Ecologic Studies Revisited

Jonathan Wakefield

Departments of Statistics and Biostatistics, University of Washington, Seattle, Washington 98195; email: jonno@u.washington.edu

Annu. Rev. Public Health 2008. 29:75–90

First published online as a Review in Advance on October 3, 2007

The *Annual Review of Public Health* is online at http://publhealth.annualreviews.org

This article's doi: 10.1146/annurev.publhealth.29.020907.090821

0163-7525/08/0421-0075$20.00

Key Words

confounding, ecological fallacy, spatial epidemiology

Abstract

Ecologic studies use data aggregated over groups rather than data on individuals. Such studies are popular because they use existing databases and can offer large exposure variation if the data arise from broad geographical areas. Unfortunately, the aggregation of data that define ecologic studies results in an information loss that can lead to ecologic bias. Specifically, ecologic bias arises from the inability of ecologic data to characterize within-area variability in exposures and confounders. We describe in detail particular forms of ecologic bias so that their potential impact on any particular study may be assessed. The only way to overcome such bias, while avoiding uncheckable assumptions concerning the missing information, is to supplement the ecologic with individual-level information, and we outline a number of proposals that may achieve this aim.

Ecological bias: the difference between associations at the individual and ecologic levels

INTRODUCTION

Ecologic studies are characterized by being based on grouped data, with the groups often corresponding to geographical areas. Such studies have a long history in many disciplines including political science (39), geography (50), sociology (59), and epidemiology and public health (48). Here we concentrate on the latter and discuss why ecologic studies are widely used and detail the unique drawbacks that lead to the potential for ecologic bias, which describes the difference between ecologic and individual associations. Ecologic data may be used for a variety of purposes including disease mapping (the geographical summarization of risk measures) and cluster detection (in which geographic anomalies are flagged); here we focus on geographical correlation studies that investigate associations between risk and exposure. In disease mapping, ecologic bias is not a problem because prediction of area-level risk summaries is the objective rather than the estimation of associations. Interesting within-area features may be masked by the process of aggregation; but although ecologic covariates may be used in disease-mapping models to improve predictions, the coefficients are not of direct interest (75).

Ecologic studies are popular for many reasons, the obvious one being the wide and increasing availability of aggregated health and population data; exposure information is usually less readily available. If the exposure is an environmental pollutant, then concentration information will rarely be aggregate in nature; it is more typical for measurements from a set of pollution monitors to be available. Nevertheless, we still refer to such nonindividual summaries as "ecologic." Improved ease of analysis also contributes to the widespread use of ecologic data. For example, geographical information systems (GIS) allow the effective storage and combination of data sets from different sources and with differing geographies (13, 14, 47, 58, 61), and recent advances in statistical methodology allow a more refined analysis of ecologic data (see References 21 and 78 for reviews).

There are numerous examples of ecologic studies in the public health and epidemiology literature. For example, **Figure 1** displays stomach cancer mortality in 1991–1993 vs. infant mortality in 1921–1923, each measured in 27 countries. One hypothesis that explains the apparent association is that stomach cancer risk is related to *H. pylori* infection, transmitted in the same way that diarrheal diseases contributed to diseases that caused the observed childhood mortality rates. However, testing this hypothesis about the cause of individual disease is complicated owing to the potential for ecological bias. Specifically, the 27 countries differ in many respects besides their rates of stomach cancer and infant mortality. The variables representing these differences may be related to both rates, and so the observed ecologic association may be due to confounding. The three highlighted countries, Japan, Russia, and Chile, "share very little in terms of their current socio-environmental conditions, and historically they are very different countries culturally, economically, and socially" (45); the implication is that confounding is not responsible for the simultaneous high values of the two rates. But confounding is harder to characterize in ecologic studies because it consists of both within-area and between-area components. For example, within each country there is variability in infant mortality rates, which may covary with confounders (see the Ecologic Bias section).

For motivation we briefly describe three additional ecological associations. Mortality rates for cervical cancer and the percentage pap test rate, both by state, are presented in figure 3 of Reference 61 as an example of an exploratory spatial analysis and to illustrate the flexibility of a GIS. In the context of income inequality and health, figure 1 of Reference 70 presents life expectancy vs. income inequality in 11 countries. The correlation between income inequality and health is −0.81, but the authors note that "data from aggregate-level studies of the effect of income

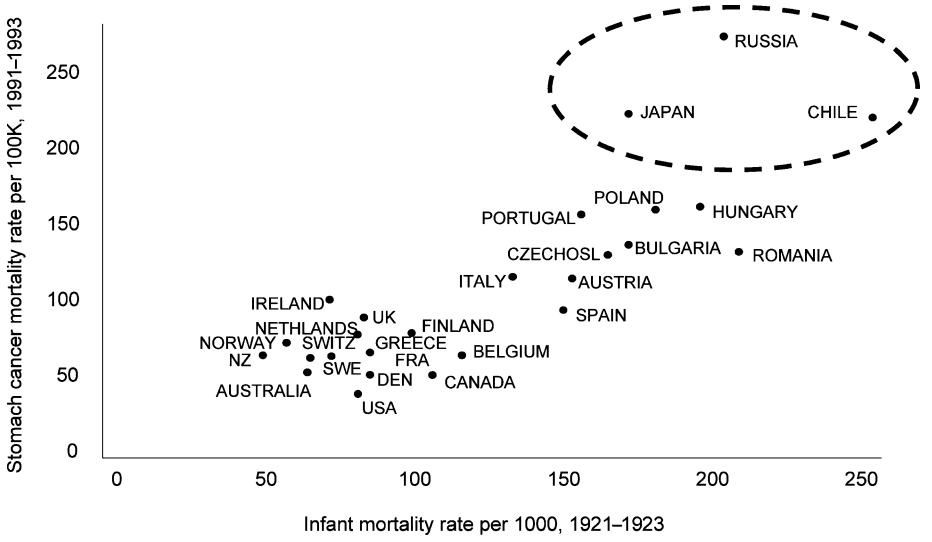


Figure 1

Stomach cancer mortality in 1991–1993 vs. infant mortality rate in 1921–1923 in 27 countries. Reprinted from the *Annual Review of Public Health*, Volume 26, 2005. Reproduced from Reference 45; data from Reference 42.

inequality on health . . . are largely insufficient to discriminate between competing hypotheses." Thus the information loss in ecologic studies leads to a fundamental identifiability problem: Many scientifically interesting models are indistinguishable from the observed aggregate data alone. Finally, the two plots of figure 5 in Reference 44 show the percentage of individuals with forced vital capacity less than 85% vs. two measures of particulate matter (<2.1 μm and 2.1–10 μm) for 22 U.S. and Canadian communities. These plots are based on semiecological data in that individual-level data on outcome and confounders are supplemented with ecological exposure information. Such studies are less susceptible to ecological bias owing to the increase in information when compared with a pure ecologic study (see Semi-Ecologic Studies section).

