# THE B FILES

**Case studies of bias in real life epidemiologic studies**

## Bias File 3. Émile Durkheim and the ecological fallacy

Compiled by

Madhukar Pai, MD, PhD

Jay S Kaufman, PhD

Department of Epidemiology, Biostatistics & Occupational Health

McGill University, Montreal, Canada

madhukar.pai@mcgill.ca & jay.kaufman@mcgill.ca

# Bias File 3. Émile Durkheim and the ecological fallacy

## The story

Émile Durkheim (1858 –1917) was a famous French sociologist and pioneer in the development of modern sociology and anthropology. In a groundbreaking book published in 1897, entitled *Le Suicide*, Durkheim explored the differing suicide rates among Protestants and Catholics.  In 19th century Europe, suicide rates were higher in countries that were more heavily Protestant. Durkheim found that suicide rates were highest in provinces that were heavily Protestant. He concluded that stronger social control among Catholics resulted in lower suicide rates. However, Durkheim's study of suicide was criticized as an example of the logical error termed the "ecological fallacy." Indeed, it is one of the most famous examples of ecological fallacy. So, what went wrong and why?

## The study

Durkheim's study of religion and suicide used data from four groups of Prussian provinces between 1883 and 1890. The groups were formed by ranking 13 provinces according to the proportion (X) of the population that was Protestant. Durkheim found that suicide rates (Y) were highest in provinces that were heavily Protestant. He concluded that stronger social control among Catholics resulted in lower suicide rates. According to Durkheim, Catholic society had normal levels of 'integration' while Protestant society has low levels.

Using ordinary least-squares linear regression on Durkheim's data, Morgenstern (1995) found a strong positive correlation (Figure below, from Morgenstern 1995) between proportion protestant and suicide rates. The estimated rate ratio, comparing Protestants with other religions, was 7.6 (i.e. suicide rates among protestants was about 8 fold higher than other religions).

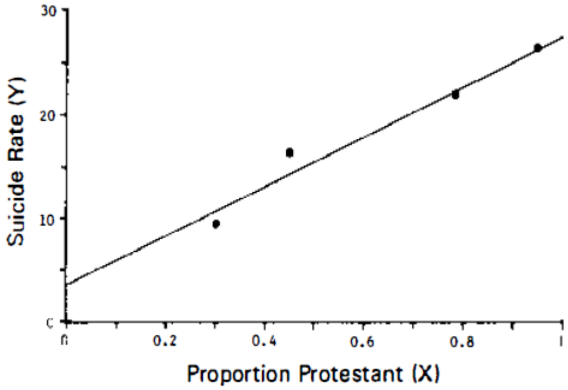


***Figure 2*** **Suicide rate ($Y$, per $10^5$/year) by proportion Protestant ($X$) for four groups of Prussian provinces, 1883–90. The four observed points ($X$, $Y$) are (0.30, 9.56), (0.45, 16.36), (0.785, 22.00), and (0.95, 26.46); the fitted line is based on unweighted least-squares regression [Source: Adapted from Durkheim (16)].**

### The bias

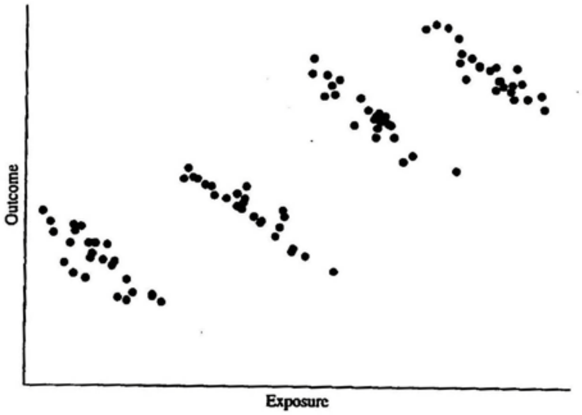
Ecological fallacy is a well recognized concept in sociology (Robinson 1950). A good description of the ecological fallacy in Durkheim's work is provided by Morgenstern (1995 & 2008). According to Morgenstern, the estimated rate ratio of 7.6 was probably not because suicide rates were nearly 8 fold higher in Protestants than in non-Protestants. Rather, because none of the regions was entirely Protestant or non-Protestant, it may have been non-Protestants (primarily Catholics) who were committing suicide in predominantly Protestant provinces. It is plausible that members of a religious minority might have been more likely to commit suicide than were members of the majority. Living in a predominantly Protestant area had a contextual effect on suicide risk among Catholics.

Interestingly, Morgenstern points out that Durkheim compared the suicide rates at the individual level for Protestants, Catholics and Jews living in Prussia, and from his data, the rate was about twice as great in Protestants as in other religious groups. Thus, when the rate ratios are compared (2 vs 8), there appears to be substantial ecological bias using the aggregate level data.

There are more striking examples. One compelling example by Robinson (1950), was the relationship between nativity (foreign vs native born) and literacy. For each of the 48 states in the USA of 1930, Robinson computed two numbers: the percent of the population who were foreign-born (i.e. immigrants), and the percent who were literate. He found the correlation between the 48 pairs of numbers was .53. This ecological correlation suggested a positive association between foreign birth and literacy: the foreign-born (immigrants) are more likely to be literate than the native-born. In reality, the association was negative: the correlation computed at the individual level was −0.11 (immigrants were less literate than native citizens). The ecological correlation gave the incorrect inference. This is because the foreign-born (immigrants) tended to migrate to and settle in states where the native-born are relatively literate. In this example by Robinson, the correlation is totally reversed.

Ecological fallacy arises from thinking that relationships observed for groups necessarily hold for individuals: if provinces with more Protestants tend to have higher suicide rates, then Protestants must be more likely to commit suicide; if countries with more fat in the diet have higher rates of breast cancer, then women who eat fatty foods must be more likely to get breast cancer. Such inferences made using group-level data may not always be correct at the individual level.

Ecological bias can be interpreted as the failure of associations seen at one level of grouping to correspond to effect measures at the grouping level of interest. For example, associations seen using country-level data may not correlate with associations that exist at the individual or neighborhood-level. The Figure (from Koepsell and Weiss 2003) illustrates this nicely. Within each of the four populations, as exposure increases, outcome decreases. But across populations, as the mean exposure level increases, the mean rate of outcome increases.



Source: Epidemiologic Methods. Thomas Koepsell & Noel Weiss. Oxford Univ Press, 2003

## The lesson

As emphasized by Morgenstern (1995 & 2008), several practical advantages make ecologic studies especially appealing for undertaking various types of epidemiologic research. Despite these advantages, ecologic analysis poses problems of interpretation when making inferences at the individual level. The correlation at the group level was valid in both Durkheim and Robinson examples. It was only invalid as a statement of individual causal effect. As pointed out by Greenland (2001), if we have predictors at the individual and the group level, and we want the causal effects at one or the other level, then our ecological level analysis could be confounded by omitted variables at the individual level.

## Sources and suggested readings*

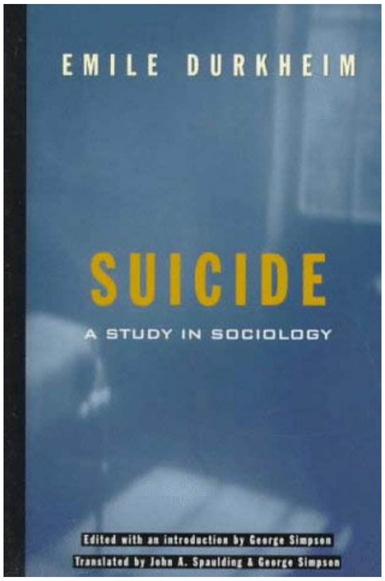

EMILE DURKHEIM
SUICIDE
A STUDY IN SOCIOLOGY
Edited with an introduction by George Simpson
Translated by John A. Spaulding & George Simpson

Image credit: Wikipedia

Durkheim E. Suicide, (1897), The Free Press reprint 1997.

Robinson WS. Ecological correlations and the behavior of individuals. Am Sociol Rev 1950;15:351-57.

Morgenstern H. Ecologic Studies in Epidemiology: Concepts, Principles, and Methods. Annual Review of Public Health 1995; Vol. 16: 61-81.

Morgenstern H. Ecological studies. In: Modern Epidemiology. 3rd Edition. Editors: Rothman, Greenland, Lash. Lippincott Williams and Wilkins, 2008.

Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol. 1989 Mar;18(1):269-74.

Greenland S. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. Int J Epidemiol. 2001 Dec;30(6):1343-50.

Koepsell T & Weiss W. Epidemiologic Methods. Oxford Univ Press, 2003.

*From this readings list, the most relevant papers are enclosed.

*Annu. Rev. Public Health 1995. 16:61–81*

# ECOLOGIC STUDIES IN EPIDEMIOLOGY: Concepts, Principles, and Methods

*Hal Morgenstern*
Department of Epidemiology and Center for Occupational and Environmental Health, University of California, Los Angeles, School of Public Health, Los Angeles, California 90024–1772

KEY WORDS: epidemiologic methods, study design, sources of bias, causal inference, aggregate studies

## Abstract

An ecologic study focuses on the comparison of groups, rather than individuals; thus, individual-level data are missing on the joint distribution of variables within groups. Variables in an ecologic analysis may be aggregate measures, environmental measures, or global measures. The purpose of an ecologic analysis may be to make biologic inferences about effects on individual risks or to make ecologic inferences about effects on group rates. Ecologic study designs may be classified on two dimensions: (*a*) whether the primary group is measured (exploratory vs analytic study); and (*b*) whether subjects are grouped by place (multiple-group study), by time (time-trend study), or by place and time (mixed study). Despite several practical advantages of ecologic studies, there are many methodologic problems that severely limit causal inference, including ecologic and cross-level bias, problems of confounder control, within-group misclassification, lack of adequate data, temporal ambiguity, collinearity, and migration across groups.

## INTRODUCTION

An ecologic or aggregate study focuses on the comparison of groups, rather than individuals. The underlying reason for this focus is that individual-level data are missing on the joint distribution of at least two and perhaps all

0163-7525/95/0501-0061$05.00

variables within each group; in this sense, an ecologic study is an incomplete design (35). Ecologic studies have been conducted by social scientists for more than a century (14a) and .have been used extensively by epidemiologists in many research areas. Nevertheless, the distinction between individual-level and group-level (ecologic) studies and the inferential implications are far more complicated and subtle than they first appear. Before 1980, ecologic studies were usually presented in the first part of epidemiology textbooks as simple descriptive analyses in which disease rates are stratified by place and/or time to generate or test hypotheses; little attention was given to statistical methods or inference (e.g. 41). The purpose of this review is to provide a methodologic overview of ecologic studies that emphasizes study design and causal inference. Although ecologic studies are easily and inexpensively conducted, the results are often difficult to interpret.

## CONCEPTS AND RATIONALE

Before discussing the design and interpretation of ecologic studies, we must first define the concepts of ecologic measurement, analysis, and inference.

### *Levels of Measurement*

The sources of data used in epidemiologic studies typically involve direct observations of individuals (e.g. age and sex), sometimes subindividual parts (e.g. intraocular pressure of each eye), and occasionally groups or regions (e.g. air pollution and social disorganization). These direct observations are then organized to measure specific variables in the study population: Individual-level variables are properties of individuals, and ecologic variables are properties of groups. To be more specific, ecologic measures may be classified into three types:

1. *Aggregate measures* are summaries (e.g. means or proportions) of observations derived from individuals in each group (e.g. the proportion of smokers or median family income).
2. *Environmental measures* are physical characteristics of the place in which members of each group live or work (e.g. air-pollution level or hours of sunlight). Note that each environmental measure has an analogue at the individual level, and these individual exposures, or doses, usually vary among members of each group, though they may remain unmeasured.
3. *Global measures* are attributes of groups or places for which there is no distinct analogue at the individual level, unlike aggregate and environmental measures (e.g. population density, level of social disorganization, or the existence of a specific law).

## *Levels of Analysis*

The unit of analysis is the common level for which the data on all variables are reduced and analyzed. In an *individual-level analysis,* a value for each variable is assigned to every subject in the study. It is possible, even common in environmental epidemiology, for one or more variables to be ecologic measures. For example, the average pollution level of each county might be assigned to every resident of that county.

