6.2 Ecological variables, ecological studies, and multilevel studies in public health research

Ana V. Diez Roux, Sharon Schwartz, and Ezra Susser

There has been much discussion in epidemiology about the utility of ecological studies in investigating the causes of disease. Epidemiology textbooks commonly appraise this study design as a crude attempt to ascertain individual-level relationships. It is argued that ecological studies should be limited to 'hypothesis generation', leaving the more esteemed process of 'hypothesis testing' to individual-level data. The limitations of ecological studies are generally attributed to the ecological fallacy, that is the well-established logical fallacy inherent in making inferences regarding individual-level associations based on group-level data (Morgenstern 1982, 1995; Piantadosi *et al.* 1988; Greenland 1992). However, in recent years, interest in potential ecological or group-level determinants of health has led epidemiologists to reconsider the utility and limitations of ecological studies (Susser 1994*a*,*b*). In this context, there has been renewed interest in the notion that group-level (or ecological) variables may provide information which is not always captured by individual-level data (Schwartz 1994; Susser 1994*a*,*b*; Koopman and Lynch 1999), and in rethinking the ways in which these group-level constructs can be examined in epidemiological analyses. More broadly, there has been growing recognition of the need to consider multiple levels of organization (for example, from molecules to society) in studying the determinants of health and disease.

Many of the conceptual and analytical issues that arise when considering the uses of ecological studies and ecological variables derive from the presence of multiple levels of organization and nested data structures more generally. For example, many problems that arise when dealing with individuals nested within groups (for example, people nested within geographical areas), are also present when dealing with groups nested within larger groups (for example, states nested within countries), people nested within families, or multiple measurements on individuals over time (in this case the 'group' is the individual, and the 'individuals' are the measurement occasions). In fact the need to deal with multiple levels of organization is the norm rather than the exception in epidemiology.

The presence of multiple levels has several implications. Firstly, the units of analysis (or observations for which independent and dependent variables are measured) can be defined at different levels. The unit of analysis determines the level at which variability is examined. For example, a study with individuals as the units of analysis (that is, where each observation is an individual) can investigate the causes of interindividual variation in the outcome. A study with groups as the units of analysis (where observations are groups), can investigate the causes of intergroup variation in the outcome. A study involving repeated measures on individuals over time in which measures at different points in time are the units of analysis can investigate the causes of variability across measures. As will be shown, the use of units of analysis at one level to make inferences about the causes of variability at a different level leads to a series of methodological problems.

The second implication of the presence of multiple levels is that both independent and dependent variables can be conceptualized (and measured) at different levels. Constructs pertaining to a higher level may be important in understanding variability at a lower level, and, conversely, constructs defined at a lower level may be important in understanding variability at a higher level. For example, characteristics of the groups to which individuals belong may be important in explaining interindividual variability, and characteristics of individuals comprising the groups may be important in explaining intergroup variability. Analogously, when looking at multiple measures on individuals over time, individual characteristics may be important in understanding variability across measures, and factors specific to measurement occasions may be important in understanding interindividual variability.

This chapter discusses the use of ecological variables, ecological studies, and individual-level and multilevel studies in epidemiology within the broader context of the implications of multiple levels of organization for understanding disease aetiology. As discussed in the sections to follow, the particular research question investigated should guide decisions regarding the most appropriate unit of analysis (and level at which variability in the outcome is examined) as well as the relevant constructs to be investigated as predictors (and the levels at which they are defined and measured). Although the discussion will focus on the simple case of individuals nested within groups, it is generalizable to many other situations involving nested data structures, as noted above.

We begin by reviewing the classical distinction between ecological and individual-level studies made in epidemiology. Next, we review the sources of the 'ecological fallacy', placing this fallacy within the context of other fallacies which may arise when the presence of multiple levels is ignored. Finally, we review the full range of study designs available to investigators based on (a) the units of analysis (and the level at which variability is examined) and (b) the levels at which relevant constructs are defined and measured.

Ecological studies and studies of individuals

Epidemiology has traditionally distinguished two types of studies based on the units of analysis: ecological studies and studies of individuals. Ecological studies are studies in which groups are the units of analysis: both the dependent and the independent variables are measured for groups, and intergroup variability (and associations between independent and dependent variables across groups), are examined. For example, common ecological studies in public health involve measuring rates of disease for different geographic areas, and relating these rates to area social or physical characteristics (for example, measures of area median income, levels of air pollution, water hardness, radiation). Ecological studies are often cross-sectional with independent and dependent variables measured at the same point in time. However, ecological studies can also involve repeated measures on a group, or on several groups, over time, as in time-trend studies (Susser 1994*b*; Morgenstern 1995). For example, an ecological study could examine yearly incidence rates of disease for different regions over a 10-year period and investigate the relation of these incidence rates to area characteristics that do and do not change over time. Although not common in the epidemiological literature, ecological studies could also use a case–control approach, in which a sample of ecological cases (groups with a certain outcome, for example a disease rate above a certain level) are compared with a sample of ecological controls. Ecological studies could also involve the analysis of groups randomized to receive or not receive an intervention, as in randomized trials. In many ecological studies the predominant analytic approach involves the estimation of correlation coefficients between the group-level exposure and the group-level outcome. However, many other analytic approaches are also possible, including the estimation of other measures of association (such as rate differences or rate ratios) using linear or log-linear models (Morgenstern 1995).

In contrast, in individual-level studies (the studies commonly referred to under headings of 'study design' in epidemiological textbooks) the units of analysis are individuals: both independent and dependent variables are measured for individuals and interindividual variation (and associations between independent and dependent variables across individuals) are investigated. Based on their design, individual-level studies can be cross-sectional (when both independent and dependent variables are measured at the same point in time), cohort (when individuals free of the outcome at baseline are followed over time to compare risk of the outcome in exposed and unexposed), or case–control (when a sample of people with the outcome is compared with a sample of controls with regard to the presence of certain exposures).

The information these two study designs provide (and the information they lack) differs because the units of analysis are not the same. Ecological studies include information on group characteristics (which may sometimes summarize the characteristics of individuals in the group), but lack information on the cross-classification of individual-level characteristics within groups. For example, an ecological study may relate the percentage of smokers in different groups to mortality rates, but have no information on whether, within groups, smokers were actually the ones more likely to die. Conversely, individual-level studies focus on interindividual variation, and have information on individual-level characteristics but often lack information on characteristics of the groups to which individuals belong.

Uses of ecological studies in public health

Descriptive ecological studies in which rates of disease or death are compared over time or across geographic areas have been a staple of epidemiology for centuries (Susser 1973). Chadwick used an ecological approach in his famous *Report on the sanitary condition of the labouring population of Great Britain* in 1842 (Chadwick 1965). Table 1 shows a portion of Chadwick's Report, in which a drained area is compared with an undrained area at three parallel points in time. By comparing mortality rates over time in both communities (one drained and the other undrained), he was able to draw inferences regarding the relationship between drainage and ill health. Drainage was the 'exposure', mortality the outcome, and communities the units of analysis. With these findings Chadwick had grounds to institute a system of sanitation nationwide. Although the miasmatic theory of disease causation which Chadwick espoused (and which he believed was supported by his data) was later shown to be mistaken, the method of sanitation that Chadwick introduced was probably as important as any other single measure in modern times (Susser 1973).

Early in his studies, Snow also used an ecological approach in comparing cholera rates for London districts, and examining whether differences in these rates were related to differences in the sources of water (Susser 1973) (Table 2). The units of analyses were districts, and both the independent variable (source of water) and the dependent variable (cholera rates) were conceptualized and measured at the district level. At the beginning of this century, Goldberger *et al.* (1920) also used an ecological approach to demonstrate that pellagra was not 'an intestinal infection transmitted in much the same way as typhoid fever', as was argued by many at the time. These findings laid the groundwork for much of Goldberger and Sydenstricker's later work relating diet and economic conditions to pellagra in the southern United States (Terris 1964).

More recently, in the chronic disease era, ecological studies relating rates of cardiovascular disease across countries to risk factor prevalences (Keys' Seven Country Study) laid the foundation for future work on the epidemiology and causes of cardiovascular disease (Keys 1980). For example, data from the Seven Country Study showed a

Table 1 Death rates compared in three successive decades in a drained and an undrained town

The following was the proportion of deaths to the population in the two towns	Beccles	Bungay
Between 1811 and 1821	1 in 67	1 in 69
Between 1821 and 1831	1 in 72	1 in 67
Between 1831 and 1841	1 in 71	1 in 59

Chadwick says in conclusion 'You will therefore see that the rate of mortality has gradually diminished in Beccles since it has been drained, whilst Bungay, notwithstanding its larger proportion of rural population, has considerably increased'

Source: Chadwick (1965) (originally published in 1842).

