Journal of Epidemiology and Community Health 1995;**49**(Suppl 2):S72–S77

Smoking behaviour can be predicted by neighbourhood deprivation measures

Immo Kleinschmidt, Michael Hills, Paul Elliott

Abstract

***Study objective* – To assess whether small area measures of socioeconomic deprivation predict variation in individual smoking behaviour. To examine the adequacy of an individual level statistical model for the analysis of data on groups of individuals who live in the same geographical area.**

***Design* – Individual level and two level logistic regression analysis of data on individual smoking from a regional health survey, and neighbourhood deprivation scores for 1991 census wards calculated from 1991 census data.**

***Setting* – The North West Thames Regional Health Authority area.**

***Participants* – Random sample of 8251 adults in North West Thames Region.**

***Main results* – There was a highly significant association between being a smoker and the neighbourhood deprivation score of the area of residence. With the two level model, after allowing for age and sex, the estimated odds ratio of being a smoker for an individual in the highest quintile of deprivation compared with someone in the lowest quintile was 1·52 (95% confidence interval 1·33, 1·74). Results obtained using the individual level model were similar. Variation between wards accounted for around 6% of the total variation in smoking behaviour after neighbourhood deprivation of the ward had been taken into account. Deprivation of the area of residence remained a significant predictor of smoking status even after the socioeconomic group of the individual had been taken into account.**

***Conclusions* – Neighbourhood deprivation of the area of residence is a predictor of smoking status of individuals. In this example the two level model was reasonably well approximated by the individual level model.**

(*J Epidemiol Comm Health* 1995;**49**(Suppl 2):S72–S77)

Since the publication of the Black report,[1] many studies have reported an association between mortality and measures of socioeconomic status, including employment grade,[2 3] social class,[4] and an index of social deprivation.[5] Often the association is greatest for smoking related diseases. For example, figure 1 shows the relationship between standardised lung cancer incidence ratios and the Carstairs deprivation score[6] for electoral wards in the North West Thames Regional Health Authority from 1975 to 1986. The higher the score, the more deprived the area. On average, there is an ap-

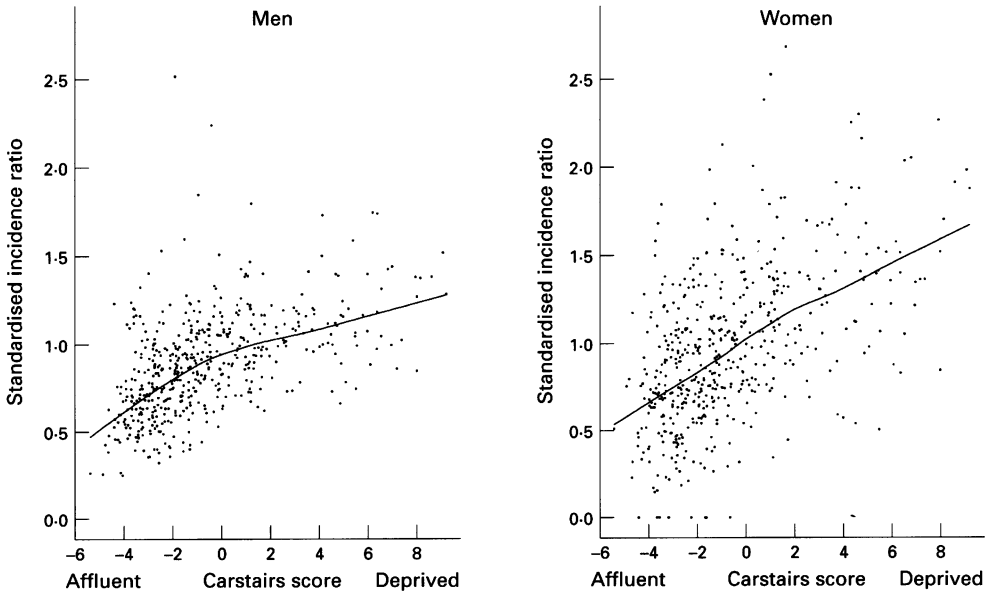

Figure 1 Standardised lung cancer incidence ratios in comparison with Carstairs deprivation scores for wards in the North West Thames Region (1975–1986). The dots are for individual wards. The smoothed lines represent median incidence ratios at each score.

