THEORY AND METHODS

Linkage of survey data with district-level lung cancer registrations: a method of bias reduction in ecological studies

Gillian A Lancaster, Mick Green, Steven Lane

J Epidemiol Community Health 2006;**60**:1093–1098. doi: 10.1136/jech.2005.043356

See end of article for authors' affiliations

Correspondence to: G Lancaster, Centre for Medical Statistics and Health Evaluation, University of Liverpool, Shelley's Cottage, Brownlow Street, Liverpool, L69 3GS, UK; g.lancaster@liv.ac.uk

Accepted 18 May 2006

Objective: To investigate a stratified ecological method for reducing ecological bias in studies that use aggregate data, by incorporating information on individual-level risk factors into the analysis.
Design: Cross-sectional study investigating associations between socioeconomic risk factors and lung cancer in the north of England, using 1991 UK Census Small Area Statistics and Sample of Anonymised Records with lung cancer registrations from three regional cancer registries for 1993–6.
Setting and patients: 92 local authority districts in the north of England containing over three million people aged 45–74 years.
Results: Generally, groups considered more socioeconomically disadvantaged had an increased risk of lung cancer across districts. In the standard ecological analysis, effects for non-car ownership, social class III manual, social class IV/V and socioeconomic inactivity were insignificant, suggesting ecological bias. In the stratified ecological analysis these effects became significant (rate ratio (RR) 2.23, 95% confidence interval (CI) 1.79 to 2.78, $p<0.001$; RR 1.35, 95% CI 1.04 to 1.74, $p=0.022$; RR 2.36, 95% CI 1.86 to 2.99, $p<0.001$; and RR 0.72, 95% CI 0.53 to 0.98, $p=0.039$, respectively), and spuriously large positive effects for the social class III non-manual (RR 20.29) and unemployment groups (RR 147.53) reduced to a more reasonable level (RR 1.92, 95% CI 1.46 to 2.52, $p<0.001$; and RR 2.36, 95% CI 1.22 to 4.55, $p=0.011$, respectively).
Conclusions: Stratified ecological analysis incorporating information on individual-level covariates reduced the bias seen in a standard ecological analysis. The method is straightforward to apply and allows the linkage of health data with data from any large-scale complex survey where district of residence is known.

Ecological studies play a useful part in establishing an initial association between potential risk factors and disease outcomes.[1] These studies use data measured at an aggregate level for geographically defined small areas, because individual-level information is not readily available or is too costly to obtain. Data are usually taken from existing disease registries, vital statistics or the national census. When the results of an ecological analysis are used to make inferences on the people living in an area, they are prone to a bias known as the ecological fallacy effect, which can also be called aggregate bias or cross-level bias.[2] As an illustration, Gatrell *et al*[3] studied access to tertiary cardiac services in east Lancashire and found that in electoral ward areas with a higher proportion of Asians, there was lower uptake of angiograms and percutaneous transluminal coronary angioplasty operations. However, although they postulated that it was the Asians living in the wards who had the lower uptake of services, it could equally well have been the white people living in those wards. The finding of an ecological association between ethnicity and uptake therefore warranted further investigation at the individual level to confirm this hypothesis. The problem of the ecological fallacy has been discussed in many publications, with possible reasons for its occurrence.[4–6]

In epidemiology and health services research, the availability of information on risk factors for ecological analysis is limited, particularly at smaller units of aggregation, such as electoral wards. The UK national census has provided socioeconomic information at the enumeration district, ward and local authority district levels, but only one measure of ill health—namely limiting long-term illness—was available in the 1991 census.[7] To overcome this, disease rates taken from different sources are often amalgamated with the census socioeconomic information for ecological analysis of other health outcomes.[8] As an exception, the Office for National Statistics Longitudinal Study is a specially commissioned survey that links census data to cancer data.[9] However, census data do not contain risk factors of lifestyle, smoking, diet and the environment. In large-scale health surveys where this information on risk factors is available—for example, the Health Surveys for England[10] or the Health and Lifestyle Survey[11]—the health outcome of interest may not have been recorded. It would therefore be beneficial to be able to link survey information to other health outcome data, which can be carried out using a stratified ecological model.[12] In this model, each areal unit is divided into strata based on age and sex groupings, and information from a survey is used to provide covariate data, cross-tabulated by age and sex, on the people in each stratum. This is possible when survey data include an aggregate-level identifier for each person as in the UK Census Sample of Anonymised Records.[13] However, the identifier is usually restricted to the local authority district or higher levels for confidentiality reasons.