In a *completely ecologic analysis,* all variables (exposure, disease, and covariates) are ecologic measures, so the unit of analysis is the group (e.g. region, worksite, school, demographic stratum, or time interval). Thus, within each group, we do not know the joint distribution of any combination of variables at the individual level (e.g. the frequencies of exposed cases, unexposed cases, exposed noncases, and unexposed noncases); all we know is the marginal distribution of each variable (e.g. the proportion exposed and the disease rate—the T frequencies in Figure 1).

In a *partially ecologic analysis* of three or more variables, we have additional information on certain joint distributions (the M and/or N frequencies in Figure 1 and/or rarely the L frequencies); but we still do not know the full joint distribution of all variables within each group (i.e. the ? cells in Figure 1 are missing). For example, in an ecologic study of cancer incidence by county, the joint distribution of age (a covariate) and disease status within each county (the M frequencies in Figure 1) might be obtained from the census and a population tumor registry.

*Multilevel analysis* is a special type of modeling technique that combines analyses conducted at two or more levels (6, 71, 72). For example, an individual-level analysis might be conducted in each group, followed by an ecologic analysis of all groups using the results from the individual-level analyses. This approach is described in a later section.

C

| | D | $\bar{D}$ | |
|---|---|---|---|
| E | ? | ? | $N_{CE}$ |
| $\bar{E}$ | ? | ? | $N_{C\bar{E}}$ |
| | $M_{CD}$ | $M_{C\bar{D}}$ | $T_C$ |

$\bar{C}$

| | D | $\bar{D}$ | |
|---|---|---|---|
| E | ? | ? | $N_{\bar{C}E}$ |
| $\bar{E}$ | ? | ? | $N_{\bar{C}\bar{E}}$ |
| | $M_{\bar{C}D}$ | $M_{\bar{C}\bar{D}}$ | $T_{\bar{C}}$ |

Total

| | D | $\bar{D}$ | |
|---|---|---|---|
| E | $L_{ED}$ | $L_{E\bar{D}}$ | $T_E$ |
| $\bar{E}$ | $L_{\bar{E}D}$ | $L_{\bar{E}\bar{D}}$ | $T_{\bar{E}}$ |
| | $T_D$ | $T_{\bar{D}}$ | T |

*Figure 1* Joint distribution of exposure status (E vs $\bar{E}$), disease status (D vs $\bar{D}$), and covariate status (C vs $\bar{C}$) in each group of a simple ecologic analysis: T frequencies are the only data available in a completely ecologic analysis of all three variables; M frequencies require additional data on the joint distribution of C and D within each group; N frequencies require additional data on the joint distribution of E and C within each group; L frequencies require additional data on the joint distribution of E and D within each group (rarely available); and ? cells are missing in an ecologic analysis.

## *Levels of Inference*

The underlying goal of a given epidemiologic study or analysis may be to make *biologic* (or *biobehavioral*) *inferences* about effects on individual *risks* or to make *ecologic inferences* about effects on group *rates* (45). The target level of causal inference, however, does not always match the level of analysis. For example, the purpose of an ecologic analysis may be to make a biologic inference about the effect of a specific exposure on disease risk. As we see later in this review, such *cross-level inferences* are particularly vulnerable to bias.

If the objective of a study is to estimate the *biologic effect* of wearing a motorcycle helmet on the risk of motorcycle-related mortality among motorcycle riders, the target level of causal inference is biologic. On the other hand, if the objective is to estimate the *ecologic effect* of helmet-use laws on the motorcycle-related mortality rate of riders in different states, the target level of causal inference is ecologic. Note that the magnitude of this ecologic effect depends not only on the biologic effect of helmet use but also on the degree and pattern of compliance with the law in each state. Furthermore, the validity of the ecologic-effect estimate depends on our ability to control for differences among states in the joint distribution of confounders, including individual-level variables such as age and amount of motorcycle riding.

We might also be interested in estimating the *contextual effect* of an ecologic exposure on individual risk, which is also a form of biologic inference (5, 64). If the ecologic exposure is an aggregate measure, we would generally want to separate its effect from the effect of its individual-level analogue. For example, we might estimate the contextual effect of living in a poor area on the risk of disease, controlling for individual poverty level (33). Similarly, in evaluating motorcycle-helmet laws in the U.S., we might want to estimate the contextual effect of living in a state that mandates helmet use on the risk of motorcycle-related mortality in riders, controlling for individual helmet use. Contextual effects are also relevant in infectious-disease epidemiology, where the risk of disease depends on the prevalence of the disease in others with whom the individual has contact (37, 65).

## *Rationale for Ecologic Studies*

There are several reasons for the widespread use of ecologic studies in epidemiology, despite frequent cautions about their methodologic limitations:

1. *Low cost and convenience* Ecologic studies are inexpensive and take little time because various secondary data sources, each involving different information needed for the analysis, can easily be linked at the aggregate level. For example, data obtained from population registries, vital records,

large surveys, and the census are often linked at the state, county, or census-tract level.

2. *Measurement limitations of individual-level studies* In environmental epidemiology and other research areas, we often cannot accurately measure relevant exposures or doses at the individual level for large numbers of subjects—at least not with available time and resources. Thus, the only practical way to measure the exposure may be ecologically (45, 46). This advantage is especially true when investigating apparent clusters of disease in small areas (66). Sometimes individual-level exposures, such as dietary factors, cannot be measured accurately because of substantial within-person variability; yet ecologic measures might accurately reflect group averages (31).
3. *Design limitations of individual-level studies* Individual-level studies may not be practical for estimating exposure effects if the exposure varies little within the study area. However, ecologic studies covering a much wider area might be able to achieve substantial variation in mean exposure across groups (e.g. 50).
4. *Interest in ecologic effects* As noted above, the stated purpose of a study may be to assess an ecologic effect, i.e. the target level of inference may be ecologic rather than biologic. Ecologic effects are particularly relevant when evaluating the impacts of population interventions such as new programs, policies, or legislation.
5. *Simplicity of analysis and presentation* In large, complex studies conducted at the individual level, it may be conceptually and statistically simpler to perform ecologic analyses and to present ecologic results than to do individual-level analyses. For example, data from large, periodic surveys, such as the National Health Interview Survey, are often analyzed ecologically by treating some combination of year, region, and demographic group as the unit of analysis.

## STUDY DESIGNS

In an ecologic study design, the planned unit of analysis is the group. Ecologic designs may be classified on two dimensions: the method of exposure measurement and the method of grouping (35, 45). Regarding the first dimension, an ecologic design is called *exploratory* if the primary exposure of potential interest is not measured, and *analytic* if the primary exposure variable is measured and included in the analysis. In practice, this dimension is a continuum, since most ecologic studies are not conducted to test a single hypothesis. Regarding the second dimension, the groups of an ecologic study may be identified by place (multiple-group design), by time (time-trend design), or by a combination of place and time (mixed design).

## Multiple-Group Study

EXPLORATORY In this type of exploratory study, we compare the rate of disease among many regions during the same period. The purpose is to search for spatial patterns that might suggest an environmental etiology or more specific etiologic hypotheses. For example, the National Cancer Institute (NCI) mapped the age-adjusted cancer mortality rates in the U.S. by county for the period 1950–69 (42). For oral cancers, they found a striking difference in geographic patterns by sex: Among men, the mortality rates were greatest in the urban Northeast, but among women, the rates were greatest in the Southeast. These findings led to the hypothesis that snuff dipping, which is common among rural southern women, is a risk factor for oral cancers (2). The results of a subsequent case-control study supported this hypothesis (70).

Exploratory ecologic studies may also involve the comparison of rates between migrants and their offspring and residents of their countries of emigration and immigration (31, 41). If the rates differ appreciably between the countries of emigration and immigration, migrant studies often yield results suggesting the influence of certain types of risk factors for the disease under study. For example, if US immigrants from Japan have rates of a disease similar to US whites but much lower than Japanese residents, the difference may be due to environmental or behavioral risk factors operating during adulthood. However, the interpretation of results from these studies is often limited by differences between countries in the classification and detection of disease or cause of death.

In mapping studies, such as the NCI investigation, a simple comparison of rates across regions is often complicated by two statistical problems. First, regions with smaller numbers of observed cases show greater variability in the estimated rate; thus the most extreme rates tend to be observed for those regions with the fewest cases. Second, nearby regions tend to have more similar rates than do distant regions (i.e. autocorrelation) because unmeasured risk factors tend to cluster in space. Statistical methods for dealing with both problems have been developed by fitting the data to an autoregressive spatial model and using empirical Bayes techniques to estimate the smoothed rate for each region (9, 44, 47). The degree of spatial autocorrelation or clustering can be measured to reflect environmental effects on the rate of disease (68, 69). The empirical Bayes approach can also be applied to data from analytic multiple-group studies (described below) by including covariates in the model (e.g. 8, 12).

ANALYTIC In this type of study, we assess the ecologic association between the average exposure level or prevalence and the rate of disease among many groups. This is the most common ecologic design; typically, the unit of analysis is a geopolitical region. For example, Hatch & Susser (29) examined the association between background gamma radiation and the incidence of childhood cancers between 1975 and 1985 in the region surrounding a nuclear power plant. Average radiation levels for each of 69 tracts in the region were estimated from a 1976 aerial survey. The authors found positive associations between radiation level and the incidence of leukemia (an expected finding) as well as solid tumors (an unexpected finding).

Data analysis in this type of multiple-group study usually involves fitting the data to a mathematical model. For example, Prentice & Sheppard (51) proposed a linear relative rate model using iteratively reweighted least-squares procedures to estimate the model parameters. Prentice & Thomas (52) also considered an exponential relative rate model, which, they argue, may be more parsimonious than the linear-form model for specifying covariates. These methods can be applied to data aggregated by place and/or time (to be discussed below). Use of ecologic modeling to estimate exposure effects is described in the next section.

## *Time-Trend Study*

EXPLORATORY An exploratory time-trend or time-series study involves a comparison of the disease rates over time in one geographically defined population. In addition to providing graphical displays of temporal trends, time-series data can also be used to forecast future rates and trends. This latter application, which is more common in the social sciences than in epidemiology, usually involves fitting the outcome data to autoregressive integrated moving average (ARIMA) models (30, 48). The method of ARIMA modeling can also be extended to evaluate the impact of a population intervention (43), to estimate associations betweens two or more time-series variables (7, 48), and to estimate associations in a mixed ecologic design (60; see below).

A special type of exploratory time-trend analysis often used by epidemiologists is age-period-cohort (or cohort) analysis. Through graphical displays or formal modeling techniques, the objective of this approach is to estimate the separate effects of three time-dependent variables on the rate of disease: age, period (calendar time), and birth cohort (year of birth) (32, 35). Because of the linear dependency of these three variables, there is an inherent statistical limitation (identification problem) with the interpretation of age-period-cohort results. The problem is that each data set has alternative explanations with respect to the combination of age, period, and cohort effects; there is no unique set of effect parameters when all three variables are considered simultaneously. The only way to decide which interpretation should be accepted is to consider the findings in light of prior knowledge and, possibly, to constrain the model by ignoring one effect.

Lee et al (40) conducted an age-period-cohort analysis of melanoma mortality among white males in the U.S. between 1951 and 1975. They concluded that the apparent increase in the melanoma mortality rate was due primarily to a cohort effect. That is, persons bom in more recent years experienced throughout their lives a higher rate than did persons bom earlier. In a subsequent paper, Lee (39) speculated that this cohort effect might reflect increases in sunlight exposure or sunbuming during youth.

ANALYTIC In this type of time-trend study, we assess the ecologic association between change in average exposure level or prevalence and change in disease rate in one geographically defined population. As with exploratory designs, this type of assessment can be done by simple graphical displays or by time-series regression modeling (e.g. 48). With either approach, however, the interpretation of findings is often complicated by two problems. First, changes in disease classification and diagnostic criteria can produce very misleading results. Second, the latency of the disease with respect to the primary exposure may be long, variable across cases, or simply unknown. Thus, employing an arbitrary lag between observations—or an empirically defined lag that maximizes the estimated association between the two trends—can also produce misleading results (28).