Table 2 Showing the mortality from cholera, and the water supply, in the districts of London in 1849; the districts are arranged in the order of their mortality from cholera

District	Population	Deaths from cholera	Deaths by cholera to 10 000 inhabitants	Annual value of house and shop room to each person (£)	Water supply
Rotherhithe	17 208	352	205	4.238	Southwark and Vauxhall Water Works, Kent Water Works, and Tidal Ditches
St Olave, Southwark	19 278	349	181	4.559	Southwark and Vauxhall
St George, Southwark	50 900	836	164	3.518	Southwark and Vauxhall, Lambeth
Bermondsey	45 500	734	161	3.077	Southwark and Vauxhall
St Saviour, Southwark	35 227	539	153	5.291	Southwark and Vauxhall
Newington	63 074	907	144	3.788	Southwark and Vauxhall, Lambeth
Lambeth	134 768	1618	120	4.389	Southwark and Vauxhall, Lambeth
Wandsworth	48 446	484	100	4.839	Pump-wells, Southwark and Vauxhall, River Wandle
Camberwell	51 714	504	97	4.508	Southwark and Vauxhall, Lambeth
West London	28 829	429	96	7.454	New River
Bethnal Green	87 263	789	90	1.480	East London
Shoreditch	104 122	789	76	3.103	New River, East London
Greenwich	95 954	718	75	3.379	Kent
Poplar	44 103	313	71	7.360	East London
Westminster	64 109	437	68	4.189	Chelsea
Whitechapel	78 590	506	64	3.388	East London
St Giles	54 062	285	53	5.635	New River
Stepney	106 988	501	47	3.319	East London
Chelsea	53 379	247	46	4.210	Chelsea
East London	43 495	182	45	4.823	New River
St George's, East	47 334	199	42	4.753	East London
London City	55 816	207	38	17.676	New River
St Martin	24 557	91	37	11.844	New River
Strand	44 254	156	35	7.374	New River
Holborn	46 134	161	35	5.883	New River
St Luke, Kensington (except Paddington)	110 491	260	33	5.070	West Middlesex, Chelsea, Grand Junction
Lewisham	32 299	96	30	4.824	Kent
Belgrave	37 918	105	28	8.875	Chelsea
Hackney	55 152	139	25	4.397	New River, East London
Islington	87 761	187	22	5.494	New River
St Pancras	160 122	360	22	4.871	New River, Hampstead, West Middlesex
Clerkenwell	63 499	121	19	4.138	New River
Marylebone	153 960	261	17	7.586	West Middlesex
St. James, Westminster	36 426	57	16	12.669	Grand Junction, New River
Paddington	41 267	35	8	9.349	Grand Junction
Hampstead	11 572	9	8	5.804	Hampstead, West Middlesex
Hanover Square and May Fair	33 196	26	8	16.754	Grand Junction
London	2280 282	14 137	62	–	

Source: Snow (1936) (originally published in 1855).

clear relationship between the average proportion of calories derived from saturated fat and coronary heart disease mortality. Another recent example is provided by psychiatric epidemiology. Ecological studies have demonstrated that the incidence of acute transient psychoses varies dramatically across sociocultural settings. In the World Health Organization (WHO) Ten Country Study, for instance, the incidence of non-affective acute remitting psychosis was 10-fold higher in developing than in developed country settings (Susser and Wanderling 1994). These studies have led to testing of specific hypotheses about causation, including antecedent fever and culturally normative life events.

Much useful public health information has been obtained from ecological studies. However, as illustrated throughout this chapter, both studies of individuals and studies with groups as the units of analysis have their limitations. The degree to which a given study design is appropriate depends on the particular research problem. Snow's research provides an illustrative example. Four years after Snow's initial investigation (illustrated in Table 2), one of the companies, the Lambeth Company, had moved its waterworks to a point higher up on the Thames, thus obtaining a supply of water free from the sewage of London. This meant that within a single district, some houses were receiving water drawn from one place on the Thames and others were receiving water drawn from a different point. Thus, the district as the unit of analysis was no longer appropriate. In Snow's words 'To turn this grand experiment into account, all that was required was to learn the supply of water to each individual house where a fatal attack of cholera might occur' (cited in Susser 1973). Snow subsequently confirmed his original findings by examining the relation between source of water and cholera risk with households, rather than districts, as the units of analysis (Table 3).

The following two sections, which focus on 'fallacies' related to the existence of multiple levels of organization, discuss some of the limitations and potentialities of both ecological studies and studies of individuals. A discussion of the ecological fallacy begins these sections because it is the fallacy most commonly mentioned in epidemiology. How traditional studies of individuals may also be subject to other types of fallacies, which are less often discussed in the epidemiological literature, is then discussed.

The ecological fallacy

The post-Second World War emphasis in epidemiology in investigating interindividual variability, and the implicit assumption that all relevant predictors can be conceptualized as individual-level constructs, has led to a critique of ecological studies. This critique is based on the well-established ecological fallacy. The ecological fallacy is the fallacy of drawing inferences at the individual level (that is, regarding variability across individuals) based on group-level data. The most common example of the ecological fallacy involves situations in which group-level variables are used as proxies for unavailable individual-level exposures. For example, in order to study the relation between exposure to substance X and cancer in the absence of individual-level data, the prevalence of exposure to X in different areas is related to cancer rates in those areas. In this case, information is unavailable on exactly who is exposed to X and who is not, and so the area prevalence of X is used as a rough approximation for the exposure of each area resident. Since information is lacking on the joint distribution of exposure and outcome at the individual level (that is, it is unknown whether people who developed cancer were actually exposed to X) it is impossible to conclude that individuals exposed to X have a higher risk of lung cancer even if it is found that areas with higher per cent exposed to X have higher cancer rates. In addition, information is unavailable on other individual-level characteristics related to cancer which may also vary between areas and may confound the ecological relationship. Another example is provided by the relation between mean income and obesity when mean income is used as a proxy for individual-level income. Suppose that a researcher finds that, at the country level, higher mean income is associated with higher prevalence of obesity (or increased body mass index). If it is inferred that within countries higher income is associated with higher body mass index, the researcher may be committing the ecological fallacy, because within countries body mass index may always be higher in low-income than in high-income people. In addition, people living in countries with high mean income may differ from those living in countries with low mean income in terms of other individual characteristics related to body mass index.

Sources of the ecological fallacy

The ecological fallacy arises when associations between two variables at the group level (or ecological level) differ from associations between analogous variables measured at the individual level. An analysis of the reasons that lead to this difference is helpful in understanding what each type of study can or cannot reveal. These differences between individual-level and group-level associations were first described for correlation coefficients (Robinson 1950) but may also be present for other measures of associations such as linear regression coefficients (Morgenstern 1982). Because the use of correlation coefficients raises additional complexities (and because of its limitations as a measure of association), the following discussion will focus on linear regression coefficients as the main measure of association estimated at both the group and the individual level.

The example on income and obesity used above is schematically illustrated in Fig. 1. At the group or country level, mean body mass index increases with increasing mean income. At the same time, for individuals within countries, body mass index decreases with increasing individual-level income. This situation arises because

Table 3 Cholera death rates by company supplying household water

Company	Number of houses	Deaths from cholera	Deaths in each 10 000 houses
Southwark and Vauxhall Company	40 046	1263	315
Lambeth Company	26 107	98	37
Rest of London	256 423	1422	59

Source: Snow (1936) (originally published in 1855).

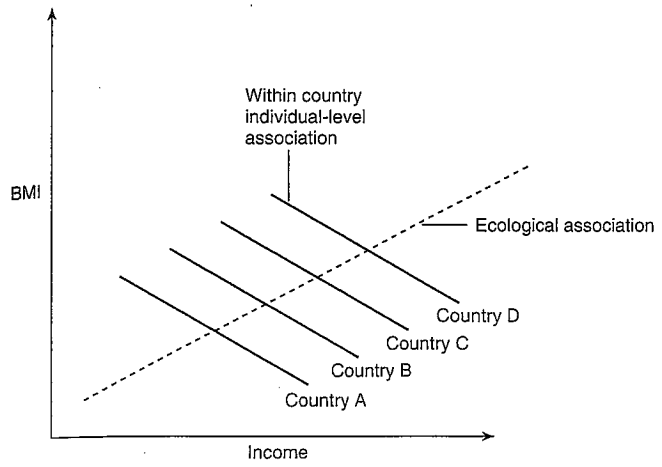

Fig. 1 Hypothetical associations of income with body mass index within and between countries. No interaction between mean country income and individual-level income.

group-level mean income is related to body mass index independently of individual-level income (or in other words because there is a group effect). People living in countries with higher mean income have generally higher body mass indexes than those living in countries with lower mean income, regardless of their individual-level income.

Another situation is depicted in Fig. 2, where the relation between individual-level income and body mass index differs by mean country income: in countries with higher mean income, individual-level income is strongly and inversely related to body mass index, whereas this relation is non-existent, or perhaps exists in the opposite direction, in countries with lower mean income. Here the group-level variable (average income of the country in which a person lives) modifies the effect of individual-level income on the outcome. In this case, group-level associations (the relation between mean country income and mean country body mass index) will also differ from individual-level associations (the relationship between individual-level income and individual-level body mass index). In addition, individual-level associations will differ from country to country, according to levels of mean country income. Thus, when a group-level variable is related to the outcome independently of the analogous variable measured at the individual level, or when the group-level variable modifies the effects of its individual-level analogue on the outcome, ecological associations (which express the relationship between group-level variables and group-level outcomes) will differ from the corresponding individual-level associations (which express the relationship between individual-level variables and individual-level outcomes) (Hammond 1973; Firebaugh 1978; Greenland and Morgenstern 1989; Levin 1995).

