29

ECOLOGICAL AND GROUP-LEVEL STUDIES

OBJECTIVES

After reading this chapter, you should be able to:

1. List the 3 major categories of variable used in ecologic models, describe their attributes, and apply these to a specific research question.

2. Describe the constructs of a linear model at the individual and group levels and the constraints on estimating incidence rate ratios at the group level.

3. Describe how within-group misclassification, group-level confounding, and group-level interaction can effect causal inferences.

4. Describe the basis of the ecologic and atomistic fallacies.

5. Identify scenarios where ecologic studies are less likely to produce cross-level inferential errors.

6. Describe how to integrate individual-level studies with ecologic studies to prevent cross-level inferential errors

7. Describe the rationale for using non-ecologic group-level studies in epidemiologic research.

29.1 INTRODUCTION

Up until this point in our text, our interest has focused on studies where the exposure, outcome, and confounders are measured and analysed at the individual level. Here, we cover studies in which groups of subjects are sampled and the exposure, outcome, and confounders are measured and analysed at the group level, but the researcher wishes to make inferences to individuals. Epidemiologists call these **ecologic studies** (see Examples 29.1 and 29.2) (Greenland and Robins, 1994; Levin, 2006). In most instances the grouping is based on geographic and/or administrative areas, in increasing size, such as townships, counties, districts, provinces, and nations, or the groupings are based on organisational units such as schools or work-place/industry. For example, in Example 29.1 the unit is the county, in Example 29.2 it is the state. In other studies, the grouping might be based on areas defined by weather, or level of air/water pollutant; the researchers then might wish to use spatial correlations to ascertain possible effects of the exposures (Blanchard *et al*, 2011). In essence, ecologic studies are cluster samples (see Chapter 2; individuals are units of concern) but all exposure, outcome, and confounder measurements are made at or summarised to the cluster (group) level, not the individual. Ecologic studies often serve as a useful step in the evolution of research, providing preliminary information that requires more definitive research designs to make progress (Savitz, 2012). Most often for epidemiologists the outcome is a health-related condition, including traditional 'disease', but also outcomes such as parasuicide (Newman and Stuart, 2005), or behaviour (Lee and Cubbin, 2009) have been used.

Between individual-level and ecologic studies, there is a collection of study designs where either the exposure or outcome is measured at the individual level, while the other variable(s) is measured at the group level and the researcher wishes to make inferences to individuals. These are called **partial ecologic studies** (Webster, 2002).

The primary analytic feature, and greatest limitation, of ecologic studies is that we do not know the joint distribution of both the risk factor(s) and the disease at the individual level within each group. Although we know the proportion exposed to each factor and the risk, or rate, of cases for each group, we do not know the proportion of exposed cases within the group. Typically this is because we lack individual-level data on the risk factor, the disease, or both (Rothman *et al*, 2008). For example in Example 29.1 we do not know the distribution of water arsenic levels to which cancer and non-cancer patients were exposed.

Ecologic studies have been sub-classified depending on whether or not the exposure of interest is measured directly. For example, the study might be called **exploratory** if there is no direct

Example 29.1 Ecologic associations between arsenic levels in ground water and cancer incidence in Idaho

County-level data on cancer incidence data (1991–2005) from the Cancer Data Registry of Idaho and arsenic levels in ground water (1991– 2005) from the Idaho Department of Environmental Quality were used (Han *et al*, 2009). The authors calculated age-adjusted incidence rates (adjusted to the US 2000 standard population) for specific cancers, and all malignant cancers combined. Multivariable linear regression analysis was applied to evaluate the relationship between arsenic levels in ground water and cancer incidence. When adjusted for the proportion of people in each county for each of race, gender, smoking, body mass index (<25), and population density/mile^2, no relationship was found between arsenic levels in ground water and cancer incidence.

Example 29.2 Bladder cancer incidence rates compared with ecologic factors among states in America

Bladder cancer mortality rates for all states from 2000 to 2003 were age adjusted to the US population for the year 2000 (Colli and Kolettis, 2010). Predictors, included age-adjusted data for the percent of the people among states who were former smokers or lacked health insurance, mean monthly UV index values, and the percent of the population who received drinking water from a surface water supply (as opposed to groundwater). A multivariable linear regression analysis was used to determine the best fit model for predicting bladder cancer incidence. Bladder cancer incidence correlated directly with smoking and inversely with solar UV radiation levels.

measure of the exposure of interest or if there is no specific exposure variable being studied. For example, if a study portrayed the rate of disease (*eg E coli* 0157 in humans) by administrative area, we might use previous knowledge of local features (*eg* cattle density in the area) as a surrogate, to explain the observed spatial variation in rates of disease, even though there was no direct measurement of cattle density in the study. Conversely, ecologic studies might be called **analytic** if the exposure factor is measured and included in the analysis.

In general, ecologic studies can be conducted using the same approaches as used for studying individuals; namely by:

(1) comparing the frequencies of exposure and disease among a number of groups at a given point in (or during a limited period of) time, similar to cross-sectional studies, or
(2) estimating the changes in both exposure and disease frequencies during a given period in one or more groups (often in just one group) as in cohort or case-control studies, or
(3) a combination of the 2 types.

If the groupings contain small numbers of people, or the groups are highly variable in population-size, the analysis should account for the different precision of disease rates by group. Spatial analysis might require adjustment for spatial correlation. Short term temporal studies might need to adjust for a lag period between exposure and disease occurrence, whereas, studies that cover an extended period of time might have to account for, and try to separate, the age, period, and cohort effects on the outcome, as well as taking account of changes in diagnostic standards. Attempts to separate age, period, and cohort effects lead to an identifiability problem as these 3 components are interlinked and cannot be assessed independently (see Osmond and Gardner (1989); Robertson *et al* (1999) for a discussion). Studies that combine both among-group and temporal approaches for identifying associations should provide a more thorough test of hypotheses than either approach alone. We begin our discussion of ecologic studies by asking ourselves, "why study groups?", especially if we want to make inferences to individuals?