The aim of this study was to investigate the potential of the stratified ecological model, compared with the standard ecological model, for reducing bias in ecological studies. Analyses were carried out to examine the associations between socioeconomic risk factors and lung cancer in the north of England.

Abbreviations: SAR, Sample of Anonymised Records; SAS, Small Area Statistics; SIR, Standardised Incidence Ratios

METHODS

Population

All people aged between 45 and 74 years (3 667 188 people) living in 92 local authority districts in Greater Manchester, Merseyside, Tyne and Wear, Cleveland, Humberside, Cheshire, Cumbria, Durham, Lancashire, Northumberland and Yorkshire in 1991 were included in the analyses.

Sources of data

The covariate data for local authority districts for the standard ecological analysis were extracted from the UK national census datasets for 1991, which are given through the Manchester Information and Associated Services national data centre at Manchester University (found at http://www.mimas.ac.uk/). The main outputs from the 1991 census are tables of aggregate data for constituent areas of Great Britain, called the Small Area Statistics and Local Based Statistics,[14] which contain information on households and people enumerated through detailed self-completed questionnaires on the day of census.

The covariate data for the stratified ecological analysis were taken from the census Sample of Anonymised Records (SAR).[13] This is a 2% sample drawn from the 1991 census, which has had identifying information removed to protect confidentiality. They are microdata files with a separate record for each person, similar to the data obtained from a sample survey. The SAR covers the full range of census topics including housing, education, health, transport, employment and ethnicity.

Population outcome data on lung cancer registrations were obtained from three regional cancer registries covering the north of England. Regional cancer registries across the UK have been collecting population-based cancer data for the past 40 years and supply data to the Office for National Statistics for the provision of national cancer statistics. All the UK registries collect information on every new diagnosis of cancer occurring in their regional populations. Their main priority has been to ensure a uniform process for registering cancers regionwide, which will deliver timely, comparable and high-quality data. The main sources of registrations are from pathology reports, medical records, radiotherapy records, hospices, independent hospitals, specialist tumour registers, screening services and death certificates.

Socioeconomic risk factors

Six socioeconomic risk factors from the 1991 Census Small Area Statistics (SAS)[14] were extracted for all residents aged between 45 and 74 years living in households. These SAS data provided the covariate information for the standard ecological analysis carried out at the local authority district level. The data were in tabulated form aggregated by district and, in some cases, data restrictions meant that the exact subgroup of residents could not be selected—for example, the age range might differ or all residents selected where data were not separately available for those living in households. Each category of a multicategory covariate was represented by a separate variable expressed as the proportion of people falling in that category (eg, proportion who were employed, proportion who were unemployed, proportion who were economically inactive, etc).[15] The six covariates, with categories expressed as proportions, were ethnicity (white or non-white), housing tenure (owner-occupier, renting privately or renting from local authority), car ownership (one car, no car, or two or more cars), social class (I+II, III non-manual, III manual or IV+V), employment status (employed, unemployed or economically inactive) and qualifications (qualified with a diploma or degree or unqualified). The reference category was taken to be the first category listed in each case.

The same covariate information was extracted at the individual level from the census 2% SAR.[13] These data were used in the stratified ecological analysis to provide more detailed information on the associations between age, sex and socioeconomic status. In the stratified analysis, each local authority district was stratified into 14 age and sex groups. The socioeconomic data were cross-tabulated with age (40–44, 45–49, 50–54, 55–59, 60–64, 65–69 and 70–74 years) and sex (male or female) to obtain a unique covariate proportion for each age and sex grouping in each district. These covariate data were then incorporated into the stratified ecological model (see Statistical analysis). Categorical covariates representing each age and sex group were also included in the model, which enabled an age and sex interaction term to be fitted.