Small Area Health Statistics Unit, Environmental Epidemiology Unit, Department of Public Health and Policy
I Kleinschmidt
P Elliott

and Medical Statistics Unit, Department of Epidemiology and Population Sciences
M Hills

London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT

Correspondence to:
I Kleinschmidt.

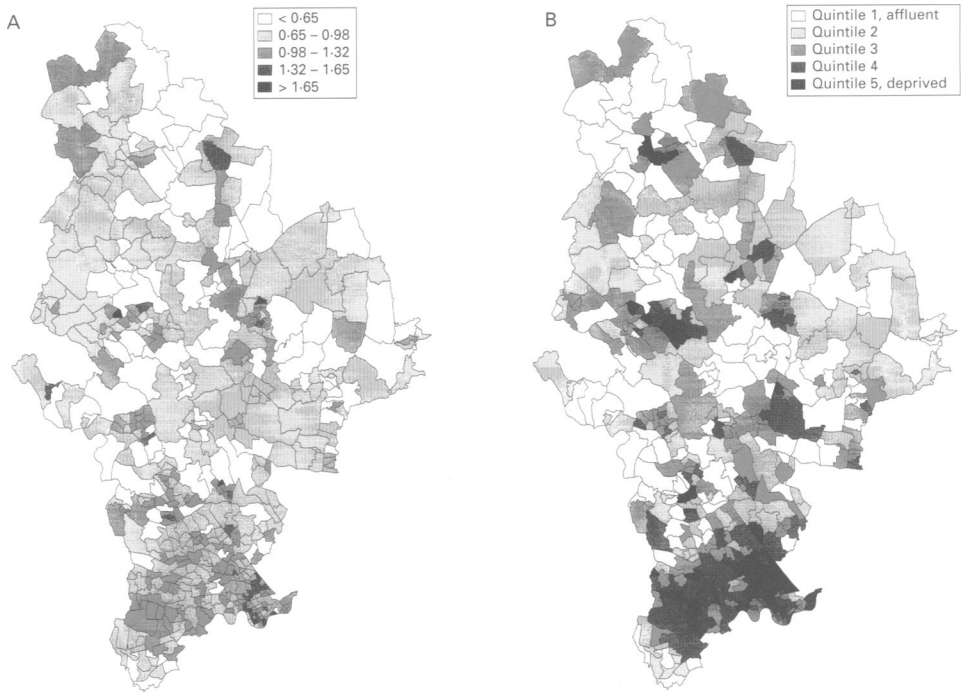

Figure 2 North West Thames Regional Health Authority. (A) Standardised incidence ratios for lung cancer by electoral ward, 1975–86 and (B) Carstairs deprivation score for 1991.

proximately threefold variation in the risk of lung cancer for the most deprived compared with the least deprived areas. This relationship is shown in map form in figure 2: the geographical distribution of lung cancer incidence by ward is shown in figure 2A, and figure 2B gives the distribution of deprivation scores.

The relationship between smoking and disease is well established.[7–9] The *Health of the Nation* report committed the British Government to reducing smoking prevalence in the population by one third, and cigarette consumption by at least 40% by the year 2000.[10] However, the prevalence of cigarette smoking varies greatly with socioeconomic group, ranging from 16% for men and women in the professional classes, to 48% for men and 36% for women in the unskilled manual group.[11] In the Whitehall study of around 17 000 civil servants studied in 1967–69, smoking prevalence varied by a factor of two between the lowest and highest employment grades.[12]

In studies of individuals, data on individual smoking behaviour can be used to adjust for the potential confounding effects of smoking, but in ecological studies data on smoking are usually unavailable. For example in small area studies of risks associated with point sources of environmental pollution carried out by the Small Area Health Statistics Unit (SAHSU), area based measures of socioeconomic deprivation are used as a means of correcting for the aggregate effect of socioeconomic and behavioural factors such as smoking.[13] It is not therefore possible to assess separately the potential impacts of smoking and social deprivation in these analyses.

