Original article

Scand J Work Environ Health 2005;31(3):184-190

doi:10.5271/sjweh.868

Model specification and unmeasured confounders in partially ecologic analyses based on group proportions of exposed

by Björk J, Strömberg U

Affiliation: Competence Centre for Clinical Research, Lund University Hospital, SE-221 85 Lund, Sweden. Jonas.Bjork@skane.se

Key terms: ecologic analysis; epidemiologic method; epidemiology bias; group proportion; model specification; noise exposure; sleeping disturbance; statistical model; unmeasured confounder

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/15999570

Original article

Scand J Work Environ Health 2005;31(3):184–190

Model specification and unmeasured confounders in partially ecologic analyses based on group proportions of exposed

by Jonas Björk, PhD,[1, 2] Ulf Strömberg, PhD [2]

Björk J, Strömberg U. Model specification and unmeasured confounders in partially ecologic analyses based on group proportions of exposed. Scand J Work Environ Health 2005;31(3):184–190.

Objectives The aim of this study was to quantify bias from a partially ecologic analysis due to (i) model misspecification and (ii) an unmeasured confounder, considering various scenarios that may occur in occupational and environmental epidemiology. A study with an aggregate exposure variable, P_E, but with outcome, group membership, and covariates assessed individually is partially ecologic. In this paper, P_E is the proportion exposed; P_E can vary across geographic areas or occupational groups.

Methods Several hypothetical scenarios were considered, varying the baseline risk, the exposure effect, the exposure distribution across groups, the impact of the (unmeasured) confounder, and the confounder distribution across groups. First, confounding within groups was introduced. Thereafter, confounding between groups was introduced by co-varying P_E and the confounder prevalence across the groups. The expected odds ratio (exposed versus unexposed) was calculated in two alternative models, the logistic regression and linear odds models, both with P_E as the independent variable. Moreover, empirical data on noise exposure and sleeping disturbances were analyzed.

Results Compared with the logistic regression model, the linear odds model yielded a markedly less biased odds ratio (OR) when the outcome was rare (≤5% baseline risk). Confounding within groups resulted in marginal bias, whereas confounding between groups resulted in more pronounced bias.

Conclusions A logistic regression analysis, with P_E as an independent variable, can yield substantial model misspecification bias. By contrast, the linear odds model is valid when the outcome is rare. Confounding between groups should be of more concern than confounding within groups in partially ecologic analyses.

Key terms bias (epidemiology); epidemiologic method; noise exposure; sleeping disturbance, statistical model.

An observational study with an aggregate, environmental, or global exposure measure (1), but with individual data on disease status, group membership, and covariates, is *partially ecologic* (2). The grouping of the study population can be based on geographic areas, occupations, or, if treatment effects are evaluated, hospitals (3–5). We restricted our attention to a specific aggregate measure, namely, the proportion exposed in each group. We assumed that such exposure data are obtained from an external database [eg, a geographic information system or a job-exposure matrix (6)]. On the individual level, the corresponding exposure variable is dichotomous or dichotomized. When the aim of a partially ecologic study is to estimate exposure–disease associations on the individual level, bias may arise that has no analogue in purely individual-level studies (7–9). Such ecologic bias, or confounding *between* groups (2), occurs when the disease rates among the unexposed or exposed varies in such a systematic way that a spurious association with the ecologic measure is produced (8). On the other hand, bias that arises in an individual-level study due to an unmeasured confounder *within* groups may be much reduced in the corresponding partially ecologic setting. Johnston et al (5) showed how failure to account for the prognosis of hospitalized patients in a simulated follow-up study of treatment effects resulted in confounding by indication and thereby bias in the individual-level analysis. The corresponding partially ecologic analysis, using a logistic regression model with the proportion of patients receiving the new treatment at each hospital as the independent variable, was robust to such confounding within groups. Only treatment practice (and not the prognosis of the patients) was assumed to vary between hospitals, implying that only limited systematic

1 Competence Centre for Clinical Research, Lund University Hospital, Lund, Sweden.
2 Department of Occupational and Environmental Medicine, Lund University, Lund, Sweden.