29.2 Rationale for group-level studies

If the grouping (*eg* census tracts, cities, or nations) are the unit of interest, these are not ecologic studies. The recent increase in the use of spatial statistics often focuses on large aggregates of people such as cities, districts, watersheds, and so forth. Providing the variables are measured at the group level and any inferences are directed towards this level, this poses no particular inferential problem. See Section 29.9 for further discussion of non-ecologic group-level studies. However, if the intent is to make inferences about individuals based on the results from the

group-level analysis, we must be very careful (reasons for this are discussed subsequently) because of a number of pitfalls of logic. Nonetheless, some advantages of studying groups are:

Measurement constraints at the individual level Often, it is difficult to measure exposure at the individual level (*eg* level of pollutants, as in our previous example) so an average for the group might suffice. In other circumstances, the variation in an exposure (*eg* diet) between individuals might be large, whereas the group average might adequately reflect exposure of the group to specific nutrients for the purposes of the study.

Exposure homogeneity If there is little variation in exposure among individuals within a group, it is difficult to assess the exposure's impact using individuals as the unit of interest. For example, if all people within a city are exposed to the same level of radon, we might need to study groups (cities with different average exposure levels of radon) to observe the apparent effect of radon exposure. Hence, using groups with a wider variation in level or type of exposure than exists within groups would be helpful.

Interest in group-level effects These arise naturally if one is studying the impact of area-wide programmes, or area-wide exposures. For example, in many circumstances, vaccines, different health care systems, and treatments (*eg* water-based fluoride) can only be delivered, or implemented practically, at the group level. Hence, groups are of interest. It is also recognized that group-level factors often impact on individual behaviour(s) and if these factors can be identified, they may be exploited in health promotion programs (*eg* in anti-smoking campaigns).

Simplicity of analysis Often it appears to be easier to display and present group-level rather than individual-level data. However, as we will point out, these group-level analyses might hide serious methodological problems when we are attempting to make inferences to individuals (see Section 29.4).

29.3 Types of ecologic variable

The categorisation of variable types within ecologic studies is still dynamic (see Diez-Roux, (1998a; 1998b); McMichael (1999) for a discussion). For our purposes, we will use 3 categories: aggregate (aka derived), environmental and global variables.

29.3.1 Aggregate variables

This type of variable is also called a **derived variable** in that it is formed, at least in part, by aggregating individual observations to form a summary variable (often the mean) for the group (*eg* proportion exposed, per cent obese, disease rate, mortality rate) such as the cancer rate in Example 29.1. Thus, aggregate variables are summaries of measurements made on individuals within the group (*eg* the proportion who smoke in Example 29.2). Aggregate variables can refer to the predictor variables, the outcome variable, or both. In the instances where disease is the outcome, the aggregate usually is measured using rates because most groups are open; if closed, then a risk-based approach can be used.

29.3.2 Environmental or contextual

Usually these are physical characteristics of the group such as local weather (*eg* UV index; Example 29.2), level of pollutants in the area, or characteristics of water supply (*eg* deep well versus surface water). Environmental/contextual variables may or may not be aggregate variables; the key feature is that they have an analogue at the individual level (*eg* level of air pollution, or level of radon exposure). Often we do not actually measure these variables at the individual-level because of practical constraints and for analysis, we assign the same value of the variable to every individual within the group. This approach becomes especially tenuous as the within-group variance in that factor increases. For example, between-individual variance for exposure to UV rays might be large, so serious misclassification can result from this process. In addition, it might well be that there is an interaction between the factor at the individual level (*eg* titre to influenza) and the contextual variable for the same factor (*eg* percentage of people with a protective titre to influenza), as in herd immunity and these need to be identified for proper inference.

29.3.3 Group or global

These variables reflect a characteristic of groups, organisations or places for which there is no analogue at the individual level (*eg* population density as in Example 29.1).

29.4 Issues related to modelling approaches in ecologic studies

We begin by noting that, at the group level, both predictor and outcome ecologic variables often are measured on a continuous scale, even though they might be dichotomous variables at the individual level; this is particularly true when aggregate variables are used. As mentioned, if the outcome at the group level is classified as dichotomous (*eg* disease present or absent in the group) and the inferences are made at the group level, the study is not an ecologic study and can be pursued with the same features and constraints as ordinary observational studies (Chapters 7–10). With aggregate variables, because the outcome reflects the average rate or risk for the group, a natural scale for modelling group level variables is the linear regression model (as outlined in Chapter 14) in which we regress the aggregate outcome variable on the aggregate exposure variables. Although this is a valid approach, many researchers opt to use correlation coefficients; whereas we would prefer the regression coefficient as a measure of association. About 33% of the ecologic studies in recent years have used correlation coefficients, instead of regression coefficients (Ojha *et al*, 2011) as their measure of association. Because of the limitations of the linear model approach, some prefer to use a Poisson model (see Example 29.3) when the outcome measures are counts of infrequent disease.

29.5 The linear model in the context of ecologic studies

As an example of the linear model approach, we can imagine the continuous outcome Y representing the annual risk or rate of hospital admission (*eg* 0.15 per person-year in city j) modelled as a linear function of the exposure to air pollution (*eg* 0.3 of the people in city j are exposed to high levels of air pollution as $X_{1)}$ and perhaps adjusting for the effects of one or more confounders (*eg* the average age of people in each city as X_2). The model could be specified as:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \epsilon_j$$

Example 29.3 A population-based ecologic study of inflammatory bowel disease

The objective of this study was to determine associations between the incidence of inflammatory bowel disease (IBD), specifically Crohn's disease and ulcerative colitis, in the Canadian province of Manitoba, and sociodemographic, geographic, and disease-related characteristics of the study population for the period 1990–2001 (Green *et al*, 2006). The unit of analysis was a 'small geographic area' (this is a defined census area in Canada) of which there were 498 in the province. For each area, potential predictors included the socioeconomic status index, the percentage of the population reporting Jewish ethnicity, the percentage of the population reporting Aboriginal ancestry, and the percentage of the population reporting 'visible minority' status, age-standardised reportable enteric disease, and small-area estimates of multiple sclerosis incidence rates. A Poisson regression model was used for analyses.

where X_{1j} is the proportion exposed to air pollution and X_{2j} is the average age in city j, respectively. Environmental or global variables might be entered and analysed as either dichotomous, ordinal, or continuous variables. The linear model would provide an incidence rate difference (ID_G) from the exposure which is estimated as β_1, conditional on the other variables in the model. In many situations, the outcome might need to be transformed to better meet the assumptions of the linear model, and a weighted regression might be needed to account for the different levels of precision by group (because of large differences in the number of study subjects per city). For example, the outcome could be weighted by the group size, the reciprocal of the within-group variance, or some function relating to the within-group homogeneity of exposure.

