Environmental Health

Research

Bias magnification in ecologic studies: a methodological investigation

Thomas F Webster*

Address: Dept. of Environmental Health, Boston University School of Public Health, 715 Albany Street, Boston, MA 02118, USA

Email: Thomas F Webster* - twebster@bu.edu

* Corresponding author

Abstract

Background: As ecologic studies are often inexpensive to conduct, consideration of the magnitude and direction of ecologic biases may be useful in both study design and sensitivity analysis of results. This paper examines three types of ecologic bias: confounding by group, effect measure modification by group, and non-differential exposure misclassification.

Methods: Bias of the risk difference on the individual and ecologic levels are compared using two-by-two tables, simple equations, and risk diagrams. Risk diagrams provide a convenient way to simultaneously display information from both levels.

Results: Confounding by group and effect measure modification by group act in the same direction on the individual and group levels, but have larger impact on the latter. The reduction in exposure variance caused by aggregation magnifies the individual level bias due to ignoring groups. For some studies, the magnification factor can be calculated from the ecologic data alone. Small magnification factors indicate little bias beyond that occurring at the individual level. Aggregation is also responsible for the different impacts of non-differential exposure misclassification on individual and ecologic studies.

Conclusion: The analytical tools developed here are useful in analyzing ecologic bias. The concept of bias magnification may be helpful in designing ecologic studies and performing sensitivity analysis of their results.

Background

Epidemiology is the study of health and disease in populations, but the standard for an observational study remains the individual level design, where we have information about outcome, exposure and covariates for each study subject [1]. This remains an ideal, although some designs mix group-level and individual-level variables in ways meant to enhance validity [2-5]. In practice, in the absence of better information, we often substitute an aggregate (group summary) value of some variable for each study subject. The extreme case is when aggregate values of exposure and outcome are used for every study variable. This is often called an ecologic study.

Resort to ecologic designs usually stems from the practical consideration that summary information is more easily obtained and more often available than individual-level data. Sometimes summary data are all that are available, and then, only in its crudest form, for example, that a certain percentage of a group of subjects is exposed (a group summary of an exposure variable) and a certain percentage of the same group has a specific health outcome (a

group summary of an outcome variable). In this case we have lost information about whether those with the outcome are the same as those who are exposed. Despite this information loss, it is tempting and plausible to say that we still have some useful information on risks of exposure.

Epidemiologists know that using ecologic designs (group level variables only) to make inferences about individual risks (individual level variables) can be seriously biased [e.g., [6,7]], but exactly how and when this bias occurs is often mysterious. In discussions of individual-level studies it is not enough to say a result might be confounded; one should consider the amount and direction of confounding. Given the potential value that ecologic studies have for obtaining information not otherwise readily available, it would seem useful to approach these studies in the same way, i.e., not dismiss them at the outset but instead try to describe the magnitude and direction of potential biases.

Here I apply this idea to ecologic studies, using individual-level studies as a reference. In particular, I will discuss the direction and extent of bias in ecologic studies compared with studies of individuals. This paper is meant to reveal underlying mechanisms with a simple model so practicing epidemiologists can begin to visualize what is happening when aggregate data are used. Among the many types of bias possible in ecologic studies [7], I will examine three of the most important: confounding by group, effect measure modification by group, and non-differential exposure misclassification.

Methods
Use of two-by-two tables
Theoretical problems are often best approached by starting simply and adding complications later. I focus here on closed cohorts with binary exposures and outcomes, using the risk difference as an effect measure. This approach allows us to see the ecologic inference problem at work using simple tools.

Individual outcome and exposure data are readily summarized by the *interior cells* of a two-by-two table, i.e., the joint distribution of exposure and outcome (Table 1). From these data we easily compute the risks of the exposed and unexposed as well as the risk difference.

The ecologic data are also visible on the *margins* of the table [8]. They provide the average exposure and average risk for the whole group but not the exposed and unexposed subjects *within* the group.
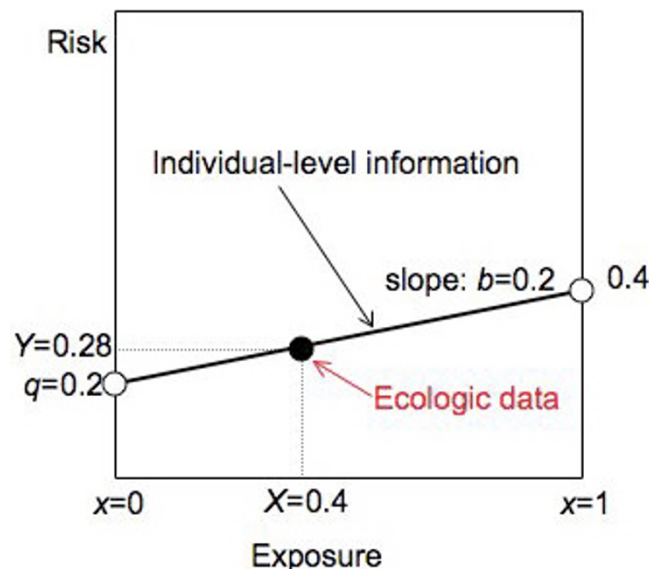
Table 1: Individual vs. ecologic data

	Individual (interior cells)			Ecologic (margins)		
	exposed	unexposed	sum	exposed	unexposed	sum
cases	16	12	28	?	?	28
noncases	24	48	72	?	?	72
total	40	60	100	40	60	100
risks	0.4	0.2		X	0.40	
RD	0.2			Y	0.28	
				n	100	

Individual-level information on exposure and outcome are shown by the interior of a two-by-two table; they are summarized by the risks in the unexposed and exposed and the risk difference (RD). Ecologic studies possess only the margins of the table, summarized by the average exposure X, average risk Y and group size n.