Correspondence to: Dr Jonas Björk, Competence Centre for Clinical Research, Lund University Hospital, SE-221 85 Lund, Sweden. [E-mail: Jonas.Bjork@skane.se]

variability in baseline risk between hospitals (ie, no pronounced confounding between groups or hospitals). This assumption may be true in settings in which the unmeasured confounder (reflecting patient prognosis) implies confounding by indication. On the other hand, in environmental or occupational exposure settings, co-variation of the proportion exposed and the (unmeasured) confounder prevalence across groups may be a more realistic assumption, implying confounding between groups.

In our study, we quantified bias of the odds ratio (exposed versus unexposed) expected from a partially ecologic analysis, considering various scenarios that may occur in occupational and environmental epidemiology. We addressed two sources of bias: (i) model misspecification and (ii) presence of an unmeasured confounder. First, we varied the disease risk in a situation with mainly confounding within groups and compared the model misspecification bias of the partially ecologic analysis using the conventional logistic regression model (5), with a linear model for the disease odds (2, 10). Thereafter, we introduced marked confounding between groups into the partially ecologic analysis and quantified the resulting bias. In addition to assessing the bias based on hypothetical scenarios, we considered an empirical example of noise exposure and sleeping disturbances.

Hypothetical examples

Models

We considered a hypothetical observational longitudinal study in which E denoted the individual exposure variable (1 = exposed, 0 = unexposed) and P_E denoted the corresponding aggregated (group level) measure (ie, the proportion of persons exposed in each group of the study population) ($0 \leq P_E \leq 1$). We assumed that E was related to disease D on the individual level by the odds ratio OR_{ED}. Moreover, there was an unmeasured binary confounder C (1 = present, 0 = absent) with group prevalence P_C, related to disease D by OR_{CD} and to exposure E by OR_{CE}. We worked with 30 exposure groups (eg, corresponding to different residential areas), defined by G = 1, 2, …, 30 and ranked in ascending order with respect to P_E. For simplicity, we assumed that the exposure groups were of equal size. All the groups could include both unexposed and exposed persons, exposure being assigned according to the formula [see Johnston et al (5)]:

$$\text{Logit[prob of exposure]} = kG + [\ln(OR_{CE})]C - m, \quad \text{(equation 1)}$$

where prob stands for probability, k is a constant that influences both the variability of the exposure between groups, and, together with the constant m, the overall exposure prevalence. We used the following three sets of values for these constants: (i) $k = 0.32$ and $m = 8$, which yield exposure distributions across groups that we refer to as *widespread* (overall exposure prevalence 25%, range 0.0–88%, assuming $OR_{CE} = 3.0$ and a confounder prevalence P_C of 50% in all groups), (ii) $k = 0.09$ and $m = 4$, which we refer to as *truncated* exposure distributions (overall exposure prevalence 15%, range 3.8–33%, if $OR_{CE} = 3.0$ and $P_C = 50\%$ in all groups), and (iii) $k = 0.21$ and $m = 8$, which we refer to as *rare* exposure distributions (overall exposure prevalence 5.2%, range 0.0–25%, if $OR_{CE} = 3.0$ and $P_C = 50\%$ in all groups).

The true individual probability of disease is expressed as follows [see Johnston et al (5)]:

$$\text{Logit[prob of disease]} = [\ln(OR_{CD})]C + [\ln(OR_{ED})]E + B, \quad \text{(equation 2)}$$

where prob stands for probability and B denotes the logit of the baseline risk. The average probability of disease, and thus the average odds of disease, in each group can be calculated using the expression for the individual probability of disease together with P_C and P_E. Two different models for the average disease odds are used for estimating OR_{ED} (ie, the OR for the individual exposure–disease association) in the *partially ecologic analysis* with data on P_E obtained from an external database: the logistic model (11):

$$\text{Logit[prob of disease]} = \ln(\text{odds of disease}) = \alpha_1 + \beta_1 P_E \quad \text{(equation 3)}$$

and the linear odds model (2, 11):

$$\text{Odds of disease} = \exp(\alpha_2) \times (1 + \beta_2 P_E). \quad \text{(equation 4)}$$