A 'nice' feature of a linear model is that, if the rate (or risk) difference is constant across groups at the individual level, assuming no other biases, the rate difference at the group level will be of the same magnitude. In contrast, if the rate ratio is constant at the individual level, a logit model of the outcome will produce biased estimates at the group level (Rothman *et al*, 2008).

Associations between predictors and dichotomous outcomes at the individual level are usually based on ratio measures (*eg OR*). However, a problem with using ratio measures at the group level in linear models is that, for aggregate variables, these estimates force us to extrapolate our inferences to groups (cities) with no exposure and to groups with 100% exposure; rarely do we have these groups in our data. For example, from a simple linear model, β_0 is the rate in non-exposed (X=0) groups and $\beta_1+\beta_0$ is the rate in exposed groups (X=1). Hence, the incidence rate ratio (IR) at the group level is:

$$IR_G = \frac{\beta_0 + \beta_1}{\beta_0} = 1 + \frac{\beta_1}{\beta_0} \qquad \textit{Eq 29.1}$$

Thus, valid inferences about ratio measures require totally exposed and non-exposed groups, and extrapolation beyond the range of the available data is not good practice.

As in linear models (Chapter 14), issues of confounding and interaction are dealt with by including these variables in the model. Control of individual level confounders in an ecologic analysis, however, is less successful than it is in an individual-level analysis because control is performed by using average or proxy data, hence attenuating associations. Also, risk factors in ecologic analysis tend to be more highly correlated with each other than they are at the individual level making it difficult to isolate the effect of individual risk factors. When other variables are included in the model, the previous estimation method for IR_G must be extended to

account for their effect. In order to accomplish this, we usually set the value of these variables (that is the *Xj*s) to their mean as shown in Eq 29.2.

$$IR_G = \frac{(\beta_0 + \beta_1 + \sum \beta \overline{X})}{\beta_0 + \sum \beta \overline{X}} \qquad \textit{Eq 29.2}$$

where $\sum \beta \overline{X}$ is the sum of the products of the other coefficients and the mean values of the other *X* variables in the model.

Some researchers prefer to use standardised outcomes, such as (standardised morbidity/ mortality ratios (*SMR*s)) to control confounding and they regress these standardised outcomes on the group-level explanatory variables. Typically age and sex, and sometimes race are included in the *SMR*. However, this approach does not prevent confounding unless the explanatory variables are also standardised in the same manner, and sufficient data to achieve this often are not available.

Interaction is usually modelled in the same manner as with individual analyses using a product term (*eg* X_1*X_2). However, creating this term based on group means is not equivalent to taking the average of the terms created at the individual level. Thus, this approach has a different (often lower) level of ability to detect an interaction. One particular type of interaction that is important to identify is a contextual effect where the group-level factor modifies the same factor's effect at the individual level. To identify this contextual effect, we create a cross-product term between the factor at the group (*eg* proportion with titres) and the individual level (presence of a titre) and test its significance.

29.6 Issues related to inferences

The major inferential problems that arise are because of heterogeneity of exposure and of confounders within the group. Thus, a finding at the group level—that exposure is associated with increases (or decreases) in the risk of disease by 3 times—does not mean that this is true at the individual level. Indeed, it could be that the exposed subjects are not the ones having the highest risk of becoming cases. This error in inference is termed the **ecologic fallacy** (see Section 29.9.2 for the opposite error, the **atomistic fallacy**). The ecologic fallacy became well known after Robinson explained it in 1950 (Robinson, 2009; Subramanian *et al*, 2009) and Webster (2002) has questioned if we are still haunted by it? In order to help avoid the fallacy, Firebaugh (2009) and Oakes (2009) point out the need for multilevel perspectives and analyses.

In addition, even without the ecologic fallacy, the group-level bias usually exaggerates the magnitude of the true association away from the null. However, occasionally it reverses the direction of the association. As a simple, hypothetical example, assume that you are investigating a disease which is caused by an infectious agent X that produces lifelong antibody titres and clinical disease which only develops if exposure occurs later in life. Early exposure does not produce clinical signs. At the individual level, disease will be positively associated with exposure to *X* (all cases will have antibodies). However, at the group level, a high prevalence of *X* will more likely result in early exposure and hence, be associated with a low level of disease.

We now examine the 3 major causes of ecologic bias—within-group bias, group-level confounding, and group-level interaction—in more detail; see Greenland and Morgenstern, 1989.

29.7 Sources of ecologic bias

29.7.1 Within-group

Within-group bias can be caused by confounding, selection bias or misclassification. Here we discuss only misclassification of individual-level exposure and its effects on observations at the group level.

As noted earlier, if aggregated exposure variables are used, the exposure level of groups is defined by combining individual exposure observations. Imperfect exposure classification of individuals in turn leads to errors in the estimates of both the individual-level association and the group-level association. As noted in Chapter 12, non-differential exposure misclassification at the individual-level biases the observed association toward the null, but, in ecologic studies, it biases the association **away** from the null (Greenland, 1992). The effect of this bias on the rate ratio derived from an ecologic linear regression model can be predicted if the necessary data are known as indicated in Eq 29.3:

$$IR_G = 1 + \frac{IR - 1}{Se + Sp * IR - IR} \qquad \textit{Eq 29.3}$$

where *Se* is the individual-level sensitivity, *Sp* is the individual-level specificity, and *IR* the true individual-level incidence rate ratio. The ID_G is also biased by the factor (*Se*+*Sp*-1). This bias can be quite large as shown in Example 29.4. Also, when exposure (or disease) prevalence of groups is based on a small sample of individuals within each group, measurement error at the individual level is compounded by sampling error (hence, the earlier referral to extreme values of outcomes with small group sizes). For more details on this bias, see Brenner *et al* (1992).