Risk diagrams and equations
We can depict the information in a two-by-two table using a *risk diagram*, a graphical device adapted from earlier work [9-11]. Figure 1 presents a risk diagram for the example of Table 1. Risk is plotted on the vertical axis, exposure on the horizontal. For binary exposures, we plot the risk in the unexposed (0.2) at $x = 0$ and the risk in the exposed (0.4) at $x = 1$. The line connecting these points has a slope equal to the risk difference: $(0.4-0.2)/(1-0) = 0.2$. This line summarizes essential individual-level information in a

Figure 1
Risk diagram illustrating Table 1. We summarize individual-level information for a group with a solid black line, ecologic data with a solid black dot. The line connects the risk in the unexposed ($q = 0.2$ at $x = 0$) with the risk in the exposed (0.4 at $x = 1$) and has slope equal to the risk difference b. The ecologic data are the average exposure X and average risk Y for the group.

two-by-two table: the risk difference, the risks in the exposed and the risks in the unexposed. We can represent the ecologic data for the table – the average exposure ($X = 0.4$) and the average risk ($Y = 0.28$) – by a large black dot. Thus risk diagrams show our knowledge about both levels, simultaneously: individual-level information is summarized by the line, ecologic data by the large dot.

For binary exposures, individual-level data only occur at exposures of zero and one, but it is convenient to think in terms of a continuous exposure. The simplest relationship would be a linear equation:

$$r_j = q + bx_j \qquad (1)$$

where r_j is the risk as a function of exposure x and j is an index for subjects. The intercept q is the risk in the unexposed, also called the background risk. The slope b is the risk difference. We call these equations (linear) risk functions. They describe risk as a function of exposure on the individual level.

The expected value of the binary outcome for individual y_j, considered as a probability, is equal to the person's risk, so that

$$y_j = q + bx_j + e_j \qquad (2)$$

where e_j is an error term. One can also think of the risks as proportions and the e_j as residuals (see appendix 1 for additional discussion of the model). Ordinary least squares regression of the individual-level data (x_j, y_j) in a two-by-two table can then be used to obtain the intercept q and risk difference b. We will see it is a useful tool for estimating ecologic bias. As discussed in appendix 2, this approach can be readily extended to rates, continuous outcomes (e.g., birth weights) and continuous exposures.

Since ecologic analyses only give us a single black dot for each two-by-two table, a collection of two-by-two tables is typically used. The idea is to extract information by examining how the outcome marginals vary as the exposure marginals change (e.g., how cancer rates change as the proportion of the population exposed to contaminated water changes across cities). This means we will usually be concerned with multiple tables, with each table describing a different group. We index the groups by the letter i:

$$y_{ij} = q_i + b_i x_{ij} + e_{ij} \qquad (3)$$

Since the background risk and risk difference may vary between groups, we must also add the index i to q and b.

Equations 1–3 describe individual-level models. In this paper, we will treat such models as a fixed reference for comparison with the results of ecologic inference. The fact that the q_i and b_i may differ between groups will prove critically important.
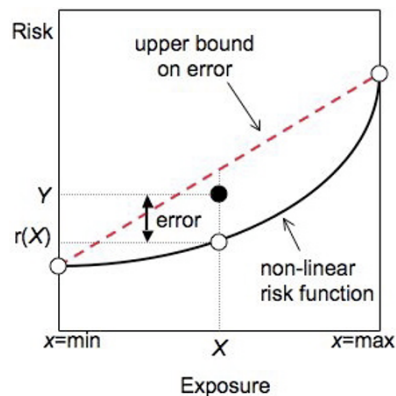
Linearity and aggregation

If the risk function is linear, as in Figure 1, then the dot describing the ecologic data must lie on the line describing the individual-level information. For binary exposures, this occurs because the dot represents a weighted average of the exposed and unexposed. More generally, this fact is a consequence of the aggregation theorem [7]. Mathematically, this means that if the risk function is linear, the group-level equations produced by aggregating individual-level equations will have the same form and the same parameters (appendix 1). For example, averaging equation 3 within each group yields

$$Y_i = q_i + b_i X_i \qquad (4)$$

where X_i and Y_i are, respectively, the average exposure and average risk in group i (the aggregate error or residual term can typically be ignored). Following Susser [12], capital letters X and Y refer to group-level variables, lower-case x and y refer to individual-level variables.