In the *individual-level analysis*, we used the following model:

$$\text{Logit[prob of disease]} = \alpha_3 + \beta_3 E. \quad \text{(equation 5)}$$

Bias calculation

In the logistic model, the OR_{ED} estimator corresponds to $\exp(\beta_1)$, and, in the linear odds model, it corresponds to $1 + \beta_2$ (ie, the disease odds for $P_E = 1$ divided by the disease odds for $P_E = 0$). The expected value of the OR_{ED} estimator is the OR that one can expect to obtain on the average if repeated studies are conducted within the same setting. Bias occurs if the expected OR differs from the true OR for the exposure–disease association. The linear odds model yields unbiased estimates of the true OR_{ED} when confounding is absent and the baseline risk is low (2, 11). Note that both the linear and the logistic model neglect the (unmeasured) confounder C, which may be a source of bias. The linearity of the linear odds model and the logistic model implies that we

Table 1. Expected odds ratio estimates for the association between exposure and disease (OR_{ED}) in the presence of confounding within groups [a], varying the true OR_{ED}, the exposure distribution across 30 groups, and the baseline risk.

Exposure distribution	Partial ecologic analysis		Individual-level analysis (OR_{ED})
	OR_{ED} in the linear odds model	OR_{ED} in the logistic model	
Widespread[b]			
True OR_{ED} 1.5			
Baseline risk 50%	1.5	1.5	1.8
Baseline risk 10%	1.5	1.5	1.8
Baseline risk 5%	1.5	1.5	1.8
Baseline risk 1%	1.5	1.5	1.8
Baseline risk 0.1%	1.5	1.5	1.8
True OR_{ED} 2.0			
Baseline risk 50%	1.9	1.9	2.4
Baseline risk 10%	2.0	2.1	2.3
Baseline risk 5%	2.0	2.1	2.4
Baseline risk 1%	2.0	2.1	2.4
Baseline risk 0.1%	2.0	2.1	2.4
True OR_{ED} 3.0			
Baseline risk 50%	2.6	2.7	3.5
Baseline risk 10%	2.9	3.2	3.5
Baseline risk 5%	3.0	3.3	3.5
Baseline risk 1%	3.0	3.4	3.5
Baseline risk 0.1%	3.0	3.4	3.5
Truncated[c]			
True OR_{ED} 1.5			
Baseline risk 50%	1.4	1.4	1.8
Baseline risk 10%	1.5	1.6	1.8
Baseline risk 5%	1.5	1.6	1.8
Baseline risk 1%	1.6	1.7	1.8
Baseline risk 0.1%	1.6	1.7	1.8
True OR_{ED} 2.0			
Baseline risk 50%	1.7	1.8	2.4
Baseline risk 10%	2.0	2.3	2.3
Baseline risk 5%	2.0	2.4	2.3
Baseline risk 1%	2.1	2.5	2.4
Baseline risk 0.1%	2.1	2.6	2.4
True OR_{ED} 3.0			
Baseline risk 50%	2.1	2.4	3.5
Baseline risk 10%	2.8	3.9	3.5
Baseline risk 5%	3.0	4.4	3.5
Baseline risk 1%	3.2	4.9	3.5
Baseline risk 0.1%	3.2	5.1	3.5
Rare[d]			
True OR_{ED} 1.5			
Baseline risk 50%	1.4	1.4	1.8
Baseline risk 10%	1.5	1.6	1.8
Baseline risk 5%	1.5	1.7	1.8
Baseline risk 1%	1.6	1.7	1.8
Baseline risk 0.1%	1.6	1.7	1.8
True OR_{ED} 2.0			
Baseline risk 50%	1.6	1.8	2.4
Baseline risk 10%	2.0	2.4	2.3
Baseline risk 5%	2.0	2.6	2.3
Baseline risk 1%	2.1	2.7	2.4
Baseline risk 0.1%	2.1	2.8	2.3
True OR_{ED} 3.0			
Baseline risk 50%	2.0	2.4	3.5
Baseline risk 10%	2.7	4.3	3.5
Baseline risk 5%	3.0	5.1	3.5
Baseline risk 1%	3.2	6.1	3.5
Baseline risk 0.1%	3.3	6.4	3.5