The paucity of exposure data has recently led a number of authors to combine ecologic population and health data with modeled exposure concentration surfaces; for a review of such modeling see Reference 36. For example, Zidek and colleagues (80) examine the association between daily hospital admissions for respiratory disease and sulfate concentrations, whereas Carlin et al. (9) examine the relationship between pediatric asthma emergency room visits and ozone, the latter modeled using kriging within a GIS. In each of these examples, great effort is placed on the modeling of the concentration surface without considering ecologic bias.

The structure of this review is to provide an illustrative ecologic study in the next section before cataloging a number of sources of bias in the Ecologic Bias section. Methods for supplementing ecologic data with individual-level information are described in the section Combining Ecologic and Individual Data.

ILLUSTRATIVE EXAMPLE: SIDS RISK IN NORTH CAROLINA

We examine data on sudden infant death syndrome (SIDS) and race; these data are available at the individual level, thus allowing the implications of aggregation to be examined. Mortality and birth data were obtained from the North Carolina State Center for Health Statistics Web

Ecological fallacy: the result of ecologic bias in which incorrect individual-level inference is drawn from ecologic data

Pure specification bias: due to nonlinear individual models changing their mathematical form under aggregation

site (**http://www.schs.state.nc.us/SCHS/**). SIDS cases, along with the number of live births, were extracted by race for each of the 100 counties of North Carolina for the years 2001–2004. There were a total of 386 cases, and **Figure 2*a*** shows the distribution of risk across the 100 areas. Race was categorized as white/nonwhite with 220 white deaths. There were 473,484 live births over the 4 years; 343,811 of them were white. **Figure 2*b*** shows the proportion of nonwhite births; across the counties, the proportion of nonwhite live births ranges between 0.006 and 0.733 with a median of 0.222 so that in most areas there are more white births than nonwhite births. The mortality rates for nonwhites and whites are 1.28 and 0.64 deaths per 1000 live births, respectively, giving a relative risk of 2.0 with asymptotic 95% confidence interval (1.64, 2.45).

We now assume that only ecological data are available. An ecologic dataset would consist of the proportion of nonwhite, $\bar{x}$, along with the number of SIDS deaths, y, and the total births, n, in each area. The top map in **Figure 3** displays the proportion of nonwhite, with areas of relative high frequency in the northeast and south, although these are not reflected in the risk map in the bottom figure. A naive ecologic model is given by

$$\text{Ecologic Risk} = e^{\alpha_e + \beta_e \bar{x}} \qquad 1.$$

and fitting this model gives an estimate of the ecologic relative risk e^{β_e}, of 0.89 (0.44–1.79), so that the risk point estimate decreases as the proportion of nonwhites increases, but the uncertainty is large and we would conclude that the ecologic data have little to say about the association. The fitted curve is superimposed on the scatterplot of y vs. $\bar{x}$ in **Figure 2*c***. *If* this point estimate was assumed to apply at the individual-level we would conclude that nonwhite babies are at lower risk than are white babies, the opposite of that found in the individual-level analysis, thus providing an example of the ecological fallacy. We return to the source of the fallacy after we discuss sources of ecological bias.

ECOLOGIC BIAS

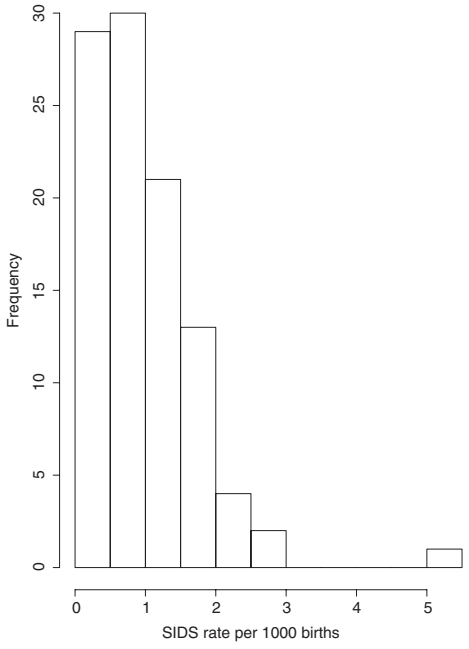
A vast literature describes sources of ecological bias (see for example, 25, 28, 29, 40, 48, 51, 52, 56, 57, 69, 73, 74, 75). The fundamental problem with ecological inference is that the aggregation process reduces information, and this information loss usually prevents identification of associations of interest in the underlying individual-level model. Ecologic bias is relative to a particular individual-level model. When trying to understand ecologic bias it is beneficial to specify an individual-level model and aggregate to determine the consequences (64, 74, 76). If there is no within-area variability in exposures and confounders, then there will be no ecological bias; so ecological bias occurs because of within-area variability in exposures and confounders, although a number of distinct consequences occur as a result of this variability. Ecologic bias is also referred to as aggregate, or cross-level, bias, the latter emphasizing the differing levels of the data and inference. Throughout this article, we assume that at the individual level the outcome is a 0/1 disease indicator.

Pure Specification Bias

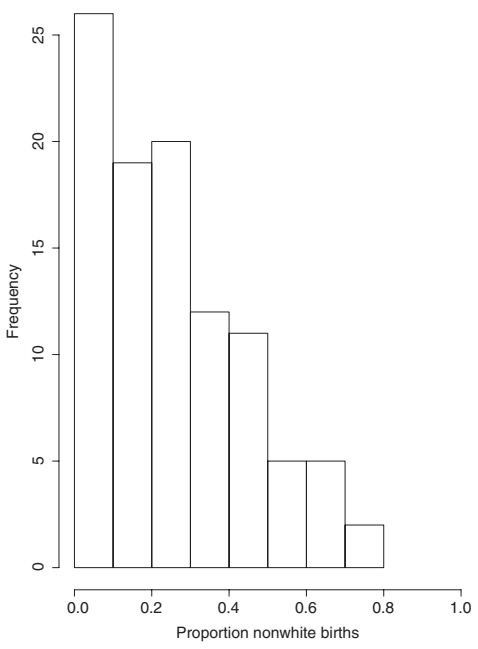
So-called pure specification bias (25) [also referred to as model specification bias (64)] arises because a nonlinear risk model changes its form under aggregation. We initially assume a single exposure x and the individual-level model