In the present study, data on smoking behaviour of individuals were used to examine the relationship of smoking to area based deprivation scores, giving an estimate of smoking prevalence from the distributions of age, sex, and deprivation scores of a population. A secondary objective was to investigate the appropriateness of the statistical model since the analysis involved the use of individual as well as aggregate variables.

Methods

Data on individual smoking behaviour were obtained from the North West Thames regional health survey. The survey, commissioned by the North West Thames Regional Health Authority, was carried out in 1990 based on a random sample of addresses selected from the postal address file. The response rate overall was 64%, giving a total of 8251 responses collected by face to face interviews. Topics covered in the questionnaire included satisfaction with local health services, self reported health, health related behaviour (smoking, drinking, eating, exercise), housing conditions, income, education, ethnicity, and language. For each respondent, the 1981 census ward of residence was known. The number sampled per ward varied from zero to 55.

The measure of deprivation used here is the Carstairs score,[6] which was calculated for all wards in North West Thames using 1991 census data. The four components of the score are the proportion of male unemployment, proportion of people living in overcrowded households, proportion of people in social classes IV and V, and proportion of people in households without access to a car. Each component of the score was standardised across Great Britain to have zero mean and unit variance.

Since the survey was carried out before the 1991 census, the residence of respondents was recorded to the 1981 electoral ward. The deprivation scores used here, however, were more appropriately based on census data for 1991. Of the 514 wards, 498 could be matched directly to 1991 wards. The remaining 16 wards (3%), could not be matched to 1991 wards, resulting in the exclusion of 385 out of the 8251 responses. Responses among the two groups, however, were similar.

STATISTICAL METHODS

Two different statistical models were used: a two level hierarchical model and a single level (individual) model. The two level model takes into account the hierarchical nature of the data, whereby some variables apply to individuals and others to the wards in which they live. Thus, two components of variance are specified – one due to variability between individuals and one to variability between wards.[14 15] In the single level model the hierarchical error structure in the data is ignored and higher level data are disaggregated to individuals.

In a logistic regression model, current smoking of an individual (yes/no) was specified as the response variable, with age and sex of the individual and deprivation of the ward as explanatory variables, using a logit link function.[16] A quadratic term for age was included as it was found to improve the fit. Specifically,

$$y_{iw} = \frac{\exp(\alpha + Age_{iw} + Age_{iw}^2 + Sex_{iw} + \beta x_w + e_w)}{1 + \exp(\alpha + Age_{iw} + Age_{iw}^2 + Sex_{iw} + \beta x_w + e_w)} + e_{iw}$$

where y_{iw} (1 = smoker, 0 = non-smoker) is the response for individual i, who lives in ward w, Age_{iw}, Age_{iw}^2 and Sex_{iw} denote terms for individual age, age^2 and sex (1 = female), x_w represents the deprivation score of the ward and β the coefficient for the deprivation score.

The first term on the right hand side of this equation represents the probability of the respondent being a smoker and the term e_{iw} is the individual level residual variation, unexplained by the model with $e_{iw} \sim N(0, \sigma_i^2)$ where σ_i^2 is the individual level variance. The second residual error term e_w is the ward-level residual (i.e. the same for all individuals in ward w), with $e_w \sim N(0, \sigma_w^2)$ where σ_w^2 is the between-ward variance. The proportion of total variation in smoking that is given by variation between wards is indicated by the fraction

$$\frac{\sigma_w^2}{(\sigma_i^2 + \sigma_w^2)},$$

known as the intraclass correlation.[17]

Socioeconomic group of the individual (a 17 level categorical variable) was also added to the model to see whether any effect of socioeconomic deprivation of area of residence could be accounted for by differences in the socioeconomic group of individuals.