Outcome measure

Population estimates of lung cancer registrations were obtained for all districts in the north of England from three regional cancer registries. The data were provided in a standard aggregate form as cross-tabulations of observed frequency counts by age, sex and district. This mirrored the usual form of outcome data for ecological analysis and required no special permission for their use. As some of the counts were small, amalgamation of some districts was carried out by the registry providers to maintain confidentiality before the data could be released. There were 52 ''super districts'' remaining after the amalgamations. These data were used to obtain age and sex-specific cancer rates for the north of England to indirectly standardise the disease rates. All new cases of lung cancer registered in the years 1993–6 were analysed in relation to the socioeconomic risk factors taken from the 1991 census. The lag between exposure and disease occurrence was to avoid ill people being socioeconomically reclassified into ''unhealthy'' categories owing to their being ill.[16] For example, in relation to employment, someone who might normally have been in employment may become economically inactive because of their illness. Lung cancer was chosen, as it is one of the major cancer sites, and because there were already known socioeconomic differentials in the incidence of lung cancer[17] that would provide an interesting illustration of the method.

Statistical analysis

As some small districts had to be amalgamated to retain the confidentiality of lung cancer registrations, the socioeconomic census data were also amalgamated into 52 super district units for the analyses. The models were fitted by maximisation of the likelihood using standard statistical procedures in STATA V.8.2.

Standard ecological model

For standard ecological analysis, the total number of observed people with lung cancer in each district were regressed on the SAS covariate proportions using a Poisson model of the form,

$$E(y_k) = \exp\left(\ln e_k + \beta_0 + \sum_j x_{jk}\beta_j\right)$$

where y_k is the frequency of developing lung cancer in district k (k = 1, …, K), x_{jk} is the j^{th} (j = 1, …, J) covariate value for district k and β_j is the parameter of the jth covariate to be estimated. Illness rates were indirectly standardised taking the north of England as the standard population, to obtain the total number of people to be expected to have lung cancer in each district k, if that district experienced the same age and sex-specific rates of illness as that in the standard

population. The log of the expected counts ($\ln e_k$) was included in the model as an offset. In this model, one observed and expected frequency per district were used. Therefore, the unit of observation was the district, and the covariate proportions for each risk factor were measured at the district level.

Stratified ecological model

In the stratified model, the observed and expected frequency counts were left expanded over each of the 14 age and sex strata used in the indirect standardisation procedure, to give 14 observations/district. The corresponding covariate information was then taken from the SAR cross-tabulations to obtain a unique covariate proportion for each strata in each district. This was carried out using the SAR individual-level data because the SAS does not provide this level of information for every covariate. A similar Poisson model was applied to the data for the stratified ecological analysis as follows

$$E(y_{ks}) = \exp\left(\ln e_{ks} + \beta_0 + \sum_j x_{jks}\beta_j\right)$$

where y_{ks} is the frequency of developing lung cancer in stratum s of district k, x_{jks} is the jth covariate value for stratum s in district k, β_j is the parameter of the j^{th} covariate to be estimated and $\ln e_{ks}$ is the offset term. When age and sex terms are included in the model, the offset could have been simplified to $\ln(n_{ks})$, where n_{ks} is the number of people in stratum s of district k. This is because the age and sex terms together with an age and sex interaction should, in theory, be able to adjust for any age and sex differences in lung cancer rates, and hence standardisation should not be necessary. However, for ease of comparison between models, $\ln(e_{ks})$ is used as the offset throughout this paper.

RESULTS

Table 1 displays the characteristics of the SAS and SAR samples. It illustrates the restrictions of the SAS data for some variables with respect to differing denominator populations and age groups, causing some slight discrepancies in prevalence between the samples with respect to housing tenure, car ownership and social class.