The concepts summarized above can also be expressed mathematically. Suppose that the 'true' relationship between country mean income, individual-level income, and individual-level body mass index is reflected in the following equation:

$$Y_{ij} = B_0 + B_1 X_{ij} + B_2 \bar{X}_j \qquad (1)$$

where Y_{ij} is the body mass index for the ith individual in the jth country, X_{ij} is the income for the ith individual in the jth country, $\bar{X}_j$ is the mean X_{ij} in country j, B_1 is the mean difference in individual-level body mass index per unit difference in individual-level income, and B_2 is the mean difference in individual-level body mass index per unit difference in mean country income. Thus, body mass index for each individual is related not only to his or her own income, but also to the mean income for the country in which he or she lives. In the equation above, B_1 reflects the individual-level relation between X_{ij} and Y_{ij}. Since it is 'adjusted' for the group effect (to the extent that between-group differences are entirely captured by B_2) it is equivalent to the within-group effect of individual-level income (or the average within-group effect if this varies across groups). B_1 is what we would like to estimate in order to quantify the individual-level effect of X_{ij} on Y_{ij}. B_2 reflects the relation between mean country income ($\bar{X}_j$) and Y_{ij}, after controlling for X_{ij} (that is, the effect of mean country income on individual-level body mass index after controlling for individual-level income). (Equation (1) can also be modified to allow interactions between X_{ij} and $\bar{X}_j$ as shown in Fig. 2. For simplicity this situation will not be illustrated, but the discussion that follows applies as well.)

Suppose that instead of this full equation showing the relationship between body mass index and individual-level income and mean country income, we fit the ecological equation

$$\bar{Y}_j = B_{e0} + B_{e1} \bar{X}_j \qquad (2)$$

where $\bar{Y}_j$ is the mean body mass index in country j, $\bar{X}_j$ is the mean income in country j, and B_{e1} is the mean difference in mean country body mass index per unit change in mean country income. (The subscript 'e' is used in this case because regression coefficients refer to ecological or group-level associations.) In this case B_{e1} reflects the ecological relation between mean country income ($\bar{X}_j$) and mean country body mass index ($\bar{Y}_j$). It is sometimes referred to as the between-group effect. Clearly, B_{e1}(the between-group effect) is not equivalent to B_1 (the within-group effect of $\bar{X}_{ij}$) or B_2 (the effect of $\bar{X}_j$) in eqn (1). In fact both the individual level within group effect (B_1) and the effect of mean $\bar{X}_j$ (B_2) are confounded in the ecological regression coefficient (B_{e1}).

Yet a third alternative is to fit a purely individual-level equation ignoring group membership:

$$Y_{ij} = B_{p0} + B_{p1} X_{ij} \qquad (3)$$

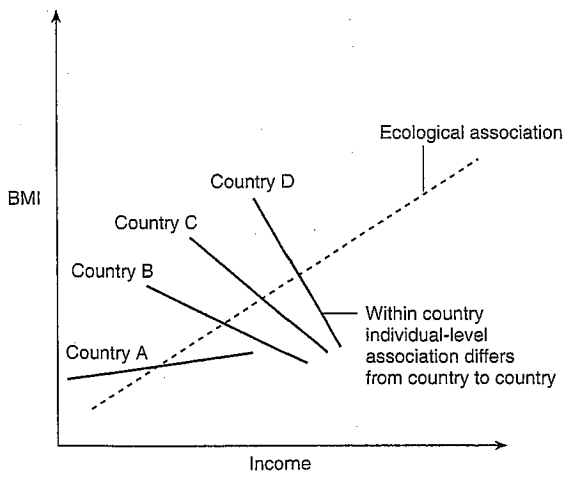

Fig. 2 Hypothetical associations of income with body mass index within and between countries. Mean country income interacts with individual-level income.

(The subscript 'p' is used in this case because regression coefficients refer to associations between individual-level variables pooled across groups.) B_{pI} reflects the effect of X_{ij} on Y_{ij} pooling across groups and ignoring group membership (it is unadjusted for any potential group effects). When group effects are present, B_{pI} will differ from B_I (the within-group effect) in eqn (1), and will also differ from B_{eI} (the ecological effect of $\overline{X}_j$) in eqn (2). Both the within-group effects of X_{ij} and the effect of $\overline{X}_j$ on Y_{ij} are confounded in B_{pI}.

In the absence of group effects (for example, when B_2 in eqn (1) is zero) and when there is no interaction between $\overline{X}_j$ and X_{ij} (the group and individual-level variables), B_I (within-group effect), B_{eI} (ecological effect), and B_{pI} (pooled individual effect) are all equivalent. However, in the presence of group effects B_e does not equal B_I or B_{pI} (source of the ecological fallacy). In addition, B_{pI} does not equal B_{eI} or B_2. When group effects are present, B_{pI} will be a weighted average of B_I and B_{eI}, and will lie between them, although the order of magnitude of B_I and B_{eI} cannot be predicted (Piantadosi *et al.* 1988). Here it is important to note that, although results for correlation coefficients generally follow those for regression coefficients outlined above, individual-level (within-group) and ecological (between-group) correlation coefficients may differ even in the absence of group effects (that is, even when $B_I = B_{eI} = B_{pI}$) (Hammond 1973; Piantadosi *et al.* 1988). This is because correlations also depend on the relative dispersion of X and Y (Piantadosi *et al.* 1988; Morgenstern 1982).

As illustrated above, ecological associations will differ from individual-level associations in the presence of group effects, that is when B_2 in eqn (1) does not equal zero (or when the interaction between X_{ij} and $\overline{X}_j$ in eqn (1) does not equal zero). The key to understanding the sources of the ecological fallacy therefore lies in specifying the conditions under which B_2 (the independent effect of $\overline{X}_j$) or the interaction between $\overline{X}_j$ and X_{ij} will differ from zero. Several different situations may result in non-zero values for B_2 or for the interaction term between X_{ij} and $\overline{X}_j$. These include situations where the group mean X_{ij} ($\overline{X}_j$) is a marker for omitted individual-level variables which individuals in a group tend to share, and situations where mean X_{ij} at the group level measures a different construct than $\overline{X}_{ij}$ at the individual level. Both situations will be discussed in more detail below.

$\overline{X}_j$ is a marker for omitted individual-level variables which individuals in a group tend to share

It is possible that $\overline{X}_j$ is a marker for omitted individual-level variables which individuals in a group tend to share, and which are causally related to the outcome. In the body mass index example above, unmeasured individual-level factors (for example, diet, exercise) may vary from country to country, and their prevalence may be associated with mean country income. These individual-level factors may affect the risk of the outcome independently of individual-level income or may modify the association of individual-level income with the outcome. If these factors are unmeasured in the analyses, their effects will be confounded in the effect of mean income in eqn (1) (B_2) (or in the mean income by individual-level income interaction), and in the ecological regression coefficient relating mean income to mean body mass index (B_{eI} in eqn (2)). Thus, differences in the distribution of these individual-level confounders or effect modifiers across groups may lead to discordances between the group-level and individual-level associations of income and body mass index. Summary ecological measures of these individual-level factors for each group are sometimes available (for example, per cent sedentary, or mean dietary fat). However, controlling for these summary measures in ecological analyses is often insufficient to account for the confounding effects of individual-level variables (although in some cases controlling for multiple summary measures of the same variable may reduce confounding to some extent) (Greenland and Morgenstern 1989; Greenland and Robins 1994*b*; Morgenstern 1995). It is even possible for this ecological adjustment to actually increase bias (Greenland and Morgenstern 1989). The limitations of the use of summary measures are especially important when the relationship between individual-level factors and the outcome is non-linear, when individual-level factors interact, or when there is measurement error (Greenland 1992; Greenland and Robins 1994*b*; Piantadosi 1994).

An additional complexity arises from the fact that even an individual-level variable that is not a confounder at the individual level (because it is not associated with exposure within groups) may still operate as a confounder when ecological associations are used to draw individual-level inferences, if it is ecologically associated with the exposure prevalence across groups (Greenland and Morgenstern 1989; Greenland 1992). Similarly, even weak effect modifiers at the individual level may lead to important differences between ecological and individual-level associations, if their prevalence differs across groups (Greenland and Morgenstern 1989). Conversely, a variable that is a confounder at the individual level (because it is associated with the exposure and the outcome within groups) may not confound the analogous ecological association if it is ecologically unassociated with the exposure across groups (Morgenstern 1995; Greenland and Robins 1994*b*) (that is, the grouping process itself may control for some confounding variables) (Morgenstern 1982). Brenner *et al.* (1992) have also shown that although non-differential misclassification of a binary variable seriously hinders the ability to control for that variable in individual-level studies, it does not always reduce the ability to control for that variable in ecological studies.

Although classical discussions of the ecological fallacy focus on omitted individual-level variables correlated with group membership as potential confounders, it is important to remember that when individual level variables are distributed differently across groups, group membership may hold clues as to the causes of the distribution of these variables. When individuals in a group tend to share individual-level characteristics (such as behavioural risk factors, for example) it may be that group membership (or certain group properties) plays a causal role in the appearance and distribution of these individual-level characteristics. Ultimately, the decision to conceptualize these individual-level variables as confounders or mediators depends on the specific research question being asked.