In addition, interaction terms were added to the model to test whether there was significant interaction between age and deprivation score, and between sex and deprivation score.

Using the individual level model implies that σ_w^2 is zero, so that random variation is restricted to binomial variation between individuals. The area based deprivation score is disaggregated to individuals. Standard logistic regression is therefore carried out.

It is possible that the outcome measure, smoking status, is spatially correlated due to some unmeasured covariate. Smoking prevalence in contiguous areas may be similar to a degree that cannot fully be accounted for by the known covariates. Such spatial pattern in the data can be examined,[18] but it requires a reasonably stable measure of smoking prevalence in each ward. In the given data set the number of observations per ward varied between none and 55, and consequently any estimate of smoking prevalence per ward is subject to large random fluctuation. For this reason no test for spatial pattern in the data was carried out.

Results

Table 1 shows results of the two logistic regression analyses. The risk of smoking was lower in females than in males and reduces with age, as has been shown in national surveys.[11] In both the individual level and two level models there was a highly significant association between individual smoking status and socioeconomic deprivation of the ward of residence. After allowing for age and sex, the odds ratio of being a smoker for an individual living in a ward at the midpoint of the top quintile of deprivation (Carstairs score = 4·1), compared with an individual living in a ward at the midpoint of the bottom quintile (score = −3·2) was 1·52 (95% C.I. 1·33, 1·74). Interactions between age and deprivation score, and sex and deprivation score were found not to be statistically significant.

Figure 3 shows smoking prevalence, as predicted by the two level model, compared with the Carstairs deprivation score at ages 20, 45, and 70 years for women and men (dotted lines show the 95% confidence interval). A comparison of the results obtained by the two models shows close agreement between the coefficients of the models, although, as expected, the standard error for the grouped variable (Carstairs) is somewhat larger for the two level model. For the Carstairs variable, the difference in standard errors was 12%. The intraclass correlation for smoking in these data was about 10% overall, which reduced to 6·3%

*Table 1 Logistic regression coefficients and p values for the individual level and two level models, and intraclass correlation coefficient**

	Individual level model		*Two level model*	
Parameter	*Value (SEM)*	*p value*	*Value (SEM)*	*p value*
Sex (F)	−0·119 (0·051)	0·02	−0·122 (0·051)	0·02
Age (y)	0·0115 (0·008)	0·15	0·0129 (0·008)	0·11
AgeSQ	−0·00033 (0·000080)	<0·0001	−0·00034 (0·000081)	<0·0001
Carstairs (per unit)	0·0558 (0·0082)	<0·00001	0·0579 (0·0093)	<0·00001
$\frac{\sigma_w^2}{(\sigma_i^2+\sigma_w^2)}$			0·063	

* For definition of terms, see text.