Table 2 gives the results of the lung cancer data regressed on socioeconomic covariate proportions for the 52 super districts, comparing the standard and stratified ecological analyses. As the cancer rates were small in all age groups, Poisson models were applied. For comparability of deviances, the standard ecological model deviance was recalculated on an expanded dataset, where each age and sex stratum had the same covariate value repeated over the 14 categories. This created a data structure similar to that used for the stratified model containing (52×14) 728 observations, and provided identical parameter estimates and standard errors to the collapsed model containing 52 observations.

The rate ratio (RR) results of the standard ecological analysis suggest that districts with a higher proportion of people living in local authority rented accommodation, with a higher proportion of people in social class III non-manual, or with a higher proportion of unemployed people, had a higher risk of lung cancer than the respective reference category. Districts with a higher proportion of non-white people, or with a higher proportion of people living in private rented accommodation, had a decreased risk of lung cancer, and districts with a higher proportion of non-car owners or two car owners, with a higher proportion of people in social classes III manual or social class IV and V, with a higher proportion of inactive or unqualified people had no increased risk of lung cancer compared with the reference category. In the stratified ecological analysis, the spuriously large effect for unemployment now reduced considerably, as did that for the social class III non-manual group, such that the social class now showed more of an increased gradient in risk. In addition, districts with a higher proportion of non-car owners now had a significantly increased risk of lung cancer and districts with a higher proportion of inactive people, a significantly decreased risk. The effects for the non-white and rent privately groups became insignificant, indicating no increased risk in lung cancer for white compared with non-white people or for those renting privately compared with owner occupiers. The effects of the two-car ownership and unqualified groups remained insignificant in both analyses.

A sensitivity analysis of men and women separately using the stratified approach identified that a social class gradient was more apparent in the analysis for men (III non-manual RR 2.28, 95% CI 1.38 to 3.75; III manual RR 1.68, 95% CI 1.24 to 2.28; IV and V RR 2.80, 95% CI 2.04 to 3.84) compared with women (III non-manual RR 1.57, 95% CI 1.08 to 2.27; III manual RR 1.03, 95% CI 0.57 to 1.87; IV and V RR 1.63, 95% CI 1.12 to 2.36). Also the risk of lung cancer in districts with a higher proportion of economically inactive men (RR 0.92, 95% CI 0.63 to 1.35) was not significantly different from the employed group, and districts with a higher proportion of women in privately rented accommodation had an increased risk of lung cancer (RR 3.71, 95% CI 1.63 to 8.45) compared with the owner occupier group. In all other respects, the results were similar to those for the combined analysis.

DISCUSSION

The relationships described by the standard ecological model show some exaggerated and spurious associations, which are counterintuitive to known socioeconomic differentials for lung cancer.[17] The results show how insignificant effects for non-car ownership, social class III manual, social class IV and V, and socioeconomic inactivity in the standard ecological analysis became significant in the stratified ecological analysis. The larger positive effects, respectively, for social class III non-manual and unemployment groups also considerably reduced in the stratified analysis. Although there is no clear explanation for these large effects in the standard model, we would suggest that they are somehow being confounded with the age and sex effects, which were separated out in the stratified model. The stratified ecological approach also enabled a sensitivity analysis by sex to compare results with the findings of Kogevinas[17] for earlier years. Kogevinas studied cancer incidence (as well as survival) using the Office for National Statistics Longitudinal Study for the years 1971–83. He showed marked socioeconomic differentials in lung cancer incidence in men more socioeconomically disadvantaged for housing tenure (standardised incidence ratios (SIR) 138 for council tenants, 116 for those privately renting and 75 for owner occupiers), social class (SIR 48, 77, 86, 105, 116, 124 for classes I, II, IIIN, IIIM, IV and V, respectively) and employment (SIR 150 for unemployed, 102 retired, and 96 for employed). No results were presented for the other socioeconomic variables considered in our study. For women, the differentials were less marked but significant for housing tenure (SIR 122 for council tenants, 111 for those privately renting and 83 for owner occupiers), non-significant in manual compared with non-manuals jobs, and economic position was not reported. Although these results were not for the same time period as our study they do show the trends in socioeconomic disadvantage that we might expect in our study and that have been shown for other disease outcomes.[18–21] The results of the sensitivity analyses in particular are broadly supported by these findings.