[a] The confounder prevalence was 50% within each group, the OR for the association between the confounder and the disease was 2.0, and the OR between the confounder and the exposure was 3.0.
[b] The overall exposure prevalence was 25% (range 0.0–88% across groups).
[c] The overall exposure prevalence was 15% (range 3.8–33% across groups).
[d] The overall exposure prevalence was 5.2% (range 0.0–25% across groups).

can assess the expected OR_{ED} estimate using the linear regression of the average disease odds (logarithmically transformed under the logistic model and untransformed under the linear odds model) in each group on P_E (12). An unweighted linear regression was used since groups of equal size were assumed. If such a linear regression results in intercept α and slope β, then the expected OR_{ED} is $exp(\beta)$ in the logistic model and $1 + \beta/\alpha$ in the linear odds model.

In the individual-level analysis, the expected value of the OR_{ED} estimator [$\exp(\beta_3)$; see the preceding text] is calculated directly from the overall average disease odds among the exposed and unexposed.

Confounding within groups—model selection

Assuming harmful effects for exposure ($OR_{ED} > 1.0$) and the confounder ($OR_{CD} = 2.0$), we introduced confounding within groups by letting $OR_{CE} = 3.0$ (table 1). When the confounder prevalence was held constant ($P_C = 50\%$) for all the exposure groups, no marked confounding between groups occurred. Irrespective of the baseline risk, the expected OR values in the partially ecologic analyses in both the linear model and the logistic model were generally close to the true value in all three exposure distributions when $OR_{ED} = 1.5$. For stronger associations between the exposure and the disease ($OR_{ED} \geq 2.0$), the linear model still generally performed well under all the exposure distributions with a baseline risk of ≤10%, but may be unduly biased for higher baseline risks. The partially ecologic analysis under the logistic model, on the other hand, performed well when $OR_{ED} = 2.0$ only if the baseline risk was high (50%) or the exposure distribution was widespread. When $OR_{ED} = 3.0$, however, the logistic model was

generally markedly biased, irrespective of the exposure distribution and baseline risk. Making the confounder less prevalent (P_C = 10%), changing OR_{CD}, or changing OR_{CE} did not change the relative performances of the two partially ecologic models (not in the tables). As expected, the individual-level analysis yielded biased OR_{ED} estimates irrespective of the baseline risk and exposure distribution (table 1).

Table 2. Expected odds ratio estimates for the association between exposure and disease (OR_{ED}) in the presence of confounding within and between groups [a], varying the exposure distribution, the association between the confounder and the exposure (OR_{CE}), and the range of the confounder prevalence (P_C) across 30 groups such that the rank correlation between P_C and the prevalence of the exposure (P_E) is one.

Exposure distribution for True OR_{ED} 1.5	Partially ecologic analysis (linear odds model) (OR_{ED})	Individual-level analysis (OR_{ED})
Widespread[b]		
OR_{CE} 1.0		
Range of P_C 0.45–0.55	1.6	1.5
Range of P_C 0.4–0.6	1.7	1.5
Range of P_C 0.3–0.7	1.9	1.5
Range of P_C 0.2–0.8	2.1	1.5
OR_{CE} 3.0		
Range of P_C 0.45–0.55	1.6	1.8
Range of P_C 0.4–0.6	1.7	1.8
Range of P_C 0.3–0.7	1.9	1.8
Range of P_C 0.2–0.8	2.0	1.8
Truncated[c]		
OR_{CE} 1.0		
Range of P_C 0.45–0.55	1.9	1.5
Range of P_C 0.4–0.6	2.3	1.5
Range of P_C 0.3–0.7	3.2	1.5
Range of P_C 0.2–0.8	4.2	1.5
OR_{CE} 3.0		
Range of P_C 0.45–0.55	1.8	1.8
Range of P_C 0.4–0.6	2.1	1.8
Range of P_C 0.3–0.7	2.5	1.8
Range of P_C 0.2–0.8	3.0	1.7
Rare[d]		
OR_{CE} 1.0		
Range of P_C 0.45–0.55	1.9	1.5
Range of P_C 0.4–0.6	2.4	1.5
Range of P_C 0.3–0.7	3.3	1.5
Range of P_C 0.2–0.8	4.2	1.5
OR_{CE} 3.0		
Range of P_C 0.45–0.55	1.8	1.8
Range of P_C 0.4–0.6	2.1	1.8
Range of P_C 0.3–0.7	2.5	1.7
Range of P_C 0.2–0.8	2.9	1.7