Darby & Doll (13) examined the associations between average annual absorbed dose of radiation fallout from weapons testing and the incidence rate of childhood leukemia in three European countries between 1945 and 1985. Although the leukemia rate varied over time in each country, they found no convincing evidence that these changes were attributable to changes in fallout radiation.

### Mixed Study

EXPLORATORY The mixed ecologic design combines the basic features of the multiple-group study and the time-trend study. Time-series (ARIMA) modeling or age-period-cohort analysis can be used to describe or predict trends in the disease rate for multiple populations. For example, to test Lee's (39) hypothesis that changes in sunlight exposure during youth can explain the observed increase in melanoma mortality in the U.S., we might conduct an age-period-cohort analysis, stratifying on region according to approximate sunlight exposure (without measuring the exposure). Assuming the amount of sunlight in the regions has not changed differentially over the study period, we might expect the cohort effect described above to be stronger for sunnier regions.

ANALYTIC In this type of mixed ecologic design, we assess the association between change in average exposure level or prevalence and change in disease rate among many groups. Thus the interpretation of estimated effects is enhanced because two types of comparisons are made simultaneously: change over time within groups and differences among groups. For example, Crawford et al (11) evaluated the hypothesis that hard drinking water (i.e. water with a high concentration of calcium and magnesium) is a protective risk factor for cardiovascular disease (CVD) mortality. They compared the absolute change in CVD mortality rate between 1948 and 1964 in 83 British towns, by water-hardness change, age, and sex. In all sex-age groups, especially for men, the authors found an inverse association between water-hardness change and CVD mortality. In middle-aged men, for example, the increase in CVD mortality was less in towns that made their water harder than in towns that made their water softer.

## EFFECT ESTIMATION

A major quantitative objective of most epidemiologic studies is to estimate the effect of one or more exposures on disease occurrence in a well-defined population at risk. A measure of effect in this context is not just any measure of association, such as a correlation coefficient; rather, it reflects a particular causal parameter, i.e. a counterfactual contrast in disease occurrence (21, 24, 27, 46, 58). In studies conducted at the individual level, effects are usually estimated by comparing the rate or risk of disease, in the form of a ratio or difference, for exposed and unexposed populations. In multiple-group ecologic studies, however, we cannot estimate effects directly in this way because of the missing information on the joint distribution within groups. Instead, we regress the group-specific disease rates (*Y*) on the group-specific exposure prevalences (*X*). For example, fitting the data to a linear model produces the following prediction equation: $\hat{Y} = B_0 + B_1X$, where $B_0$ and $B_1$ are the estimated intercept and slope, using ordinary least-squares methods. The estimated biologic effect of the exposure (at the individual level) can be derived from the regression results (1, 19). The predicted disease rate ($\hat{Y}$) in a group that is entirely exposed is $B_0 + B_1(1) = B_0 + B_1$, and the predicted rate in a group that is entirely unexposed is $B_0 + B_1(0) = B_0$. Therefore, the estimated rate difference is $B_1$ and the estimated rate ratio is $1 + B_1/B_0$. Note that this ecologic method of effect estimation requires rate predictions be extrapolated to both extreme values of the exposure variable (i.e. $X = 0$ and 1), which are likely to lie well beyond the observed range of the data. It is not surprising, therefore, that different model forms (e.g. log-linear vs linear) can lead to very different estimates of effect (22). Fitting a linear model, in fact, may lead to negative, and thus meaningless, estimates of the rate ratio.

As an illustration of rate-ratio estimation in an ecologic study, consider Durkheim's (16) examination of religion and suicide in four groups of Prussian provinces between 1883 and 1890 (see Figure 2). The groups were formed by

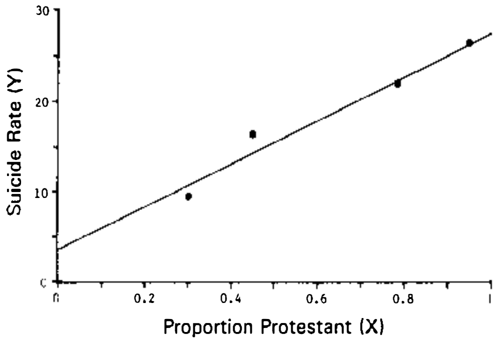


*Figure 2* Suicide rate (*Y*, per $10^5$/year) by proportion Protestant (*X*) for four groups of Prussian provinces, 1883–90. The four observed points (*X*, *Y*) are (0.30, 9.56), (0.45, 16.36), (0.785, 22.00), and (0.95, 26.46); the fitted line is based on unweighted least-squares regression [Source: Adapted from Durkheim (16)].

ranking 13 provinces according to the proportion (*X*) of the population that was Protestant. Using ordinary least-squares linear regression, we estimate the suicide rate ($\hat{Y}$, per $10^5$/year) in each group to be 3.66 + 24.0(*X*). Therefore, the estimated rate ratio, comparing Protestants with other religions, is 1 + (24.0/3.66) = 7.6. Note in Figure 2 that the fit of the linear model is excellent ($R^2 = 0.97$).

There are two methods used to control for confounders in multiple-group ecologic analyses. The first is to treat ecologic measures of the confounders as covariates (**Z**) in the model, e.g. percent male and percent white in each group. If the individual-level effects of the exposure and covariates are additive (i.e. if the disease rates follow a linear model), then the ecologic regression of *Y* on *X* and **Z** will also be linear with the same coefficients (22, 38). That is, the estimated coefficient for the exposure variable can be interpreted as the rate difference adjusted for other covariates, analogously to the crude estimate discussed above.

The second method used to control for confounders in ecologic analyses is rate standardization for these confounders (57), followed by regression of the standardized rates as the outcome variable. Note that this method requires additional data on the joint distribution of the covariate and disease within each group (i.e. the M frequencies in Figure 1). Nevertheless, it cannot be expected to reduce bias unless all predictors in the model (*X* and **Z**) are mutually standardized for the same confounders (22, 25, 56). Standardization of the exposure prevalences, for example, requires data on the joint distribution of the covariate and exposure within groups (i.e. the N frequencies in Figure 1); however, this information is not often available in ecologic studies.

As in individual-level analyses, product terms (e.g. $XZ$) are often used in ecologic analyses to model interaction effects, i.e. to assess effect modification. In ecologic analyses, however, the product of $X$ and $Z$ (both group averages) is not, in general, equal to the average product of the exposure ($x$) and covariate ($z$) at the individual level within groups. Assuming a linear model, $XZ$ will be equal to the mean $xz$ in each group only if $x$ and $z$ are uncorrelated within groups (22). Thus, as pointed out in the next section, interaction (nonadditive) effects at the individual level complicate the interpretation of ecologic results.

## METHODOLOGIC PROBLEMS

Despite the many practical advantages of ecologic studies mentioned previously, there are several methodologic problems that may severely limit causal inference, especially biologic inference.

### *Ecologic Bias*

The major limitation of ecologic analysis for making causal inferences is ecologic bias, which is the failure of expected ecologic effect estimates to reflect the biologic effect at the individual level (18, 19, 25, 45, 54). In addition to the usual sources of bias that threaten individual-level analyses (35, 57), the underlying problem of ecologic analyses for estimating biologic effects is heterogeneity of exposure level and/or covariate levels within groups; as noted earlier, this heterogeneity is not fully captured with ecologic data because of missing information on joint distributions (see Figure 1). Robinson (55) was the first to describe mathematically how ecologic associations could differ from the corresponding associations at the individual level within groups of the same population. He expressed this relationship in terms of correlation coefficients; this relationship was later extended by Duncan et al (15) to regression coefficients in a linear model. The phenomenon became widely known as the *ecologic(al) fallacy* (61), and the magnitude of the ecologic bias may be severe in practice (10, 17, 54, 62, 63).

As an illustration of ecologic bias, consider again Durkheim's data on religion and suicide (Figure 2). The estimated rate ratio of 7.6 in the ecologic analysis may not mean that the suicide rate was nearly eight times greater in Protestants than in non-Protestants. Rather, since none of the regions were entirely Protestant or non-Protestant, it may have been non-Protestants (primarily Catholics) who were committing suicide in predominantly Protestant provinces. It is certainly plausible that members of a religious minority might have been more likely to take their own lives than were members of the majority. The implication of this alternative explanation is that living in a predominantly Protestant area has a contextual effect on suicide risk among non-Protestants, i.e. there is an interaction effect at the individual level between religion and religious composition of one's area of residence.

Interestingly, Durkheim (16) compared the suicide rates (at the individual level) for Protestants, Catholics, and Jews living in Prussia. From his data, we find that the rate was about twice as high in Protestants as in other religious groups. Thus, there appears to be substantial ecologic bias (i.e. comparing rate-ratio estimates of about 2 vs 8). Durkheim, however, failed to notice this quantitative difference because he did not actually estimate the magnitude of the effect in either analysis.

Greenland & Morgenstern (25) showed that ecologic bias can arise from three sources when using simple linear regression to estimate the crude exposure effect: The first may operate in any type of study; the latter two are unique to ecologic studies (i.e. *cross-level bias*), but are defined in terms of individual-level associations.

1. *Within-group bias* The exposure effect within groups may be biased by confounding, selection methods, or misclassification (35, 57). Thus, for example, if there is positive net bias in every group, we would expect the ecologic estimate to be biased as well.
2. *Confounding by group* Ecologic bias may result if the background rate of disease in the unexposed population varies across groups, specifically if there is a nonzero ecologic (linear) correlation between mean exposure level and the background rate.
3. *Effect modification by group* Ecologic bias may also result if the rate difference for the exposure effect at the individual level varies across groups.

Confounding and effect modification by group (the sources of cross-level bias) can arise in three ways: (*a*) Extraneous risk factors (confounders or modifiers) are differentially distributed across groups; (*b*) the ecologic exposure variable has an effect on risk separate from the effect of its corresponding individual-level analogue, e.g. living in a predominantly Protestant area vs being Protestant (in the suicide example); or (*c*) disease risk depends on the prevalence of that disease in other members of the group, which is true of many infectious diseases (37).

Unfortunately, those conditions that produce ecologic bias cannot be observed in ecologic data. Furthermore, the fit of the ecologic regression model, in general, gives no indication of the presence, direction, or magnitude of ecologic bias. Thus, a model with excellent fit may yield substantial bias (e.g. Figure 2), and one model with a better fit than another model may yield more bias.

A potential strategy for reducing ecologic bias is to use smaller units in an ecologic study (e.g. counties instead of states) in order to make the groups more homogeneous with respect to the exposure. On the other hand, this strategy might not be feasible because of the lack of available data aggregated at the same level, and it might lead to two other problems: greater migration between groups (see below) and less precise estimation of disease rates (45, 67).

### *Problems of Confounder Control*

As already indicated, covariates are included in ecologic analyses to control for confounding, but the conditions for a covariate being a confounder are different at the ecologic and individual levels (25, 26). At the individual level, a risk factor must be associated with the exposure to be a confounder. In a multiple-group ecologic study, in contrast, a risk factor may produce ecologic bias (i.e. it may be an ecologic confounder) even if it is unassociated with the exposure in every group, especially if the risk factor is ecologically associated with the exposure across groups (22, 25). Conversely, a risk factor that is a confounder within groups may not produce ecologic bias if it is ecologically unassociated with the exposure across groups.

Control for confounders is more problematic in ecologic analyses than in individual-level analyses (22, 25, 26). Even when all variables are accurately measured for all groups, adjustment for extraneous risk factors may not reduce the ecologic bias produced by these risk factors. In fact, it is possible for such ecologic adjustment to increase bias. It follows from the principles presented in the previous section (25) that there will be no ecologic bias in a multiple-linear-regression analysis if the following conditions are met:

1. There is no residual within-group bias in exposure effect in any group because of confounding by unmeasured risk factors, selection methods, or misclassification.
2. There is no ecologic correlation between the mean value of each predictor and the background rate of disease in the joint reference (unexposed) level of all predictors.
3. The rate difference for each predictor is uniform across levels of the other predictors within groups (i.e. the effects are additive), and each rate difference is uniform across groups (i.e. group does not modify the effect of each predictor at the individual level).