$$\text{Individual Risk} = e^{\alpha + \beta x}, \qquad 2.$$

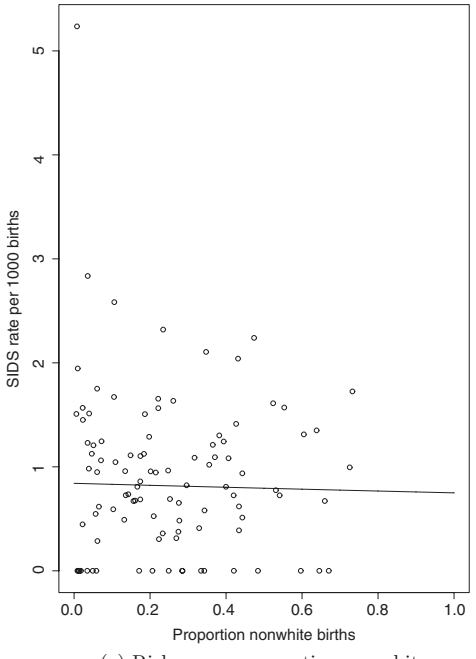
which is often used for a rare disease; e^{α} is the risk associated with $x = 0$ (baseline risk) and e^{β} is the relative risk corresponding to an increase in x of one unit. We concentrate on this model but will also comment on linear forms. Unfortunately the logistic model,



(a) SIDS risk (× 1000)



(b) Nonwhite proportion



(c) Risk versus proportion nonwhite

Figure 2

Proportion of nonwhite births and risk of SIDS (× 1000) across 100 counties of North Carolina in the years 2001–2004.

which is often used for nonrare outcomes, is not amenable to analytical study, and so the effects of aggregation are difficult to discern (63).

We consider a generic area containing n individuals with exposures x_i, $i = 1, \ldots, n$. Aggregation of Model 2 yields

$$\text{Ecologic Risk} = \frac{1}{n}\sum_{i=1}^{n} e^{\alpha+\beta x_i} \qquad 3.$$

so that the ecologic risk is the average of the risks of the constituent individuals. We let $\bar{x}$ represent the proportion of exposed individuals, i.e., $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. A naive ecologic model would assume

$$\text{Ecologic Risk} = e^{\alpha_e+\beta_e\bar{x}}, \qquad 4.$$

where the ecologic parameters α_e, β_e have been subscripted with "e" to distinguish them from the individual-level parameters in Model 2. Model 4 is a contextual effects model because risk depends on the proportion of exposed individuals in the area (see Contextual Effects section for further discussion). Interpreting e^{β_e} as an individual association would correspond with a belief that average exposure is causative and that individual exposure is irrelevant, or that the difference between the aggregated and individual exposures is negligible (so that within-area variability in exposure is small). The difference between Models 3 and 4 is clear: Whereas the former averages the risks across all exposures, the latter is the risk corresponding to the average exposure. We have $e^{\beta} = e^{\beta_e}$ only when there is no within-area variability in exposure so that $x_i = \bar{x}$ for all $i = 1, \ldots, n$ individuals. Hence pure specification bias is reduced in size as homogeneity of exposures within areas increases so that small areas are advantageous. Unfortunately data aggregation is usually carried out according to administration groupings and not to obtain areas with constant exposure.

For a binary exposure Model 2 can be written

$$e^{\alpha+\beta x} = (1-x)e^{\alpha} + xe^{\alpha+\beta},$$

which is linear in e^{α} and $e^{\alpha+\beta}$. This form simply yields the aggregate form

$$\text{Ecologic Risk} = (1-\bar{x})e^{\alpha} + \bar{x}e^{\alpha+\beta}, \qquad 5.$$

which shows that with a linear risk model there is no pure specification bias. If model 4 is fitted using a binary proportion $\bar{x}$, there will be no correspondence between e^{β} and e^{β_e} because they are associated with completely different comparisons. The extension to general categorical exposures is straightforward, and the parameters of the disease model are identifiable so long as we have the aggregate proportions in each category.

For continuous exposure, pure specification bias is dominated by the relationship between the within-area mean and variance of the exposure and will be small if the within-area variability is unrelated to the mean; if the variance increases with the mean (which will often be the case for environmental exposures), then overestimation of a harmful exposure ($\beta > 0$) will occur (73). Unfortunately this condition is impossible to assess without individual-level data on the exposure. If β is close to zero, pure specification bias is also likely to be small (because then the exponential model will be approximately linear for which there is no bias), although in this case confounding is likely to be a serious worry.

Pure specification bias will result unless we have a categorical variable and we know the within-area proportions in each category, except when the exposure is constant within areas or the risk model is linear. If the exposure is heterogeneous within areas, we need information on the variability within each area to control the bias. Such information may come from a sample of individuals within each area; how to use this individual-level data is the subject of the Combining Ecologic and Individual Data section.

Example Revisited

Returning to the North Carolina example, the discrepancy between the individual-level relative risk estimate of 2.0 and the ecologic

association of 0.89 derived from Model 1 is explained by pure specification bias; we fitted the contextual effects Model 4 and not the aggregate form in Model 5. Unfortunately fitting the latter model produces an estimate of 0.91 for these data. The reason for this discrepancy is that Model 5 is very unstable statistically and produces a likelihood surface that is highly irregular. In particular an asymptotic confidence interval is not appropriate here. This phenomenon has been observed elsewhere (30), which suggests that great care should be taken in fitting Model 5.