[a] The baseline risk was 0.1%. The OR for the association between the confounder and the disease was 2.0.
[b] The overall exposure prevalence, which was dependent on OR_{CE} and the range of P_C, varied between 20% and 27% in the different scenarios under the widespread exposure distribution.
[c] The overall exposure prevalence, which was dependent on OR_{CE} and the range of P_C, varied between 9% and 16% in the different scenarios under the truncated exposure distribution.
[d] The overall exposure prevalence, which was dependent on OR_{CE} and the range of P_C, varied between 2.9% and 6.1% in the different scenarios under the rare exposure distribution.

Confounding both within and between groups

For a baseline risk of 0.1%, we introduced marked confounding between groups by varying the confounder prevalence P_C between the groups according to the following equation:

$$P_C = a + b(G\text{-}1), \qquad \text{(equation 6)}$$

where G is group number (G = 1, 2,…, 30) and a and b are positive constants that determine the range of P_C variability (a = minimum prevalence, $a + 29b$ = maximum prevalence). As the groups were ranked in ascending order according to exposure prevalence P_E, this equation implies a perfect positive rank correlation between P_C and P_E. As expected, the variation of P_C across groups had no important impact on the expected OR values in the individual-level analysis (table 2). For the partially ecologic analysis, we restricted our attention to the linear odds model, which generally outperforms the logistic model when a disease is rare. The bias of the partially ecologic analysis increased as the variability of P_C increased and soon exceeded the bias of the corresponding individual-level analysis produced by the confounding within the groups (scenarios with OR_{CE} = 3.0 in table 2); the bias of the partially ecologic analysis was less pronounced when the exposure distribution was widespread rather than truncated or rare. Furthermore, the partially ecologic analysis was even more vulnerable to bias from confounding *between* groups when confounding *within* groups was absent (OR_{CE} = 1.0, table 2). If we, instead, introduced a perfect negative rank correlation between P_C and P_E, negative bias occurred in the partially ecologic analysis (not in the tables). Confounding between groups may produce marked bias of the partially ecologic analysis also when exposure has no effect (table 3).

Table 3. Expected odds ratio estimates for the association when the exposure has no effect (OR_{ED} = 1) in the presence of confounding within and between groups under the truncated exposure distribution [a], varying the range of the confounder prevalence (P_C) across 30 groups such that the rank correlation between P_C and the prevalence of the exposure (P_E) is one.

	Partially ecologic analysis (linear odds model) OR_{ED}	Individual-level analysis OR_{ED}
True OR_{ED} 1.0		
Range of P_C 0.45–0.55	1.2	1.2
Range of P_C 0.4–0.6	1.4	1.2
Range of P_C 0.3–0.7	1.8	1.2
Range of P_C 0.2–0.8	2.2	1.2

[a] The baseline risk was 0.1%. The OR for the association between the confounder and the disease was 2.0, and the OR between the confounder and the exposure was 3.0. The overall exposure prevalence, which was dependent on the range of P_C, varied between 15% and 16% in the different scenarios under the truncated exposure distribution.