Table 1 Characteristics of the Small Area Statistics (SAS) and the Sample of Anonymised Records (SAR) samples used in the standard and stratified ecological analyses of lung cancer

	SAS sample used in standard ecological analysis (n = 3 667 188)		SAR sample used in stratified ecological analysis (n = 76 217)	
Variable	**Number**	**Percentage**	**Number**	**Percentage**
Age group (years)*				
45–49	707 832	19.3	14 461	19.0
50–54	655 019	17.9	13 654	17.9
55–59	622 235	17.0	12 960	17.0
60–64	618 267	16.9	12 820	16.8
65–69	597 007	16.3	12 445	16.3
70–74	466 828	12.7	9877	13.0
Sex*				
Male	1 752 166	47.8	36 559	48.0
Female	1 915 022	52.2	39 658	52.0
Ethnicity†				
White	3 738 016	98.2	74 818	98.2
Non-white	67 942	1.8	1399	1.8
Housing tenure‡				
Home owner	3 196 541	65.7	53 837	71.8
Rent privately	373 657	7.7	3938	5.3
Rent LA	1 294 811	26.6	17 233	23.0
Car ownership‡				
No car	1 900 163	39.0	24 154	32.2
One car	2 050 989	42.1	33 854	45.1
⩾2 cars	925 212	19.0	17 000	22.7
Social class§				
I and II	271 694	34.9	17 144	29.8
IIIN	95 621	12.3	10 373	18.0
IIIM	247 882	31.8	14 158	24.6
IV and V	164 101	21.1	15 923	27.6
Economic position†				
Employed	1 622 881	42.6	32 394	42.5
Unemployed	164 362	4.3	3321	4.4
Inactive	2 018 715	53.1	40 502	53.1
Qualifications¶				
Yes	108 517	11.7	7902	10.4
No	817 476	88.3	68 315	89.6

*For SAS data: People in households aged 45 to 74 years.
†All people aged 45–74 years.
‡All people in households.
§10% sample of all people in households, social class of head of the household.
¶10% sample of all persons aged ⩾18 years.

The datasets used in this study were generated from population registers, where completeness of ascertainment can be an issue. Coverage of the national population in the 1991 census is estimated to be 98%. Although people in less advantaged socioeconomic groups tend not to answer inquiries such as this, and therefore prevalence estimates may be underestimated, their omission would probably not have greatly affected the results, as the focus here was on associations between variables, and the findings showed trends in the socioeconomically disadvantaged that were consistent with previous work. The discrepancies between the two census samples shown in Table 1, due to differences in denominator populations, were small and therefore were also likely to have had minimal effect on our findings. This was confirmed in a sensitivity analysis of social class, where it was calculated using SAR data for both the head of the household's social class and the person's own social class, and the findings remained robust. Cancer registries in the UK provide the best source of population data for the study of specific cancers. Incompleteness of cancer registry data, when apparent, is generally due to a breakdown in reporting procedures and not to individual patient attributes. In general, ascertainment is high, and there is no tendency for inaccurate registration to occur in specific regions.[17]

In this analysis, the covariate information was limited to the variables available in the census SAS and SAR. No data on smoking were available—for example, a known risk factor for lung cancer. This highlights the advantage of the stratified ecological method in incorporating individual-level data from other large-scale surveys in which information on smoking may have been collected. In this respect, it is likely that some of the socioeconomic variables here acted as a proxy for smoking—for example, with those in the lower social classes being more likely to smoke than those from less disadvantaged groups.[17] In this study, we only fitted a simple age and sex interaction term. It could be argued that this interaction was not needed, because an offset term was included in the model and most of these effects were non-significant. However, there were significant interaction effects for the two older age groups, indicating the necessity for a correction factor to adjust for variation in the socioeconomic variables by age and sex not accounted for by the offset. The inclusion of the interaction term also illustrates the potential of the method to fit more complex interactions with other risk factors if required.