The absence of information on individual-level confounders (which may differ from group to group) and the limitations inherent in using ecological summaries (and interactions between ecological summaries) to control for individual-level confounders or non-linearities in the individual-level effects are the most common critiques of ecological studies in epidemiology. A key underlying assumption of these critiques (which often goes unstated) is that the ecological measures (for example, mean country income) and the individual-level measures (individual-level income) are indicators of the same construct; that is, if it were somehow possible to control for the unavailable individual-level information on other confounders and to fully capture non-linearities, it would be perfectly legitimate to infer things about how individual-level income is related to the outcome based on how mean country income is associated with it. But this assumption is not always true. Even if all possible individual-level

confounders are controlled, ecological associations may differ from individual-level associations because the ecological measure of exposure and its individual-level namesake may be tapping into different constructs. This brings us to another important source of the ecological fallacy, discussed below.

$\bar{X}_j$ and X_{ij} measure different constructs

An alternative interpretation of Fig. 1 (or Fig. 2) is that country mean income and individual-level income are measuring different constructs, and country mean income is related to body mass index independently of its individual level analogue. Firebaugh (1978) has noted that 'The demystification of cross-level bias begins with the recognition that an aggregate variable often measures a different construct than its name-sake at the individual-level'. In this case, country mean income is a measure of a truly group-level attribute and not a proxy for individual-level data. Living in countries with high mean income places individuals at greater risk of having high body mass index regardless of their individual-level income (or modifies the relation between individual-level income and body mass index, as in Fig. 2). Thus mean country income is said to exert a contextual effect on body mass index. Both country-level and individual-level income provide distinct information, and both are needed to understand completely the distribution of body mass index. In this case, the origin of the ecological fallacy lies in assuming that the group-level measure is tapping into the same construct as the individual-level measure, when in fact it is not. Mean country income is associated with increased body mass index, after controlling for individual-level income, but within countries individual-level income is negatively associated with body mass index. In this situation the ecological fallacy can be thought of as a problem of construct validity; it arises because the aggregate measure is assumed to be measuring an individual-level construct when in fact it is measuring a group-level construct (Schwartz 1994).

The contextual effect of mean income may be mediated through a variety of different mechanisms. For example, countries with higher levels of mean income may rely to a greater extent on mass food production, and this may in turn be associated with higher fat contents, and higher fat in the diets of individuals. In addition the higher standard of living may be related to more sedentary occupations, higher frequency of food consumption outside the home, exposure to food advertisements, dieting behaviours, and so on. Since disease is defined in terms of dysfunctions of the body, in order to affect health, contextual effects must ultimately be mediated through individual-level processes (that is, through processes defined at a lower level of organization), just as the effects of individual-level behaviours, for example, are mediated through biological mechanisms. Mediators of a contextual effect do not necessarily involve the individual-level namesake of the group-level variable. In our body mass index income example, for instance, the contextual effect of mean income may work through individual level variables other than income. As in any epidemiological analysis, the extent to which contextual effects should be estimated before or after adjustment for individual-level factors depends on whether the latter are conceptualized as confounders or mediators.

A variant of the example above is when the ecological variable investigated is actually associated with another group-level attribute which exerts a contextual effect on the outcome. Thus, the ecological effect is actually the result of confounding by another group-level or ecological variable. For example, mean income may be a proxy for level of industrialization. It may be that level of industrialization, rather than mean income, that exerts an effect on body mass index. In this case, the observed contextual effect of mean income is the result of confounding by level of industrialization. The discordance between the group and individual effects of income arises because mean income is a proxy for level of industrialization, which is related to body mass index independently of individual-level income. This is an example of confounding, and is not only specific to ecological variables. It is directly analogous to the situation that arises when an individual-level association between a factor and a disease is due to confounding by a third factor.

The previous discussion (like typical discussions of the ecological fallacy) focuses on the situation where the independent variable at the group level is an aggregate of characteristics of individuals in the group. Thus the group-level variable (mean income) has an individual-level namesake (individual-level income). However, other ecological variables do not involve aggregates of individual-level data (for example, the existence of a certain law) and the problem of separating out contextual effects from individual-level effects of the variable (for example, the effects of mean country income from that of individual-level income) is not an issue. Ecological studies relating these types of predictors to outcomes may still be limited in their ability to draw individual-level inferences because of the absence of information on individual-level confounders or effect modifiers which may differ from group to group (similar to what we have discussed earlier).

In addition to the conceptual issues discussed above, there are statistical considerations which may lead to a discrepancy between estimates from ecological and individual-level studies. The example used throughout this section (country mean income, individual-level income, and body mass index) is based on a continuous individual-level dependent variable, and on the limitations inherent in using a linear ecological model as a proxy for a linear individual-level model. Additional complexities arise when the individual-level outcome is binary. The use of a linear regression ecological model as a proxy for a non-linear individual-level regression model may not always be appropriate. Fitting the aggregate or ecological regression model that directly corresponds to a non-linear individual-level regression model is not always simple or possible with available ecological data (see Greenland (1992) for details). Failure to specify correctly the ecological model to be used as a proxy for the individual-level model may be yet another source of differences between individual-level and aggregate-level regression coefficients. However, all the sources of the ecological fallacy described above may still be present even if the form of the ecological model is appropriate for the individual-level model it is attempting to proxy.

It is important to emphasize that the degree to which ecological and individual-level coefficients differ may vary from situation to situation. The logical possibility of the ecological fallacy should not be taken as evidence that the fallacy necessarily occurs in all cases (Greenland and Robins 1994*a*) or that when it occurs it has a critical impact. Some researchers have proposed strategies which may sometimes help reduce the ecological fallacy when drawing individual-level inferences from aggregate data, including selecting regions so as to minimize within-region variability and maximize between-region variability in individual-level exposure, comparing groups with similar covariate distributions, conducting sensitivity

analyses, and comparing results based on different specifications of the ecological model (Greenland 1992), as well as other recently proposed statistical approaches (King 1997). Nonetheless the use of ecological associations as proxies for individual-level relationships is often problematic.

Other fallacies related to the existence of multiple levels

The ecological fallacy is only one of a set of possible 'fallacies' that derive from the existence of multiple levels of organization (Diez Roux 1998). Because, at least recently, epidemiologists have been mostly concerned with drawing inferences regarding the causes of interindividual variability, the ecological fallacy has received much more attention than its counterpart, the atomistic fallacy. The atomistic fallacy is the fallacy of drawing inferences regarding variability across groups based on individual-level data. The effect of an individual-level predictor on an outcome in a study of individuals is not necessarily the same as the effect of its group-level namesake on group-level outcomes. Thus the use of individual-level associations to draw inferences regarding group-level associations may also lead to incorrect inferences. In the body mass index example above, B_{pl} (the relation between individual-level income and body mass index pooling individuals across groups) does not equal B_{el} (the ecological relation between country mean income and country mean body mass index). In addition, B_l (the within-group effect of individual-level income) does not equal B_{el} either. Moreover, the body mass index example includes multiple groups (countries) and individuals within them, but many individual-level studies only include individuals from a single group. Factors that explain variability across individuals within groups are not necessarily the same as those that explain variability across groups. For example, if stress levels are relatively invariant within groups (for example, communities or countries), stress may not be important in explaining variability in coronary heart disease within groups, but may be strongly associated with differences in coronary heart disease rates across groups. This is another reason why the use of individual-level data to infer group-level associations may lead to incorrect inferences.

Both the ecological and atomistic fallacies can be thought of as methodological problems inherent in drawing inferences at one level when the data are collected at another level. These fallacies arise when the conceptual model being tested corresponds to one level, but the data are collected at another level, or in Riley's words when 'the methods fail to fit the model' (Riley 1963). We have seen that the sources of these problems lie in (a) the lack of information on constructs pertaining to another level of organization and (b) the failure to realize that a variable defined and measured at one level of organization may tap into a different construct than its namesake at another level, and that constructs at both levels may be relevant to the outcome studied. Both of these issues indicate a more general problem which is that even when making inferences about a given level, other levels of organization may need to be taken into account. For example, the failure to consider group characteristics in drawing inferences regarding individuals, and the failure to consider individuals in drawing inferences regarding groups gives rise to another set of fallacies, which are closely related to the ecological and atomistic fallacies described above. In these fallacies (which have been termed the psychologistical or individualistic and sociologistical fallacies), although the level at which data are collected may fit the conceptual model being investigated, important facts pertaining to other levels have been ignored, in Riley's words 'the methods may fail to fit the facts' (Riley 1963).

Ignoring relevant group-level variables in a study of individual-level associations may lead to what Riley has termed the 'psychologistical fallacy', that is assuming that individual-level outcomes can be explained exclusively in terms of individual-level characteristics. For example, a study based on individuals might find that immigrants are more likely to develop depression than natives. But suppose this is only true for immigrants living in communities where they are a small minority. A researcher ignoring the contextual effect of community composition might attribute the higher overall rate in immigrants to the psychological effects of immigration *per se* or even to genetic factors, ignoring the importance of community-level factors and thus committing the psychologistic fallacy (Riley 1963; Valkonen 1969). (The term 'psychologistical fallacy' is not the most appropriate because the individual-level factors used to explain the outcome are not always exclusively psychological. Other authors have used the term 'individualistic fallacy', (Valkonen 1969) but because the term has also been used as a synonym of the 'atomistic fallacy' (Alker 1969; Scheuch 1969) it will be avoided here.)