Empirical example

In a large public health survey (N=13 715) conducted in the 33 municipalities of the region Skåne (Scania region) in southern Sweden in 2000, detailed questions regarding, for example, health status, smoking habits, alcohol consumption, social network, work environment, education, and residence were asked in a postal questionnaire (13). Here, we focused on a potential association between perceived interference from aircraft noise at home and sleeping disturbances. Severe sleeping disturbances during the last 2 weeks, which we regarded as the outcome variable, were reported by 7.9% of all the respondents.

Individual-level analysis

Some interference (including interference perceived as mild) from aircraft noise was reported by 13.4% of all the respondents, but only 1.6% reported fairly severe or severe interference. With adjustment for age in four broad categories (<30, 30–44, 45–64, and ≥65 years) and gender, the OR for the association between some interference from aircraft noise and sleeping disturbances was 1.3 [95% confidence interval (95% CI) 1.0–1.5], when the individual-level data were used. When only those reporting fairly severe or severe interference from aircraft noise were considered to be exposed, the effect estimate was less precise (OR 1.5, 95% CI 0.98–2.3). However these associations were confounded at the individual-level by perceived interference from noise from other sources. In particular, there was a strong association between at least mild interference from aircraft noise and from road noise (OR 4.4). When interference from road noise was adjusted in four categories (none, mild, fairly severe, severe), together with age and gender as before, the association between some interference from aircraft noise and sleeping disturbances was not evident (OR 1.1, 95% CI 0.91–1.3), and, furthermore, when the exposure definition was restricted to those reporting fairly severe or severe interference from aircraft noise, no effect was discerned (OR 1.0, 95% CI 0.63–1.6).

Partially ecologic analysis

The prevalence of some interference from aircraft noise at home varied between 4.1% and 38.8% across the 33 municipalities (mean 13.4%), but there was no obvious association between the prevalences of some interference from aircraft noise and road noise (Spearman rank correlation = 0.0). Thus a confounder that merely operates within groups does not have to be accounted for in a partially ecologic analysis. Adjustment for other covariates, age in four categories, and gender can be done in the partially ecologic analysis by extending the simple linear odds model to an additive-relative model (2, 11) as follows.

$$\text{Odds of disease} = \exp(\alpha_2) \times (1 + \beta_2 P_E) \times \exp(\textstyle\sum_j \gamma_j s_j), \qquad \text{(equation 7)}$$

where $s_1,\ldots, s_J$ are indicators for the different strata of age and gender and γ_j is the log-transformed OR associated with stratum j. An approximate 95% CI for OR $= 1 + \hat{\beta}_2$ was calculated as follows:

$$\exp\{\ln(1 + \hat{\beta}_2) \pm 1.96[\mathrm{SE}(\hat{\beta}_2)/(1 + \hat{\beta}_2)]\}, \qquad \text{(equation 8)}$$

where $\hat{\beta}_2$ is the maximum likelihood estimate of β_2 with standard error $\mathrm{SE}(\hat{\beta}_2)$ (2). The OR for the association between some interference from aircraft noise and sleeping disturbances in the additive-relative model was more imprecise than in the corresponding individual-level analysis, but it indicated no effect (OR 0.92, 95% CI 0.40–2.1). Restricting the exposure definition to fairly severe or severe interference from aircraft noise at home implied more limited variability in the exposure prevalence (range 0.0–6.5% across the 33 municipalities, mean 1.6%). This limited exposure variability across the municipalities under the more restrictive exposure definition made it impossible to obtain an effect estimate with reasonable precision in the partially ecologic analysis (OR 2.1, 95% CI 0.19–23).

Discussion

Biases in individual-level and corresponding ecologic, or partially ecologic, studies can indeed diverge (9). Nondifferential misclassification of a binary individual-level exposure variable yields bias towards the null (14). The corresponding partially ecologic analysis is biased away from the null when the aggregate exposure measure is established such that the misclassification has the same sensitivity and specificity in each group (15), but may also be biased towards the null under other error structures of the aggregate measure (2). Although seldom called "ecologic", many epidemiologic studies actually use some degree of aggregation, and such studies are, in practice, not necessarily inferior to truly individual-level studies (16), which may suffer from, for example, participation bias (17). Including aggregate measures from an exposure database for the nonparticipants may however reduce participation bias (18). We also addressed another aspect on using aggregate exposure measures, namely, how a partially ecologic analysis can affect the bias produced by an unmeasured confounder that may be present on the individual level.