Confounding

We assume a single exposure x, a single confounder z, and the individual-level model

$$\text{Individual Risk} = e^{\alpha+\beta x+\gamma z}. \qquad 6.$$

As with pure specification bias, the key to understanding sources of, and correction for, ecological bias is to aggregate the individual-level model to give

$$\text{Ecologic Risk} = \frac{1}{n}\sum_{i=1}^{n} e^{\alpha+\beta x_i+\gamma z_i}. \qquad 7.$$

To understand why controlling for confounding is generally impossible with ecologic data, we consider the simplest case of a binary exposure and a binary confounder (which for ease of explanation we refer to as gender). **Table 1** shows the distribution of the exposure and confounder within a generic area. The complete within-area distribution of exposure and confounder can be described by three frequencies, but the ecologic data usually consist of only the proportion exposed, $\bar{x}$, and the proportion male, $\bar{z}$. From Model 7 the aggregate form is

$$\begin{aligned}&\text{Ecologic Risk}\\ &= p_{00}e^{\alpha} + p_{10}e^{\alpha+\beta} + p_{01}e^{\alpha+\gamma} + p_{11}e^{\alpha+\beta+\gamma}\\ &= (1-\bar{x}-\bar{z}+p_{11})e^{\alpha}\\ &\quad + (\bar{x}-p_{11})e^{\alpha+\beta} + (\bar{z}-p_{11})e^{\alpha+\gamma}\\ &\quad + p_{11}e^{\alpha+\beta+\gamma}, \qquad 8.\end{aligned}$$

showing that the marginal prevalances $\bar{x}$, $\bar{z}$ alone are not sufficient to characterize the joint distribution unless x and z are independent, in which case z is not a within-area confounder. This scenario has been considered in detail elsewhere (41), where it was argued that if the proportion of exposed males (p_{11}) is missing it should be estimated by the marginal prevalences ($\bar{x} \times \bar{z}$); however, we cannot determine the accuracy of this approximation without individual-level data. This is a recurring theme in the analysis of ecologic data. Bias can be reduced under model assumptions, but estimation is crucially dependent on the appropriateness of these assumptions, which are uncheckable without individual-level data.

We now turn to the situation with a binary exposure and a continuous confounder. Let the confounders in the unexposed be denoted, z_i, $i = 1, \ldots, n_0$, and the confounders in the exposed, z_i, $i = n_0 + 1, \ldots, n_0 + n_1$. In this case the ecologic form corresponding to Model 6 is

$$\text{Ecologic Risk} = q_0 \times r_0 + q_1 \times r_1,$$

where $q_0 = n_0/n$ and $q_1 = n_1/n$ are the probabilities of being unexposed and exposed, and

$$r_0 = \frac{e^{\alpha}}{n_0}\sum_{i=1}^{n_0} e^{\gamma z_i}, \qquad r_1 = \frac{e^{\alpha+\beta}}{n_1}\sum_{i=n_0+1}^{n_0+n_1} e^{\gamma z_i}$$

so that r_0 and r_1 are the aggregated risks in the unexposed and exposed. Thus we need the confounder distribution *within* each exposure category, unless z is not a within-area confounder. The requirement for stratum-defined exposure distributions is closely related to mutual standardization as described

Table 1 Exposure and gender distribution in a generic area: $\bar{x}$ is the proportion exposed and $\bar{z}$ is the proportion male; p_{00}, p_{01}, p_{10}, and p_{11} are the within-area cross-classification frequencies

	Female	Male	
Unexposed	p_{00}	p_{01}	$1-\bar{x}$
Exposed	p_{10}	p_{11}	$\bar{x}$
	$1-\bar{z}$	$\bar{z}$	1.0

in Reference 60, which requires exposure distributions to be standardized with respect to a confounder, if risk has been standardized to this confounder. Again, if we fit the model

$$\text{Ecologic Risk} = e^{\alpha_e + \beta_e \bar{x} + \gamma_e \bar{z}},$$

where $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$, then the coefficient β has no relation to β_e in the naive ecologic model.

Often one can attempt to control for confounding via expected numbers E using the regression model:

$$\text{Ecologic Risk} = E \times e^{\alpha_e + \beta_e \bar{x}}$$

(17, 18, 22). This approach implicitly assumes, however, that there is no within-area confounding (75). For example, the expected numbers are often calculated on the basis of the age and gender distribution, but this controls only for between-area confounding and will provide confounder control only if the within-area exposure distribution is the same across age and gender stratum. This is not likely to hold for age in particular. Whenever an ecologic study is considered, the ability to control for known confounders for the disease/exposure under investigation should be considered. For most chronic diseases, known lifestyle risk factors include one or more of the following: smoking, alcohol, and diet. In an ecologic study, individual-level information on these variables is not available, and it has become popular to attempt to control for these variables using area-level measures of socioeconomic status (e.g., 46). Although these measures may be strongly correlated with lifestyle variables (20), they cannot pick up the subtleties of within-area confounding. Therefore, unless the association of interest is strong, ecologic results controlled for confounding in this way should be interpreted with great caution.

The extension to general exposure and confounder scenarios is obvious from the discussion above. If we have true confounders that are constant within areas (for example, access to health care), then they are analogous to conventional confounders because the area is the unit of analysis. Hence, the implications are relatively easy to understand and adjustment is straightforward.

Without an interaction between exposure and confounder, the parameters of a linear model are estimable from marginal information only, although if an interaction is present within-area information is required.

Contextual Effects

A contextual variable represents a characteristic of individuals in a shared neighborhood, and in some scenarios (for example, the measurement of health disparities) such effects are of great interest. For example, the mean income in an area, in addition to individual income, has been hypothesized as being predictive of health (37). We consider the simple individual-level linear model

$$E[Y_i | x_i, \bar{x}] = \alpha + \beta_W (x_i - \bar{x}) + \beta_B \bar{x}, \quad 9.$$

where β_B is the between-area (contextual) effect and β_W is the within-area individual effect. The aggregate form is

$$E[\bar{Y} | \bar{x}] = \alpha + \beta_B \bar{x},$$

which shows that both individual and contextual effects cannot be simultaneously estimated without individual-level data. In a nonlinear model, both effects may be estimable with ecologic data, but the amount of information concerning β_W is small (64). More importantly, the derivation for the linear model reveals that estimation is crucially dependent on the form of the nonlinear model, and we cannot check the form of the model from only the ecologic data. Hence, although sensitivity analyses to identify both parameters may be carried out, inference is totally unreliable with only ecologic data. Investigators (26) have also noted that when contextual effects are of interest they are susceptible to cross-level bias when estimated from ecologic data.

Some have argued, in dietary and environmental contexts, that the contextual exposure

$\bar{x}$ may be a better estimate of exposure for an individual than x_i would be when individual-level measurement error is large. For example, Navidi et al. (49) propose a design that combines individual-level regression with ecologic comparisons to combine the best aspects of each data source; individual-level analyses are free of ecologic bias but may have poor power and measurement error in exposures, each of which may be rectified in ecologic data.