Analogously, ignoring the role of individual-level factors in a study of groups may lead to what has been termed the sociologistical fallacy (Riley 1963). Suppose that a researcher finds that communities with higher rates of transient population have higher rates of schizophrenia, and he or she concludes that higher rates of transient population lead to social disorganization, breakdown of social networks, and increased risk of schizophrenia among all community inhabitants. But suppose that schizophrenia rates are only elevated for transient residents (because transient residents tend to have fewer social ties, and individuals with few social ties are at greater risk of developing schizophrenia). That is, rates of schizophrenia are high for transient residents and low for non-transient residents, regardless of whether they live in communities with a high or a low proportion of transient residents. If this is the case, the researcher would be committing the sociologistical fallacy in attributing the higher schizophrenia rates to social disorganization affecting all community members rather than to differences across communities in the percentage of transient residents, who are at higher risk because of individual characteristics.

Both the psychologistical and the sociologistical fallacies arise because relevant variables pertaining to other levels have been excluded from the model which led to an inappropriate explanation for the association. Although it is didactically useful to distinguish both sets of fallacies (ecological and atomistic versus psychologistical and sociologistical), they are closely interrelated and are essentially different manifestations of the same phenomenon: the failure to recognize that constructs defined at different levels may be important in understanding variability within a given level, and the failure to adequately distinguish constructs defined at different levels.

The types of fallacies are summarized in Table 4.

Table 4 Types of fallacies

Unit of analysis	Level of inference	Type of fallacy
Group	Individuals	Ecological
Individual	Groups	Atomistic[a]
Individual (relevant group-level variables excluded)	Individuals	Psychologistical[a]
Group (relevant individual-level variables excluded)	Groups	Sociologistical

[a]Also called individualistic by some authors.

Source: Diez-Roux (1998).

The full range of epidemiological studies

In considering the most appropriate study design to answer a given research question, investigators need to consider two issues. The first issue is the level of organization about which inferences are to be made. For example, is the interest in drawing inferences regarding causes of variation in the outcome among groups or among individuals? The answer to this question will determine the most appropriate unit of analysis. The second issue is the level of organization at which the constructs of interest in explaining the outcome are conceptualized and measured. The answer to this second question will determine the predictors to be investigated and the level at which they are conceptualized. Constructs relevant to health may be conceptualized, and measured, at different levels of organization (for example countries, states, neighbourhoods, peer groups, families, couples, people, measurement occasions). Factors defined at a higher level may be important in understanding variability at a lower level, and vice versa factors defined at a lower level may be important in understanding variability at a higher level. Decoupling the unit of analysis from the level of organization of the constructs investigated may be helpful in discussing the full range of studies available to epidemiologists and the advantages and disadvantages of each for a particular research question.

Clearly specifying the constructs of interest (as well as how they will be measured) is an important requirement of any study. Lack of clarity on exactly what constructs group-level or ecological variables are actually measuring underlies an important part of the confusion generated by the use of ecological studies in epidemiology, and the interpretation of group-level or ecological effects. The next section reviews the use of group-level variables in epidemiology based on the constructs which they are intended to be measuring. The study designs with different units of analysis are then considered in terms of the types of inferences that can be drawn from them and the types of constructs they are best suited to investigating.

Group-level variables in epidemiology

In this discussion of group-level variables the focus is on the generic example of 'individuals' nested within 'groups'. However, as previously noted, the issues discussed may pertain to many other situations (for example, neighbourhoods nested within states, countries nested within regions, siblings nested within families, repeated measures nested within individuals) where the 'higher' level can be thought of as the group (and higher-level variables as 'group' variables) and the lower level can be thought of as the 'individual' (and lower-level variables as 'individual-level' variables).

Group-level (higher-level) variables as proxies for individual-level (lower-level) variables

One of the most common uses of ecological variables in epidemiological studies is as proxies for individual-level variables, either because individual-level data are unavailable or because individual-level measurements are prone to measurement error. For example, in the absence of detailed information on smoking for individuals, median smoking levels in the area in which an individual lives may be used as a proxy. Of course, the use of these group-level proxies implies loss of information. It is not known whether a given person smokes or not; mean smoking levels in the area are used as an approximation. In this case the group-level measure is a second class alternative to the ideal individual-level measurement. The relevant construct (smoking) is defined at the individual level, but a group-level measurement is used as a proxy for it because direct individual-level measurements are unavailable. If a valid and reliable measure of individual-level exposure were available, it would be used instead.

Group-level variables are also used as proxies for individual-level data in cases in which individual-level measures are subject to a lot of measurement error, or when intraindividual variability makes a single measure a poor marker for the person's true exposure. For example, the mean yearly hours of sunlight in an area may be used as a proxy for individual-level exposure to sunlight and average per capita fat consumption in a group may be used as a proxy for the fat consumption of each member (due to limitations in characterizing an individual's fat intake based on a single one-day measurement). In these situations, the ecological measure is believed to be a better indicator of the 'true' individual-level exposure than the individual-level measure itself, because the ecological measure reduces the 'noise' associated with measurement error or intraindividual variation.

Group-level proxies for individual-level variables can be used in studies with individuals or groups as the units of analysis. Regardless of the study design in which they are used, the key assumption in the use of these variables is that the group-level measure is an adequate proxy for the individual-level construct; that is, even if there is measurement error, the construct that is being tapped into by the measure is an individual-level property rather than a group-level property. But this may not always be true; it is possible that the group-level measure is capturing information about a group-level attribute rather than (or in addition to) information about the individual-level characteristics of people comprising the group. For example, it is theoretically possible that areas with a higher percentage of smokers have higher cancer rates not only because smokers develop cancer, but also because there is something about areas with a high percentage of smokers (for example, the percentage of smokers may exert a contextual effect on cancer risk mediated, for example, through

exposure to passive/second-hand smoke) which places everyone in the area at higher risk regardless of whether they smoke or not. Unless individual-level data are available, it may not be possible to differentiate the effects of individual-level smoking from the contextual effect of the percentage of smoking in the area (or other group-level properties associated with it). Thus, in considering the use of group-level variables as proxies of individual-level data, researchers may need to consider two issues: (a) the degree of measurement error in the individual-level construct inherent in using the group-level variable (for example, misclassification of smokers as non-smokers), and (b) the degree to which the group-level measure is tapping into a group-level construct rather than the individual-level construct it purports to proxy.

Group-level (higher-level) variables as measures of group-level (higher-level) constructs

Another application of ecological variables, much less common in epidemiology, is to measure group-level constructs. Variables that reflect the characteristics of groups have been classified into two basic types (Valkonen 1969; Lazarsfeld and Menzel 1971; Blalock 1984; Von Korff *et al.* 1992; Morgenstern 1995): derived variables and integral variables. Derived variables (also termed analytical or aggregate variables) summarize the characteristics of individuals in the group (means, proportions, and so on) (for example, percentage of people with incomplete high school, median household income). Although created through the aggregation of information from the individual members of a group, derived variables are group characteristics and may provide information distinct from their individual-level analogue. As discussed above, mean neighbourhood income and individual-level income are indicators of two distinct constructs, each of which may be important to health. Mean neighbourhood income may be a marker for neighbourhood-level factors potentially related to health (such as recreational facilities, school quality, road conditions, environmental conditions, types of foods available and their cost), and these factors may affect everyone in the community regardless of their individual-level income. Similarly, community unemployment levels may affect everyone in the community regardless of whether they are unemployed or not.

A special subset within derived variables is the average of the dependent variable within the group (Susser 1994*a*). As noted by Ross (1911) in his theory of happenings, for some types of events, the frequency of occurrence may depend on the number of individuals already affected. The prevalence of a given infection in the group to which a person belongs will affect his or her risk of infection, or may modify the relation between individual-level risk factors and the risk of disease (Halloran and Struchiner 1991; Koopman *et al.* 1991*b*; Koopman and Longini 1994). The classic concept of herd immunity is a variant of this notion: the prevalence of immunity in a community will determine whether an epidemic of disease does or does not occur, and will therefore influence a non-immune individual's risk of acquiring disease (Susser 1973). The contextual effect of the dependent variable's prevalence within a group may also be important in understanding other health outcomes. For example, the prevalence of obesity in a community may influence the likelihood that an individual is obese. This effect may operate through several different mechanisms. The prevalence of obesity may itself generate societal norms regarding acceptability and desirability of obesity, which may influence an individual's risk of being obese. In addition, the probability of adopting behaviours conducive to obesity (for example, certain types of diet or physical activity patterns) may be higher in situations where the behaviour is highly prevalent in the community. Although infrequently considered, these types of contextual effects may be of crucial importance in understanding the distribution (and causes) of health-related behaviours (smoking and alcoholism are two common examples).