29.7.2 Confounding by group

If both the level of exposure and the background rate of disease in the unexposed individuals varies across groups, this sets up a group-level correlation of exposure and outcome. Such confounding can arise from the differential distribution of extraneous individual-level risk factors across groups (note that these risk factors need not (although they can) be confounders at the individual level (*ie* within groups)), or from the occurrence of group-level confounders (*ie* here the covariates are associated with both exposure and disease at the group level). One example of this is research on the contribution of influenza to hospital admissions or mortality; the contribution occurs in the fall and winter but there are numerous confounders that vary seasonally including the weather and human behaviour. Jackson (Jackson, 2009) examines some approaches to research on factors that vary seasonally, including the use of time-series methodology. Example 29.5 demonstrates group-level confounding.

29.7.3 Effect modification (interaction) by group

In a linear model, bias will occur at the group level if the rate difference at the individual level varies across groups. We should recall that although we use a logit scale (usually) at the individual level, we often use a linear model at the group level. This introduces a non-linearity into the comparison of the results which might evidence itself as interaction in the linear scale. Such variation can arise from the differential distribution of individual level effect modifiers

Example 29.4 Effect of individual-level exposure misclassification on group-level results

For simplicity we will use schools as our group and childhood respiratory disease (CRD) as our outcome. We begin with the correctly classified study population structures in 2 schools (j=1,2).

	School 1			School 2		
Correctly classified	**Exposed**	**Non-exposed**	**Totals**	**Exposed**	**Non-exposed**	**Totals**
Number of cases	50	40	90	100	30	130
Person-time (t_j)	200	800	1000	400	600	1000
CRD Rate (I_j)	0.250	0.050	**0.090**	0.250	0.050	**0.130**
Group proportion exposed		**0.20**			**0.40**	

The data in **bold typeface** are the numbers we would use for the analysis at the group level if there was no misclassification. At the individual level of analysis, the IR=5 and the ID=0.20 in both schools. Note that in school 1, 20% of the person-time is exposed (200/1000), while in school 2, this is 40% (400/1000). The regression coefficients for the group level analysis are obtained by solving the 2 equations for the 2 unknowns: $0.09=\beta_0+\beta_1*0.2$ and $0.13=\beta_0+\beta_1*0.4$ which gives the following model Y=0.050+0.2X. The ID_G=0.20 and

$$IR_G = 1 + \frac{0.2}{0.05} = 1 + 4 = 5$$

Now, with an imperfect test of exposure, having a sensitivity of 0.8 and a specificity of 0.9 (see Section 12.7,) we would observe the data below.

	School 1			School 2		
Incorrectly classified	**Exposed**	**Non-exposed**	**Overall rate**	**Exposed**	**Non-exposed**	**Overall rate**
Number of cases	44	46	90	83	47	130
Person-time (t_j)	240	760	1000	380	620	1000
Rate (I_j)	0.183	0.061	**0.090**	0.218	0.076	**0.130**
Group proportion exposed		**0.24**			**0.38**	

At the individual level, (based on the misclassified data pooled over the schools) the IR=3.04 and the ID=0.137. These are biased towards the null (*ie* 0). However, the exposure misclassification leads to biased group-level estimates of the proportion of exposed person-time in each school; the difference between the 2 schools becomes smaller and hence, the apparent effect of exposure becomes larger. Using the same approach to obtain the regression coefficients, the model is Y=0.0214+0.286X. At the group level, the misclassified IR_G is 14.3 and the ID_G is 0.29. Thus, a non-differential misclassification at the individual level has biased the group IR_G and ID_G away from the null.

across groups, or due to effect modification by a group-level factor (see Example 29.6).

29.7.4 Summary of confounding and interaction at the group level

To summarise the previous discussion, cross-level (*ie* ecologic) bias will **not** occur if :

- the incidence rate difference, within groups, is uniform across groups, and
- if there is no correlation between the group-level exposure and the rate of the outcome in the unexposed.

The only (but huge) drawback to these criteria is that individual-level data are required to evaluate them and these data rarely are available. Glynn and Wakefield (2010) demonstrate how to conduct sensitivity analysis of ecologic data and also the value of small amounts of individual level data to prevent cross-level bias.

On the other hand, if individual-level effect modifiers are differentially (*ie* unequally) distributed across groups, ecologic bias will occur as a result of the consequent group-level effect modification. If extraneous risk factors are differentially distributed across groups, ecologic bias will occur as a result of group-level confounding, **regardless** of whether the extraneous risk factor is a confounder at the individual level or not. Controlling for the extraneous risk factor in the ecologic analysis will generally remove only part of the bias.

Example 29.5 Effects of confounding on group-level results

In this example, E_1 is the exposure of interest at the individual level and E_2 is the potential individual-level confounder (both binary). At the group level, these are represented by the variables X_1 and X_2, respectively (for simplicity, we omit subscripts for schools), both measured on the continuous scale (data in **bold typeface** in table). Consider these data from 3 schools:

School A	**E_2+**		**E_2-**		**E_2 pooled**	
	E_1+	**E_1-**	**E_1+**	**E_1-**	**E_1+**	**E_1-**
Cases	52	74	5	7	57	81
Person-time t_a	260	740	260	740	520	1480
I_a	0.20	0.10	0.02	0.01	0.11	0.055
IR_a		2		2		2
	X_1: p(E_1+)=0.26		**X_2: p(E_2+)=0.50**		**Y: p(D+)=0.068**	
School B	**E_2+**		**E_2-**		**E_2 pooled**	
	E_1+	**E_1-**	**E_1+**	**E_1-**	**E_1+**	**E_1-**
Cases	56	52	8	8	64	60
t_b	280	520	420	780	700	1300
I_b	0.20	0.10	0.02	0.01	0.09	0.046
IR_b		2		2		2
	X_1: p(E_1+)=0.35		**X_2: p(E_2+)=0.40**		**Y: p(D+)=0.062**	
School C	**E_2+**		**E_2-**		**E_2 pooled**	
	E_1+	**E_1-**	**E_1+**	**E_1-**	**E_1+**	**E_1-**
Cases	60	30	14	7	74	37
t_c	300	300	700	700	1000	1000
I_c	0.20	0.10	0.02	0.01	0.74	0.037
IR_c		2		2		2
	X_1: p(E_1+)=0.50		**X_2: p(E_2+)=0.30**		**Y: p(D+)=0.056**	