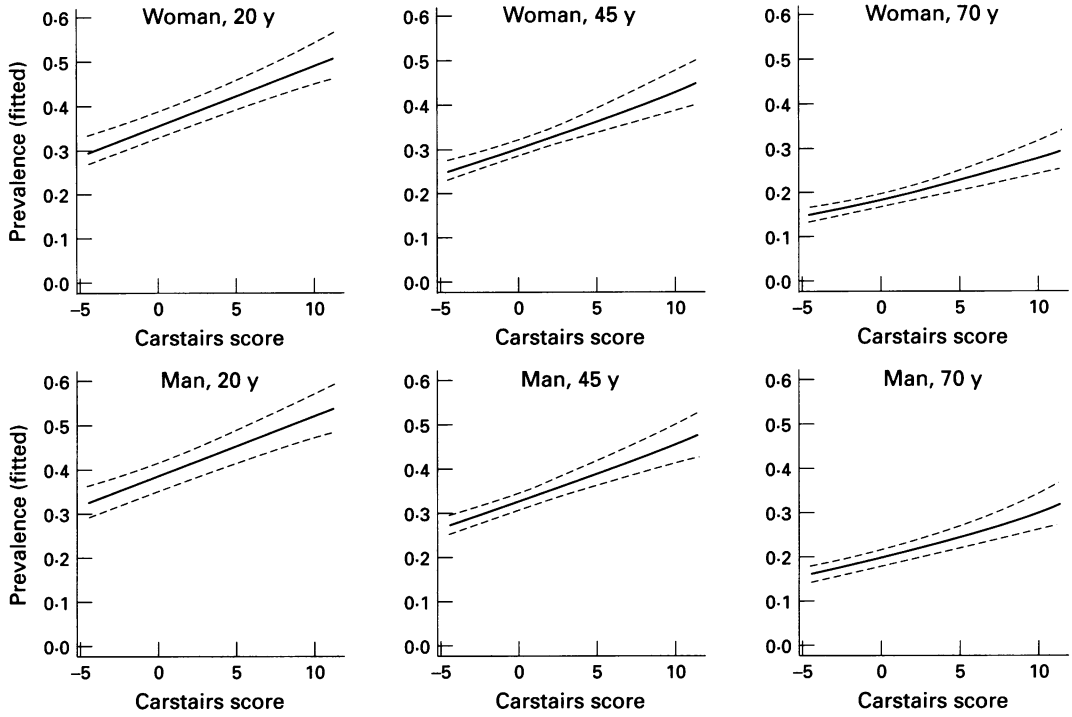

Figure 3 Predicted smoking prevalence in comparison with Carstairs deprivation score in relation to sex and age.

after modelling indicating that there was some residual, unexplained interward variability.

Table 2 shows that the socioeconomic group of the respondent provided significant additional explanation for the variation in smoking behaviour after allowing for the Carstairs deprivation score. Although there is need for caution in adding socioeconomic group of the individual to a model already containing deprivation as a covariate, there was no evidence here to suggest collinearity, since the standard error of the deprivation term was virtually unchanged after adding socioeconomic group. With socioeconomic group in the model, the estimate for the Carstairs coefficient was reduced, but both terms are highly significant.

The coefficients of the categorical socioeconomic group terms showed, for example, that the odds ratio of being a smoker for an unskilled manual worker (group 11) compared with a person in the professional group (group 4) was 3·5 after allowing for age, sex, and Carstairs deprivation score.

A random coefficients regression model[14] showed no significant random variation between wards in the coefficients for age, sex, or socioeconomic group.

Table 3 illustrates the combined predicted effect of ward level Carstairs score and individual level socioeconomic group on smoking prevalence in men and women aged 40 years. Within each quintile of Carstairs score there is a strong effect of socioeconomic group with lowest smoking rates found among professional workers. Likewise, within each socioeconomic group, there is higher smoking prevalence in the most deprived compared with the least deprived areas.

*Table 2 Logistic regression coefficients and p values for the individual level and two level models, with inclusion of term for individual level socioeconomic group (SEG), and intraclass correlation coefficient**

	Individual level model		*Two level model*	
Parameter	*Value (SEM)*	*p value*	*Value (SEM)*	*p value*
Sex (F)	−0·130 (0·052)	0·012	−0·132 (0·052)	0·012
Age (y)	0·0187 (0·008)	0·022	0·0194 (0·008)	0·018
AgeSQ	−0·000418 (0·00008)	<0·00001	−0·000425 (0·00008)	<0·00001
Carstairs	0·0387 (0·0085)	<0·00001	0·0408 (0·0093)	0·00001
SEG (17 level factored)		<0·00001		<0·00001
$\frac{\sigma_w^2}{(\sigma_i^2+\sigma_w^2)}$			0·0445	

* For definition of terms, see text.