These conditions are sufficient, but not necessary, for the ecologic estimate to be unbiased, i.e. there might be little or no bias even if none of these conditions are met. On the other hand, minor deviations from these conditions can produce substantial ecologic bias (22). Since the sufficient conditions for no ecologic bias cannot be checked with ecologic data alone, the unpredictable and potentially severe nature of such bias makes biologic inference from ecologic analyses particularly problematic. Prentice & Sheppard (51) have suggested that ecologic data be supplemented with individual-level data from each group (or a representative sample) to enhance biologic inference.

Lack of additivity at the individual level (see #3 above) is common in epidemiology, but unmeasured modifiers do not bias results at the individual level if they are unrelated to the exposure (21). Furthermore, interactions may be handled readily at the individual level by including product terms as predictors in the model (e.g. $xz$). In ecologic analyses, however, lack of additivity within groups is a source of ecologic bias, and this bias cannot be eliminated or reduced by the inclusion of product terms (e.g. $XZ$) unless the effects are exactly multiplicative and the two variables are uncorrelated within groups (53).

Another source of ecologic bias is misspecification of confounders (26). Although this problem can also arise in individual-level analyses, it is more difficult to avoid in ecologic analyses because the relevant confounder may be the distribution of covariate histories for all individuals within each group. In ecologic studies, therefore, adjustment for covariates derived from available data (e.g. proportion of current smokers) may be inadequate to control confounding. It is preferable, whenever possible, to control for more than a single summary measure of the covariate distribution (e.g. the proportions of the group in each of several smoking categories). In addition, since it is usually necessary to control for several confounders (among which the effects may not be linear and additive), the best approach for reducing ecologic bias is to include covariates for categories of their joint distribution within regions. For example, to control ecologically for race and sex, the investigator might adjust for the proportions of white women, nonwhite men, and nonwhite women (treating white men as the referent), rather than the conventional approach of adjusting for the proportions of men (or women) and whites (or nonwhites).

### *Within-Group Misclassification*

The principles of misclassification bias with which epidemiologists are familiar when interpreting the results of analyses conducted at the individual level do not apply to ecologic analyses. At the individual level, for example, nondifferential misclassification of exposure nearly always leads to bias toward the null. In multiple-group ecologic studies, however, this principle does not hold when the exposure variable is an aggregate measure. Brenner et al (4) have shown that nondifferential misclassification of a binary exposure within groups usually leads to bias away from the null and that the bias may be severe. Greenland & Brenner (23) have provided a simple method to correct for nondifferential misclassification of exposure or disease in ecologic analyses, based on estimates of sensitivity and specificity.

In studies conducted at the individual level, misclassification of a covariate, if nondifferential with respect to both exposure and disease, will usually reduce our ability to control for that confounder (20, 59). That is, adjustment will not completely eliminate the bias due to the confounder. In ecologic studies, however, nondifferential misclassification of a binary confounder within groups does not affect our ability to control for that confounder, provided there is no cross-level bias (3).

If all but one variable (e.g. the exposure or a covariate) in a given analysis is measured at the individual level, this partially ecologic analysis may also be regarded as nonecologic with the ecologic variable misclassified. Thus, the resulting bias may be understood in terms of misclassification bias operating at the individual level.

## *Other Problems*

LACK OF ADEQUATE DATA Certain types of data, such as medical histories, may not be available in aggregate form; or available data may be too crude, incomplete, or unreliable, such as sales data for measuring behaviors (45, 67). In addition, secondary sources of data from different administrative areas or from different periods may not be comparable. For example, disease rates may vary across countries because of differences in disease classification or case detection. Furthermore, since many ecologic analyses are based on mortality rather than incidence data, causal inference is further limited (35).

TEMPORAL AMBIGUITY In a well-designed cohort study of disease incidence, we can usually be confident that disease occurrence did not precede the exposure. In ecologic studies, however, use of incidence data provides no such assurance against this temporal ambiguity (45). The problem is most troublesome when the disease can influence exposure status in individuals or when the disease rate can influence the mean exposure in groups (through the impact of population interventions designed to change exposure levels in areas with high disease rates).

The problem of temporal ambiguity in ecologic studies (especially time-trend studies) is further complicated by an unknown or variable latent period between exposure and disease occurrence (28, 67). The investigator can only attempt to deal with this problem in the analysis by examining associations for which there is a specified lag between observations of average exposure and disease rate. Unfortunately, there may be little prior information about latency on which to base the lag, or appropriate data may not be available to accommodate the desired lag.

COLLINEARITY Another problem with ecologic analyses is that certain predictors, such as sociodemographic and environmental factors, tend to be more highly correlated with each other than they are at the individual level (10, 62).

The implication of such collinearities is that it is very difficult to separate the effects of these variables statistically; analyses yield model coefficients with very large variances, so effect estimates may be severely distorted. In general, collinearity is most problematic in multiple-group ecologic analyses involving a small number of large, heterogeneous regions (15, 64).

MIGRATION ACROSS GROUPS Migration of individuals into or out of the source population can produce selection bias in a study conducted at the individual level because migrants and nonmigrants may differ on both exposure prevalence and disease risk. Although it is clear that migration can also cause ecologic bias (36, 49), little is known about the magnitude of this bias or how it can be reduced in ecologic studies (46).

## CONTEXTUAL AND MULTILEVEL ANALYSES

Knowing the severe methodologic limitations of ecologic analysis for making biologic inferences, many epidemiologists who report ecologic results argue that there can be no cross-level bias because their primary objective is to estimate an ecologic effect. For example, we might want to estimate the ecologic effect (effectiveness) of state laws requiring smoke detectors by comparing the fire-related mortality rate in those states with the law vs other states without the law (45). Although this is a reasonable objective, the interpretation of observed ecologic effects is complicated by two issues:

First, biologic inference may be implicit to the objectives of an ecologic study unless the underlying biologic and contextual effects are already known from previous research. Can smoke detectors placed appropriately in homes reduce the risk of fire-related mortality in those homes by providing an early warning of smoke? Does living in an area where most homes are properly equipped with smoke detectors reduce the risk of fire-related mortality in homes with and without smoke detectors? The first question refers to a possible biologic (biobehavioral) effect; the second question refers to a possible contextual effect. Even if these effects exist, the ecologic effect of smoke-detector laws also depends on other factors, e.g. the level of enforcement, the quality of smoke-detector design and construction, the cost and availability of smoke detectors, and their proper placement, installation, operation, and maintenance. In an ecologic study without additional information, the ecologic effect is completely confounded with biologic and contextual effects.

The second complicating issue in interpreting observed ecologic effects is the need to control for confounders measured at the individual level. Even if the exposure is a global measure, such as a law, groups are seldom completely homogeneous or comparable with respect to confounders. To make a valid comparison between states with and without smoke-detector laws, for example, we would need to control for differences among states in the joint distribution of extraneous risk factors, such as socioeconomic status of residents, firefighter availability and access, building design, and construction (see also *Problems of Confounder Control*).

Perhaps the best solution to these problems is to incorporate both individual-level and ecologic measures in the same analysis. This approach might include different measures of the same factor; e.g. each subject would be characterized by his/her own exposure level as well as the average exposure level for all members of the group to which s/he belongs (aggregate measure). Not only would this approach help to clarify the sources and magnitude of ecologic and cross-level bias, but it would also allow us to separate biologic, contextual, and ecologic effects. It is especially appropriate in social epidemiology, infectious-disease epidemiology, and the evaluation of population interventions.

There are two statistical methods for including both individual-level and ecologic measures in the same analysis. The first method, often called *contextual analysis* in the social sciences, is a simple extension of conventional modeling such as multiple linear regression or logistic regression (5, 34). The model, which is fit to the data at the individual level, includes both individual-level and ecologic predictors. For example, suppose we wanted to estimate the effect of "herd immunity" on the risk of an infectious disease. The risk ($y$) of disease might be modeled as a function of the following linear component: $b_0 + b_1x + b_2\bar{x} + b_3x\bar{x}$, where x is the individual's immunity status and $\bar{x}$ is the prevalence of immunity in the group to which that individual belongs (65). Therefore, $b_2$ represents the contextual effect of herd immunity, and $b_3$ represents the interaction effect, which allows the herd-immunity effect to depend on the individual's immune status. The interaction term is needed in this application, since we would expect no herd-immunity effect among immune individuals. Note, however, that the interpretation of the interaction effect depends on the form of the model (35, 57).

An important limitation of contextual analysis is that observations for individuals within groups are not likely to be independent, which is a basic assumption of conventional modeling. If there are contextual effects, then the outcomes for individuals in the same group are more likely to resemble each other than are the outcomes for individuals in different groups. To handle this problem of within-group clustering, we treat the sampling of individuals from groups as random effects; this approach is called *multilevel modeling, hierarchical regression,* or *random-effects modeling* (6, 71, 72).

Multilevel modeling is a powerful technique with many applications; it can be used to estimate contextual and ecologic effects and to derive improved (empirical Bayes) estimates of biologic effects. At the first level of analysis, we might predict individual risk or health status within each group as a function of several individual-level variables. At the second (ecologic) level, we predict the estimated regression parameters (e.g. the intercept and slopes) from the first level as a function of several ecologic variables. For example, Humphreys & Carr-Hill (33) used multilevel modeling to estimate the contextual effect of living in a poor area on several health outcomes, controlling for the individual's income and other covariates. In a conventional ecologic analysis, the effects of living in a poor area and income would be confounded, and ecologic estimates of effect would be susceptible to cross-level bias.

## CONCLUSIONS

Several practical advantages make ecologic studies especially appealing for undertaking various types of epidemiologic research. Despite these advantages, however, ecologic analysis poses major problems of interpretation when making ecologic inferences and especially when making biologic inferences (due to ecologic bias, etc). From a methodologic perspective, it is best to have individual-level data on as many relevant nonglobal measures as possible. Just because the exposure variable is measured ecologically, for example, does not mean that other variables should be as well.

Even when the stated purpose of the study is to estimate an ecologic effect, biologic inference is usually implicit in epidemiology. Thus, to address the underlying research questions, we typically would want to estimate and/or control for biologic and contextual effects, preferably using multilevel analysis. In contemporary epidemiology, the "ecologic fallacy" reflects the failure of the investigator to recognize the need for biologic inference and thus for individual-level data.

### ACKNOWLEDGMENTS

The author would like to thank Drs. Sander Greenland and Matthew Longnecker for their helpful comments.

### *Literature Cited*

1. Beral V, Chilvers C, Fraser P. 1979. On the estimation of relative risk from vital statistical data. *J. Epidemiol. Community Health* 33:159–62
2. Blot WJ, Fraumeni JF Jr. 1977. Geographic patterns of oral cancer in the United States: etiologic implications. *J. Chron. Dis.* 30:745–57
3. Brenner H, Greenland S, Savitz DA. 1992. The effects of nondifferential confounder misclassification in ecologic studies. *Epidemiology* 3:456–59
4. Brenner H, Savitz DA, Jöckel K-H, Greenland S. 1992. Effects of nondifferential exposure misclassification in ecologic studies. *Am. J. Epidemiol.* 135: 85–95
5. Boyd LH Jr, Iversen GR. 1979. *Contex-*

*tual Analysis: Concepts and Statistical Techniques*. Belmont, CA: Wadsworth

6. Bryk AS, Raudenbush SW. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage
7. Catalano R, Serxner S. 1987. Time series designs of potential interest to epidemiologists. *Am. J. Epidemiol.* 126:724–31
8. Clayton DG, Bernardinelli L, Montomoli C. 1993. Spatial correlation in ecological analysis. *Int. J. Epidemiol.* 22:1193–202
9. Clayton D, Kaldor J. 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 43:671–81
10. Connor MJ, Gillings D. 1984. An empiric study of ecological inference. *Am. J. Public Health* 74:555–59
11. Crawford MD, Gardner MJ, Morris JN. 1971. Changes in water hardness and local death-rates. *Lancet* 2:327–29
12. Cressie N. 1993. Regional mapping of incidence rates using spatial Bayesian models. *Med. Care* 31:YS60–65 (Suppl.)
13. Darby SC, Doll R. 1987. Fallout, radiation doses near Dounreay, and childhood leukaemia. *Br. Med. J.* 294:603–7
14. Dogan M, Rokkan S, eds. 1969. *Social Ecology*. Cambridge, MA: MIT Press