(*continued on next page*)

Example 29.5 (*continued*)

Examining these data from the individual's perspective, we observe that the true (individual) *IR*s for E_1 and E_2 are 2 and 10, respectively. Both ratios are constant across schools so there is no interaction at the individual level. Also, there is no confounding by E_1 or E_2 within schools (as E_1 and E_2 are independent). However, because the prevalence of E_2 varies by school, this results in an association of school with Y that is independent of E_1. Consequently, the group-level estimate of the effect of E_1 (*ie* using X_1) may be biased. At the school level, a simple linear regression of Y of X_1 yields Y=0.080-0.049X_1 and the ecological estimate of IR_G is (0.031/0.080)=0.39 suggesting that exposure is sparing. Controlling for exposure 2 in the analysis does not prevent the bias with the equation being Y=0.038+0.000X_1+0.060X_2. The ID_G is zero, and using the mean prevalence of exposure for X_2 of 0.40, when X_1 changes from 0 to 1, we have (based on Eq 29.2)

$$IR_G = \frac{(.038 + .000 + .4 * .06)}{(.038 + .4 * .06)} = 1.00$$

This adjustment brings the IR_G for exposure 1 to the null value suggesting 'no effect.' Unfortunately, because we rarely have sufficient information to know whether or not the group- and individual-level results agree, relating group findings to individuals is fraught with difficulty.

Based on this summary, it is clear that we need to be careful when making inferences about individuals using group-level analyses; yet, group-level analyses will continue to be used. So, how can we help avoid some of these problems? Well, the misclassification issue is best resolved by reducing the level of errors, but the bias away from the null is still a reality and needs to be considered in all group-level studies. With respect to confounding and interaction, again these are real problems. But, both the confounding and effect modification examples used here are taken from scenarios where group-level analyses are unlikely to be rewarding because most of the variation is at the individual level. Because the outcome varies little across groups, research should focus on the individual level.

In general, ecologic bias will be less of a problem when:

(1) The observed range of exposure level across groups is large (Savitz, 2012). Linear regression analysis of ecologic data is especially sensitive to problems of limited among-group exposure variation. If this is the situation you are faced with, consider using other model forms, such as exponential and log-additive models (*eg* Poisson);
(2) The within-group variance of exposure is small; therefore in selecting study populations minimise the within-group and maximise the among-group exposure variation (sometimes using smaller, more homogeneous, groupings helps accomplish this);
(3) Exposure is a strong risk factor and varies in prevalence across groups (hence, the group-to-group variation in incidence is large), and
(4) The distribution of extraneous risk factors is similar among groups (*ie* little group-level confounding). Because confounding is notoriously difficult to control statistically in ecologic studies, research settings should be identified in which confounding is unlikely or modest in magnitude.

The opportunity to include health conditions that can serve as positive and negative controls strengthens ecologic evidence for the health outcome of interest (Savitz, 2012).

Despite the pitfalls, we should continue our struggle to gain valid knowledge from group level studies (Webster, 2002). While the biases discussed very likely occur frequently, the effects might be small and need not prevent us making valid inferences to individuals. In this regard,

Example 29.6 Effect modification by group

Consider the following data from 3 schools:

	School A		School B		School C		Total	
	E+	E-	E+	E-	E+	E-	E+	E-
CRD Cases	120	30	120	36	120	42	360	108
Person-time (t)	1000	1000	800	1200	600	1400	2400	3600
I	0.12	**0.03**	0.15	**0.03**	0.20	**0.03**	0.15	**0.03**
IR		4.0		5.0		6.7		5.0
ID		0.09		0.12		0.17		0.12
X_1 = p(E+)		**0.5**		**0.4**		**0.3**		
Y = p(D+)		**0.075**		**0.078**		**0.081**		

First let's examine the data from the perspective of the individual. We observe that the effect of the exposure *E* (as denoted by *IR*, or the *ID*) varies by school. Thus, some school-level factor is interacting with the exposure *E*, and with a large enough sample, this might be declared as significant interaction on either the additive or the multiplicative scale (see Chapter 13). Note that there is no confounding by group (*ie* school level) at the individual level because p(*D*+|*E*-)=0.03 in all 3 schools. Thus, school *per se* is not a cause of disease at the individual level (although we would argue against presenting a single estimate of effect when interaction is present). Also, because there is no confounding, the crude *IR* of 5.0 provides an unbiased estimate of the effect at the individual level. There is, however, interaction because some factor at the school level is making the impact of exposure (whether measured by *IR* or *ID*) vary across schools, and this effect increases as the prevalence of *E*+ decreases.

An ecologic analysis at the school level would only use the aggregated summary data (**bold typeface**) from the table. The ecologic linear regression of *Y* on *X* yields:

$$Y = 0.09 - 0.03\,X$$

and the ecologic estimate of IR_G would be:

$$1 + (-0.03/0.09) = 0.67$$

Clearly this is not anywhere near the individual-level *IR* of 5. Thus, the effect modification by group has led to an ecologic bias that actually reversed the direction of the association at the individual level.

we should treat these potential biases in the same manner as in individual-level studies; try to understand, quantify and minimise them.