In scenarios not untypical for clinical epidemiology, with a common outcome, relatively high exposure (treatment) prevalences in the groups (hospitals), a fairly modest treatment effect (OR 0.75), and data missing on an important confounder, Johnston et al (5) obtained valid OR estimates of the partially ecologic analysis using the logistic model. Bouyer & Hémon (11) have noted that the OR estimates in the logistic model are not unduly biased in partially ecologic analyses of rare diseases provided that there is no confounding between groups, the true OR ≤3, and the exposure prevalence varies over the entire range [0,1] between groups. According to our bias calculation for the partially ecologic analysis, the linear and the logistic models perform similarly when the effect of exposure is modest (here, OR 1.5), irrespective of the overall exposure prevalence, which varied between 5% and 25%, and baseline risk, which varied between 0.1% and 50% in our scenarios. For stronger exposure effects, however, the linear model generally outperforms the logistic model except when the baseline risk is very high (50% in our calculations). Consistent with our bias calculations, Bouyer & Hémon (11) observed that the logistic model may perform poorly in situations in which all the subjects are in groups with low exposure prevalences. Such exposure distributions are often encountered in occupational epidemiology.

We described the true exposure–disease associations in terms of OR values in order to make a fair comparison with the expected OR estimate in the logistic model and in the linear odds model. These two models can be used in case–control settings with partially ecologic data but also in follow-up settings with a follow-up period of fixed length. Depending on the design and baseline risk, the estimated OR may or may not be interpreted as an estimate of the risk ratio (19).

Our bias calculations further elucidated the robustness of the partially ecologic analysis with respect to uncontrolled confounding within groups, as observed by Johnston et al (5). However, lacking or inadequate individual-level data on a factor that acts as a confounder *between* groups is of much more concern in the partially ecologic analysis. If we introduce an association across groups between the prevalence of the uncontrolled confounder and the group-level exposure measure, the resulting confounding between groups may well lead to bias in the partially ecologic analysis that far exceeds the bias in the corresponding individual-level analysis. The bias of the partially ecologic analysis can be substantial even when the exposure has no effect. Furthermore, the partially ecologic analysis tends to be even more vulnerable to confounding between groups when the corresponding confounding within groups is weak or nonexistent. On the other hand, an exposure distribution with large variability in exposure prevalence across groups (eg, range 0.0–88% across groups in our bias calculations) tends to reduce somewhat the impact of confounding between groups. It is also worth noting that we performed our bias calculations in situations with a perfect rank correlation between the prevalence of the unmeasured confounder and the group-level exposure measure and with no other covariates available. In practice, a detailed grouping of the study population (20) and the inclusion of available covariates, such as age and gender (2), are likely to reduce considerably the rank correlation and thereby weaken the confounding between groups.

Group-level data on the confounder cannot generally replace individual-level data in order to achieve adequate confounding control (21). However, if both the exposure and the confounder are dichotomous on the individual-level and prevalence data are available for the confounder on the group level, Lasserre et al (22) showed that the bias due to confounding between groups can be much reduced if a cross-product term of the marginal prevalences is included in the regression model.

Another concern with the partially ecologic analysis is the markedly reduced precision when compared with that of the corresponding individual-level analysis. Bouyer et al (10) reported substantially reduced precision for the partially ecologic analysis in situations in which the exposure variability is limited and few subjects (here, ≤5%) are in the groups in which everyone is exposed. In the partially ecologic analysis in our empirical example, we obtained an acceptable loss of precision when compared with the result of the individual-level analysis when the overall exposure prevalence was above 10% and had a reasonable variability across groups. By contrast, for an exposure distribution with both a much lower overall exposure prevalence and a more-limited exposure variability, an effect estimation with acceptable precision may not be possible in the partially ecologic analysis.