14a. Dogan M, Rokkan S. 1969. Introduction. See Ref. 14, pp. 1–15

15. Duncan OD, Cuzzort RP, Duncan B. 1961. *Statistical Geography: Problems in Analyzing Areal Data*, pp. 64–67. Westport, CT: Greenwood Press
16. Durkheim E. 1951. *Suicide: A Study in Sociology*, pp. 153–54. New York: Free Press
17. Feinleib M, Leaverton PE. 1984. Ecological fallacies in epidemiology. In *Health Information Systems*, ed. PE Leaverton, L Massö, pp. 33–61. New York: Praeger
18. Firebaugh G. 1978. A rule for inferring individual-level relationships from aggregate data. *Am. Sociol. Rev.* 43:557–72
19. Goodman LA. 1959. Some alternatives to ecological correlation. *Am. J. Sociol.* 64:610–25
20. Greenland S. 1980. The effect of misclassification in the presence of covariates. *Am. J. Epidemiol.* 112:564–69
21. Greenland S. 1987. Interpretation and choice of effect measures in epidemiologic analysis. *Am. J. Epidemiol.* 125: 761–68
22. Greenland S. 1992. Divergent biases in ecologic and individual-level studies. *Stat. Med.* 11:1209–23
23. Greenland S, Brenner H. 1993. Correcting for non-differential misclassification in ecologic analyses. *Appl. Statist.* 42: 117–26
24. Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. 1991. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology* 2:387–92
25. Greenland S, Morgenstern H. 1989. Ecological bias, confounding, and effect modification. *Int. J. Epidemiol.* 18:269–74
26. Greenland S, Robins J. 1994. Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. *Am. J. Epidemiol.* 139:747–60
27. Greenland S, Schlesselman JJ, Criqui MH. 1986. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am. J. Epidemiol.* 123:203–8
28. Gruchow HW, Rimm AA, Hoffman RG. 1983. Alcohol consumption and ischemic heart disease mortality: are time-series correlations meaningful? *Am. J. Epidemiol.* 118:641–50
29. Hatch M, Susser M. 1990. Background gamma radiation and childhood cancers within ten miles of a US nuclear plant. *Int. J. Epidemiol.* 19:546–52
30. Helfenstein U. 1991. The use of transfer function models, intervention analysis and related time series methods in epidemiology. *Int. J. Epidemiol.* 20:808–15
31. Hiller JE, McMichael AJ. 1991. Ecological studies. In *Design Concepts in Nutritional Epidemiology*, ed. BM Margetts, M Nelson, pp. 323–53. Oxford: Oxford Univ. Press
32. Holford TR. 1991. Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annu. Rev. Public Health* 12:425–57
33. Humphreys K, Carr-Hill R. 1991. Area variations in health outcomes: artefact or ecology. *Int. J. Epidemiol.* 20:251–58
34. Iversen GR. 1991. *Contextual Analysis*. Newbury Park, CA: Sage
35. Kleinbaum DG, Kupper LL, Morgenstern H. 1982. *Epidemiologic Research: Principles and Quantitative Methods*, pp. 77–81, 130–34, 184–280. New York: Van Nostrand Reinhold
36. Kliewer EV. 1992. Influence of migrants on regional variations of stomach and colon cancer mortality in the western United States. *Int. J. Epidemiol.* 21:442–49
37. Koopman JS, Longini IM Jr. 1994. The ecological effects of individual exposures and nonlinear disease dynamics in populations. *Am. J. Public Health* 84: 836–42

38. Langbein LI, Lichtman AJ. 1978. *Ecological Inference.* Beverly Hills, CA: Sage
39. Lee JAH. 1982. Melanoma and exposure to sunlight. *Epidemiol. Rev.* 4:110–36
40. Lee JAH, Petersen GR, Stevens RG, Vesanen K. 1979. The influence of age, year of birth, and date on mortality from malignant melanoma in the populations of England and Wales, Canada, and the white population of the United States. *Am. J. Epidemiol.* 110:734–39
41. MacMahon B, Pugh TF. 1970. *Epidemiology: Principles and Methods,* pp. 137–98, 175–84. Boston: Little, Brown & Co.
42. Mason TJ, McKay FW, Hoover R, Blot WJ, Fraumeni JF Jr. 1975. *Atlas of Cancer Mortality for US Counties: 1950–1969,* pp. 36, 37. DHEW Publ. No. (NIH) 75–780. Washington, DC: US GPO
43. McDowall D, McCleary R, Meidinger EE, Hay RA Jr. 1980. *Interrupted Time Series Analysis.* Beverly Hills, CA: Sage
44. Mollie A, Richardson S. 1991. Empirical Bayes estimation of cancer mortality rates using spatial models. *Stat. Med.* 10:95–112
45. Morgenstern H. 1982. Uses of ecologic analysis in epidemiologic research. *Am. J. Public Health* 72:1336–44
46. Morgenstern H, Thomas D. 1993. Principles of study design in environmental epidemiology. *Environ. Health Perspect.* 101:23–38 (Suppl. 4)
47. Moulton LH, Foxman B, Wolfe RA, Port FK. 1994. Potential pitfalls in interpreting maps of stabilized rates. *Epidemiology* 5:297–301
48. Ostrom CW Jr. 1990. *Time Series Analysis: Regression Techniques.* Newbury Park, CA: Sage. 2nd ed.
49. Polissar L. 1980. The effect of migration on comparison of disease rates in geographic studies in the United States. *Am. J. Epidemiol.* 111:175–82
50. Prentice RL, Kakar F, Hursting S, Sheppart L, Klein R, Kushi LH. 1988. Aspects of the rationale for the Women's Health Trial. *J. Natl. Cancer Inst.* 80: 802–14
51. Prentice RL, Sheppard L. 1989. Validity of international, time trend, and migrant studies of dietary factors and disease risk. *Prev. Med.* 18:167–79
52. Prentice RL, Thomas D. 1993. Methodologic research needs in environmental epidemiology: data analysis. *Environ. Health Perspect.* 101:39–48 (Suppl. 4)
53. Richardson S, Hémon D. 1990. Ecological bias and confounding (letter). *Int. J. Epidemiol.* 19:764–66
54. Richardson S, Stücher I, Hémon D. 1987. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *Int. J. Epidemiol.* 16:111–20
55. Robinson WS. 1950. Ecological correlations and the behavior of individuals. *Am. Sociol. Rev.* 15:351–57
56. Rosenbaum PR, Rubin DB. 1984. Difficulties with regression analyses of age-adjusted rates. *Biometrics* 40:437–43
57. Rothman KJ. 1986. *Modern Epidemiology,* pp. 41–49, 82–94. Boston: Little, Brown & Co.
58. Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–58
59. Savitz DA, Baron AE. 1989. Estimating and correcting for confounder misclassification. *Am. J. Epidemiol.* 129:1062–71
60. Sayrs LW. 1989. *Pooled Time Series Analysis.* Newbury Park, CA: Sage
61. Selvin HC. 1958. Durkheim's "Suicide" and problems of empirical research. *Am. J. Sociol.* 63:607–19
62. Stavraky KM. 1976. The role of ecologic analysis in studies of the etiology of disease: a discussion with reference to large bowel cancer. *J. Chron. Dis.* 29: 435–44
63. Stidley C, Samet JM. 1994. Assessment of ecologic regression in the study of lung cancer and indoor radon. *Am. J. Epidemiol.* 139:312–22
64. Valkonen T. 1969. Individual and structural effects in ecological research. See Ref. 14, pp. 53–68
65. Von Korff M, Koepsell T, Curry S, Diehr P. 1992. Multi-level analysis in epidemiologic research on health behaviors and outcomes. *Am. J. Epidemiol.* 135:1077–82
66. Walter SD. 1991. The ecologic method in the study of environmental health. I. Overview of the method. *Environ. Health Perspect.* 94:61–65
67. Walter SD. 1991. The ecologic method in the study of environmental health. II. Methodologic issues and feasibility. *Environ. Health Perspect.* 94:67–73
68. Walter SD. 1992. The analysis of regional patterns in health data: I. Distributional considerations. *Am. J. Epidemiol.* 136:730–41
69. Walter SD. 1992. The analysis of regional patterns in health data: II. The power to detect environmental effects. *Am. J. Epidemiol.* 136:742–59
70. Winn DM, Blot WJ, Shy CM, Pickle LW, Toledo A, Fraumeni JF Jr. 1981. Snuff dipping and oral cancer among

women in the southern United States. *N. Engl. J. Med.* 304:745–49

71. Wong GY, Mason WM. 1985. The hierarchical logistic regression model for multilevel analysis. *J. Am. Statist. Assoc.* 80:513–24

72. Wong GY, Mason WM. 1991. Contextually specific effects and other generalizations for the hierarchical linear model for comparative analysis. *J. Am. Statist. Assoc.* 86:487–503

© International Epidemiological Association 2001 Printed in Great Britain

# Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects

Sander Greenland

A number of authors have attempted to defend ecologic (aggregate) studies by claiming that the goal of those studies is estimation of ecologic (contextual or group-level) effects rather than individual-level effects. Critics of these attempts point out that ecologic effect estimates are inevitably used as estimates of individual effects, despite disclaimers. A more subtle problem is that ecologic variation in the distribution of individual effects can bias ecologic estimates of contextual effects. The conditions leading to this bias are plausible and perhaps even common in studies of ecosocial factors and health outcomes because social context is not randomized across typical analysis units (administrative regions). By definition, ecologic data contain only marginal observations on the joint distribution of individually defined confounders and outcomes, and so identify neither contextual nor individual-level effects. While ecologic studies can still be useful given appropriate caveats, their problems are better addressed by multilevel study designs, which obtain and use individual as well as group-level data. Nonetheless, such studies often share certain special problems with ecologic studies, including problems due to inappropriate aggregation and problems due to temporal changes in covariate distributions.

Keywords Aggregate studies, confounding, contextual studies, ecologic fallacy, ecologic studies, environmental health, epidemiology, multilevel studies, relative risk, risk assessment

Accepted 19 March 2001

Studies limited to characteristics of aggregates (groups) of individuals are usually termed *ecologic studies*, a usage that will be adopted here.[1–5] This usage is perhaps unfortunate, for the word 'ecologic' suggests that such studies are especially appropriate for studying the impact of environmental factors, including societal characteristics. I will here review some criticisms of this notion, arguing that it arises from confusion of an ecologic perspective (addressing relations at the environmental or social level) with ecologic studies. As a number of authors have pointed out,[6–12] overcoming this confusion requires adoption of a *multilevel* perspective, which allows integration of theory and observations on all available levels: physiological (which examines exposures and responses of systems within individuals), individual (which examines exposures and responses of individuals), and aggregate or contextual (which examines exposures and responses of aggregates or clusters of individuals, such as locales or societies).

Defences of ecologic studies argue (correctly) that many critics have presumed individual-level relations are the ultimate target of inference of all ecologic studies, when this is not always so,[9,13,14] and that contagious outcomes necessitate group-level considerations in modelling regardless of the target level.[15] They also point out that an ecologic summary may have its own direct effects on individual risk beyond that conferred by the contributing individual values; for example, average economic status of an area can have effects on an individual over and above the effects of the individual's economic status.[16,17] Unfortunately, some defences go on to make implicit assumptions to 'prove' that entire classes of ecologic studies are valid, or at least no less valid than individual-level analyses; see Greenland and Robins,[18,19] Morgenstern,[5] and Naylor[20] for critical commentaries against such arguments in the health sciences. Some ecologic researchers are well aware of these problems and explicate the assumptions they use,[21,22] but still draw criticism because of the sensitivity of inferences to those assumptions.[23–25] Thus I will review some controversial assumptions that appear common in ecologic analyses of epidemiological data. Finally, I will briefly discuss multilevel methods that represent both individual-level and ecologic data within a single model.

The present paper relies on simple illustrations designed to make the points transparent to non-mathematical readers, and focuses on problems of confounding and specification bias;

Department of Epidemiology, UCLA School of Public Health, and Department of Statistics, UCLA College of Letters and Science, 22333 Swenson Drive, Topanga, CA 90290, USA.

a companion paper[12] provides an overview of the underlying mathematical theory. Many other issues have been raised in the ongoing ecologic-study controversy; see the references for details, especially those in the Discussion section.