Integral variables (also termed primary or global variables) describe group characteristics that are not derived from characteristics of its members (for example, the existence of certain types of laws, availability of health care, political system, or population density). Integral variables do not have analogues at the individual level. They may be discrete and dichotomous (for example, an intervention or a disaster, presence of a certain law), scaled and polychotomous (for example, social disorganization, intensity of newborn care), or continuous (for example, doctors per capita). A special type of integral variable refers to patterns and networks of contacts or interactions between individuals within groups. These patterns are derived from how individuals are connected to each other, and yet they are more than aggregates of individual characteristics. Lazarsfeld and Menzel (1971) have referred to these variables as structural variables, although the term structural effects has also been used to refer to the effects of group-level properties more generally (Blau 1960). Patterns of interconnections among individuals may be important determinants of individual risk, particularly for infectious diseases, but possibly for many other health outcomes as well (Koopman *et al.* 1991*a*; Koopman and Longini 1994; Koopman and Lynch 1999). In addition, these patterns of interconnections may modify the relation between certain individual-level attributes and risk of disease (Koopman *et al.* 1991*a*). These patterns of interactions can be summarized in the form of group-level attributes such as network size or structure (Lazarsfeld and Menzel 1971; van den Eeden and Huttner 1982). Just as other 'group-level' variables can refer to groups of various sizes, these patterns of interactions may characterize a whole continuum of groups depending on the particular research problem: large groups, smaller groups within larger groups, or even pairs of individuals.

When group-level variables are used to characterize group-level properties there is no ambiguity in defining whether individuals are or are not exposed (as there is when group variables are used as proxies for individual-level exposure data). The group-level variable (whether derived or integral) applies equally to all individuals within the group; for example, all are 'exposed' to living in a neighbourhood with high unemployment regardless of whether they themselves are employed, and all are 'exposed' to existing laws regarding seatbelt use. Thus the measurement error problem which may be present when group measures are used as proxies for measures of individual-level constructs is not present (although there may be measurement error in the measure of the group-level construct itself, as there may be for any measure). As discussed below, such group-level constructs can be investigated in studies with either individuals or groups as the units of analysis.

Study designs based on the units of analysis

Studies with groups as units of analysis

Studies with groups as the units of analysis (traditional ecological studies) are most appropriate when investigators are interested in explaining variation between groups and the constructs of interest can

be conceptualized as group-level properties. As discussed above, these group-level properties may be derived or integral variables. Over the past few years numerous studies have examined the relationship between area measures of deprivation (a derived variable) and area mortality rates with geographic regions as the units of analysis (Townsend *et al.* 1988; Carstairs and Morris 1991). This analytical approach is most appropriate if the research question is formulated at the area (group) level and the main construct investigated (deprivation) is conceptualized as an area- or group-level attribute. Similar studies have examined the relation between area socio-environmental characteristics and the decline of coronary heart disease mortality rates in the United States (Wing *et al.* 1988): area socio-environmental characteristics are conceptualized as group attributes that affect all individuals living within the community and the interest lies in drawing inferences regarding differences between areas.

An example involving integral group-level variables is provided by studies relating national legislation restricting tobacco advertising in different countries to country smoking rates: a country-level construct is examined and the interest lies in drawing country-level inferences. Ecological designs may also be appropriate for the evaluation of the effects of group-level interventions on group-level outcomes (Morgenstern 1982). For example, a study may want to investigate the relationship between the introduction of a mass media campaign to prevent teenage smoking (an integral variable) and the prevalence of teenage smoking in the area. Because the mass media campaign may affect all community inhabitants (regardless of whether they actually saw the advertisements or not) through mechanisms involving diffusion, peer pressure, and so on, the intervention can be conceptualized as a group-level attribute.

As discussed above in the section on the ecological fallacy, these studies are limited in their examination of the role of individual-level constructs—as confounders, mediators, or effect modifiers of the group-level associations. In the example above, differences in the effects of the mass media campaign by individual-level characteristics (effect modification) could not be investigated. Neither could the impact of differences in individual exposures to the mass media campaign. In addition, in the case of ecological variables with individual-level analogues (for example, area unemployment and individual-level unemployment) studies with groups as the units of analysis cannot differentiate the contextual effect of the variable from its individual-level effect. For example, an ecological study relating area unemployment to rates of mental health outcomes could not differentiate whether the increased rate of mental illness is seen only in the unemployed or is present in all area inhabitants regardless of whether they are unemployed or not. However, from a public health perspective, the group-level association may itself be of great interest. Decreasing the unemployment rate would decrease the rate of mental illness whether the effect resulted from a group-level or individual-level construct or both. Similarly, the country-level relationship between income inequality and health may have important policy implications regardless of whether it is due to a contextual effect of income inequality, or to the fact that more unequal countries tend to have more people in the lower-income categories.

It is often argued that there is one situation in which ecological studies may be particularly useful, even when the intention is to draw individual-level inference (despite the potential for the ecological fallacy). This situation involves individual-level attributes with little within-group variation but large between-group variation. For example, if dietary fat is homogeneous within countries but varies greatly from country to country, an individual-level study restricted to individuals from a single country may find no association between dietary fat and cardiovascular disease, but an ecological study comparing country rates with country average fat intake may find a strong relationship. From this perspective, the advantage of the ecological study results purely from the fact that it is able to include more variability in the exposure of interest. (The same research question could be addressed in a study of individuals that included individuals from different countries and thus ensured sufficient variation in the exposure. But often this option is not feasible, whereas country-level measures may be available from standard sources.) Of course, the presence of significant between-country differences in diet raises the important question of why countries differ in diet to begin with, and suggests that the diet of individuals has important country-level determinants. Diet may be simply a mediator in the causal pathway relating country characteristics to cardiovascular disease. A study which concluded that differences between countries in cardiovascular disease are reducible to differences in the diets of individuals could be missing important disease antecedents.

Although they will not be discussed here in detail, studies with groups as the units of analysis are subject to many of the same analytical issues as individual-level studies with respect to bias, confounding, and establishment of temporality. In addition, these types of analyses raise additional methodological issues, such as the need to have adequate numbers of groups as well as enough individuals per group, the need to account for differences in the variability of outcomes (for example, rates) for the different groups due to the fact that they may be based on different numbers of observations, and the possibility of multicolinearity between the predictors examined (which is often more of a problem in ecological studies than in individual-level studies) (Morgenstern 1982). In addition, studies in which the units of analysis are geographic areas may need to use statistical methods to account for the fact that areas geographically closer to each other may tend to be more similar (in outcomes) than those more distant from each other (due to unmeasured factors that cluster in space), which leads to violation of the assumption of independence of observations (for example, see Clayton *et al.* 1993). This is identical to situations involving individuals nested within groups or repeat measures on individuals over time, in which residual correlation between outcomes may be present. Time-trend studies also raise additional methodological issues related to time-series analyses generally (Morgenstern 1995).

Studies with individuals as units of analysis

Traditional individual-level studies are most appropriate when investigators are interested in drawing inferences regarding variability across individuals, and all potential constructs of interest can be conceptualized as individual-level properties. The most common 'classic' epidemiological studies (for example, case–control studies of lung cancer and cohort studies of cardiovascular disease) are of this type. The assumption is that all constructs relevant to the outcome being studied are individual-level constructs.

Studies with individuals as the units of analysis and with information limited to individual-level constructs cannot examine the role of group-level constructs as antecedents of individual-level variables, as independent predictors of outcomes, or as confounders of individual-level associations. They cannot determine whether the

effect of a given individual-level variable is only present in certain group contexts, or varies from group to group, as a function of group characteristics. In summary, they are unable to investigate the role of group-level factors in explaining variability in the outcome across individuals. In order to answer these questions, other types of analyses are needed.

Studies limited to individuals from a single group are clearly unable to examine the role of group-level constructs in causing the outcome (or in interacting with individual-level variables), because group-level properties are invariant within groups (Schwartz and Carpenter 1999). If group-level factors are important in causing the outcome, studies focused on a single group may fail to detect important disease determinants. In the dietary fat example mentioned above, it was noted that a study based on individuals from a single country would not detect dietary fat as a risk factor if it were invariant within countries. More fundamentally, the country-level factors that influence the range of fat intake is the salient variable that the individual-level study could not capture.

If the study involves individuals from many different groups, relevant group-level properties may be included in individual-level analyses. For example, group-level variables can be included in regression equations with individuals as the units of analysis. These types of analyses have been called contextual analyses (Blalock 1984; Iversen 1991). Susser (1994*a*) has referred to studies which investigate the effects of group-level variables on individual-level outcomes as mixed studies. This approach is still relatively uncommon in epidemiology perhaps partly because it is sometimes difficult to measure the group-level construct adequately, and cover the variation in this construct, in studies designed with individuals as the units of analysis.

A simple example of the type of regression model fitted in contextual analysis (for the linear case) is

$$Y_{ij} = b_0 + b_1 C_j + b_2 X_{ij} + e_{ij} \qquad (4)$$

where Y_{ij} is the outcome for the ith individual in the jth group, C_j is the group-level variable, X_{ij} is the individual-level variable, and e_{ij} is the error term for the ith individual in the jth group. Contextual models can include multiple group-level and individual-level variables as well as their interactions. In the model shown above, b_1 estimates the effect of the group-level characteristic on the individual-level outcome (after adjustment for X_{ij}) and b_2 estimates the effect of the individual-level variable on the outcome (after adjustment for C_j). Contextual models can be used, for example, to investigate the effects of neighbourhood context on fertility outcomes by including characteristics of the neighbourhoods where individuals live (derived or integral variables) together with individual-level characteristics in individual-level regression models. Special methods may be required to account for non-independence of the outcomes within groups. Although contextual analysis can be used for simultaneous investigation of the effects of group-level and individual-level constructs in shaping individual-level outcomes, it does not allow examination of group-to-group variability *per se*, or of the factors associated with it. The unit of analysis remains the individual and only interindividual variation is examined.