In general, multilevel models have provided a popular framework for analyzing associations at different geographical scales (for example, to estimate neighborhood effects), but these models cannot control for confounding due to unmeasured variables, and the interpretation of parameters is not always straightforward. The usual interpretation of a parameter associated with a particular variable is revealed by increasing the variable by one unit, while keeping all other variables fixed. Consideration of Model 9 illustrates the difficulties in applying this approach in cases where the variable appears at more than one level. Suppose we wish to interpret β_W; if we increase x_i by one unit, the mean also increases by $1/n$. To interpret β_W we must keep the mean in the area constant, for example, by reducing everyone else's x by $1/(n-1)$. Further discussion may be found in Reference 27, and interpretation of more complex models is provided in References 2 and 67.

Semi-Ecologic Studies

Table 2 summarizes four distinct scenarios in terms of data availability (40, 64). In a semiecologic study, sometimes more optimistically referred to as a "semi-individual study" (40), individual-level data are collected on outcome and confounders, with exposure information arising from another source. In the Harvard six-cities study (16), for example, the exposure was city specific and an average of pollution monitors the study's follow-up.

We consider the risk for an individual in confounder stratum c; under aggregation we have

$$\text{Semi-Ecologic Risk in stratum } c = e^{\alpha+\gamma_c} \sum_{i=1}^{n_c} e^{\beta x_{ci}},$$

where x_{ci} are the exposures of individuals within stratum c, $i = 1, \ldots, n_c$, and γ_c is the baseline risk in stratum c. A naive semiecologic model is

$$\text{Semi-Ecologic Risk in stratum } c = e^{\alpha_e+\gamma_e c+\beta_e x}, \quad 10.$$

where x is some summary exposure measure. Kunzli & Tager (40) argue that semiecologic studies are free of ecologic bias, but there are two possible sources of bias here: The first is that we have pure specification bias because we have not acknowledged within-area variability in exposure; the second is that we have not allowed the exposure to vary by confounder stratum, so we have not controlled for within-area confounding. In an air pollution study in multiple cities, x may correspond to a monitor average or an average over several monitors. In this case, Model 10 will provide an approximately unbiased estimate of β if there is small exposure variability in cities and if this variability is similar across confounder strata.

Semiecologic studies frequently have survival as an endpoint, but there has been less focus on the implications of aggregation in the context of survival models. References 1 and 33 discuss some of the implications.

Table 2 Study designs by level of outcome and exposure data

		Exposure	
		Individual	Ecologic
Outcome	Individual	Individual	Semi-ecologic
	Ecologic	Aggregate	Ecologic

Semiecologic study: individual-level data are available on outcome and confounders, with an ecologic exposure assessment

Spatial Dependence and Hierarchical Modeling

When data are available as counts from a set of contiguous areas, we might expect residual dependence between the counts, particularly

for small-area studies, owing to the presence of unmeasured variables with spatial structure. The use here of the word "residual" acknowledges that variables known to influence the outcome have already been adjusted for in the mean model. Analysis methods that ignore the dependence are strictly not applicable, with inappropriate standard errors being the most obvious manifestation. A great deal of work has focused on models for spatial dependence (3, 5, 10–12, 15, 38, 43); Richardson (55) provides an excellent review of this literature. Regarding ecological bias, the most important message is that unless the mean model is correct, adjustment for spatial dependence is a pointless exercise (75).

In a much-cited book, King (39) proposed a hierarchical model to analyze ecologic data in a political science context as "a solution to the ecological inference problem." Identifiability in this model is imposed through the random effects prior, however, and we cannot check the appropriateness of this prior from the ecological data alone (23, 74).

COMBINING ECOLOGIC AND INDIVIDUAL DATA

As we saw in the previous section, the only solution to the ecologic inference problem that does not require uncheckable assumptions is the supplementation of ecologic-level with individual-level data. We stress that ecologic data can also supplement already available individual data to improve power. Here we briefly review some of the proposals for such an endeavor. The obvious approach is to collect a random sample of individuals within areas. For a continuous outcome, Raghunathan et al. (54) show that moment and maximum likelihood estimates of a common within-group correlation coefficient will improve when aggregate data are combined with individual data within groups, and Glynn et al. (24) derive optimal design strategies for the collection of individual-level data when the model is linear. With a binary nonrare outcome, the benefits have also been illustrated (68, 74).

For a rare disease few cases will be present in the individuals within the sample, and so only information on the distribution of exposures and confounders will be obtained via a random sampling strategy (which is therefore equivalent to using a survey sample of covariates only). This prompted the derivation of the so-called aggregate data method of Prentice & Sheppard (53, 65, 66) (**Table 2**). Inference proceeds by constructing a model based on the sample of $m \leq n$ individuals in each area and estimates the mean (which is given by Model 3 for the case of a single exposure), based on the empirical averages. This is an extremely powerful design because estimation is not based on any assumptions with respect to the within-area distribution of exposures and confounders [though this distribution may not be well characterized for small samples (62)]. Ecologic bias is reduced to a greater extent than in the semiecologic study because within-area variability in exposures and confounders is acknowledged.

An alternative approach is to assume a parametric distribution for the within-area distribution of exposures and confounders (57, 76), although this implicitly assumes that a sample of these is available (see also 34, 35). As an example, if we assume that exposures in an area are normally distributed with mean $\bar{x}$ and variance s^2, then the implied ecologic risk is $e^{\alpha+\beta\bar{x}+\beta^2 s^2/2}$, and this model may be fitted to ecologic data if $\bar{x}$ and s^2 are available in each area (4). More recently investigators have suggested an approach that takes the mean as a combination of these two approaches, with the parametric approach dominating for small samples (for which the aggregate data method can provide unstable inference) (62).

A different approach in the context of a rare disease is outcome-dependent sampling, which avoids the problem of zero cases encountered in random sampling. Inferential approaches have been developed for when ecologic data are supplemented with individual case-control information gathered within

the constituent areas (30–32). The case-control data remove ecologic bias, whereas the ecologic data provide increased power and constraints on the sampling distribution of the case-control data, which improves the precision of estimates.

Two-phase methods have a long history in statistics and epidemiology (7, 8, 77, 79) and are based on an initial crossclassification by outcome and confounders and exposures; this classification provides a sampling frame within which additional covariates may be gathered via the sampling of individuals. Such a design may be used in an ecologic setting, where the initial classification is based on one or more area(s), confounder strata, and possibly error-prone measures of exposure (J. Wakefield & S. Haneuse, submitted manuscript).

In all these approaches it is clearly vital to avoid response bias in the survey samples or selection bias in outcome-dependent sampling, and establishing a relevant sampling frame is essential.