## How Ecologic Confounding Depends on Joint Individual-level Distributions

There are two major types of measurements on aggregates: Summaries of distributions of individuals within aggregates, such as mean age and per cent female; and purely ecologic (contextual) variables that are defined directly on the aggregate, such as whether there is a needle-exchange programme in an area. The causal effects of the latter purely contextual variables are the focus of much social research and ecosocial epidemiology.[9,10,13,26,27] Nonetheless, most outcome variables of public-health importance are summaries of individual-level distributions, such as prevalence, incidence, mortality, and life expectancy, all of which can be expressed in terms of average individual outcomes.[28] Furthermore, many contextual variables are measured by surrogates that are summaries over individuals; for example, neighbourhood social class is often measured by average income and average education.

The presence of summary measures in an ecologic analysis introduces a major source of uncertainty in ecologic inference: Effects on summaries depend on the joint individual-level distributions within aggregates, but distributional summaries do not fully determine (and sometimes do not even seriously constrain) those joint distributions. This problem corresponds to the 'information lost due to aggregation', and is a key source of controversy about ecologic studies.[29]

Panel 1 of Table 1 illustrates this problem. For simplicity, just two areas are used here, but examples with many areas have also been given.[18] Suppose we wish to assess a contextual effect, i.e. the impact of an ecologic difference between areas A and B (such as a difference in laws or social programmes) on the rate of a health outcome, and we measure this effect by the amount $RR_A$ that this difference multiplies the rate (the true effect of being in A versus being in B). One potential risk factor X differs in distribution between the areas; an effect of X (measured by the rate ratio $RR_X$ comparing X = 1 to X = 0 within areas) may be present, but we observe no difference in rates between the areas.

Do the ecologic data in Panel 1 of Table 1 demonstrate no contextual effect? That is, do they correspond best with $RR_A = 1$? Unfortunately, the ecologic (marginal) data on X distributions and the outcome rates are mathematically equally compatible with a broad spectrum of possibilities, two of which are given in Panels 2 and 3 of Table 1: In the first, area A has benefited from the contextual difference ($RR_A < 1$), but this fact has been obscured by area A's higher prevalence of X, which is harmful ($RR_X > 1$); in the second, area A has been harmed by the contextual difference ($RR_A > 1$), but this fact has been obscured by area A's higher prevalence of X, which is beneficial ($RR_X < 1$). One could obtain the correct answers in either possibility by comparing the area rates after they had been standardized directly to a common X distribution; such standardization would however require the X-specific rates *within* the areas, which are not available in the example. Furthermore, an ecologic regression could not solve the problem because it would only regress the crude area rates on the proportion with X = 1 in each area: Because both crude area rates are 5.6, the ecologic X-coefficient would be zero, and so the regression would produce no X-adjustment of the area rates.

Lacking within-area data on the joint distribution of X and the outcome, an ecologic analysis must necessarily rely on external (prior) information to make inferences about the contextual effect, although the margins may impose some bounds on the possibilities.[21,29,30] If one were willing to assume that the X-specific rates in each area were proportional to those in some external reference population with known X-specific rates, one could use those external rates to construct and compare standardized morbidity ratios (SMR) for the areas (indirect adjustment). Unfortunately, such external rates are rarely available for all important covariates, and so one must fall back on other external (prior) information to produce an effect estimate.

The results of such an analysis can be disappointing if the prior information is ambiguous. If X indicates (say) regular cigarette use and the outcome is total mortality, we might be

**Table 1** An example demonstrating the complete confounding of contextual and individual effects in ecologic data: The ecologic data cannot identify the effect of group (A versus B) on the rate of the outcome Y = 1 when only a marginal summary of the individual-level covariate X is available. N = denominator (in thousands of person-years); $RR_A$ and $RR_X$ are the rate ratios for the true effects of A versus B and of X = 1 versus X = 0, respectively

| | Group A | | | Group B | | |
|---|---|---|---|---|---|---|
| | X = 1 | X = 0 | Total | X = 1 | X = 0 | Total |
| **1. Ecologic (marginal) data:** | | | | | | |
| Y = 1 | ? | ? | 560 | ? | ? | 560 |
| N | 60 | 40 | 100 | 40 | 60 | 100 |
| Rate | ? | ? | 5.6 | ? | ? | 5.6 |
| **2. Possibility 1 ($RR_X = 2$, $RR_A = 7/8$):** | | | | | | |
| Y = 1 | 420 | 140 | 560 | 320 | 240 | 560 |
| N | 60 | 40 | 100 | 40 | 60 | 100 |
| Rate | 7.0 | 3.5 | 5.6 | 8.0 | 4.0 | 5.6 |
| **3. Possibility 2 ($RR_X = ½ RR_A = 8/7$):** | | | | | | |
| Y = 1 | 240 | 320 | 560 | 140 | 420 | 560 |
| N | 60 | 40 | 100 | 40 | 60 | 100 |
| Rate | 4.0 | 8.0 | 5.6 | 3.5 | 7.0 | 5.6 |

confident that $RR_X$ is well above one and hence that the contextual effect (i.e. the A-B rate ratio) is protective. If however X indicates regular alcohol consumption we might feel justified in ruling out scenarios involving values for the relative risk $RR_X$ that are very far from 1, but, because alcohol may be protective at moderate levels and causative at higher levels, we could not be sure of the direction of $RR_X$; that would depend on the relative proportion of moderate and heavy drinkers in the areas. As a consequence, we could not be sure of the direction (let alone degree) of confounding in the ecologic estimate of the contextual effect (i.e. the ecologic A-B rate ratio of 5.6/5.6 = 1).

The problem of cross-level confounding just illustrated has been discussed extensively since the early 20th century (Achen and Shively[29,Ch.1]) and is a mathematically trivial consequence of the fact that marginals do not determine joint distributions. Yet this non-identification problem, which is an absolute demarcation between ecologic and individual-level studies, continues to be misunderstood or ignored by many ecologic researchers, so much that Achen and Shively[29,p.8] remark: 'A cynic might conclude that social scientists tend to ignore logical problems and contradictions in their methods if they do not see anything to be done about them'.

Their remark applies to the health sciences as well, as illustrated by this quote: 'In practice, it may be that confounding usually poses a more intractable problem for ecological than for individual-level studies. But this is due to the greater reliance on secondary data and proxy measures in ecological studies, *not to any problem inherent in ecological studies*'[13,p.820] (emphasis added).

While ecologic studies do suffer from greater reliance on secondary data and proxy measures, this passage is typical of defences that overlook the non-identifiable aspects of confounding inherent in ecologic studies; other examples include Cohen,[31–33] Susser,[34] and Pearce.[14,p.682] Table 1 illustrates that, given confounding by a measured risk factor X, the individual-level data allows one to control the confounding in the most straightforward way possible: Stratify by X. In contrast, control of confounding by X cannot be accomplished using only the ecologic data, despite the fact that the effect under study is contextual (the effect of the social differences between areas A and B on outcome rates). Because contextual and individual effects are completely confounded in ecologic data,[6,12] the only solutions to this problem are either to obtain individual data within the ecologic units, or else resort to using assumptions that are untestable with the ecologic data and liable to strong dispute.[18,19,29]

Another fallacy in some defences of ecologic studies is the claim that individual-level information is not needed if one is interested only in contextual (ecologic) effects. Examples like that above show that such 'holistic' arguments are incorrect, especially in health studies in which the outcome measure is an individual-level summary, because individual-level factors can confound the ecologic results even if the study factor is contextual.[6,12] Holistic arguments also tend to overlook that ecologic data usually refer to arbitrary administrative groupings, such as counties, that are often poor proxies for social context or environmental exposures.[29,pp.20–22] The severity of this problem is illustrated by the potential for great sensitivity of ecologic relations to the grouping definition.[35]

The non-identification problem illustrated in Table 1 applies symmetrically to ecologic estimates of average individual-level effects (cross-level inferences from the ecologic to individual level).[12,29,36] For example, if Table 1 represented a contrast of two areas A and B with a goal to estimate the impact of differences in the X distribution, we see that very small contextual effects can obscure substantial X effects in the ecologic data. My emphasis here, however, is that even if the overall goal is to estimate contextual effects, ecologic manifestations of those effects (Panel 1 of Table 1) can be confounded due to individual-level relations, and are not estimable without information about those relations.

To summarize: Observational ecologic data alone tell us little about either contextual or individual-level effects on summary measures of population health, precisely because (by definition) they lack data on individual-level associations. Thus, methods that purport to adjust ecologic results for the confounding problem just described either must employ external data about non-identified individual relations, or must invoke assumptions about those relations. The non-identified nature of the relations means that neither approach can be fully tested (validated) against the ecologic data.

## Some Assumptions and Fallacies in Ecologic Analyses

All too often, identification is forced by making fairly arbitrary modelling assumptions. Controversy then arises surrounding the credibility or strength of the assumptions used to derive effect estimates from ecologic data, the sensitivity of those estimates to changes in assumptions, and failures of proposed methods in real or simulated data. For examples, compare Freedman *et al.*[37] versus their discussants; Greenland and Morgenstern[38,39] and Richardson and Hemon[40] versus Cohen;[31] Greenland and Robins,[18,19] Piantadosi,[41] Stidley and Samet[42] and Lagarde and Pershagen[43,44] versus Cohen;[32,33] King[21,22] versus Rivers,[23] Cho,[45] Freedman *et al.*,[24,25] and the example in Stoto;[46] and Wen and Kramer[47] versus Naylor.[20]

All causal inferences from observational epidemiological data must rely on restrictive assumptions about the distributions of errors and background causes. Most estimates also rely on parametric models for effects. Thus, validity of inferences depend on examining the robustness of the estimates to violations of the underlying assumptions and models.

### Randomization assumptions

Interpretation of an association as a causal effect must depend on some sort of non-confounding or ignorability assumption, which in statistical methodology becomes operationalized as a covariate-specific randomization assumption.[48,49] Such causal inferences are usually not robust to violations of those assumptions, and this non-robustness is a major source of controversy in most non-experimental research.

Suppose we are to study K communities. The distinction between ecologic and individual-level confounding may become clearer by contrasting two levels of randomized intervention to reduce sexually transmitted diseases (STD):

(Trial C) A community-health programme (e.g. establishment of free STD clinics) is provided at random to half of the communities (K is assumed to be even).

(Trial W) Within community k, a treatment (e.g. a series of free STD-clinic visits) is randomly provided to a proportion $p_k$ of individuals, and $p_k$ varies across communities.

Trial C is a cluster-randomized design. In this trial, the ecologic data would comprise a community treatment-status indicator plus the outcome measures (e.g. subsequent STD rates). These data would support randomization-based inferences about what the average causal effect the programme would be for the K communities (the communities would be the analysis units, and the sample size for the inferences would be K). Nonetheless, analysing the individual-level data from trial C as a K-stratum study with fixed treatment margin (the usual individual-level analysis) would support no inferences at all about treatment effects because each stratum (community) would have a zero margin. Put another way, community and treatment effects would be completely confounded within the standard individual-level model. Analysis of the individual-level data would instead require use of methods for cluster-randomized trials.

In trial W, the ecologic data would comprise the proportion treated ($p_k$) in each community, along with outcome measures. Unless the $p_k$ had been randomly assigned across communities (as in trial C, in which $p_k = 0$ or 1), the ecologic data would *not* support randomization-based inferences about treatment effects: If the $p_k$ were constant, there would be no data information for such an analysis; if the $p_k$ varied, the community and treatment effects would be completely confounded. (This observation is essentially a contrapositive version of Goodman's identification condition for ecologic regression,[50] translated to the present causal setting.) Nonetheless, the individual-level data from any of or all the communities with $0 < p_k < 1$ would support the usual randomization-based inferences (e.g. exact tests stratified on community). Taking X as the treatment indicator and k = A, B, Panels 1 and 2 can be used as an example of trial W with $p_A = 0.6$ and $p_B = 0.4$; it then exhibits complete confounding in the ecologic data and no confounding of the individual-level data within community.