When the risk function is not linear, the equation describing aggregated data will generally not have the same form, and the ecologic data point will not lie on the risk function. Even if there are no other sources of bias, nonlinearity can thus cause trouble for ecologic studies – a problem called pure specification bias [7,9]. For example, suppose that exposure is a continuous variable and that the subjects in a group have the exposures denoted by the three open circles in Figure 2. The aggregate data, shown by the large dot, must then lie above the risk function (and below the line connecting the risks at the minimum and maximum exposures [11]). The average risk Y in the group is larger than the risk at the average exposure $r(X)$ when

Figure 2
Nonlinearity causes error during aggregation. The ecologic data point (X, Y) will generally not fall on the risk function when the latter is non-linear as shown here. The amount of error depends on the curvature of the risk function and the exposure distribution, but is bounded above by the line connecting the risks at the minimum and maximum exposures. This error can lead to pure specification bias.

the risk function is concave up. The difference between Y and $r(X)$ depends on the shape of the risk function and the exposure distribution within groups [7,9]. Log-linear models introduce bias terms that can be approximated using within-group variances (appendix 2).

We continue our exposition using linear risk functions, showing how confounding and effect modification between groups and exposure misclassification increase bias when variables are aggregated.

Loss of information and ecologic inference
The traditional goal of ecologic inference is to draw conclusions about individuals based on group-level data, or equivalently: to deduce the interior of a two-by-two table from its margins, to obtain the line in a risk diagram from the dot, or to estimate the risk difference from the ecologic data (the average risks Y_i and average exposures X_i). This goal runs into the fundamental problem that ecologic studies suffer from a loss of information [10]. In terms of two-by-two tables, many tables with very different interior cell contents can have the same margins. In terms of risk diagrams, many lines can go through the same ecologic data point (Figure 3).

A single dot is insufficient to determine a line. What if there were two or more dots, i.e., several two-by-two
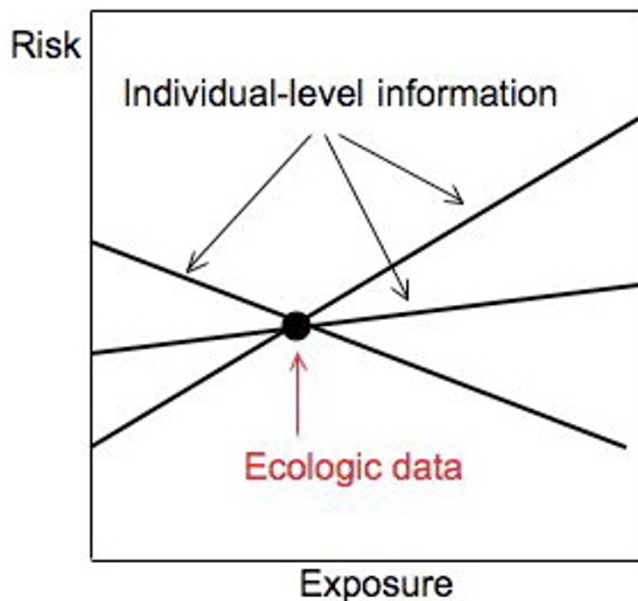
Figure 3
Loss of information is a fundamental problem of ecologic studies. Many sets of individual-level information (lines, interiors of two-by-two tables) generate the same ecologic data (dot, table margins). Only some possible lines are shown.

tables? Could we then recover the individual level information? The answer is "Yes," but only by making some very strong assumptions. If the assumptions are violated, large biases can occur.

Results
Epidemiologists often use regression of data from a number of groups for ecologic inference, regressing the average risk Y_i in each group against the average exposure X_i in each group. This approach is sometimes called ecologic or Goodman regression [10,13] (For a more formal treatment of ecologic regression and other methods, as well as ecologic bias, see [14]). We will use weighted least squares, weighting each group by its population n_i. Unweighted regression of ecologic data can cause an additional source of bias relative to individual-level analysis (appendix 3).

Ecologic regression *can* produce unbiased results. One way is to assume the individual-level model has the same background risk (intercept q) and risk difference (slope b) in every group:

$$y_{ij} = q + bx_{ij} + e_{ij} \qquad (5)$$

Aggregating yields the equation

$$Y_i = q + bX_i \qquad (6)$$

Ecologic regression then yields the correct estimate of the risk difference b. Assuming $q_i = q$ and $b_i = b$ is not the only way to achieve unbiased results, but it is the easiest to understand. In terms of risk diagrams, the lines describing the individual-level information in every group coincide. Since the dots representing the ecologic data all lie on this line, the ecologic regression reproduces the individual-level result.

But if q and b differ between the groups, things can go wrong. This difference corresponds to confounding and effect measure modification between groups. In an important paper, Greenland and Morgenstern [15] described these sources of ecologic bias. We use the analytic framework described above, the risk diagram and the elegant work of Palmquist [16] to show how the magnitude and direction of the ecologic bias from these sources affects biases present at the individual-level.