Table 3 Predicted smoking prevalence (%) for three different socioeconomic groups (SEG) within Carstairs deprivation quintiles (age = 40 y)

	Carstairs quintile	*SEG 4 (employed professional workers)*	*SEG 7 (personal service workers)*	*SEG 11 (unskilled manual workers)*
Men	1	14·9	43·8	37·8
	5	19·0	51·2	44·9
Women	1	13·3	40·6	34·7
	5	17·1	47·9	41·7

Discussion

The main finding is that the Carstairs score, a measure of neighbourhood deprivation, was strongly predictive of smoking status of individuals. This association was independent of differences in individual socioeconomic status. Although such an association is often assumed,[19] to our knowledge it had not previously been examined directly. Our findings add further weight to small area analyses of environmental pollution and health, that, in the absence of available data on individuals,

attempts to adjust for socioeconomic surrounding by use of areal deprivation scores derived from census variables.

It would have been surprising if an association between deprivation and smoking were not found. Other studies that have measured deprivation of individuals, such as low income, lone parenthood, or lacking educational qualifications, have shown that variations in smoking prevalence mirror variations in social and material deprivation.[20] Census based socioeconomic deprivation scores are a cruder measure, however, even at the small area level; they nevertheless explain some of the geo-demographic variation in smoking prevalence. As an illustration, our results predict a smoking prevalence of 44% among men aged 20 years living in areas in the highest quintile of deprivation compared with a prevalence of 16% among women aged 70 years in areas in the lowest quintile of deprivation. In addition to its relevance to ecological studies of health and the environment, knowledge of predicted smoking prevalence by age and deprivation could usefully inform efforts to reduce smoking to meet the declared aim of a prevalence of 20% nationally by the year 2000.[10]

Similar results were obtained using the underprivileged area score (UPA)[21] for wards indicating that the findings were not an artefact of the choice of a particular deprivation index.

Conclusions from this study must be interpreted against a background of a 64% response rate. Deprivation of area of residence of the non-respondents was not available. If disproportionately more smokers who live in the less deprived areas of the region than those in the more deprived areas were non-respondents, the association between smoking and deprivation would have been overestimated. (The same would be true if disproportionately more non-smokers in deprived areas were non-respondents). Our data cannot provide evidence for or against such a bias. The overall smoking prevalence in the survey is similar to that found in national surveys[11] and that calculated for the North West Thames Region using the Acorn (CACI Ltd. Smoking data by electoral district type were aggregated to wards for the North West Thames area.) prevalence figures for small areas.

The study also addressed the question of the appropriateness of the statistical method when analysing data on groups of individuals who live in the same area – that is, is it reasonable to ignore the higher (area) level variance component? The geographical grouping of individuals studied here was electoral wards in and around London, which would tend to have a weaker identity than, say, in a small town or village. If instead of the ward the grouping were, for example, families or households, the analysis would almost certainly need to take into account intraclass correlations as members of the same family share many factors including genetic, dietary, lifestyle, housing, indoor pollution etc. In the case of electoral wards, most people probably have only vague notions about a shared identity with others who live in their area. Nonetheless, the type of area may determine conditions such as housing, deprivation, ethnicity or pollution, although these do not respect arbitrary geographical boundaries on moving from one ward to the next.

Although the results presented here showed a degree of intraclass correlation, it was not enough to materially alter the findings. For the Carstairs variable, the individual-level model underestimated the standard error by 12%. More generally, ignoring the hierarchical structure of the data will tend to overestimate the statistical significance of aggregate variables in the statistical model. Thus, it is recommended that some measure of the intraclass correlation is made to assess whether the approximation of zero within-cluster correlation is justified in a particular case.[17 22]

In summary, this study confirms the assumption that neighbourhood deprivation measures are predictors of the smoking status of individuals. Furthermore, it shows that when analysing data on individuals who are grouped into areas, the choice of an appropriate statistical model depends on the intraclass correlation coefficient. Earlier studies of area deprivation and individual morbidity have either assumed implicitly that the individual level model is a reasonable approximation,[23] or have opted for an area level analysis, which avoids the assumption of zero variance between wards.[24] The results presented here provide some guidance on when use of the simplifying assumptions that lead to a standard individual level model are justified.