CONCLUDING REMARKS

The use of ecological data is ubiquitous. This article has concentrated on area-aggregated data, but many other variables can be collapsed. For example, it is common practice to collapse continuous age into five-year age bands; this results in a loss of information, but within each age bands the changes in risk are small and so ecologic bias will be ignorable.

A skeptic might conclude from the litany of potential biases described in the ecologic bias section, above, that ecologic inference should never be attempted, but this would be too pessimistic a view. A useful starting point for all ecologic analyses is to write down an individual-level model for the outcome-exposure association of interest, including known confounders. Ecologic bias may be small when within-area variability in exposures and known confounders is small, and for small-area studies in particular this may be approximately true. A serious source of bias is that caused by confounding because ecologic data on exposure are rarely stratified by confounder strata within areas. If a small area study has been carried out with a correctly aggregated individual-level model, then parameter estimates can be cautiously interpreted at the individual level and compared with other studies at the individual level, and hence add to the totality of evidence for a hypothesis. When comparing ecologic and semiecologic estimates with individual-level estimates, it is clearly crucial to have a common effect measure (e.g., a relative risk or a hazard ratio). So, for example, it will be difficult to compare an ecologic correlation coefficient, which is an often-reported measure, with an effect estimate from an individual-level study.

Less well-designed ecologic studies can be suggestive of hypotheses to investigate if strong ecologic associations are observed. An alternative to the pessimistic view often applied to ecological analyses is that when a strong ecological association, such as that observed in **Figure 1**, is seen, one should attempt to explain how such a relationship could have arisen if it is not due to the ecologic predictor.

We have not discussed a number of issues. Care should be taken in determining the effects of measurement error in an ecologic study because the directions of bias may not be predictable. For example, in the absence of pure specification and confounder bias for linear and log-linear models, if there is nondifferential measurement error in a binary exposure, there will be overestimation of the effect parameter, in contrast with individual-level studies (6). We refer interested readers to alternative sources (19, 71) for other issues such as consideration of migration, latency periods, and the likely impacts of inaccuracies in population and health data.

Studies that investigate the acute effects of air pollution are another common situation in which ecologic exposures are used. For example, daily disease counts in a city are often regressed against daily and/or lagged concentration measurements taken from a monitor or the average of a collection of monitors to

estimate the acute effects of air pollution. If day-to-day exposure variability is greater than within-city variability, then we would expect ecologic bias to be relatively small.

With respect to data availability, exposure information is generally not aggregate in nature (unless the "exposure" is a demographic or socioeconomic variable), and in an environmental epidemiological setting the modeling of pollutant concentration surfaces will undoubtedly grow in popularity. However, an important insight is that in a health-exposure modeling context it may be better to use measurements from the nearest monitor, rather than model the concentration surface, because the latter approach may be susceptible to large biases particularly when, as is usually the case, the monitoring network is sparse (72). A remaining challenge is to diagnose when the available data are of sufficient abundance and quality to support the use of complex models.

We have described a number of proposals for the combination of ecologic and individual data. Such endeavors will no doubt increase, and will hopefully allow the reliable exploitation of ecologic information.

SUMMARY POINTS

1. Ecologic bias, defined as the difference between associations obtained from individual and ecologic data, occurs because of within-group variability in exposures and/or confounders.
2. To understand the implications of the use of ecologic data in any setting, it is useful to write down the individual-level model that would be fitted if individual-level data were available. Aggregation of an individual-level model allows the characterization of ecologic bias and reveals the individual-level data that would reduce the chance of ecologic bias.
3. Ecologic bias can be safely removed only by combining ecologic- and individual-level data.
4. Semiecologic studies are less susceptible to ecologic bias because some components of bias are not possible, but again the implications of aggregation should be carefully examined.

DISCLOSURE STATEMENT

The author is not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENT

This work was supported by grant R01 CA095994 from the National Institutes of Health.

LITERATURE CITED

1. Abrahamowicz M, du Berger R, Krewski D, Burnett R, Bartlett G, et al. 2004. Bias due to aggregation of individual covariates in the Cox regression model. *Am. J. Epidemiol.* 160:696–706
2. Berhane K, Gauderman WJ, Stram DO, Thomas DC. 2004. Statistical issues in studies of the long-term effects of air pollution: the Southern California Children's Health study. *Stat. Sci.* 19:414–34

3. Besag J, York J, Mollié A. 1991. Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Stat. Math.* 43:1–59
4. Best N, Cockings S, Bennett J, Wakefield J, Elliott P. 2001. 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* 164:155–74
5. Best N, Ickstadt K, Wolpert R. 2000. Ecological modelling of health and exposure data measured at disparate spatial scales. *J. Am. Stat. Assoc.* 95:1076–88
6. Brenner H, Savitz D, Jockel K-H, Greenland S. 1992. Effects of nondifferential exposure misclassification in ecologic studies. *Am. J. Epidemiol.* 135:85–95
7. Breslow NE, Cain KC. 1988. Logistic regression for two-stage case-control data. *Biometrika* 75:11–20
8. Breslow NE, Chatterjee N. 1999. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Appl. Stat.* 48:457–68
9. Carlin BP, Xia H, Devine O, Tolbert P, Mulholland J. 1999. Spatio-temporal hierarchical models for analyzing Atlanta pediatric asthma ER visit rates. In *Case Studies in Bayesian Statistics*, Vol. IV, ed. C Gatsonis, RE Kass, B Carlin, A Carriquiry, A Gelman, et al., pp. 303–20. New York: Springer
10. Christensen OF, Waagepetersen R. 2002. Bayesian prediction of spatial count data using generalised linear mixed models. *Biometrics* 58:280–86
11. Clayton D, Bernardinelli L, Montomoli C. 1993. Spatial correlation in ecological analysis. *Int. J. Epidemiol.* 22:1193–202
12. Cressie N, Chan NH. 1989. Spatial modelling of regional variables. *J. Am. Stat. Assoc.* 84:393–401
13. Cromley EK. 2003. GIS and disease. *Annu. Rev. Public Health* 24:7–24
14. Croner CM. 2003. Public health, GIS and the Internet. *Annu. Rev. Public Health* 24:57–82
15. Diggle PJ, Tawn JA, Moyeed RA. 1998. Model-based geostatistics (with discussion). *Appl. Stat.* 47:299–350
16. Dockery D, Pope CA III, Xiping X, Spengler J, Ware J, et al. 1993. An association between air pollution and mortality in six U.S. cities. *N. Engl. J. Med.* 329:1753–59
17. Eaton N, Shaddick G, Dolk H, Elliott P. 1997. Small-area study of the incidence of neoplasms of the brain and central nervous system among adults in the West Midlands. *Br. J. Cancer* 75:1080–83
18. Elliott P, Shaddick G, Kleinschmidt I, Jolley D, Walls P, et al. 1996. Cancer incidence near municipal solid waste incinerators in Great Britain. *Br. J. Cancer* 73:702–7
19. Elliott P, Wakefield JC. 1999. Small-area studies of environment and health. In *Statistics for the Environment 4: Health and the Environment*, ed. V Barnett, A Stein, KF Turkman, pp. 3–27. New York: Wiley
20. Elliott P, Wakefield JC. 2000. Bias and confounding in spatial epidemiology. See Elliott et al., pp. 68–84
21. Elliott P, Wakefield JC, Best NG, Briggs DJ. 2000. *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford Univ. Press
22. Elliott P, Westlake AJ, Kleinschmidt I, Hills M, Rodrigues L, et al. 1992. The small area health statistics unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom. *J. Epidemiol. Community Health* 46:345–49
23. Freedman DA, Klein SP, Ostland M, Roberts MR. 1998. A solution to the ecological inference problem (book review). *J. Am. Stat. Assoc.* 93:1518–22
24. Glynn A, Wakefield J, Handcock M, Richardson T. 2008. Alleviating linear ecological bias and optimal design with subsample data. *J. R. Stat. Soc. Ser. A.* 171:In press