Confounding by group
Suppose two groups have the same risk difference b but different background risks, $q_0 \neq q_1$:

$$y_{0j} = q_0 + bx_{0j} + e_{0j} \qquad (7)$$
$$y_{1j} = q_1 + bx_{1j} + e_{1j}$$

Figure 4A is a risk diagram illustrating the example of Table 2. The lines describing the individual-level informa-
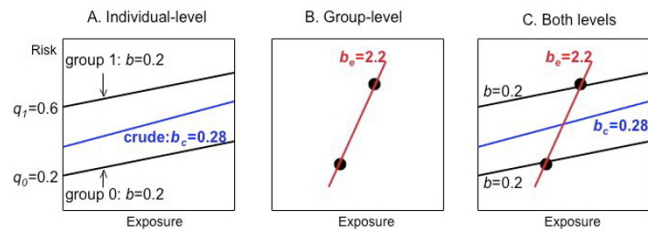
tion for the two groups are parallel because they have the same risk difference, but have different intercepts because the background risks are not equal. As the exposure distributions of the two groups also differ (since $X_0 \neq X_1$), ignoring groups causes confounding on the individual level.

The line describing the crude individual-level information in Table 2 (the table obtained from combining both groups) has a somewhat higher slope, i.e., confounding by group biased the crude risk difference upward (we use the word bias in an epidemiologic sense, the difference between an estimate and the correct value, b). If we know the individual-level data, including the variable describing group, we can prevent confounding by controlling for group, either by stratifying or adding group as a covariate in a regression.

Figure 4B shows the ecologic data (X_i, Y_i) for the two groups and the result of an ecologic regression. We know something has gone wrong, since a risk difference cannot exceed one, but we cannot determine the source of the problem from ecologic data alone. Unlike the individual case, we cannot control for group by stratifying or including an indicator variable in the regression: the ecologic data provide insufficient information for using these techniques (e.g., with only two ecologic data points, we cannot add a covariate to the ecologic regression).

Figure 4C plots the ecologic and crude individual-level results on the same risk diagram. The biases are in the same direction, but much larger for the ecologic study. Indeed, the ecologic bias is 25 times larger than the crude individual-level bias:

$$\frac{b_e - b}{b_c - b} = \frac{2.2 - 0.2}{0.28 - 0.2} = 25 \qquad (8)$$

Figure 4
Confounding by group, illustrating Table 2. **A)** Individual level: The solid black lines describing the individual-level information in the two groups are parallel (same risk differences b) but have different intercepts (different background risks $q_0 \neq q_1$). The crude estimate of the risk difference b_c is confounded (blue line). **B)** Group level: The ecologic estimate of the risk difference b_e is the slope of the red line through the two ecologic data points. Massive confounding has occurred, but we can't tell this from the ecologic data alone. **C)** Comparison of results on the two levels: The ecologic estimate of the risk difference b_e is much more biased than the crude individual-level estimate b_c. Both biases are in the same direction.

b_e and b_c are, respectively, the ecologic and crude individual-level estimates of the risk difference b. In appendix 3 we show that the relative amounts of bias due to confounding by group equals the exposure variance on the individual level divided by the exposure variance on the group level:

$$\frac{b_e - b}{b_c - b} = \frac{\text{var}[x_{ij}]}{\text{var}_B[X_i]} = 25 \qquad (9)$$

$\text{var}[x_{ij}]$ is the total exposure variance on the individual level and $\text{var}_B[X_i]$ is the exposure variance on the ecologic level (the between-group variance) weighted using the population of each group.

Table 2: Confounding by group

	Group 0			Group 1			Crude		
	expose	unexposed	sum	expose	unexpose	sum	expose	unexpose	sum
case	16	12	28	48	24	72	64	36	100
noncase	24	48	72	12	16	28	36	64	100
total	40	60	100	60	40	100	100	100	200
risk	0.40	0.20		0.80	0.60		0.64	0.36	
RD	0.20			0.20			0.28		
X_i	0.40			0.60					
Y_i	0.28			0.72					
n_i	100			100					

Since background risks and exposure distributions differ between the two groups, the crude individual-level estimate of the risk difference is confounded. The ecologic estimate of the risk difference, $(Y_1-Y_0)/(X_1-X_0) = 2.2$, is much more confounded. RD = risk difference; X_i and Y_i are the average exposure and average risk in group i; n_i = size of group i.

Bias magnification of confounding by group
We can rewrite equation 9 as

$$(b_e - b) = (b_c - b) M \qquad (10)$$

The amount of ecologic confounding by group (b_e - b) equals the amount of individual-level confounding by group (b_c - b) times a *magnification factor M*:

$$M = \frac{\mathrm{var}[x_{ij}]}{\mathrm{var_B}[X_i]} \qquad (11)$$

See Palmquist [16] for a closely related result.

Applying equations 10–11 to our example (Table 2 and Figure 4) shows that a moderate amount of confounding on the individual level (0.28-0.2 = 0.08) is magnified 25 times, producing a huge amount of confounding (2.2-0.2 = 2) on the ecologic level:

$$M = \frac{\mathrm{var}[x_{ij}]}{\mathrm{var_B}[X_i]} = \frac{0.25}{0.01} = 25$$
$$(2.2 - 0.2) = (0.28 - 0.2)25 \qquad (12)$$
$$2 = (0.08)25$$

Several conclusions follow immediately from equation 10. If there is no confounding by group on the individual level (b_c - b = 0), there is no confounding by group on the ecologic level: b_e - b = 0. Furthermore, since M is always positive, both biases are in the same direction. Suppose, as in the example, that we use the mean exposure in each group as the ecologic measure of exposure. M is then always *at least* one, i.e., the amount of confounding by group on the ecologic level equals or exceeds the amount on the individual level (appendix 4). (Note that the derivation of (10) assumes that $\mathrm{var_B}[X_i]$ is non-zero. When this assumption is violated, as occurs when the X_i are equal in all groups, there is no confounding by group on the individual level. However, ecologic regression is uninformative since division by zero makes b_e and M undefined).