Therefore, there are several advantages in using the stratified ecological model. Firstly, it provides a more detailed analysis that takes into account population differences in the age and sex structure of the area through the strata, which is not possible in a standard ecological model where only the overall disease rate ratio for each area is known. Secondly, it is able to incorporate individual-level survey information into an ecological analysis that opens up the way for taking additional covariate information from large-scale surveys, such as those held on the Economic and Social Research Council's Data Archive, which contain a district-level

Table 2 Rate ratios of lung cancer for years 1993–6 by standard and stratified ecological regression

	Standard ecological model using SAS covariate data		Stratified ecological model using SAR covariate data	
	Rate ratio (95% CI)	p Value	Rate ratio (95% CI)	p Value
Age group (years)				
45–49	—	—	1.0	—
50–54	—	—	0.88 (0.80 to 0.98)	0.024
55–59	—	—	0.86 (0.77 to 0.96)	0.007
60–64	—	—	0.93 (0.80 to 1.08)	0.341
65–69	—	—	1.02 (0.78 to 1.33)	0.898
70–74	—	—	0.92 (0.70 to 1.22)	0.581
Sex				
Male	—	—	1.0	—
Female	—	—	0.88 (0.74 to 1.05)	0.165
Age (years) by sex				
All males and females 45–49	—	—	1.0	—
Females 50–54	—	—	1.04 (0.88 to 1.22)	0.670
Females 55–59	—	—	1.02 (0.87 to 1.19)	0.782
Females 60–64	—	—	0.99 (0.85 to 1.16)	0.938
Females 65–69	—	—	0.84 (0.72 to 0.97)	0.022
Females 70–74	—	—	0.83 (0.71 to 0.96)	0.015
Ethnicity				
White	1.0	—	1.0	—
Non-white	0.41 (0.25 to 0.65)	<0.001	1.33 (0.75 to 2.36)	0.328
Housing tenure				
Home owner	1.0	—	1.0	—
Rent privately	0.29 (0.14 to 0.60)	0.001	0.93 (0.61 to 1.42)	0.726
Rent LA	1.78 (1.26 to 2.54)	0.001	1.55 (1.29 to 1.87)	<0.001
Car ownership				
No car	2.00 (0.63 to 6.31)	0.238	2.23 (1.79 to 2.78)	<0.001
One car	1.0	—	1.0	—
Two cars	1.03 (0.22 to 4.82)	0.966	0.93 (0.68 to 1.27)	0.646
Social class				
I and II	1.0	—	1.0	—
IIIN	20.29 (3.24 to 127)	0.001	1.92 (1.46 to 2.52)	<0.001
IIIM	0.91 (0.33 to 2.50)	0.851	1.35 (1.04 to 1.74)	0.022
IV and V	1.58 (0.42 to 5.87)	0.497	2.36 (1.86 to 2.99)	<0.001
Economic position				
Employed	1.0	—	1.0	—
Unemployed	147.53 (17.69 to 1230)	<0.001	2.36 (1.22 to 4.55)	0.011
Inactive	0.55 (0.23 to 1.30)	0.172	0.72 (0.53 to 0.98)	0.039
Qualifications				
Yes	1.0	—	1.0	—
No	0.77 (0.20 to 2.93)	0.698	1.34 (0.87 to 2.06)	0.184
Log likelihood	−2180		−2337	
Deviance	1157		1472	

identifier. This type of analysis will typically only be feasible at the district level, as access to survey risk factor information at smaller geographical units will usually breach confidentiality. Smaller aggregate units are preferred whenever possible to further reduce bias.[6 12] Thirdly, by leaving the illness rates expanded over the age and sex strata, interaction terms can be incorporated into the model between age and sex, and the other risk factors, giving a more flexible model. The socioeconomic variables can even be summarised into a ''deprivation'' score to facilitate interpretation of more complex interactions.[15] It is important to note, however, that the model has the potential to reduce ecological bias but not totally irradicate it, and therefore results should still be treated with caution as some ecological bias will remain. Associations suggested at the aggregate level can only be confirmed through large-scale epidemiological studies, such as cross-sectional surveys, case–control or cohort studies, conducted on individual people.