## Observational studies

Observational studies suffer from a fundamental weakness in interpreting estimated associations as causal effects: The validity of such interpretations depend on assumptions that natural or social circumstances have miraculously randomized the study exposure, e.g. by carrying out trial C or W for us. For individual-level studies, it is widely recognized that the number of individuals in the study tells us nothing about the validity of this or other such 'no-confounding' assumptions. Larger size only increases the precision of randomization-based inferences by reducing the chance that randomization left large imbalances of uncontrollable covariates across treatment groups. This benefit of size stems from and depends on an assumption of exposure randomization, as in trial W. Systematic imbalances within groups are by definition violations of that assumption.

The same argument applies to ecologic studies. The number of (say) ecologic groups involved tells us nothing about the validity of an assumption that the exposure distribution ($p_k$ in the above binary-treatment trials) was randomized across the groups. The benefit of a large number of groups stems from and depends on an assumption that those distributions were randomized, as in trial C. Systematic imbalances across groups are by definition violations of that assumption. Despite these facts, defences of ecologic studies have appeared based on the circular argument that large numbers of areas would reduce the chance of ecologic confounding;[31] this circularity arises because the large-number effect assumes randomization across areas, which is precisely the assumption in doubt.

## Covariate control

To achieve some plausibility in causal inferences from observational data, researchers attempt to 'control' for covariates that affect the outcome but are not affected by the exposure (potential confounders). In individual-level studies, the traditional means of control is to stratify the data on these covariates, because within such strata the exposed and unexposed units cannot have any imbalance on the covariates beyond the stratum boundaries, e.g. within a 65–74-year-old age stratum the exposed and unexposed could not be more than 10 years apart in age. Causal inferences then proceed by assuming randomization *within* these strata; however implausible it may be, in the face of observed imbalances this assumption is always more plausible than the assumption of simple (unstratified) randomization.

The stratification process can be applied in ecologic analyses, but usually faces serious data limitations. With few exceptions, the ecologic exposures and covariates in public-use databases are insufficient in detail and accuracy to create strata with assured balance on key covariates. For example, in ecologic studies of radon levels and lung-cancer rates across US counties, the key covariates are the county-specific distributions of age, sex, race, and smoking habits. To lay to rest concerns about bias from possible relations of these strong lung-cancer risk factors to radon exposure, one would have to stratify the county data by age, sex, race, and smoking behaviour (note that smoking behaviour is multidimensional, as it includes intensity, duration, and type of cigarette use). One would then examine the relation of radon distributions to lung-cancer rates across the stratum-specific county data. This stratified analysis requires the county-specific joint distributions of age, sex, race, smoking behaviour and radon, and age, sex, race, smoking behaviour and lung cancer. Unfortunately, to date no such data have been available. Although data on the age-sex-race-lung cancer distributions in counties are published, their joint distributions with radon and smoking are unobserved; only marginal distributions of radon are surveyed, and only crude summaries of cigarette sales are available.

The limitations of the ecologic data may be better appreciated by considering an analogous problem in an individual-level study of residential radon and lung cancer. One might attempt to 'control' for smoking by using cigarette sales in a subject's county of residence as a surrogate for smoking behaviour (as in Cohen[31]). Presumably, few epidemiologists would regard this strategy at providing adequate control of smoking, especially upon considering that it would impute an identical 'smoking' level to every subject in the county, regardless of their age, sex, or lung-cancer status. The shortcomings of this control arise precisely because smoking behaviour varies to an extreme among individuals within any given county, much more so than average smoking behaviour varies across counties.[18]

Because different randomization assumptions underly causal inferences from individual-level and ecologic studies, it can happen that these two study types require control of different (though overlapping) sets of covariates for valid inferences.

See Greenland and Robins[18] and Robins *et al.*[51] for discussions of this point.

### Modelling assumptions

To get around the ecologic data limitations described above, ecologic-study investigators have employed analysis models under which the available ecologic data (comprising simple marginal summaries of crude exposure and covariate measures) are sufficient for valid effect estimation. As mentioned earlier, these models are restrictive and no more supported by data than randomization assumptions. For example, a common assumption in ecologic analyses is that effects follow a multiple-linear regression model. This assumption is both natural and somewhat misleading, because a multiple-linear model for individual-level effects induces a multiple-linear ecologic model, but this parallel relation between the individual and ecologic regressions fails in the presence of non-additive or non-linear effects within the ecologic groups.[18,52–56]

Not even the functional form of individual-level effects (which can confound ecologic associations, as in Table 1) is identified by the marginal data in ecologic studies. For example, suppose individual risk R structurally depends on the covariate vector X (which may contain contextual variables) via $R = f(X)$, and A indexes contexts (such as geographical areas). The ecologic observations identify only relations of average risks $E_A(R)$ to average covariate levels $E_A(X)$ across contexts. These ecologic relations will generally not follow the same functional form as the individual relations because $E_A(R) = E_A[f(X)] \neq f[E_A(X)]$ except in some very special cases, chiefly those in which f is additive and linear in all the X components.

Most analyses of individual-level epidemiological studies assume a multiplicative (loglinear) model for the regression of the outcome on exposure and analysis covariates, in which $f(X) = \exp(X\beta)$. Such non-linearities in individual regressions virtually guarantee that simple ecologic models will be misspecified, and thus further diminish the effectiveness of ecologic control of confounding, although the problem can be mitigated somewhat by expanding the ecologic model to include more detailed covariate summaries if those are available,[18,57] and by including higher-order covariate terms.[40,53,56]

Unfortunately, some authors have attempted to deny the misspecification problem by claiming that a linear regression is justified by Taylor's theorem.[31] This justification is circular because approximate linearity of f(X) over the within-context (area-specific) range of X is required for a first-order Taylor approximation of f(X) to be accurate.[18,54] Furthermore, in most applications this requirement is known to be violated. For example, the dependence of risk on age is highly non-linear for nearly all cancers. One may attempt to circumvent the latter problem by using age-specific outcomes, but will then face the problem that one lacks age-context-specific measures of potential confounders such as smoking. Use of age-standardized rates also fails to solve the problem for that requires one use age-standardized measures of the covariates in the regression model (see Discussion) and such measures are rarely available.

## Multilevel Methods

The vital statistics and registry data used in ecologic health studies are collected at great expense and so it seems imperative to exploit them fully. Furthermore, these data often describe outcomes across a much broader spectrum of exposures than found in most individual-level studies, suggesting greater power to detect effects could be achieved if confounding were controlled. For example, individual-level dietary studies are usually conducted in restricted populations with little dietary variation relative to international variation, which limits their power and suggests much could be learned from international comparisons.[58] A major problem of international ecologic comparisons, however, is the presence of numerous other differences across countries that could confound the results.

To address this ecologic confounding problem, one may apply individual-level risk models to within-region survey data and aggregate the resulting individual risk estimates for comparison to observed ecologic rates.[54,59] For example, suppose we have a covariate vector X measured on $N_k$ surveyed individuals in region k, a rate model $r(x; \beta)$ with a $\beta$ estimate $\hat{\beta}$ from individual-level studies (e.g. a proportional-hazards model derived from cohort-study data), and the observed rate $\bar{r}_k$ in region k. Then we may compare $\bar{r}_k$ to the area rate predicted from the model applied to the survey data, $\Sigma_i r(x_i;\hat{\beta})/N_k$, where the sum is over the surveyed individuals $i = 1, \ldots, N_k$ and $x_i$ is the value of X for survey individual i. This approach is a regression analogue of indirect adjustment: $\hat{\beta}$ is the external information, and so corresponds to the reference rates used to construct expectations in SMR.

Unfortunately, fitted models generalizable to the regions of interest are rarely available. Thus, starting from the individual-level model $r_k(x;\beta) = r_{0k}\exp(x\beta)$, Prentice and Sheppard[55] and Sheppard and Prentice[60] proposed estimating the individual parameters $\beta$ by directly regressing the $\bar{r}_k$ on the survey data using the induced aggregate model $r_k = r_{0k}E_k[\exp(x\beta)]$, where $E_k[\exp(x\beta)]$ is the average of $\exp(x\beta)$ over the individuals in region k. Prentice and Sheppard[55] show how the observed rates $\bar{r}_k$ and the survey data (the $x_i$) can be used to fit this model. As do earlier authors, they estimate region-specific averages by the sample averages, but in the absence of external data on $\beta$ they impose identifying constraints on the region-specific baseline rates $r_{0k}$ (e.g. by treating them as random effects); see Cleave *et al.*[3] and Wakefield[61] for related approaches.

Prentice and Sheppard call their method an 'aggregate-data study'; however, much of the social-science literature has long used this term as a synonym for 'ecologic study',[16,26] and so I would instead call it an incomplete multilevel study ('incomplete' because, unlike standard multilevel analyses,[62] individual-level outcomes are not obtained). Prentice and Sheppard conceived their approach in the context of cancer research, in which few cases would be found in modest-sized survey samples. For studies of common acute outcomes, Navidi *et al.*[8] propose a complete multilevel strategy in which the within-region surveys obtain outcome as well as covariate data on individuals, which obviates the need for indentifying constraints.

Multilevel studies can combine advantages of both individual-level and ecologic studies, including the confounder control achievable in individual-level studies, and the exposure variation and rapid conduct achievable in ecologic studies.[57] These advantages are subject to a number of assumptions that must be carefully evaluated,[55] several of which they share with ecologic studies. For example, multilevel studies based on recent individual surveys must assume stability of exposure and covariate distributions over time to ensure that the survey distributions are representative of the distributions that determined the observed ecologic rates; this assumption will be suspect when there were individual behavioural trends or important degrees of migration following the exposure period relevant to the observed rates.[63,64] They also can suffer from the problem, mentioned earlier, that the aggregate-level (ecologic) data usually concern arbitrary administrative or political units, and so can be poor contextual measures. Furthermore, multilevel studies face a major practical limitation in requiring data from representative samples of individuals within ecologic units, which can be orders of magnitude more expensive to obtain than the routinely collected data on which most ecologic studies are based.

In the absence of valid individual-level data, multilevel analysis can still be applied to the available ecologic data via non-identified random-coefficient regression. As in applications to individual-level studies,[65] this approach begins with specification of a hierarchical prior distribution for parameters not identified by available data. The distribution for the exposure effect of interest is then updated by conditioning on the available data. This approach is a multilevel extension of random-effects ecologic regression to allow constraints on $\beta$ (including a distribution for $\beta$) in the aggregate model, in addition to constraints on the $r_{0k}$. It is essential to recognize that these constraints are what identify exposure effects in ecologic data; hence, (as with all ecologic results), a precise estimate should always be traced to the constraints that produced the precision.

## Discussion

The present review has focused on confounding problems in ecologic studies. These are not the only such problems faced by ecologic studies. Two others are especially noteworthy for their divergence from individual-level study problems.

### Non-comparability among ecologic analysis variables

Non-comparable restriction and standardization of variables remains common in ecologic analyses, despite the fact that it can lead to considerable bias.[66] Typical examples involve restricted standardized rates regressed on crude ecologic variables, such as sex-race-specific age-standardized mortality rates regressed on contextual variables (e.g. air-pollution levels) and unrestricted unstandardized summaries (e.g. per-capita cigarette sales). If, as is usual, only unrestricted unstandardized regressor summaries are available, less bias might be incurred by using the *crude* (rather than standardized) rates as the outcome and controlling for ecologic demographic differences by entering multiple age-sex-race-distribution summaries in the regression[66] (e.g. proportions in different age-sex-race categories). More work is needed to develop methods for coping with non-comparability.

### Measurement errors

Effects of measurement errors on ecologic and individual-level analyses can be quite different. For example, Brenner *et al.*[67] found that independent non-differential misclassification of a binary exposure could produce bias away from the null and even reverse estimates from a standard linear or log-linear ecologic regression analysis, even though the same error would produce only bias toward the null in a standard individual-level analysis; analogous results were obtained by Carroll[68] for ecologic probit regression with a continuous exposure. Results in Brenner *et al.*[67] also indicate that ecologic regression estimates can be extraordinarily sensitive to errors in exposure measurement. On the other hand, Brenner *et al.*[69] found that independent non-differential misclassification of a single binary confounder produced no increase in bias in an ecologic linear regression. Similarly, Prentice and Sheppard[55] and Sheppard and Prentice[60] have found robustness of their incomplete multilevel analysis to purely random measurement errors.