Dufault and Klar (2011) have examined the quality of recent cross-sectional ecologic studies and note the following:

- Much cross-sectional research potentially has its data dictated by either the convenience or the necessity of using pre-existing information. Only 18% of authors implicitly or explicitly justified their choice of ecologic units
- The vast majority (97%) of outcomes were aggregate in nature. Non-aggregate outcomes included global measures such as regional socioeconomic inequality and governmental health-care expenditures.
- Fifty-four per cent of studies relied on fewer than 100 group-level observations.
- The most common analytical approaches were to use either ordinary least squares

regression or Poisson regression,

- Ninety per cent of studies used adjusted outcome rates, unfortunately in over half of these studies, the investigators failed to adjust covariates for age or sex when the outcomes had been standardised for these potential confounders and were at risk of publishing biased inferences.
- Only 42% of studies adequately and explicitly justified an ecologic analysis or why the design was either necessary or desirable.
- In the majority of the reviewed studies, investigators did not sufficiently inform the reader about the possibility of cross-level (ecologic) bias. The authors concluded that a STROBE-type list (standardised reporting guidelines; see Section 7.4.2) for ecologic studies was needed to provide guidance for researchers.

29.8 Analysis of ecologic data

Wakefield (2004) described ecological inference with 2X2 tables; his paper is a very detailed review of the analysis and interpretation of ecologic data. Wakefield concludes "We end with the obvious statement that the solution to the ecological inference problem is to supplement the aggregate data with representative and accurate survey sample information on individuals within the areas of interest. The use of prior information concerning the within-area probabilities and the extent of contextual effects will also be beneficial."

More recently, Wakefield (2009) authored an instructive paper on multilevel modelling (MLM). We will not pursue the details of multilevel modelling here, but would pick up on 2 of Wakefield's suggestions. First, it is important to validate the assumptions of MLM and Wakefield provides some examples of exploratory data analyses. For example, serious bias can arise when the random effects are correlated with covariates. Second, he stresses that valid interpretation of MLM coefficients requires considerable care (and gives examples). He concludes that although MLM is very helpful, largely because of the 'vast' list of potential confounders, most of which have not been measured, it remains very difficult to make causal inferences. Subramanian *et al* (2009) stress the need for multilevel analyses to resolve the the cross-level analytic and inferential issues we have described (see Chapters 21–23).

Wakefield & Haneuse (2008) describe how to link individual-level data on individuals with the ecologic data, using a two-phase design (similar to that discussed in Section 10.8) to prevent ecologic bias. One suggested approach is to use outcome-dependent (case-control) sampling within groups to obtain individual-level data (Haneuse and Wakefield, 2008). Another approach is to extend this and sub-sample on both outcome and exposure, as well as confounders of importance (Wakefield and Haneuse, 2008). The advantage of the 2-phase approach is that the sampling is more efficient since it incorporates exposure and confounder variables as well as outcome. Equal numbers of individuals can be sampled from each of the categories designated by the cross-classification of the outcome, exposure and confounding variables as well as 'group'. If there are a large number of groups (*eg* schools) it might be necessary to aggregate groups into larger units such as regions. The analysis of the data must take account of the outcome dependent sampling and the authors provide example software code (in R) on their website. In order to maintain efficiency of sub-sampling, it is important to limit the stratification to factors likely to be important predictors. However, if unimportant variables are included although efficiency is lowered, the procedure still prevents ecologic bias.

29.9 NON-ECOLOGIC GROUP-LEVEL STUDIES

A number of epidemiologists have noted that epidemiology initially focused on groups as the unit of interest, and only recently has it shifted that emphasis to individuals. In general, it is their view that we should strive to refocus on groups given the recognition that health is impacted by individual-level characteristics (including behaviours) as well as the physical, biological, and social environments. If the individual is really the level of interest, then multilevel models allow us to include core information from higher levels of organisation, and investigate contextual effects. However, there is also a need to focus inferences on groups *per se* (Diez-Roux, 1998a; 1998b; McMichael, 1999). Unfortunately (for us), the terminology can be very confusing. We will continue to refer to these as group-level studies, but we must recognise that others call these 'ecological perspective', 'ecological model(s)', and multilevel model(s)—these are often used interchangeably (Richard *et al*, 2011). In thinking about studying groups and whether we should be making inferences to groups or individuals, Rose (2001) stated that it is helpful to distinguish between 2 questions.

1. What is the etiology of a case?
2. What is the etiology of incidence?

Both questions emphasise that there is more than one cause of a given disease or condition. However, the first question about causes of cases requires that we conduct our study at the individual level. With individual people as our principal or only level of interest, we identify causes of disease in individuals. In this context, within a defined population (group), the use of the ratio measures of association to identify potential causes, and measure their strength, assumes a heterogeneity of exposure within the study population. In the extreme, if every study subject is exposed to a necessary cause, then the distribution of cases (in individuals) would be wholly determined by individual susceptibility determined by the other components of the sufficient causes (for example, a genetic component, not the widespread (albeit essential) exposure). In general, Rose notes that the more widespread or prevalent a risk factor is, the less it explains the distribution of cases within that population. Hence, we might even conclude that a prevalent necessary cause was of little causal importance—it might even be considered normal background exposure.

In addition to this inferential problem, when we focus on individuals, we often treat group-level factors as nuisance variables, whether through using a fixed-effect or a random-effect modelling approach. In this context, we do not try to understand, or explain, the group-to-group variation, just to deal with it. As was discussed in Chapter 20, in choosing the appropriate aggregation level to study, it is useful to examine the proportion of variance that can be attributed to the individual and to the group because this is a useful guide for focusing future investigations (see Section 20.2.2). Even if our focus is on individuals, it is also useful to investigate if the effect of an exposure factor on individuals depends mainly on that individual-level risk factor, or others at the group level (the contextual effects). Herd immunity is one example where we know this to be a real biological phenomenon; the prevalence of disease in a group might have a similar important effect on the nature of the disease (*eg* timing and/or dosage of first exposure) in individuals.

To address the question about causes of incidence in populations, we must investigate the determinants of group or population means (*eg* why is the disease more common in group 'A' than in group 'B'?). To do so, we need to study the characteristics of groups to identify factors

that act by shifting the distribution of disease of the entire group. For their success, group-level studies require either a large variance of exposure levels across groups, a large study size (*ie* number of groups), or a combination of the two. Obtaining a sufficient number of groups (*eg* census areas or counties) to give a study reasonable power has often been a practical limitation of group-level studies. Nonetheless, especially in many public-health activities, we have a particular need to know the determinants of incidence, be they families, schools, or geographic areas, in order to help prevent disease in the population.