We thank the North Thames Health Authority for permission to use the data from the North West Thames Regional health survey and Chris Grundy for producing the maps.

1 Townsend P, Davidson N. *Inequalities in health: The Black report.* Harmondsworth: Penguin, 1982.
2 Davey Smith G, Shipley M. Rose G. Magnitude and causes of socio-economic differentials in mortality: further evidence from the Whitehall study. *J Epidemiol Community Health* 1990;**44**:265–70.
3 Rose G. Marmot M. Social class and coronary heart disease. *Br Heart J* 1981;**45**:13–9.
4 Blane D, Davey Smith G, Bartley M. Social class differences in years of potential life lost: size, trends and principal causes. *BMJ* 1990;**301**:429–32.
5 Eames M, Ben-Shlomo Y, Marmot M. Social deprivation and premature mortality: regional comparison across England. *BMJ* 1993;**307**:1097–102.
6 Carstairs V, Morris R. *Deprivation and health in Scotland.* Aberdeen: Aberdeen University Press, 1991.
7 Doll R, Peto R. Mortality in relation to smoking: 20 years observation on male British doctors. *BMJ* 1976;**2**:1525–36.
8 IARC. Monographs on the evaluation of the carcinogenic risk of chemicals to humans. Vol 38. *Tobacco smoking.* Lyon: IARC, 1986.
9 Doll R, Peto R, Wheatley K, Gray R, Sutherland I. Mortality in relation to smoking: 40 years' observations on male British doctors. *BMJ* 1994;**309**:901–11.
10 Department of Health. *The health of the nation: a strategy for health in England.* London: HMSO, 1992.
11 OPCS. *General household survey: cigarette smoking 1972 to 1990.* OPCS Monitor SS 91/3. London: HMSO, 1991.
12 Davey Smith G, Shipley M. Confounding of occupation and smoking: its magnitude and consequences. *Soc Sci Med* 1991;**32**(11):1297–300.
13 Elliott P, Hills M, Beresford J, *et al.* Incidence of cancers of the larynx and lung near incinerators of waste solvents and oils in Great Britain. *Lancet* 1992;**392**:854–8.
14 Goldstein, H. *Multilevel models in educational and social research.* London: Griffin, 1987.
15 Goldstein, H. Non-linear multilevel models, with an application to discrete response data. *Biometrika* 1991;**78**(1): 45–51.
16 McCullagh P, Nelder J. *Generalised Linear Models.* 2nd ed. London: Chapman and Hall, 1989.
17 Kendall M, Stuart A. *The advanced theory of statistics.* Vol 2. London: Griffin, 1961;302–4.
18 Walter SD. A simple test for spatial pattern in regional health data. *Stat Med* 1994;**13**:1037–44.
19 Jolley D, Jarman B, Elliott P. Socio-economic confounding. In: Elliott P, Cuzick J, English D, Stern R, eds. *Geographical*

and environmental epidemiology: methods for small area studies. Oxford. Oxford University Press, 1992:115–24.
20 Marsh A, McKay S. *Poor smokers.* London: Policy Studies Institute, 1994.
21 Jarman B, Identification of underprivileged areas. *BMJ* 1983;**286**:1705–9.
22 Katz J, Carey V, Zeger S, Sommer A. Estimation of design effects and diarrhoea clustering within households and villages. *Am J Epidemiol* 1993;**138**(11):994–1006.
23 Curtis SE. Use of survey data and small area statistics to assess the link between individual morbidity and neighbourhood deprivation. *J Epidemiol Community Health* 1990;**44**:62–68.
24 Jessop EG. Individual morbidity and neighbourhood deprivation in a non-metropolitan area. *J Epidemiol Community Health* 1992;**46**:543–6.