Several other methods for reducing bias in ecological studies have been proposed in the literature. In particular, Cleave *et al*[22] reviewed four methods using examples of voting transitions between two different elections at the ward level, and advocated the aggregated compound multinomial model. Lancaster *et al*[12] evaluated this method in comparison to two other potential methods, endorsing the stratified approach used in this study. They also reviewed the aggregated individual-level model, proposed by Prentice and Sheppard.[23] This model is appealing, as it too can combine data from population disease registries with individual-level survey data. However, most examples found in the literature have been carried out on simulated data,[24 25] and where empirical results have been obtained they have been shown to have convergence problems.[12] Tranmer and Steel[26] presented methods using SAR data to provide adjusted correlation coefficients at the aggregate level, with adjustments made using individual-level variables that explained much of the within-area homogeneity. A Bayesian hierarchical modelling approach has also been implemented for modelling spatial dependence in disease rates in ecological regression.[27] However, these are fairly complex procedures and are not routinely used by epidemiologists.

In conclusion, stratified ecological analysis incorporating individual-level covariate information reduced the bias seen in a standard ecological analysis. It is straightforward to apply and allows the linkage of health data with data from any large-scale complex survey, where district of residence is known. Further empirical examples are needed to verify its potential in ecological regression.

What is already known

It is well known that standard ecological regression is prone to ecological bias when results from this type of analysis are used to make inferences about the people living in geographically defined areas.

What this study adds

- Stratified ecological regression is a method for reducing bias in ecological studies.
- It has the advantage of being able to link area-level health outcome data with individual-level information on risk factors from large-scale surveys that include an area-level identifier; it allows age and sex interaction terms to be fitted, it is straightforward to apply and reduces ecological bias in our example.

Policy implications

- Valuable health service resources and interventions are targeted at people in most need, but identification of vulnerable groups is difficult.
- Ecological analysis is a useful first step at identifying associations between areas containing a higher percentage of people who are socioeconomically disadvantaged people and disease outcomes, but these analyses may misconstrue the relationships.
- Methods for bias reduction including the one reported in this paper therefore make an important contribution to eliminating spurious associations and to identifying target groups within areas for further study.

Authors' affiliations
G A Lancaster, S Lane, Centre for Medical Statistics and Health Evaluation, University of Liverpool, Liverpool, UK
M Green, Centre for Applied Statistics, Fylde College, Lancaster University, Lancaster, UK

Funding: This project was funded by the Economic and Social Research Council (ESRC) grant number RES-000-22-0143. Lung cancer data were provided by three regional cancer registries covering the Northern and Yorkshire region, the North West, and Merseyside and Cheshire. The SAS and SAR are Crown copyright and supplied by the Census Microdata Unit at the University of Manchester, with the support of the ESRC Joint Information Systems Committee.

Competing interests: None.