29.9.1 The group as the aggregate-scale of interest

Virtually all epidemiologists are aware of the hierarchical organisation of the populations we study. These levels of organisation range from subcellular units, to cells, organs, body systems, individuals, aggregates of individuals such as households, families, neighbourhoods, census tracts, states, nations *etc*. The key point is that each higher level of organisation subsumes all the properties of lower levels, but has additional unique properties of its own (Diez-Roux, 1998a; 1998b; Krieger, 1994; Susser, 1973). From this principle, it would seem crucial that risk-factor identification is conducted in the light of the appropriate population level context, but with an awareness of risk factors at other levels of organisation.

Moving beyond the primarily biologic individual-based explanations of disease causation does not imply denying biology, but rather involves viewing biologic phenomenon within their global and social-environmental contexts. For example, Richard *et al* (2011) discuss the ecologic (group level) approaches that have been used over the last two decades to increase physical activity and to encourage the addition of more fruit and vegetables to our diet. Lee & Cubbin (2009) discuss the impact of social environment in the context of ecologic approaches on the effectiveness of our programs designed to prevent obesity. When focusing on issues that have a large impact on our health, it is crucial to accept that many of the factors that impact on these behaviours occur at levels beyond the individual. Nonetheless, these factors might alter the behaviour of individuals both directly and indirectly. Preventing the negative health-impacts caused by a single exposure (*eg* low levels of physical activity), would 'prevent' a large number of specific diseases. Further, focusing on the single exposure often is a more efficient and effective approach to improving health than focusing on the numerous specific outcomes they may cause (*ie* diseases). Similar arguments are advanced for continuing to emphasise the health benefits of a non-smoking society, and that these programs need to focus on group-level as well as, or more so than, individual-level factors. Richard *et al* (2011) note that unfortunately, "in-depth analyses of various sets of programs have empirically confirmed that ... individual and interpersonal determinants continue to be the favourite or preferred levels targeted by planners and practitioners". Further, "although it would be logical to integrate community-based participatory research and planning principles ... few ecological models have explicitly made this conceptual link". Interestingly, based largely on ecologic studies which indicate that the prevalence of obesity is negatively correlated with physician numbers on the county level within the USA, Gaglioti *et al* (2009) argue for continued support of family physicians (of course Gaglioti *et al* might be correct, but the reader should recall the issue of ecologic fallacy, as explained by Robinson in 1950, (see Firebaugh, 2009) when making inferences such as these.

As well as the need to conduct research at the population level to help resolve many endemic diseases, collective experience has been that disease control programmes for contagious or exotic diseases need to be directed more at the population than at the individual level. Despite

our most advanced tests for identifying infected individuals, at the end stages of many national-level infectious disease control programmes, the optimal strategy for disease control is almost always to focus prevention and control on groups, not on individuals.

29.9.2 The group as the level of inference

The desired level of inference links to the level of analysis. In some studies the intent is to identify causal factors of cases by investigating individual-level risk factors, whereas in others it might be to make inferences about causal factors of incidence by focusing on the group level. However, as noted in earlier sections, if one is trying to make inferences about one level (a lower level) from data collected at a higher level, then such cross-level inferences are open to considerable bias. If we are interested in the interaction between person-level and group-level variables, then that aspect can be studied using analyses aimed at individuals but with an appropriate group-level variable (contextual effect (Section 21.4)—*eg* prevalence of exposure) included to allow the interaction to be identified.

Previously, we examined some of the features that can help us avoid the ecologic fallacy when making inferences about the effect of an exposure on individuals when we use group-level or ecologic studies. In that context, correct meant that the group-level findings were consistent with the findings at the individual level. However, despite our discussion on this point, given the pervasiveness of reductionism in biomedical science, it is likely that the **atomistic fallacy** (using data from lower levels to make inferences about higher levels) is undoubtedly the more common of the 2 errors. We certainly risk making this error if our explanations of disease in populations are based primarily on what we know about disease in individuals. However, little is written about this fallacy. The difference in our assessments of these errors likely reflects the prevailing scientific view about what constitutes valid causal inferences. It seems that ecologic fallacies are viewed as serious problems because the associations, while true at the aggregate level, are not true at the individual level; whereas in the atomistic fallacy, the facts at the cellular or individual level are deemed to be correct, regardless of how correct, or useful (or useless) that knowledge is for efficient and effective disease prevention in populations.

In addition to the atomistic fallacy, a long-held axiom is that if one is interested in populations, one must study populations (McMichael, 1999). This axiom arises in part because the physical, chemical, biological, and sociological/managerial properties at the higher level likely differ from those at the lower level, and in part because there are a host of sociological/managerial factors and some biological factors which operate principally at the group level. A simple physical-chemical example is that the properties of oxygen and hydrogen tell us very little about the properties of water. Also as Schwartz (1994) observes, we should not confuse characteristics of a group with that of its individuals, “a hung jury might be indecisive but its members might be anything but indecisive”.

In our research endeavours, we should not look at group-level studies as only crude attempts to uncover individual-level relationships. Many criticisms of ecologic studies are based on the questionable assumption that the individual level of analysis is the most appropriate (Schwartz, 1994). In fact, the health status of an individual, is itself an aggregated measure, because it is body cells/systems, not individuals that become diseased. The threshold for disease being present in an individual usually is based on a set of criteria, some quantitative, some qualitative. Most often, as epidemiologists, we define the cutpoint(s) for ‘having the disease’ and then ignore the tremendous variance in severity and effects of that disease in most of our studies

(because these are not our primary interest). In a similar vein, we need to study disease at the group level, where a family or group (say a village) might be categorised as diseased or not and we might ignore the actual proportion of people with disease. However, in other studies the dichotomisation of disease presence or absence (or presence beyond a specified cutpoint of frequency) might be too crude an approach because we are forced to discard valuable information about the extent or severity of disease at the group level. In this situation, it might be preferable to retain the level of disease (or outcome) as a quantitative statement about disease frequency, even though there is no intent on making inferences below the group level.