Open discussion

DIGGLE – I liked this analysis because it considers the possibility of variation between wards as opposed to variation between individuals. Although I do not believe it would make much difference here, I think it is worth making the general point that ward residuals can be considered as recognising that all the right covariates are not all included in the model, which induces apparently random variation between wards. The implicit assumptions in this study is that whatever those unobserved covariates are, they are not spatially correlated – yet what *is* in the model is highly spatially structured. So if I could give another little plug for Breslow and Clayton,[1] you need more flexible correlation structures for your electoral wards, which would be provided by the more general machinery of generalised linear mixed models which are certainly close cousins to multilevel models.

KLEINSCHMIDT – That would be an additional step which was not attempted here.

ELLIOTT – It would be fairly simple and possibly worthwhile to test for spatial autocorrelation between model residuals to see how serious the problem might be.

BITHELL – Did you fit the age as it is or did you standardise it by subtracting the mean or something similar?

KLEINSCHMIDT – We used age unmodified.

BITHELL – I am astonished that you get such a significant quadratic relationship with age, but no linear relationship. Your model showed no effect of linear age?

KLEINSCHMIDT – If linear age is in the model by itself it is significant. If both age squared and linear age are in the model, only age squared is significant. It is customary to retain lower order terms in the model, even if they are not significant.

BEN SHLOMO – I was interested that the response rate was 64%, which is slightly less than we usually regard as acceptable. Did you look at the response rate in relation to the deprivation ward score because it could be predicted that the most deprived wards have the worst response rate, and of course we know from other studies[2] that people who do not respond are more likely to be smokers as well. I wonder how much underestimation is hidden by non-response: I am sure there would be even stronger effects with better response rates.

KLEINSCHMIDT – Unfortunately, there was no information about the non-respondents and whether they were mainly from deprived wards.

JOSHI – How do you take this forward? You show that deprived areas are more likely to have people who smoke in them. Everything else shows that deprived areas are more likely to show people who are sick and who die. How does one disentangle whether this is a smoking induced problem? How do you avoid your results being interpreted that way?

KLEINSCHMIDT – I am not sure what you mean.

JOSHI – Someone might look at your results and say all we need to do in a deprived area is not give the GPs extra inducements to improve health but to intensify a stamping out smoking campaign – a locally targeted campaign.

KLEINSCHMIDT – Well, I think that if you are saying that smoking and bad health have nothing to do with each other, that is not true, they do. Obviously, we could remove some bad health if people were persuaded not to smoke.

JOSHI – We have to put the two bits of behaviour and outcomes together.

ELLIOTT – We were particularly interested in this sort of analysis, again because it tries to answer the focussed question, "does pollution affect health, with or without deprivation?" We want to know what the deprivation index is doing and whether it is related to known causes of ill health. We do not generally have smoking data for small areas, although we would like to have this. It is reassuring from that perspective therefore to find that deprived areas, which are strongly related to ill health, are associated with factors which, from individual medical models, are also related to ill health, such as smoking. There is the bigger issue which was discussed earlier – what is it about deprived areas in their totality that is associated with ill health? One cannot derive that sort of implication from these sorts of data.

KLEINSCHMIDT – Of course it is well known that the relationship between smoking and individual markers of deprivation is highly significant. Deprivation scores are rather blunt in a way and I would like to resolve that.

GORDON – You are trying to predict small area smoking rates from deprivation but we do know how this varies with class. Have you looked at class in the census as a predictor?

KLEINSCHMIDT – I included socioeconomic group, which is similar to class, into the model. It is significant, and deprivation still remains significant. The coefficient is less, but even if we had individual socioeconomic grouping in the model, area deprivation still explains a lot of the variation.

1 Breslow NE, Clayton DG. Approximate inference in generalised mixed models. *J Am Stat Assoc* 1993;**88**:9–25.
2 Macera CA, Jackson KL, Davis DR, Kronenfeld JJ, Blair SN. Patterns of non-response to a mail survey. *J Clin Epidemiol* 1990;**43**:1427–30.