REFERENCES

1 **Che D**, Decludt B, Campese C, *et al*. Sporadic cases of community acquired legionnaires' disease: an ecological study to identify new sources of contamination. *J Epidemiol Community Health* 2003;**57**:466–9.
2 **Wakefield J**. A critique of statistical aspects of ecological studies in spatial epidemiology. *Environ Eco Stat* 2004;**11**:31–54.
3 **Gatrell A**, Lancaster G, Chapple A, *et al*. Variations in use of tertiary cardiac services in part of North-West England. *Health Place* 2002;**8**:147–53.
4 **Robinson WS**. Ecological correlations and the behavior of individuals. *Am Sociol Rev* 1950;**15**:351–7.
5 **Greenland S**, Robins J. Ecological studies—biases, misconceptions, and counterexamples. *Am J Epidemiol* 1994;**139**:747–60.
6 **Morgenstern H**. Ecologic studies in epidemiology: concepts, principles, and methods. *Annu Rev Public Health* 1995;**16**:61–81.
7 **Dale A**. The content of the 1991 census: change and continuity. In: Dale A, Marsh C, eds. *The 1991 census user's guide*. London: HMSO, 1993.
8 **Gunnell D**, Shepherd M, Evans M. Are recent increases in deliberate self-harm associated with changes in socio-economic conditions? An ecological analysis of patterns of deliberate self-harm in Bristol 1972-3 and 1995-6. *Psychol Med* 2000;**30**:1197–203.
9 **Hattersley L**, Creeser R. *Longitudinal study 1971–1991 history, organisation and quality of data, LS no. 7*. London: HMSO, Office of Population Censuses and Surveys, 1995.
10 **Sproston K**, Primatesta P, eds. *Health survey for England 2002*, Vols 1–3. London: The Stationery Office, 2003.
11 **Cox BD**, Blaxter M, Buckle ALJ, *et al*. The Health and Lifestyle Survey. Preliminary report of a nationwide survey of the physical and mental health, attitudes and lifestyle of a random sample of 9,003 British adults. London: The Health Promotion Research Trust, 1987.
12 **Lancaster GA**, Green M, Lane S. Reducing bias in ecological studies: an evaluation of different methodologies. *J R Stat Soc [Ser A]* 2006;**169**(4):681–700.
13 **Marsh C**. The Sample of Anonymised Records. In: Dale A, Marsh C, eds. *The 1991 census user's guide*. London: HMSO, 1993.
14 **Cole K**. The 1991 Local Base and Small Area Statistics. In Dale A, Marsh C, eds. *The 1991 census user's guide*. London: HMSO, 1993.
15 **Lancaster GA**, Green M. Deprivation, ill-health and the ecological fallacy. *J R Stat Soc [Ser A]* 2002;**165**:263–78.
16 **Fox AJ**, Goldblatt P, Jones D. Social class mortality differentials: artifact, selection or life circumstances? In: Goldblatt P, eds. *Longitudinal study: mortality and social organization, LS no. 6*. London: HMSO, Office of Population Censuses and Surveys, 1990.
17 **Kogevinas E**. 1971–1983 Longitudinal study: socio-demographic differences in cancer survival, LS no. 5. London: HMSO, Office of Population Censuses and Surveys, 1990.
18 **Davey-Smith G**, Leon D, Shipley M, *et al*. Socio-economic differentials in cancer among men. *Int J Epidemiol* 1991;**20**:339–45.
19 **Gould MI**, Jones K. Analyzing perceived limiting long-term illness using U.K. census micro data. *Soc Sci Med* 1996;**42**:857–69.
20 **Brown J**, Harding S, Bethune A, *et al*. Incidence of health of the nation cancers by social class. *Popul Trends* 1997;**90**:40–7.
21 **Sloggett A**, Joshi H. Deprivation indicators as predictors of life events 1981–1992 based on the UK ONS longitudinal study. *J Epidemiol Community Health* 1998;**52**:228–33.
22 **Cleave N**, Brown PJ, Payne CD. Evaluation of methods for ecological inference. *J R Stat Soc [Ser A]* 1995;**158**:55–72.
23 **Prentice RL**, Sheppard L. Aggregate data studies of disease risk factors. *Biometrika* 1995;**82**:113–25.
24 **Sheppard L**, Prentice RL, Rossing MA. Design considerations for estimation of exposure effects on disease risk, using aggregate data studies. *Stat Med* 1996;**15**:1849–58.
25 **Guthrie KA**, Sheppard L. Overcoming biases and misconceptions in ecological studies. *J R Stat Soc [Ser A]* 2001;**164**:141–54.
26 **Tranmer M**, Steel DG. Using census data to investigate the causes of the ecological fallacy. *Environ Plann A* 1998;**30**:817–31.
27 **Best N**, Cockings S, Bennett J, *et al*. Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *J R Stat Soc [Ser A]* 2001;**164**:155–74.