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

29.1 INTRODUCTION

The initial part of this chapter deals with studies in which groups of subjects are sampled, and analyses are conducted at the group level when the researcher wishes to make inferences to individuals. These are called **ecologic studies**. The primary analytic feature of an ecologic study is that we do not know the joint distribution of the risk factor(s) and the disease within each group. In other words, although we know the proportion exposed and the risk or rate of cases within each group, we do not know the proportion of exposed cases, typically because we lack individual-level data on the risk factor, the disease, or both (Rothman *et al*, 2008).

For example, in an ecologic study of the role of selected micro-organisms as potential causes of respiratory disease (BRD) in pens of feedlot cattle, we would know the pen-level incidence of BRD and the pen-level frequency of infection with each organism; however, we would not know the joint distribution of BRD and each organism. The lack of this piece of information can lead to inferential problems. Thus, given a positive association between infection with a micro-organism and higher rates of BRD, it is possible that the animals developing BRD are those that are not infected with the organism in question.

Ecologic studies might be called **exploratory** if there is no direct 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* O157 in humans) by administrative area on a map, we might use previous knowledge of local features (*eg* cattle density) to explain the observed spatial variation in rates of disease, even though there was no direct measurement of this factor in the study. 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 groups are small, the analysis should account for the different precision of disease rates by group. Spatial analysis might require adjustment for spatial correlation. Temporal studies might need to adjust for a lag period and inferences might need to take account of changes in diagnostic standards. Studies that include an extended period of time might have to account for, and try to separate, the age, period, and cohort effects on the outcome. This leads to an identifiability problem as these 3 components are interlinked and cannot be assessed independently (Osmond and Gardner, 1989; Robertson *et al*, 1999 for a discussion). Studies that combine both among-group and temporal approaches might provide a more thorough test of the hypothesis than either approach alone. We begin our discussion by asking ourselves why we might study groups, especially if we want to make inferences to individuals?

29.2 RATIONALE FOR GROUP LEVEL STUDIES

Particularly in veterinary medicine, the group (*eg* the herd) is often the sampling unit as well as the unit of interest; these are not ecologic studies (Carver et al, 2002). The aggregate level, for

example, litters of animals, hives of bees, sea-pens of fish, flocks/barns of poultry, mobs of sheep *etc* is often of more interest than the elements or components (*ie* individual piglets, bees, chickens, fish, sheep *etc*) of the group. The recent increase in the use of spatial statistics often focuses on even larger aggregates 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 problems**. See Section 29.7 for further discussion of non-ecologic group-level studies. It is often the intent, however, to make inferences about individuals based on the results from the group-level analysis, and in doing this, one must be very careful (reasons for this are discussed subsequently). Nonetheless, the major 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, dietary intake) so an average for the group might suffice. In other circumstances, the variation in diet within an individual might be large, whereas the group average might adequately reflect exposure to specific nutrients for the purposes of the study.

Exposure homogeneity If there is little variation in exposure among individuals within a group, it might be difficult to assess the exposure's impact on them. For example, if all animals within a group are managed in the same way, one might need to study groups to observe the apparent effect of different management schemes. 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 rations, types of housing, and treatments (*eg* water or feed-based antimicrobials) can only be delivered, or implemented practically, at the group level. Hence, farms or groups are of interest.

Simplicity of analysis Often it appears to be easier to display and present group-level rather than individual-level data. However, group-level analyses might hide serious methodological problems if 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,b and McMichael, 1999, for a discussion). For our purposes, we will use 3 categories: aggregate, environmental and global variables.

29.3.1 Aggregate

Aggregate variables are summaries of measurements made on individuals within the group such as: the proportion exposed, the average age, average nutrient intakes *etc*. They can relate to the predictor variables, the outcome variable, or both. When a disease is the outcome, it is usually measured using rates because most groups are open; if closed, then a risk-based approach can be used. This type of variable is also called a **derived variable**. The type of derived variable used in ecologic studies is that which is formed, at least in part, by aggregating individual observations to form a summary variable (usually the mean) for the group (*eg* proportion exposed, feed conversion ratio, average daily gain, average somatic cell count, disease rate, mortality rate *etc*).