Equation (10) tells us that the relative amount of confounding by group on the ecologic and individual level stems from the reduction of exposure variance caused by aggregation. The information loss from discarding within-group exposure variance magnifies the bias already present on the crude individual level. However, if exposure within groups is homogeneous, i.e., everyone within a group has the same exposure, M equals one and the amount of confounding by group is equal on the ecologic and individual levels. This formalizes, for one source of bias, the simple idea that ecologic studies with homogeneous exposures are really just individual-level studies.

Changes in the background risks (q_i) and average exposures (X_i) have different implications for confounding on the individual and group levels. As shown in Figure 5, keeping average exposures the same but making the background risks more similar decreases confounding by group on both the individual and group levels. However, keeping background risks the same but making average exposures more similar (decreasing $\mathrm{var_B}[X_i]$) decreases confounding by group on the individual level but increases it on the ecologic level: the decrease in b_c - b is outweighed by the increase in M. Thus even a little confounding on the individual level can produce a lot on the ecologic level.

Effect modification of the risk difference by group
Suppose two groups have the same background risk (q) but different risk differences ($b_0 \neq b_1$):

$$\gamma_{0j} = q + b_0 x_{0j} + e_{0j} \qquad (13)$$
$$\gamma_{1j} = q + b_1 x_{1j} + e_{1j}$$

The lines describing the two groups in Figure 6 have the same intercept but different slopes. As Figure 6 and Table 3 illustrate, the crude individual-level risk difference b_c lies between the b_i of the two groups, as it must for binary exposures (see appendix 5). In contrast, the ecologic estimate of the risk difference b_e is wildly biased, not even having the same sign.

We can use equation 10 with one small change to describe the implications of effect modification of the risk difference by group. Since b is no longer constant, we use b_w, a weighted average of the risk differences b_i in the groups with weights depending on the within-group exposure variances (b_w is also obtained by regressing the individual-level data while adjusting for group; see appendix 3):

$$(b_e - b_w) = (b_c - b_w) M \qquad (14)$$

In the example, b_w is approximately 0.296 (appendix 6). Allowing for rounding error, applying equation 14 yields

$$M = \frac{\mathrm{var}[x_{ij}]}{\mathrm{var_B}[X_i]} = \frac{0.2475}{0.0025} = 99$$
$$(-1.5 - 0.296) = (0.278 - 0.296)99 \qquad (15)$$
$$-1.796 = (-0.018)99$$

