear no-threshold theory and a procedure for resolving them. *Health Phys.* 72:623-628; 1997.
- Greenland, S. Divergent biases in ecologic and individual-level studies. *Stat. Med.* 11:1209-1223; 1992.
- Greenland, S.; Robins, J. Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. *Am. J. Epidemiol.* 139:747-760; 1994.
- Hornung, R. W.; Meinhardt, T. J. Quantitative risk assessment of lung cancer in U.S. uranium miners. *Health Phys.* 52:417-430; 1987.
- Kosary, C. L.; Ries, L. A. G.; Miller, B. A.; Hankey, B. F.; Harras, A.; Edwards, B. K. SEER Cancer Statistics Review, 1973-1992: Tables and graphs. Bethesda, MD: National Cancer Institute; NIH Pub. No. 96-2789; 1995.
- Lubin, J. H.; Boice Jr., J. D. Lung cancer risk from residential radon: meta-analysis of eight epidemiologic studies. *J. Natl. Cancer Inst.* 89:49-57; 1997.
- Lubin, J. H.; Boice Jr., J. D.; Edling, C.; Hornung, R. W.; Howe, G.; Kunz, E.; Kusiak, R. A.; Morrison, H. I.; Radford, E. P.; Samet, J. M.; Tirmarche, M.; Woodward, A.; Yao, S. X.; Pierce, D. A. Lung cancer in radon-exposed miners and estimation of risk from indoor exposure. *J. Natl. Cancer Inst.* 87:817-827; 1995.

- Marcinowski, F.; Lucas, R. M.; Yeager, W. M. National and regional distributions of airborne radon concentrations in U.S. homes. *Health Phys.* 66:699-706; 1994.
- Morgenstern, H. Ecologic studies in epidemiology: concepts, principles, and methods. *Annual Rev. Public Health* 16:61-81; 1995.
- Muirhead, C. R.; Butland, B. K.; Green, B. M. R.; Draper, G. J. Childhood leukaemia and natural radiation (letter). *Lancet* 337:503-504; 1991.
- National Research Council. Committee on the Biological Effects of Ionizing Radiation. Health effects of radon and other internally deposited alpha emitters (BEIR IV). Washington, DC: National Academy Press; 1988.
- National Research Council. Committee on the Biological Effects of Ionizing Radiation. Health effects of exposure to radon (BEIR VI). Washington, DC: National Academy Press; 1998.

- Piantadosi, S.; Byar, D. P.; Green, S. B. The ecologic fallacy. *Am. J. Epidemiol.* 127:893-904; 1988.
- Prentice, R. L.; Sheppard, L. Aggregate data studies of disease risk factors. *Biometrika* 82:113-125; 1995.
- Sheppard, L.; Prentice, R. L.; Rossing, M. A. Design considerations for estimation of exposure effects on disease risk, using aggregate data studies. *Stat. Med.* 15:1849-1858; 1996.
- Stidley, C. A.; Samet, J. M. Assessment of ecologic regression in the study of lung cancer and indoor radon. *Am. J. Epidemiol.* 139:312-322; 1994.
- Swedjemark, G. A.; Buren, A.; Mjones, L. Radon levels in Swedish homes: Comparison of the 1980s with the 1950s. In: Radon and its daughter products: the importance, properties, and health effects. Washington, DC: American Chemical Society; Symposium Series 331; 1987: 84-96.

## APPENDIX

IN THIS appendix, we present the conditions for reversal of the regression of lung cancer risk for individuals on radon progeny exposure, eqn (1), when data are aggregated at the county-level, eqn (2), assuming the background risk  $r_0$  and probability of smoking  $P_1$  are the same in both counties. Using first principles, the correlation coefficient for radon level ( $W$ ) and smoking ( $S$ ) in county  $c$ , denoted  $\text{corr}^c[W, S]$ , can be expressed as:

$\text{corr}^c[W, S]$

$$= \frac{\bar{W}_1^c - \bar{W}_0^c}{\sqrt{\frac{\text{Var}[W_0^c]}{P_1} + \frac{\text{Var}[W_1^c]}{P_0} + \{\bar{W}_1^c - \bar{W}_0^c\}^2}}, \quad (\text{A1})$$

where  $\text{Var}[W_s^c]$  is the variance of radon exposure for smoking status  $s$  within county  $c$ .

Condition (1),  $\bar{W}^1 < \bar{W}^2$ , and condition (2),  $\bar{W}^1 > \bar{W}^2$ , can be reexpressed as the two inequalities

$$\frac{\bar{W}_0^1 - \bar{W}_0^2}{\lambda_1} < (\bar{W}_1^2 - \bar{W}_0^2) - (\bar{W}_1^1 - \bar{W}_0^1) < \frac{\bar{W}_0^1 - \bar{W}_0^2}{P_1}. \quad (\text{A2})$$

Setting  $K^c$  equal to the denominator in  $\text{corr}^c[W, S]$  for county  $c$  above, the conditions can be rewritten as

$$\frac{\bar{W}_0^1 - \bar{W}_0^2}{\lambda_1} < K^2 \text{corr}^2[W, S] - K^1 \text{corr}^1[W, S] < \frac{\bar{W}_0^1 - \bar{W}_0^2}{P_1}. \quad (\text{A3})$$

Abstr  
cont  
indiv  
disea  
unit  
meas  
each  
not t  
disea  
inclu  
atten  
no-th  
gate  
that,  
cause  
relati  
linear  
lung  
from  
the n  
ual a  
by di  
the p  
conce  
of the  
bias a  
therm  
rates  
large  
cance  
Healt  
Key

OVER  
since

\*  
Enviro  
City,  
F  
above  
(  
16 Oc  
00  
C