In conclusion, the logistic regression model, with the proportion exposed in each group as the independent variable, can yield substantial model misspecification bias. By contrast, the linear odds model is valid when the outcome is rare. In a partially ecologic analysis, confounding between groups should be of more concern than confounding within groups.

Acknowledgments

We are grateful to Per-Olof Östergren, who provided access to the empirical data set.

The Swedish Council for Working Life and Social Research contributed with financial support.

References

1. Morgenstern H. Ecologic studies. In: Rothman K, Greenland S, editors. Modern epidemiology. 2nd edition. Philadelphia (PA): Lippincott-Raven, 1998:459–80.
2. Björk J, Strömberg U. Effects of systematic exposure assessment errors in partially ecologic case-control studies. Int J Epidemiol 2002;31:154–60.
3. Künzli N, Tager IB. The semi-individual study in air pollution epidemiology: a valid design as compared to ecologic studies. Environ Health Perspect 1997;105:1078–83.
4. Björk J, Albin M, Welinder H, Tinnerberg H, Mauritzson N, Kauppinen T, et al. Are occupational, hobby, or lifestyle exposures associated with Philadelphia chromosome-positive chronic myeloid leukemia? Occup Environ Med 2001;58:722–7.
5. Johnston SC, Henneman T, McCulloch CE, van der Laan M. Modeling treatment effects on binary outcomes with grouped-treatment variables and individual covariates. Am J Epidemiol 2002;156:753–60.
6. Kauppinen T, Toikkanen J, Pukkala E. From cross-tabulations to multipurpose exposure information systems: a new job-exposure matrix. Am J Ind Med 1998;33:409–17.
7. Robinson W. Ecological correlations and the behavior of individuals. Am Soc Rev 1950;15:351–7.
8. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 1989;18:269–74.
9. Greenland S. Divergent biases in ecologic and individual-level studies. Stat Med 1992;11:1209–23.
10. Bouyer J, Dardenne J, Hémon D. Performance of odds ratios obtained with a job-exposure matrix and individual exposure assessment with special reference to misclassification errors. Scand J Work Environ Health 1995;21(4):265–71.
11. Bouyer J, Hémon D. Comparison of three methods of estimating odds ratios from a job exposure matrix in occupational case-control studies. Am J Epidemiol 1993;137:472–81.
12. Hastie TJ, Tibshirani RJ. Generalized additive models. London: Chapman and Hall; 1990. p 97.
13. Kommunförbundet Skåne and Skåne Läns Allmänna Försäkringskassa. Hälsoförhållanden i Skåne [Public health conditions in the Scania region]. Malmö/Kristianstad (Sweden): Region Skåne, Kommunförbundet Skåne and Skåne Läns Allmänna Försäkringskassa; 2001.
14. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 1977;105:488–95.
15. Brenner H, Savitz DA, Jockel KH, Greenland S. Effects of nondifferential exposure misclassification in ecologic studies. Am J Epidemiol 1992;135:85–95.
16. Webster T. Commentary: does the spectre of ecologic bias haunt epidemiology? Int J Epidemiol 2002;31:161–2.
17. Olson SH, Voigt LF, Begg CB, Weiss NS. Reporting participation in case-control studies. Epidemiology 2002;13:123–6.
18. Strömberg U, Björk J. Incorporating group-level exposure information in case-control studies with missing data on dichotomous exposures. Epidemiology 2004;15:494–503.
19. Rothman KJ. Epidemiology: an introduction. New York (NY): Oxford University Press; 2002.
20. Morgenstern H. Uses of ecologic analysis in epidemiologic research. Am J Public Health 1982;72:1336–44.
21. Greenland S, Robins J. Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. Am J Epidemiol 1994;139:747–60.
22. Lasserre V, Guihenneuc-Jouyaux C, Richardson S. Biases in ecological studies: utility of including within-area distribution of confounders. Stat Med 2000;19:45–59.

Received for publication: 13 March 2004