In addition to assuming very simple models for individual-level errors, the foregoing results also assume that the ecologic covariates in the analysis are direct summaries of the individual measures. Often, however, the ecologic covariates are subject to errors beyond random survey error or individual-level measurement errors, as for example when per-capita sales data are used as a proxy for average individual consumption, and when area measurements such as pollution levels are subject to errors. Some work has been done examining the impact of such *ecologic* measurement error on cross-level inferences under simple error models,[8,70] but more research is needed, especially for the situation in which the grouping variable is an arbitrary administrative unit serving as a proxy for a contextual variable.

## Conclusion

The validity of any inference can only benefit from explication and critical scrutiny of the assumptions used to derive the inferences. My focus on ecologic-study problems stems solely from what I perceive as a common blindness to (or even denial of) the special assumptions needed to derive effect estimates from ecologic data alone, and the profound sensitivity of ecologic estimates to those assumptions (even if the estimate is of a contextual effect). The fact that individual-level studies have complementary limitations does not excuse this oversight.

Nonetheless, despite the critical tone of my remarks here and in earlier articles, I believe that ecologic data are worth examining, as demonstrated by careful ecologic analyses[53,58] and by methods that combine individual and ecologic data.[8,11,55,60] Furthermore, it is important to remember that the *possibility* of bias does not demonstrate the presence of bias, and that a conflict between ecologic and individual-level estimates does not by itself demonstrate that the ecologic estimates are the more biased.[13,18,19,71] This is because (1) the two types of estimates are subject to overlapping but distinct sets of biases, and it can happen that the individual-level estimates are the more biased; and (2) the effects measured by the two types of estimates are overlapping but distinct, with ecologic estimates incorporating a contextual component that is frequently absent from the individual estimates due to contextual (population) restrictions on individual-level studies. Indeed, the contextual component may be viewed as both a key strength and weakness of ecologic studies, for it is often of greatest substantive importance even as it is especially vulnerable to confounding. Thus, in the absence of good multilevel studies, ecologic studies will no doubt continue to fuel controversy, and so inspire the conduct of potentially more informative studies.

**KEY MESSAGES**

- Though it is commonly recognized that ecological studies can suffer from special biases in estimating individual effects, it is rarely acknowledged that the same biases affect ecologic estimates of contextual effects.
- Individual-level data are required to address these problems without resorting to controversial assumptions.

## References

[1] Langbein LI, Lichtman AJ. *Ecological Inference*. Series/No. 07–010, Thousand Oaks, CA: Sage, 1978.

[2] Piantadosi S, Syar DP, Green SB. The ecological fallacy. *Am J Epidemiol* 1988;127:893–904.

[3] Cleave N, Brown PJ, Payne CD. Evaluation of methods for ecological inference. *J Roy Stat Soc Ser A* 1995;158:55–72.

[4] Plummer M, Clayton D. Estimation of population exposure in ecological studies. *J Roy Stat Soc Ser B* 1996;58:113–26.

[5] Morgenstern H. Ecologic studies. In: Rothman KJ, Greenland S (eds). *Modern Epidemiology. 2nd Edn.* Philadelphia: Lippincott, 1998, pp.459–80.

[6] Firebaugh G. Assessing group effects. In: Borgatta EF, Jackson DJ (eds). *Aggregate Data: Analysis and Interpretation.* Beverly Hills: Sage, 1980, pp.13–24.

[7] Von Korff M, Koepsell T, Curry S, Diehr P. Multi-level analysis in epidemiologic research on health behaviors and outcomes. *Am J Epidemiol* 1992;135:1077–82.

[8] Navidi W, Thomas D, Stram D, Peters J. Design and analysis of multilevel analytic studies with applications to a study of air pollution. *Environ Health Persp* 1994;102(Suppl.8):25–32.

[9] Susser M, Susser E. Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. *Am J Public Health* 1996;86:674–77.

[10] Duncan C, Jones K, Moon G. Health-related behaviour in context: a multilevel modeling approach. *Soc Sci Med* 1996;42:817–30.

[11] Duncan C, Jones K, Moon G. Context, composition and heterogeneity: using multilevel models in health research. *Soc Sci Med* 1998;46: 97–117.

[12] Greenland S. A review of multilevel theory for ecologic analysis. *Stat Med* 2001;20:to appear.

[13] Schwartz S. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *Am J Public Health* 1994;84: 819–24.

[14] Pearce N. Traditional epidemiology, modern epidemiology, and public health. *Am J Public Health* 1996;86:678–83.

[15] Koopman JS, Longini IM Jr. The ecological effects of individual exposures and nonlinear disease dynamics in populations. *Am J Public Health* 1994;84:836–42.

[16] Firebaugh G. A rule for inferring individual-level relationships from aggregate data. *Am Soc Rev* 1978;43:557–72.

[17] Hakama M, Hakulinen T, Pukkala E, Saxen F, Teppo L. Risk indicators of breast and cervical cancer on ecologic and individual levels. *Am J Epidemiol* 1982;116:990–1000.

[18] Greenland S, Robins J. Ecologic studies—biases, misconceptions, and counterexamples. *Am J Epidemiol* 1994;139:747–60.

[19] Greenland S, Robins JM. Accepting the limits of ecologic studies. *Am J Epidemiol* 1994;139:769–71.

[20] Naylor CD. Ecological analysis of intended treatment effects: caveat emptor. *J Clin Epidemiol* 1999;52:1–5.

[21] King G. *A Solution to the Ecological Inference Problem.* Princeton: Princeton University Press, 1997.

[22] King G. The future of ecological inference (letter*). J Am Stat Assoc* 1999;94:352–54.

[23] Rivers D. Review of 'A solution to the ecological inference problem.' *Am Pol Sci Rev* 1998;92:442–43.

[24] Freedman DA, Klein SP, Ostland M, Roberts MR. Review of 'A solution to the ecological inference problem.' *J Am Stat Assoc* 1998;93:1518–22.

[25] Freedman DA, Ostland M, Roberts MR, Klein SP. Reply to King (letter). *J Am Stat Assoc* 1999;94:355–57.

[26] Borgatta EF, Jackson DJ (eds.). *Aggregate Data: Analysis and Interpretation.* Beverly Hills: Sage, 1980.

[27] Iversen GR. *Contextual Analysis.* Thousand Oaks, CA: Sage, 1991.

[28] Rothman KJ, Greenland S. *Modern Epidemiology. 2nd Edn.* Philadelphia: Lippincott, 2000.

[29] Achen CH, Shively WP. *Cross-Level Inference.* Chicago: University of Chicago Press, 1995.

[30] Duncan OD, Davis B. An alternative to ecological correlation. *Am Soc Rev* 1953;18:665–66.

[31] Cohen BL. Ecological versus case-control studies for testing a linear-no-threshold dose-response relationship. *Int J Epidemiol* 1990;19: 680–84.

[32] Cohen BL. In defense of ecological studies for testing a linear no-threshold theory. *Am J Epidemiol* 1994;139:769–68.

[33] Cohen BL. Re: Parallel analyses of individual and ecologic data residential radon, cofactors, and lung cancer in Sweden (letter). *Am J Epidemiol* 2000;152:194–95.

[34] Susser M. The logic in ecological. *Am J Public Health* 1994;84:825–35.

[35] Openshaw S, Taylor PH. The modifiable area unit problem. In: Wrigley N, Bennett RJ (eds). *Quantitative Geography.* London: Routledge, 1981, Ch. 9.

[36] Sheppard L. Insights on bias and information in group-level studies. *Biostatistics* 2002;to appear.

[37] Freedman DA, Klein SP, Sacks J, Smyth CA, Everett CG. Ecological regression and voting rights (with discussion). *Eval Rev* 1998;15: 673–816.

[38] Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol* 1989;18:269–74.

[39] Greenland S, Morgenstern H. Neither within-region nor cross-regional independence of covariates prevents ecological bias (letter). *Int J Epidemiol* 1991;20:816–18.

[40] Richardson S, Hémon D. Ecological bias and confounding (letter). *Int J Epidemiol* 1990;19:764–66.

[41] Piantadosi S. Ecologic biases. *Am J Epidemiol* 1994;139:761–64.

[42] Stidley C, Samet JM. Assessment of ecologic regression in the study of lung cancer and indoor radon. *Am J Epidemiol* 1994;139:312–22.

[43] Lagarde F, Pershagen, G. Parallel analyses of individual and ecologic data on residential radon, cofactors, and lung cancer in Sweden. *Am J Epidemiol* 1999;149:268–74.

[44] Lagarde F, Pershagen, G. The authors reply (letter). *Am J Epidemiol* 2000;152:195.

[45] Cho WTK. If the assumption fits: a comment on the King ecologic inference solution. *Pol Anal* 1998;7:143–63.

[46] Stoto MA. Review of 'Ecological inference in public health.' *Pub Health Rep* 1998;113:182–83.

[47] Wen SW, Kramer MS. Uses of ecologic studies in the assessment of intended treatment effects. *J Clin Epidemiol* 1999;52:7–12.

[48] Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990;1:421–29.

[49] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;14:29–46.

[50] Goodman LA. Some alternatives to ecological correlation. *Am J Sociol* 1959;64:610–25.

[51] Robins JM, Murphy S, Greenland S. Towards a formal theory of causation in ecologic and multilevel studies. *J Roy Stat Soc Ser A*, In Press.

[52] Vaupel JW, Manton KG, Stallard, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; 16:439–54.

[53] Richardson S, Stücker I, Hémon D. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *Int J Epidemiol* 1987;16:111–20.

[54] Dobson AJ. Proportional hazards models for average data for groups. *Stat Med* 1988;7:613–18.

[55] Prentice RL, Sheppard L. Aggregate data studies of disease risk factors. *Biometrika* 1995;82:113–25.

[56] 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.

[57] Guthrie KA, Sheppard L. Overcoming biases and misconceptions in ecologic studies. *J Roy Stat Soc Ser A* 2001;164:141–54.

[58] Prentice RL, Sheppard L. Validity of international, time trend, and migrant studies of dietary factors and disease risk. *Prev Med* 1989; 18:167–79.

[59] Kleinman JC, DeGruttola VG, Cohen BB, Madans JH. Regional and urban-suburban differentials in coronary heart disease mortality and risk factor prevalence. *J Chron Dis* 1981;34:11–19.

[60] Sheppard L, Prentice RL. On the reliability and precision of within- and between-population estimates of relative rate parameters. *Biometrics* 1995;51:853–63.

[61] Wakefield J. Ecological inference for 2 × 2 tables. *J Roy Stat Soc* 2002; to appear.

[62] Goldstein H. *Multilevel Statistical Models*. New York: Edward Arnold, 1995.

[63] Stavraky KM. The role of ecologic analysis in studies of the etiology of disease: a discussion with reference to large bowel cancer. *J Chron Dis* 1976;29:435–44.

[64] Polissar L. The effect of migration on comparison of disease rates in geographic studies in the United States. *Am J Epidemiol* 1980;111: 175–82.

[65] Greenland S. When should epidemiologic regressions use random coefficients? *Biometrics* 2000;56:915–21.

[66] Rosenbaum PR, Rubin DB. Difficulties with regression analyses of age-adjusted rates. *Biometrics* 1984;40:437–43.

[67] Brenner H, Savitz DA, Jöckel K-H, Greenland S. Effects of nondifferential exposure misclassification in ecologic studies. *Am J Epidemiol* 1992;135:85–95.

[68] Carroll RJ. Some surprising effects of measurement error in an aggregate data estimator. *Biometrika* 1997;84:231–34.

[69] Brenner H, Greenland S, Savitz DA. The effects of nondifferential confounder misclassification in ecologic studies. *Epidemiology* 1992; 3:456–59.

[70] Wakefield J, Salway R. A statistical framework for ecological and aggregate studies. *J Roy Stat Soc Ser A* 2001;164:119–37.

[71] Morgenstern H. Ecologic study. In: Armitage P, Colton T (eds). *Encyclopedia of Biostatistics. Vol. 2*. Chichester: Wiley, 1998, pp.1255–76.