Studies with both groups and individuals as the units of analysis (multilevel studies)

Recently, multilevel studies and multilevel analysis have emerged as new analytic strategies in several fields including education, sociology, demography, and public health (Mason *et al.* 1983; Bryk and Raudenbush 1992; Von Korff *et al.* 1992; DiPrete and Forristal 1994; Paterson and Goldstein 1995; Wu 1995; Duncan *et al.* 1998; Kreft and de Leeuw 1998). Multilevel studies simultaneously examine groups (or samples of groups) and individuals within them (or samples of individuals within them). Variability at both the group level and the individual level can be simultaneously examined and the role of group-level and individual-level constructs in explaining variation between individuals and between groups can be investigated. For example, a study may have information on a series of country-level characteristics (for example, gross national product, inequality in the distribution of income) and on the individual-level characteristics of a sample of individuals within each country (including health outcomes). Researchers may be interested in investigating how country-level and individual-level factors are related to health outcomes, as well as the extent to which between-country and between-individual variability in the outcomes are explained by variables defined at both levels. Thus, multilevel analysis allows researchers to deal with the microlevel of individuals and the macrolevel of groups or contexts simultaneously (Duncan *et al.* 1998). Multilevel models can be used to draw inferences regarding the causes of interindividual variation and the extent to which it is explained by individual-level or group-level variables, but inferences can also be made regarding intergroup variation, whether it exists in the data, and to what extent it is accounted for by group- and individual-level characteristics.

In the case of multilevel analysis involving two levels (for example, individuals nested within groups), the model can be conceptualized as a two-stage system of equations. The case for a normally distributed dependent variable is illustrated below. For reasons of simplicity the illustration will focus on the case of only one independent variable at the individual and one independent variable at the group level (although models can of course be extended to include as many independent variables as needed).

In the first stage, a separate individual-level regression is defined for each group:

$$Y_{ij} = b_{0j} + b_{1j} I_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim n(0, \sigma^2) \qquad (5)$$

where Y_{ij} is the outcome variable for the ith individual in the jth group (or context), I_{ij} is the individual-level variable for the ith individual in the jth group (or context), b_{0j} is the group-specific intercept, and b_{1j} is the group-specific effect of the individual-level variable. Individual-level errors (ε_{ij}) within each group are assumed to be independent and identically distributed with a mean of zero and a variance of σ^2. The same regressors are generally used in all groups, but regression coefficients (b_{0j} and b_{1j}) are allowed to vary from one group to another.

In a second stage, each of the group- or context-specific regression coefficients defined in eqn (5) (b_{0j} and b_{1j} in this example) are modelled as a function of group-level variables:

$$b_{0j} = \gamma_{00} + \gamma_{01} C_j + U_{0j} \qquad U_{0j} \sim n(0, \tau_{00}) \qquad (6)$$

$$b_{1j} = \gamma_{10} + \gamma_{11} C_j + U_{1j} \qquad U_{1j} \sim n(0, \tau_{11}) \qquad (7)$$

$$\mathrm{cov}(U_{0j}, U_{1j}) = \tau_{10}$$

where C_j is the group-level or contextual variable, γ_{00} is the common intercept across groups, γ_{01} is the effect of the group-level predictor on

the group-specific intercepts, γ_{10} is the common slope associated with the individual-level variable across groups, and γ_{11} is the effect of the group-level predictor on the group-specific slopes.

The errors in the group-level equations (U_{0j} and U_{1j}), sometimes called 'macro errors', are assumed to be normally distributed with mean zero and variances τ_{00} and τ_{11} respectively. τ_{01} represents the covariance between intercepts and slopes; for example, if τ_{01} is positive, as the intercept increases the slope increases. Thus, multilevel analysis summarizes the distribution of the group-specific coefficients in terms of two parts—a 'fixed' part which is unchanging across groups (γ_{00} and γ_{01} for the intercept, and γ_{10} and γ_{11} for the slope) and a 'random' part (U_{0j} for the intercept and U_{1j} for the slope) which is allowed to vary from group to group.

By including an error term in the group-level equations (eqns (6) and (7)), these models allow for sampling variability in the group-specific coefficients (b_{0j} and b_{1j}) and also for the fact that the group-level equations are not deterministic (that is, the possibility that not all relevant macrolevel variables have been included in the model) (Wong and Mason 1985). The underlying assumption is that group-specific intercepts and slopes are random samples from a normally distributed population of group-specific intercepts and slopes (or equivalently, that the groups or macro errors are 'exchangeable') (Diprete and Forrister 1994).

An alternative way to present the model fitted in multilevel analysis is to substitute eqns (6) and (7) in (5) to obtain

$$Y_{ij} = \gamma_{00} + \gamma_{01}C_j + \gamma_{10}I_{ij} + \gamma_{11}C_jI_{ij} + U_{0j} + U_{1j}I_{ij} + \varepsilon_{ij}. \quad (8)$$

This final model is a random effects model (Hox and Kreft 1994). The model includes the effects of group level variables (γ_{01}), individual-level variables (γ_{10}), and their interaction (γ_{11}) on the individual-level outcome Y_{ij} (often called the 'fixed effects'). It also includes a random intercept component (U_{0j}), and a random slope component (U_{1j}), which, together with the individual-level errors ε_{ij}, comprise a complex error structure. Because of the presence of this complex error structure, special estimation methods must be used. Although multilevel or random effects models were first developed for continuous dependent variables, analogous methods have been developed or are under development other types of outcomes (Wong and Mason 1985; Goldstein 1995).

Multilevel models allow investigation of a variety of interrelated research questions. They allow separation of the effects of context (that is, group characteristics) and of composition (characteristics of the individuals in groups) (Duncan *et al.* 1998). Do groups differ in average outcomes after controlling for the characteristics of individuals within them? Are group-level variables related to outcomes after controlling for individual-level variables? Multilevel models can also be used to examine the effects of individual-level variables. Are individual-level variables related to the outcome after controlling for group-level variables? Do individual-level associations vary from group to group, and is this partly a function of group-level variables? Multilevel models also allow quantification of variation at different levels (for example, within group and between group) and the degrees to which variation is 'statistically explained' by individual-level and group-level variables. For example, is there significant variation in group-specific intercepts or slopes (do τ_{00} and τ_{11} differ significantly from 0)? How does this variability change as individual-level or group-level variables are added? What percentage of the variability in individual-level outcomes is between and within groups?

Although the terms contextual analysis and multilevel analysis have often been used synonymously (Van den Eeden and Huttner 1982; Mason *et al.* 1983), today's multilevel models are more general than early contextual models. Contextual effects models do not allow examination of group-to-group variability in outcomes as multilevel models do. In multilevel analysis, inferences can be drawn regarding differences among groups as well as differences among individuals, and the role of both group-level and individual-level constructs can be examined. In a sense, both groups and individuals are units of analysis, and both group-to-group and individual-to-individual variation are examined. Thus, multilevel studies provide a link between traditional ecological and individual-level studies. The advantages and limitations of multilevel models have been reviewed in several publications (DiPrete and Forristal 1994; Duncan *et al.* 1998; Diez Roux 2000).

The types of studies, the levels at which variability is examined, and the types of constructs which they are more suited to investigate are summarized in Table 5.

Conclusion

In public health research, both predictors and outcomes may be conceptualized at different levels of organization, and understanding outcomes at a given level may require taking into account information pertaining to levels above or below it. Each system can be thought of as nested within another level and dynamically interrelated with the levels above and below it. In addition, each level may acquire 'emergent' properties, unique characteristics confined to that level,

Table 5 Types of study designs used in public health based on unit of analysis, level at which variability is examined, and constructs most appropriately investigated

Type of study	Unit of analysis	Level at which variability is examined	Constructs investigated as potential 'causes' of variability	
			Group-level	Individual-level
Ecological	Groups	Groups (utility for interindividual variability limited)	Yes	Only group-level proxies
Individual level	Individuals	Individuals (utility for intergroup variability limited)	No (Yes in contextual)	Yes
Multilevel	Groups and individuals	Groups and individuals	Yes	Yes

which are different from the properties of its components. The selection of the appropriate study design should be based on the specific research question investigated, including the level of organization about which inferences are to be made, as well as the levels of organizations of the constructs of interest (including the main independent variables as well as potential confounders or effect modifiers of the association).

As noted above, many of the issues discussed in relation to generic 'individuals' and 'groups' are generalizable to other situations in public health involving 'lower levels' nested within 'higher levels'. Generic issues which are present across the continuum of 'levels' include limitations of using measures at a higher level as proxies for lower-level measures, the problems in making inferences regarding causes of variability at one level based on data collected at another level, and the need to consider multiple levels in drawing inferences at a given level. For example, in a study involving multiple measures on individuals over time the use of 'average' outcome and exposure measurements for individuals could lead to problems analogous to those described for the ecological fallacy. Similarly, both individual characteristics and occasion-specific (or time-specific) constructs may be important in understanding the outcome being studied. Indeed, the multilevel analysis methods described above as an alternative analytical approach which allows simultaneous examination of intergroup and interindividual variation, are suited to situations involving nested data structures generally (for example, multiple measures on individuals over time, patients nested within providers, neighbourhoods nested within regions). It is interesting to point out that analogous issues related to the existence of multiple levels of organization may also exist within individuals, for example, when looking at cells within tissues.