29.3.2 Environmental or contextual

Usually these are physical characteristics of the group such as local weather, level of pollutants in the area, or herd characteristics such as bulk-tank somatic cell count, characteristics of water supply (*eg* deep well versus surface water), and management strategy (*eg* teat-dipping strategy or colostrum-feeding protocol). The key feature of these variables is that they have an analogue at the individual level (*eg* the colostrum-feeding protocol might state that every calf gets a litre of colostrum within 4 hours of birth; whereas the individual-level factor would indicate whether this particular calf received that amount of colostrum within that time period). 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, a farmer might say that all calves get adequate colostrum, but in fact, only a small proportion actually receives it in the appropriate time or manner so serious misclassification results. In addition, it might well be that there is an interaction between the factor at the individual level (*eg* titre to agent *X*) and the contextual variable for the same factor (*eg* percentage of animals with a protective titre), as in herd immunity and these need to be identified for proper inference. (**Note** Contextual variables described in Section 21.4 are usually an aggregate variable because they are derived from the individual level data.)

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). Global variables include farmer characteristics, and herd characteristics or management strategies such as herd size, open versus closed herd policy, density of housing, reproductive strategies, and some disease prevention programmes.

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 factors might be dichotomous 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) and the inferences are 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 grouped outcome variable on the grouped exposure variables. Some prefer to use a Poisson model (see Rothman *et al* (2008), pp 517-518 for other examples of analytic approaches. Ducrot *et al* (1996) also discuss these in the context of veterinary medicine).

As an example of the linear model approach, we can imagine the continuous outcome Y representing the risk or rate of disease (*eg* 0.15 per animal-year in herd j) modelled as a linear function of the exposure (*eg* 0.3 of the calves in herd j do not receive early adequate colostrum) and perhaps adjusting for the effects of one or more confounders (*eg* the average age of calves in each herd). The model could be specified as:

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

where X_{1j} is the proportion receiving adequate colostrum and X_{2j} is the average age in herd 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 analyses, 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 differences in the number of study subjects). The outcome often should 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 not produce unbiased estimates at the group level (Rothman *et al*, 2008, p 468).

Associations between predictors and dichotomous outcomes at the individual level are usually based on ratio measures. However, a problem with using ratio measures at the group level in linear models is that, for aggregate variables, these estimates often force us to extrapolate our inferences to groups 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:

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

Hence, valid inferences about ratio measures requires totally exposed and non-exposed groups.

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 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 Xjs) 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 in the model and the mean values of the other X variables.

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, sex, and breed are included in the *SMR*. However, this approach does not prevent confounding unless the explanatory variables are also standardised in the same manner, and usually sufficient data to achieve this 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 and the individual level and test its significance.

29.5 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 increases (or decreases) the risk of disease by, for example, 3 times, does not mean that this is true at the individual level. Indeed, it might not mean that the exposed subjects are the ones having the highest individual risk of becoming cases. This error in inference is termed the **ecologic fallacy** (see section 29.7.2 for atomistic fallacy). In addition, even without the ecologic fallacy, the group-level bias almost always exaggerates the magnitude of the true association away from the null, but occasionally it reverses the direction of the association. As a simple, hypothetical example, assume that you are investigating a disease which is only caused by an infectious agent X (which produces lifelong antibody titres) and clinical disease 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.

29.6 Sources of ecologic bias

29.6.1 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.

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. 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 29.1. 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).

29.6.2 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). Example 29.2 explains this phenomenon.

29.6.3 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 across groups, or due to effect modification by a group-level factor (Example 29.3).

29.6.4 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.

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.

It is clear we need to be careful when making inferences about individuals based on 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,

Example 29.1 Effect of individual-level exposure misclassification on group-level results

We begin with the correctly classified study population structures in 2 farms (j=1,2).

	Farm 1			Farm 2		
Correctly classified	**Exposed**	**Non-exposed**	**Totals**	**Exposed**	**Non-exposed**	**Totals**
Number of cases	50	40	90	100	30	130
Animal-time (t_j)	200	800	1000	400	600	1000
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 one would use for the analysis at the group level if there was no misclassification. Note that in Farm 1, 20% of the animal-time is exposed (200/1000), while in Farm 2, this is 40% (400/1000). At the individual level, the IR=5 and the ID=0.20. 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

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

Based on an exposure sensitivity of 0.8 and an exposure specificity of 0.9, and using the general approach shown in section 12.6, we would observe the data below.