In order to optimally interpret some of our group-level studies, a major issue is to differentiate the causal inferences we make about associations at the group level from inferences we might make relative to the effect of that same (or apparently similar) variable at the individual level (Diez-Roux, 1998a; 1998b; Schwartz, 1994). For example, if variable X_1 at the individual level indicates seroconversion to a specific agent, then $X_2=(\Sigma X_1/n)$ at the group level inherently carries more information than just the proportion that seroconverted; by its nature a group with a low level of X_2 likely has different dynamics of infection than one with a high level of X_2. For example, as noted, the frequency of exposure in the group could influence the timing of initial exposure to an agent by individuals, and this is often an important factor in the type of clinical syndrome that might result.

In conclusion, it is clear that there are numerous problems in using aggregated data to make inferences about events in individuals. Multilevel analyses (Chapters 20–22) allow us to include important factors from higher levels of organisation when studying individuals, including contextual effects. However, appropriately designed studies that focus on groups are necessary to identify factors of importance in the distribution of health and disease in populations.

References

Blanchard KS, Palmer RF, Stein Z. The value of ecologic studies: mercury concentration in ambient air and the risk of autism. Rev Environm Health. 2011;26(2):111-8.

Brenner H, Savitz DA, Jockel KH, Greenland S. Effects of nondifferential exposure misclassification in ecologic studies. Am J Epidemiol. 1992;135(1):85-95.

Colli JL, Kolettis PN. Bladder cancer incidence and mortality rates compared to ecologic factors among states in America. Int J urology and nephrology. 2010;42(3):659-65.

Diez-Roux AV. On genes, individuals, society, and epidemiology. Am J Epidemiol. 1998a;148(11):1027-32.

Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. Am J Public Health. 1998b;88(2):216-22.

Dufault B, Klar N. The quality of modern cross-sectional ecologic studies: a bibliometric review. Am J Epidemiol. 2011;174(10):1101-7.

Firebaugh G. Commentary: 'Is the social world flat? W.S. Robinson and the ecologic fallacy'. Int J Epidemiol. 2009;38(2):368-70; author reply 70-3.

Gaglioti A, Petterson SM, Bazemore AW, Phillips RL, Jr., Dodoo MS, Zhang X. Primary care's ecologic impact on obesity. Am Family Physician. 2009;79(6):446.

Glynn A, Wakefield J. Ecological Inference in the Social Sciences. Statist method. 2010;7(3):307-22.

Green C, Elliott L, Beaudoin C, Bernstein CN. A population-based ecologic study of inflammatory bowel disease: searching for etiologic clues. Am J Epidemiol. 2006;164(7):615-23; discussion 24-8.

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

Greenland S. Divergent biases in ecologic and individual-level studies. Stat Med. 1992;11(9):1209-23.

Greenland S, Robins J. Invited commentary: ecologic studies--biases, misconceptions, and counterexamples. Am J Epidemiol. 1994;139(8):747-60.

Han YY, Weissfeld JL, Davis DL, Talbott EO. Arsenic levels in ground water and cancer incidence in Idaho: an ecologic study. Int Arch Occup Environ Health. 2009;82(7):843-9.

Haneuse SJ, Wakefield JC. The Combination of Ecological and Case-Control Data. J R Stat Soc B. 2008;70(1):73-93.

Jackson ML. Confounding by season in ecologic studies of seasonal exposures and outcomes: examples from estimates of mortality due to influenza. Annals Epidemiol. 2009;19(10):681-91.

Krieger N. Epidemiology and the web of causation: has anyone seen the spider? Social science & medicine. 1994;39(7):887-903.

Lee RE, Cubbin C. Striding toward social justice: the ecologic milieu of physical activity. Exercise and sport sciences reviews. 2009;37(1):10-7.

Levin KA. Study design VI - Ecological studies. Evidence-based Dent. 2006;7(4):108.

McMichael AJ. Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am J Epidemiol. 1999;149(10):887-97.

Newman SC, Stuart H. An ecologic study of parasuicide in Edmonton and Calgary. Can J Psychiatry. 2005;50(5):275-80.

Oakes JM. Commentary: Individual, ecological and multilevel fallacies. Int J Epidemiol. 2009;38(2):361-8; author reply 70-3.

Ojha RP, Offutt-Powell TN, Evans EL, Singh KP. Correlation coefficients in ecologic studies of environment and cancer. Archives Environ & Occup Health. 2011;66(4):241-4.

Osmond C, Gardner MJ. Age, period, and cohort models. Non-overlapping cohorts don't resolve the identification problem. Am J Epidemiol. 1989;129(1):31-5.

Richard L, Gauvin L, Raine K. Ecological models revisited: their uses and evolution in health promotion over two decades. Annual Rev Public Health. 2011;32:307-26.

Robertson C, Gandini S, Boyle P. Age-period-cohort models: a comparative study of available methodologies. J Clin Epidemiol. 1999;52(6):569-83.

Robinson WS. Ecological correlations and the behavior of individuals. Int J Epidemiol. 2009;38(2):337-41.

Rose G. Sick individuals and sick populations. Int J Epidemiol. 2001;30(3):427-32; discussion 33-4.

Rothman KG, Greenland S, Lash T. Modern Epidemiology. Philadelphia: Lippincott; 2008.

Savitz DA. Commentary: A niche for ecologic studies in environmental epidemiology. Epidemiol. 2012;23(1):53-4.

Schwartz S. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. Am J Public Health. 1994;84(5):819-24.

Subramanian SV, Jones K, Kaddour A, Krieger N. Revisiting Robinson: the perils of individualistic and ecologic fallacy. Int J Epidemiol. 2009;38(2):342-60; author reply 70-3.

Susser M. Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology. Anonymous B, editor: Oxford University Press, Toronto; 1973.

Wakefield J. Ecological inference for 2 X 2 tables. J R Stat Soc A. 2004;167:385-445.

Wakefield J, Haneuse SJ. Overcoming ecologic bias using the two-phase study design. Am J Epidemiol. 2008;167(8):908-16.

Wakefield J. Multi-level modelling, the ecologic fallacy, and hybrid study designs. Int J Epidemiol. 2009;38(2):330-6.

Webster T. Commentary: does the spectre of ecologic bias haunt epidemiology? Int J Epidemiol. 2002;31(1):161-2.