As Figure 6 illustrates, the tiny discrepancy between the crude risk difference b_c and the weighted average b_w is multiplied by a large magnification factor of 99, producing a large bias on the ecologic level. The difference between b_c and b_w is not usually considered a bias on the individual level. For individual-level studies, some epidemiologists might report the b_i if they consider the variation between groups important; others might ignore it. While b_w is not
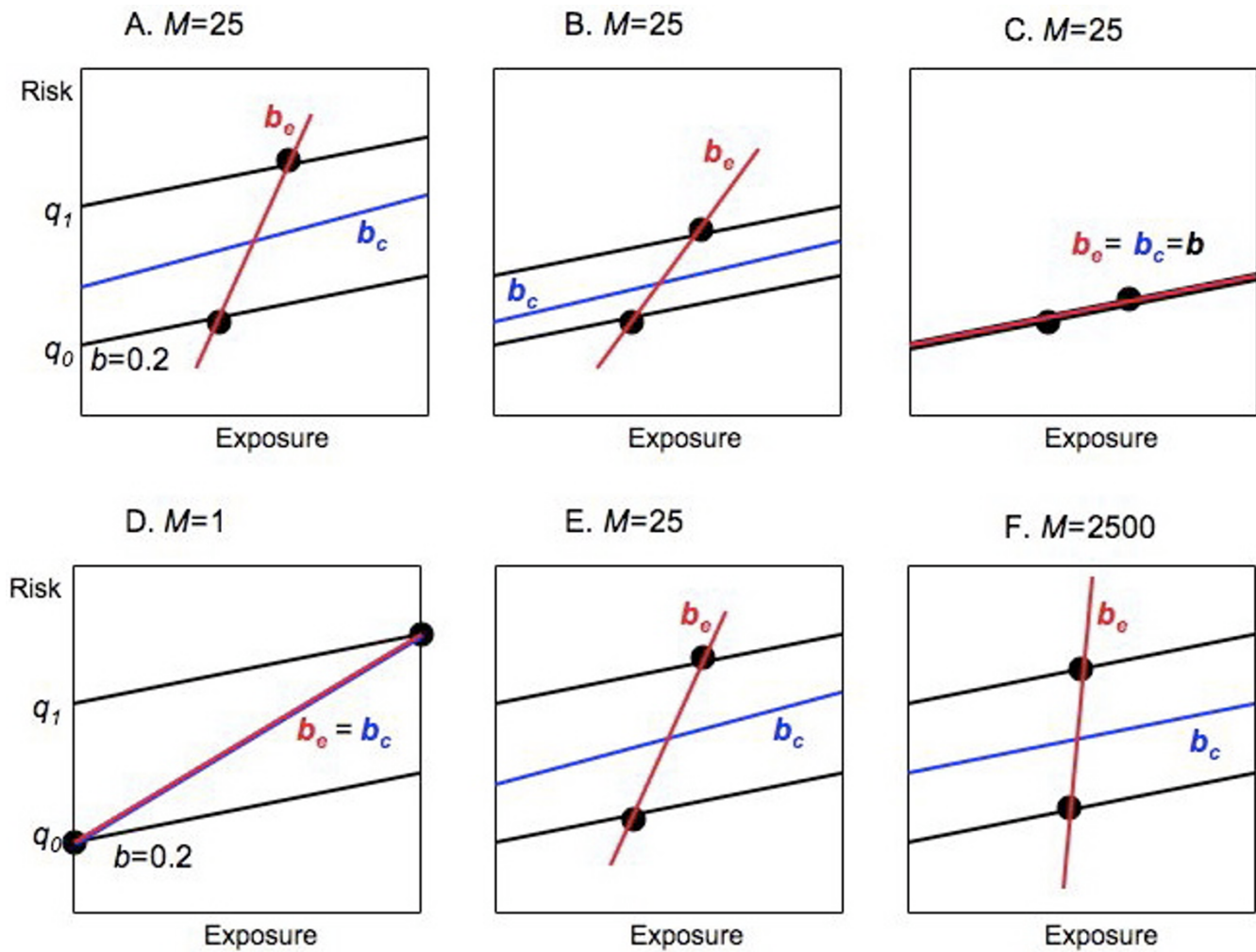
Figure 5
Confounding by group on the individual and group level. **A, B, C)** Suppose average exposures are the same, but the difference between the background risks (q_i) decreases. Confounding by group decreases on both the individual and group levels with constant proportionality factor *M*. **D, E, F)** Suppose background risks (q_i) are the same, but the difference between the average exposures decreases. Confounding by group decreases on the individual level, but increases on the ecologic level because of the large increase in *M*.

a commonly used data summary of the b_i, it helps us understand a source of ecologic bias when there is effect modification of the risk difference by group.

Magnification factor
The magnification factor is the linchpin of the mechanism. As Figure 5D–F illustrates, if the exposure distribution changes so that *M* increases, the amount of ecologic bias can increase dramatically. Another example may provide a better feel for the magnification factor. In Figure 7, we keep the average exposures (X_i) the same in all three cases; the between-group variance ($\text{var}_B[X_i]$) thus remains constant. Changing the within-group exposure distribution from binary to homogeneous reduces the within-

group variance ($\text{var}[x_{ij}]$). As a result, *M* decreases from 25 to 1.

Bias magnification
The *bias magnification equation* (equation 14), governs bias from both sources – confounding by group and effect modification of the risk difference by group – in an additive fashion, i.e., it can be applied to both sources of bias separately or together [11]. Application of the bias magnification equation to these sources of ecologic bias brings together two lines of research. Greenland and Morgenstern showed that both confounding by group and effect measure modification by group could cause ecologic bias [15]. The bias magnification equation can be derived by
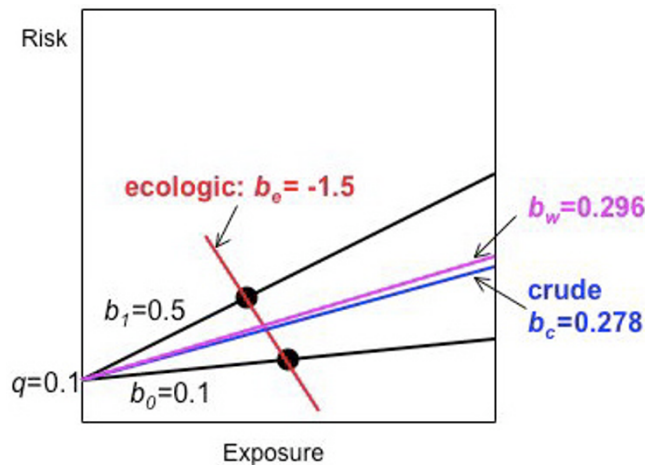
Table 3: Effect modification of the risk difference by group

|	Group 0			Group 1			Crude		
	expose	unexpose	sum	expose	unexpose	sum	expose	unexpose	sum
case	20	10	30	48	12	60	68	22	90
noncase	80	90	170	32	108	140	112	198	310
total	100	100	200	80	120	200	180	220	400
risk	0.2	0.1		0.6	0.1		0.378	0.1	
RD	0.1			0.5			0.278		
X_i	0.5			0.4					
Y_i	0.15			0.3					
n_i	200			200					