	Farm 1			Farm 2		
Incorrectly classified	**Exposed**	**Non-exposed**	**Overall rate**	**Exposed**	**Non-exposed**	**Overall rate**
Number of cases	44	46	90	83	47	130
Animal-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 farms) the IR=3.04 and the ID=0.137. Here, the exposure misclassification leads to biased estimates of the proportion of animal-time exposed on each farm; the difference between these 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 at the group level.

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. 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;

(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).

Example 29.2 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 farms), both measured on the continuous scale (data in **bold typeface** in table). Consider these data from 3 farms:

Farm 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
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**	
Farm 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**	
Farm 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 29.2 (*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 farms so there is no interaction at the individual level. Also, there is no confounding by E_1 or E_2 within farms (as E_1 and E_2 are independent). However, because the prevalence of E_2 varies by farm, this results in an association of farm 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 farm 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.

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

29.7 NON-ECOLOGIC GROUP-LEVEL STUDIES

A number of epidemiologists have noted that our discipline 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. If the individual is really the level of interest, then multilevel models (Chapters 21-23) allow us to include core information from higher levels of organisation, and investigate any contextual effects. However, there is also a need to focus inferences on groups *per se* (McMichael, 1995, 1999; Diez-Roux, 1998a,b).

In thinking about studying groups and whether we should be making inferences to groups or individuals, Rose (1985) 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. The first question about causes of cases requires that we conduct our study at the individual level. With individual animals 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 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

Example 29.3 Effect modification by group

Consider the following data from 3 farms:

	Farm A		Farm B		Farm C		Total	
	E+	**E-**	**E+**	**E-**	**E+**	**E-**	**E+**	**E-**
Cases	120	30	120	36	120	42	360	108
Animal-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 farm. Thus, some farm-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 any group (*ie* farm level) factor at the individual level because p(*D*+|*E*-)=0.03 in all 3 farms. Thus, farm *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 farm level is making the impact of exposure (whether measured by *IR* or *ID*) to vary, across farms, and this effect increases as the prevalence of *E*+ decreases.

An ecologic analysis at the farm 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.

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 any group-level factors that are present as nuisance variables, whether through using a fixed-effect or a random-effect modelling approach. In this context, we have not tried to 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. Even if our focus is on individuals, it is also useful to investigate if the effect of

an exposure factor on individuals depends on that, or other factors, 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 causally 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* herds) to give a study reasonable power has often been a practical limitation of group-level studies. Nonetheless, in both herd-health management, and veterinary public-health activities, we have a particular need to know the determinants of incidence, be they groups, herds or geographic areas, in order to help prevent disease in the population.

29.7.1 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 (households of people, families, litter mates, pens and herds of non-human animals), neighbourhoods, 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 (Susser, 1973; Krieger, 1994; Diez-Roux, 1998a; Ducrot *et al*, 1996). From this, 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 environmental contexts.

A natural level of aggregation as the unit of interest for veterinarians is the farm (or kennel) as veterinary clinicians are often required to be responsible for the healthcare of all animals within that farm. The reason(s) we emphasise aggregates of animals as the unit of concern could, in large part, reflect the relative economic value of the individual; the single fish in a sea pen, the broiler chicken in a poultry house, or a single sheep in a mob is of little economic importance to the group, and therefore to its owner. The same is true to a decreasing extent of individual pigs and beef cattle. Individual dairy cattle are of more relative economic value and perhaps because of this, the majority of epidemiologic studies in dairy cattle have tended to focus on individuals. Studies of health problems in horses and companion animals are usually focused at the individual level, and a logical level on which to aggregate them for population approaches is not easily apparent. However, an obvious need when considering population control in pets is to move beyond the simple individual-animal-oriented approach of spaying the pet or constraining contact, to examining the social and biological contexts of domestic and feral pets. Similarly, in vaccination programmes, if we are principally vaccinating (or prophylactically medicating) the low-risk group, we will have little impact on the disease in the population, even when a significant proportion of the population is vaccinated.