25. Greenland S. 1992. Divergent biases in ecologic and individual level studies. *Stat. Med.* 11:1209–23
26. Greenland S. 2001. Ecologic vs individual-level sources of bias in ecologic estimates of contextual health effects. *Int. J. Epidemiol.* 30:1343–50
27. Greenland S. 2002. A review of multilevel theory for ecologic analyses. *Stat. Med.* 21:389–95
28. Greenland S, Morgenstern H. 1989. Ecological bias, confounding and effect modification. *Int. J. Epidemiol.* 18:269–74
29. Greenland S, Robins J. 1994. Ecological studies: biases, misconceptions and counterexamples. *Am. J. Epidemiol.* 139:747–60
30. Haneuse S, Wakefield J. 2007. Geographic-based ecological correlation studies using supplemental case-control data. *Stat. Med.* 26:In press
31. Haneuse S, Wakefield J. 2007. Hierarchical models for combining ecological and case-control data. *Biometrics* 63:128–36
32. Haneuse S, Wakefield J. 2008. The combination of ecological and case-control data. *J. R. Stat. Soc. Ser. B.* 70:In press
33. Haneuse S, Wakefield J, Sheppard L. 2007. The interpretation of exposure effect estimates in chronic air pollution studies. *Stat. Med.* 26:3172–87
34. Jackson C, Best N, Richardson S. 2008. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *J. R. Stat. Soc. Ser. A.* 171:In press
35. Jackson CH, Best NG, Richardson S. 2006. Improving ecological inference using individual-level data. *Stat. Med.* 25:2136–59
36. Jerrett M, Afrain A, Kanaroglou P, Beckerman B, Potoglou D, et al. 2005. A review and evaluation of intraurban air pollution exposure models. *J. Expo. Anal. Environ. Epidemiol.* 15:185–204
37. Judge K, Mulligan J, Benzeval M. 1998. Income inequality and population health. *Soc. Sci. Med.* 46:565–79
38. Kelsall JE, Wakefield JC. 2002. Modeling spatial variation in disease risk: a geostatistical approach. *J. Am. Stat. Assoc.* 97:692–701
39. King G. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton Univ. Press
40. Künzli N, Tager IB. 1997. The semi-individual study in air pollution epidemiology: a valid design as compared to ecologic studies. *Environ. Health Perspect.* 10:1078–83
41. Lasserre V, Guihenneuc-Jouyaux C, Richardson S. 2000. Biases in ecological studies: utility of including within-area distribution of confounders. *Stat. Med.* 19:45–59
42. Leon DA, Smith GD. 2000. Infant mortality, stomach cancer, stroke, and coronary heart disease: ecological analysis. *Br. Med. J.* 320:1705–6
43. Leroux BG, Lei X, Breslow N. 1999. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical Models in Epidemiology. The Environment and Clinical Trials*, ed. ME Halloran, DA Berry, pp. 179–92. New York: Springer
44. Lippmann M, Schlesinger RB. 2000. Toxicological bases for the setting of health-related air pollution standards. *Annu. Rev. Public Health* 21:309–33
45. Lynch J, Smith GD. 2005. A life course approach to chronic disease epidemiology. *Annu. Rev. Public Health* 26:1–35
46. Maheswaran R, Morris S, Falconer S, Grossinho A, Perry I, et al. 1999. Magnesium in drinking water supplies and mortality from acute myocardial infaction in North West England. *Heart* 82:455–60

47. McLafferty SL. 2003. GIS and health care. *Annu. Rev. Public Health* 24:25–42
48. Morgenstern H. 1995. Ecologic studies in epidemiology: concepts, principles, and methods. *Annu. Rev. Public Health* 16:61–81
49. Navidi W, Thomas D, Stram D, Peters J. 1994. Design and analysis of multilevel analytic studies with applications to a study of air-pollution. *Environ. Health Perspect.* 102(Suppl. 8):25–32
50. Openshaw S. 1984. *The Modifiable Areal Unit Problem*. Norwich: Geo Books
51. Piantadosi S, Byar DP, Green SB. 1988. The ecological fallacy. *Am. J. Epidemiol.* 127:893–904
52. Plummer M, Clayton D. 1996. Estimation of population exposure. *J. R. Stat. Soc. Ser. B* 58:113–26
53. Prentice RL, Sheppard L. 1995. Aggregate data studies of disease risk factors. *Biometrika* 82:113–25
54. Raghunathan TE, Diehr PK, Cheadle AD. 2003. Combining aggregate and individual level data to estimate an individual level correlation coefficient. *J. Educ. Behav. Stat.* 28:1–19
55. Richardson S. 2003. Spatial models in epidemiological applications. In *Highly Structured Stochastic Systems*, ed. PJ Green, NL Hjort, S Richardson, pp. 237–59. Oxford: Oxford Stat. Sci. Ser.
56. Richardson S, Montfort C. 2000. Ecological correlation studies. See Elliott et al., pp. 205–20
57. Richardson S, Stucker I, Hémon D. 1987. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *Int. J. Epidemiol.* 16:111–20
58. Ricketts TC. 2003. Geographic information systems and public health. *Annu. Rev. Public Health* 24:1–6
59. Robinson WS. 1950. Ecological correlations and the behavior of individuals. *Am. Sociol. Rev.* 15:351–57
60. Rosenbaum PR, Rubin DB. 1984. Difficulties with regression analyses of age-adjusted rates. *Biometrics* 40:437–43
61. Rushton G. 2003. Public health, GIS, and spatial analytic tools. *Annu. Rev. Public Health* 24:43–56
62. Salway R, Wakefield J. 2007. A hybrid model for reducing ecological bias. *Biostatistics* 8:In press
63. Salway RA, Wakefield JC. 2005. Sources of bias in ecological studies of nonrare events. *Environ. Ecol. Stat.* 12:321–47
64. Sheppard L. 2003. Insights on bias and information in group-level studies. *Biostatistics* 4:265–78
65. Sheppard L, Prentice RL. 1995. On the reliability and precision of within- and between-population estimates of relative rate parameters. *Biometrics* 51:853–63
66. Sheppard L, Prentice RL, Rossing MA. 1996. Design considerations for estimation of exposure effects on disease risk, using aggregate data studies. *Stat. Med.* 15:1849–58
67. Sheppard L, Wakefield J. 2004. Discussion of: Statistical issues in studies of the long-term effects of air pollution: the Southern California Children's Health Study. *Stat. Sci.* 19:438–41
68. Steel DG, Beh EJ, Chambers RL. 2004. The information in aggregate data. In *Ecological Inference: New Methodological Strategies*, ed. G King, O Rosen, M Tanner, pp. 51–68 Cambridge, UK: Cambridge Univ. Press