The background risks in each group are the same but the risk differences vary. The crude individual-level risk difference lies between the two extremes, as it must when exposure is binary. The ecologic estimate of the risk difference, $(Y_1-Y_0)/(X_1-X_0)$= -1.5, is extremely biased. RD = risk difference; X_i and Y_i are the average exposure and average risk in group i; n_i = size of group i.

partitioning covariance and variance within and between groups (appendix 3). This approach has a distinguished history, only some of it mentioned here. Robinson's landmark 1950 paper [17] discussed such partitions in terms of correlation coefficients. Duncan *et al.* discussed regression coefficients [18]. Piantadosi *et al.* derived the bias magnification equation, but they did not emphasize the magnification factor or discuss the role of effect measure modification by group [19]. Palmquist derived a general-ized form of a closely related equation using matrix methods [16]. Palmquist's insightful work, discussed by King [10], stresses the role of the inflation factor – the magnification factor minus one – and its effect on individual-level bias (appendix 3). Palmquist and King do not discuss the individual-level bias (which they call the specification shift) in terms of confounding and effect measure modification, in part because these authors are social scientists. King considers b_c as his parameter of interest, biased *by* grouping. In our context, which is epidemiology, the

Figure 6
Effect modification of the risk difference by group, illustrating Table 3. The solid black lines describing the individual-level information for the two groups have the same intercept (background risk q) but different slopes (risk differences $b_0 \neq b_1$). The crude estimate of the risk difference b_c (blue line) lies between these two extremes. Relative to b_w, the ecologic estimate of the risk difference b_e (red line) is far more biased than the crude individual-level estimate b_c. Both biases are in the same direction. b_w (purple line) is the weighted average of the risk differences used in the bias magnification equation.
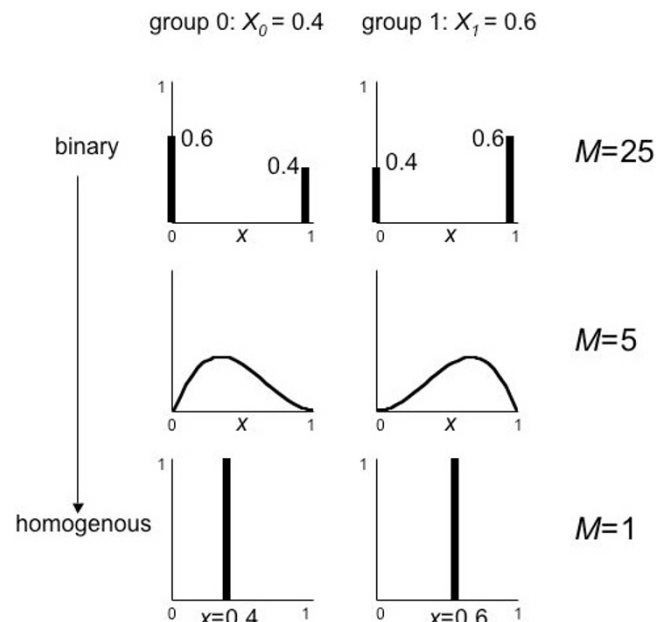
Figure 7
Effect on M of different within-group exposure distributions. The magnification factor M decreases when the within-group exposure variance is reduced, keeping the between-group variance constant (the X_i do not change between rows).

crude risk difference b_c is considered biased by *ignoring* groups.

Non-differential misclassification of binary exposure

Here, non-differential exposure misclassification (NDEM) means that the proportion of people misclassified by exposure does not depend on disease status. Sensitivity s refers to the proportion of exposed people classified as exposed; specificity t means the proportion of non-exposed people classified as non-exposed (More general definitions may regard s and t as probabilities). NDEM causes bias towards the null in individual-level studies, but away from the null in ecologic studies [20]. This difference has been called one of the most significant problems of ecologic studies [6], so we conclude this exposition with an explanation of the mechanism in this simple case.

We compute misclassified individual and group level data for a two by two table as shown in Table 4. Since the misclassification is non-differential, the average exposure is affected but the average risk Y_i is not changed. This is the key to the effect. The average exposure U_i in the misclassified table is given by

$$U_i = \lambda X_i + (1 - t) \quad (16)$$

$$\lambda = s + t - 1 \quad (17)$$

where λ is *Youden's index* [1] (for details of this section, see appendix 7). Since sensitivity and specificity must be between zero and one, λ must be between -1 and 1. Sensitivity and specificity are typically greater than 0.5 (i.e., better than random), so we assume λ is between 0 and 1.

With this background, it is easy to show that ecologic studies are biased away from the null when sensitivity and specificity are the same in every group [20]. The ecologic estimate of the risk difference for the misclassified data (b_e') equals

$$b_e' = \frac{b_e}{\lambda} \quad (18)$$

When λ is between 0 and 1, b_e' is farther away from the null than the true ecologic estimate and has the same sign.