The previous ideas relating to focusing on levels beyond the individual would suggest that when researching, for example, food safety issues, while it might be necessary to include

features of individual micro-organisms such as *E. coli* O157, and/or factors which influence its survival at the individual/farm/flock level, one must also understand the operation of modern farms and modern meat-processing plants, as well as the impact of the industry structure, and the centralisation of food processing that has been under way recently in the food industry. The same comments apply to researching large-scale disease outbreaks in the food-animal industries such as BSE in cattle; regardless of its origin, one cannot deny that the spread of this disease was aided and abetted by the structure of the animal feed-stuff industries. Wing, 1998, as an example, has commented on the need to work at the large scale in resolving many of our current important problems, especially those relating to farming and the environment.

In addition to the need to conduct research at the population level to help resolve 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 control on groups not individuals.

29.7.2 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 animal-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 disease) 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 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, 1995). 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 herd might be categorised as diseased or not and we might ignore the proportion of animals with disease (*eg* if one is attempting to establish disease-free groups, then this approach is workable). However, in other studies the dichotomisation of disease presence or absence (or presence beyond a specified cutpoint) might be too crude an approach because one is forced to discard valuable information about the extent or severity of disease at the herd 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 (Schwartz, 1994; Diez-Roux, 1998a). For example, if variable X_1 at the individual level indicates seroconversion to a specific agent, then $X_2=(\sum X_1/n)$ 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, it could influence the timing of initial exposure to an agent, and this is often an important factor in the type of 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 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 needed to identify factors of importance in the distribution of health and disease in populations.

REFERENCES

Brenner H, Greenland S, Savitz DA. The effects of non-differential confounder misclassification in ecologic studies. Epidemiology 1992; 3: 456-459.

Carver DK, Fetrow J, Gerig T, Krueger T, Barnes HJ. Hatchery and transportation factors associated with early poult mortality in commercial turkey flocks. Poult Sci 2002; 81: 1818-1825.

Diez-Roux AV. Bringing context back into epidemiology: Variables and fallacies in multilevel analyses. Am J Pub Hlth. 1998a; 88: 216-222.

Diez-Roux AV. On genes, individuals, society and epidemiology. Am J Epidemiol. 1998b; 148: 1027-1032.

Ducrot C, Legay J, Grohn Y, Envoldsen C, Calavas D. Approach to complexity in veterinary epidemiology; example of cattle reproduction. Natures-Sciences-Societies. 1996; 4: 23-33.

Greenland S. Divergent biases in ecologic and individual-level studies. Stat Med 1992; 11: 1209-1223.

Greenland S, Morgenstern H. Ecological bias, confounding and effect modification. Int J Epidemiol. 1989; 18: 269-274.

Greenland S, Robins J. Ecologic studies: Biases, misconceptions, and counter examples. Am J Epidemiol. 1994; 139: 747-760

Krieger N. Epidemiology and the causal web: Has anyone seen the spider? Soc Sci Med. 1994; 39: 887-903.

McMichael AJ. The health of persons, populations, and planets: epidemiology comes full circle. Epidemiology. 1995; 6: 633-636.

McMichael AJ. Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am J Epidemiol. 1999; 149: 887-897.

Morgenstern, H. Ecologic studies in Rothman KJ and Greenland S. Modern epidemiology, 2 ed. Lippincott-Raven; Philadelphia. 1998.

Osmond C, Gardner MJ. Age, period, and cohort models. Non-overlapping cohorts don't resolve the identification problem. Am J Epidemiol. 1989; 129: 31-35.

Robertson C, Gandini S, Boyle P. Age-period-cohort models: a comparative study of available methodologies. J Clin Epidemiol. 1999; 52: 569-583.

Rose G. Sick individuals and sick populations. P.A.H.O. Epidemiological Bulletin. 6: 1-8, 1985.

Rothman KJ, Greenland S, Lash T. Modern Epidemiology, 3rd ed. Lippincott-Raven; Philadelphia. 2008.

Schwartz, S. The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences. Am Jour Pub Hlth. 1994; 84: 819-824.

Susser M. Causal Thinking in the health sciences: concepts and strategies of epidemiology.

Oxford University Press; Toronto. 1973.

Webster T. Commentary: Does the spectre of ecologic bias haunt epidemiology? Int J Epidemiol. 2002; 31: 161-162.

Wing S. Whose epidemiology, whose health. Int Jour Hlth Serv. 1998; 28: 241-252.