69. Steel DG, Holt D. 1996. Analysing and adjusting aggregation effects: the ecological fallacy revisited. *Int. Stat. Rev.* 64:39–60
70. Wagstaff A, van Doorslaer E. 2000. Income inequality and health: What does the literature tell us? *Annu. Rev. Public Health* 21:543–67
71. Wakefield J, Elliott P. 1999. Issues in the statistical analysis of small area health data. *Stat. Med.* 18:2377–99
72. Wakefield J, Shaddick G. 2006. Health-exposure modelling and the ecological fallacy. *Biostatistics* 7:438–55
73. Wakefield JC. 2003. Sensitivity analyses for ecological regression. *Biometrics* 59:9–17
74. Wakefield JC. 2004. Ecological inference for 2 × 2 tables (with discussion). *J. R. Stat. Soc. Ser. A* 167:385–445
75. Wakefield JC. 2007. Disease mapping and spatial regression with count data. *Biostatistics* 8:158–83
76. Wakefield JC, Salway RE. 2001. A statistical framework for ecological and aggregate studies. *J. R. Stat. Soc. Ser. A* 164:119–37
77. Walker AM. 1982. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics* 38:1025–32
78. Waller LA, Gotway CA. 2004. *Applied Spatial Statistics for Pubic Health Data*. New York: Wiley
79. White JE. 1982. A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am. J. Epidemiol.* 115:119–28
80. Zidek JV, White R, Lee ND, Sun W, Burnett RT. 1998. Imputing unmeasured explanatory variables in environmental epidemiology with application to health impact analysis of air pollution. *Environ. Ecol. Stat.* 5:99–115

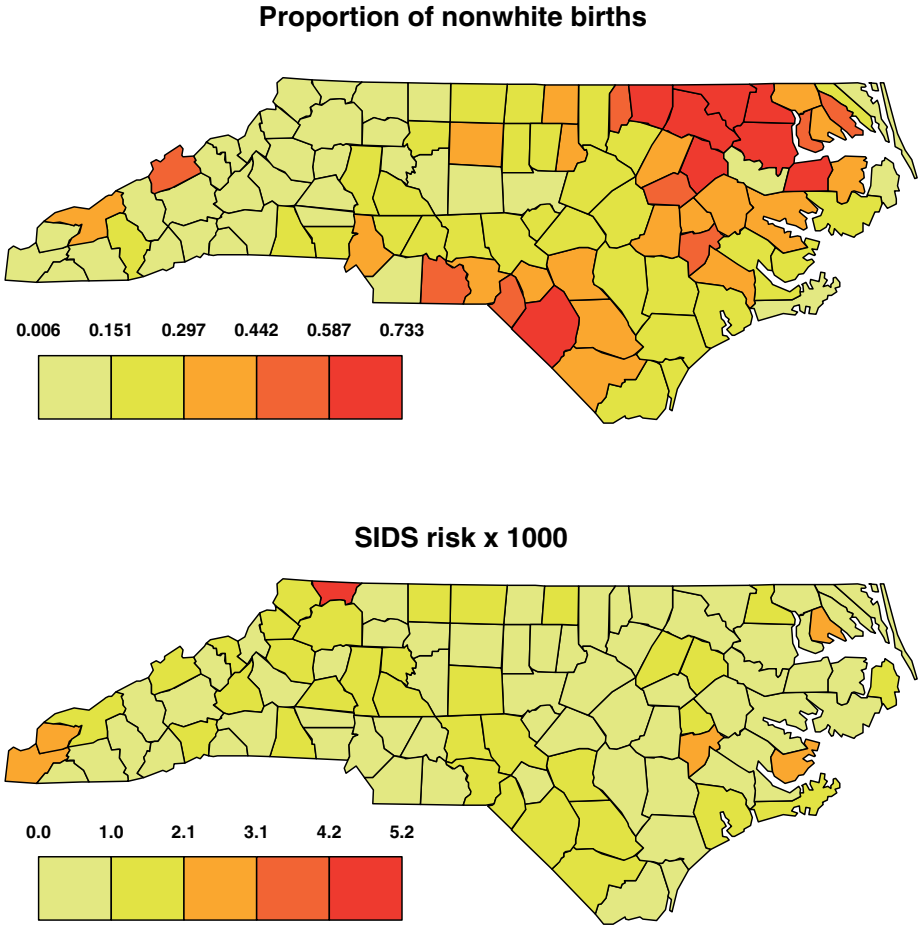


Figure 3

Maps of the proportion of nonwhite births and risk of SIDS (× 1000) across 100 counties of North Carolina in the years 2001–2004.

Annual Review of
Public Health

Volume 29, 2008

Contents

Environmental and Occupational Health

Public Health Practice

Social Environment and Behavior

Health Services

Indexes

Errata

An online log of corrections to *Annual Review of Public Health* articles may be found at http://publhealth.annualreviews.org/