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 β estimate $\hat{\beta}$ from individual-level studies (e.g. a proportional-hazards model derived from cohort-study data), and the observed rate $\tilde{r}_k$ in region k. Then we may compare $\tilde{r}_k$ 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 β by directly regressing the $\tilde{r}_k$ 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 $\tilde{r}_k$ 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 β 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 β (including a distribution for β) 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.