Problems related to the use of ecological studies and ecological variables in epidemiology often result from confusion regarding the level of organization to which the research question pertains, the level of organization at which the constructs of interest are defined and measured, and the sometimes inappropriate use of variables defined and measured at one level to proxy constructs defined at another level. Researchers must necessarily focus on certain aspects of the continuum of levels of organization and not all studies need (or can) span all levels. Rather than defending or criticizing one study design in favour of another, it is more useful to evaluate whether the level of analysis investigated and the constructs examined are appropriate for the specific question being asked. Because 'ideal' study designs are often not possible, the key lies in determining whether the particular design employed is 'good enough' for the question being asked. The issues reviewed in this chapter may be helpful in making this judgement.

References

Alker, H.R. (1969). A typology of ecological fallacies. In *Social ecology* (ed. M. Dogan and M.S. Rokkam), pp. 69-86. MIT Press, Cambridge, MA.

Blalock, H.M. (1984). Contextual-effects models: theoretical and methodological issues. *Annual Review of Sociology*, **10**, 353–72.

Blau, P.M. (1960). Structural effects. *American Sociology Review*, **25**, 178–93.

Brenner, H., Greenland, S., and Savitz, D. (1992). The effects of nondifferential confounder misclassification in ecologic studies. *Epidemiology*, **3**, 456–9.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchichal linear models: applications and data analysis methods.* Sage, Newbury Park, CA.

Carstairs, V. and Morris, R. (1991). *Deprivation and health in Scotland.* Aberdeen University.

Chadwick, E. (1965). *Report on the sanitary conditions of the labouring population of Great Britain.* Edinburgh University Press (originally published 1842).

Clayton, D.G., Bernardinelli, L., and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, **22**, 1193–202.

Diez Roux, A.V. (1998). Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *American Journal of Public Health*, **88**, 216–22.

Diez Roux, A.V. (2000). Multilevel analysis in public health research. *Annual Review of Public Health*, **21**, 171–92.

DiPrete, T.A. and Forristal, J.D. (1994). Multilevel models: methods and substance. *Annual Review of Sociology*, **20**, 331–57.

Duncan, C., Jones, K., and Moon, G. (1998). Context, composition, and heterogeneity: using multilevel models in health research. *Social Science and Medicine*, **46**, 97–117.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociology Review*, **43**, 557–72.

Goldberger, J., Wheeler, G.A., Sydenstricker, E., and Tarbett, R.E. (1920). A study of the relation of factors of a sanitary character to pellagra incidence in seven cotton-mill villages of South Carolina in 1916. *Public Health Report*, **35**, 1701–24.

Goldstein, H. (1995). *Multilevel statistical models.* Halsted Press, New York.

Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine*, **11**, 1209–23.

Greenland, S. and Morgenstern, H. (1989). Ecological bias, confounding, and effect modification. *International Journal of Epidemiology*, **18**, 269–74.

Greenland, S. and Robins, J. (1994*a*). Accepting the limits of ecologic studies: Drs Greenland and Robins reply to Drs Piantadosi and Cohen. *American Journal of Epidemiology*, **139**, 769–71.

Greenland, S. and Robins, J. (1994*b*) Ecologic studies—biases, misconceptions, and counter-examples. *American Journal of Epidemiology*, **139**, 747–60.

Halloran, M.E. and Struchiner, C.J. (1991). Study design for dependent happenings. *Epidemiology*, **2**, 331–8.

Hammond, J.L. (1973). Two sources of error in ecological correlations. *American Sociology Review*, **38**, 764–77.

Hox, J.J. and Kreft, I.G. (1994). Multilevel analysis methods. *Sociological Methods and Research*, **22**, 283–99.

Iversen, G. (1991). *Contextual analysis.* Sage, Newbury Park, CA.

Keys, A. (1980). *Seven countries: a multivariate analysis of death and coronary heart disease.* Harvard University Press, Cambridge, MA.

King, G. (1997). *A solution to the ecological inference problem. Reconstructing individual behavior from aggregate data.* Princeton University Press.

Koopman, J.S. and Longini, I.M. (1994). The ecological effects of individual exposures and nonlinear disease dynamics in populations. *American Journal of Public Health*, **84**, 836–42.

Koopman, J.S. and Lynch, J. (1999). Individual causal models and population system models in epidemiology. *American Journal of Public Health*, **89**, 1170–4.

Koopman, J.S., Longini, I.M., Jacquez, J.A., *et al.* (1991*a*). Assessing risk factors for transmisson of infection. *American Journal of Epidemiology*, **133**, 1199–209.

Koopman, J.S., Prevots, D.R., Vaca Marin, M.A., *et al.* (1991*b*). Determinants and predictors of dengue infection in Mexico. *American Journal of Epidemiology*, **133**, 1168–78.

Kreft, I. and deLeeuw, J. (1998). *Introducing multilevel modeling.* Sage, London.

Lazarsfeld, P.F. and Menzel, H. (1971). On the relation between individual and collective properties. In *A sociological reader on complex organizations* (ed. A. Etzioni), pp. 499–516. Holt, Rinehart, and Winston, New York.

Levin, B. (1995). Annotation: accounting for the effects of both group- and individual-level variables in community-level studies. *American Journal of Public Health*, **85**, 163–4.

Mason, W., Wong, G., and Entwisle, B. (1983). Contextual analysis through the multilevel linear model. In *Sociological methodology 1983–1984* (ed. S. Leinhardt), pp. 72–103. Jossey Bass, San Francisco, CA.

Morgenstern, H. (1982). Uses of ecologic analysis in epidemiologic research. *American Journal of Public Health*, **72**, 1336–44.

Morgenstern, H. (1995). Ecologic studies in epidemiology: concepts, principles, and methods. *Annual Review of Public Health*, **16**, 61–81.

Paterson, L. and Goldstein, H. (1991). New statistical methods for analysing social structures: an introduction to multilevel models. *British Educational Research Journal*, **17**, 387–93.

Piantadosi, S. (1994). Invited commentary: ecologic biases. *American Journal of Epidemiology*, **139**, 761–71.

Piantadosi, S., Byar, D.P., and Green, S.B. (1988). The ecological fallacy. *American Journal of Epidemiology*, **127**, 893–903.

Rice, N. and Leyland, A. (1996). Multilevel models: applications to health data. *Journal of Health Service Research Policy*, **1**, 154–64.

Riley, M.W. (1963). Special problems of sociological analysis. In *Sociological research I: a case approach*, pp. 700–25. Harcourt, Brace, and World, New York.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociology Review*, **15**, 351–7.

Ross, R. (1911). *The prevention of malaria* (2nd edn). John Murray, London.

Scheuch, E.K. (1969). Social context and individual behavior. In *Social ecology* (ed. M. Dogan and S. Rokkam), pp. 133–55. MIT Press, Cambridge, MA.

Schwartz, S. (1994). The fallacy of the ecological fallacy: the potential misuse of a concept and its consequences. *American Journal of Public Health*, **84**, 819–24.

Schwartz, S. and Carpenter, K. (1999). The right answer for the wrong question: consequences of type III error for public health research. *American Journal of Public Health*, **89**, 1175–80.

Snow, J. (1936). On the mode of communication of cholera. In *Snow on cholera* (2nd edn). Commonwealth Fund, New York (originally published 1855).

Susser, M. (1973). *Causal thinking in the health sciences.* Oxford University Press, New York.

Susser, M. (1994*a*). The logic in ecological. I: The logic of analysis. *American Journal of Public Health*, **84**, 825–9.

Susser, M. (1994*b*). The logic in ecological. II: The logic of design. *American Journal of Public Health*, **84**, 830–5.

Susser, E. and Wanderling, J. (1994). Epidemiology of nonaffective acute remitting psychosis versus schizophrenia: sex and sociocultural setting. *Archives of General Psychiatry*, **51**, 294–301.

Terris, M. (1964). *Goldberger on pellagra.* Louisiana State University Press, Baton Rouge, LA.

Townsend, P., Phillimore, P., and Beattie, A. (1988). *Health and deprivation. Inequality and the North.* Routledge, London.

Valkonen, T. (1969). Individual and structural effects in ecological research. In *Social ecology* (ed. M. Dogan and S. Rokkam), pp. 53–68. MIT Press, Cambridge, MA.

van den Eeden, P. and Huttner, H.J. (1982). Multi-level research. *Current Sociology*, **30**, 1–178.

Von Korff, M., Koepsell, T., Curry, S., and Diehr, P. (1992). Multi-level research in epidemiologic research on health behaviors and outcomes. *American Journal of Epidemiology*, **135**, 1077–82.

Wing, S., Casper, M., Riggan, W., *et al.* (1988). Socioenvironmental characteristics associated with the onset of decline of ischemic heart disease mortality in the United States. *American Journal of Public Health*, **78**, 923–6.

Wong, G. and Mason, W. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistics Association*, **80**, 513–24.

Wu, Y.-W. (1995). Hierarchical linear models: a multilevel data analysis technique. *Nursing Research*, **44**, 123–6.