Figure 8 shows how NDEM works in this simple model. Suppose we have an ecologic study of two groups with no other sources of bias (Table 5). The true ecologic data lie on top of the line describing the underlying individual level information. NDEM has no effect on the average risks Y_i, so the height of the dots describing the ecologic data stays the same. NDEM causes the misclassified average exposures U_i to move closer together and towards the center compared with the true average exposures X_i. Consequently, the regression line through the ecologic data is steeper, i.e., biased away from the null. The different effects of NDEM on the individual and group levels is ultimately due to the loss of information caused by aggregation. While average risks are unaffected, aggregation brings the average exposures closer together, i.e., the same risks are produced by a narrower range of exposures, resulting in a steeper slope.

It is important to note that not all forms of exposure measurement error will bias ecologic studies away from the null. We examined a particular error model above: NDEM of binary exposure at the individual level with sensitivities and specificities not changing between groups. Other error models lead to other results [11]. For example, application of the classical error model to a continuous exposure variable biases results toward the null in ecologic studies.

Discussion

Roughly speaking, the bias magnification equation says that ecologic bias equals individual-level bias magnified. More precisely, the reduction in exposure variance caused

Table 4: Non-differential exposure misclassification in a 2 × 2 table

	Correct			Misclassified		
	expose	unexpose	sum	expose	unexpose	sum
Case	a_i	b_i	a_i+b_i	$sa_i + (1-t)b_i$	$(1-s)a_i+tb_i$	a_i+b_i
noncase	c_i	d_i	c_i+d_i	$sc_i + (1-t)d_i$	$(1-s)c_i+td_i$	c_i+d_i
Total	a_i+c_i	b_i+d_i	n_i	$s(a_i+c_i) + (1-t)(b_i+d_i)$	$(1-s)(a_i+c_i) + t(b_i+d_i)$	n_i
X_i	$(a_i+c_i)/n_i$			$s(a_i+c_i)/n_i + (1-t)(b_i+d_i)/n_i$		
Y_i	$(a_i+b_i)/n_i$			$(a_i+b_i)/n_i$		

Assume that the proportion of people misclassified does not depend on disease status. Sensitivity s is the fraction of exposed people classified as exposed; specificity t is the fraction of unexposed people classified as unexposed. On the group level, the average risk remains the same but the average exposure changes. X_i and Y_i are the average exposure and average risk in the group; n_i = size of the group.
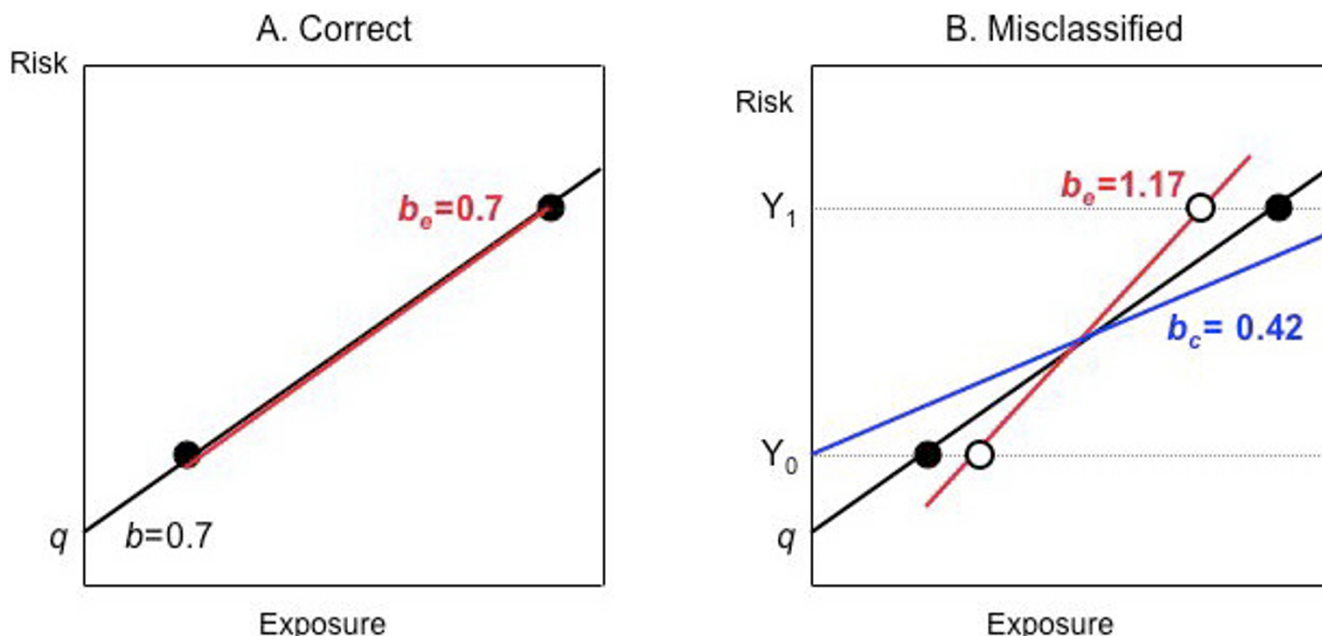
Figure 8
Non-differential exposure misclassification (NDEM), illustrating Table 5. **A**. If there are no other sources of bias, the ecologic- and individual-level analyses of the correct data are the same. **B**. Suppose the dichotomous exposure data are misclassified with the same sensitivity and specificity in each group. Then the individual-level result (blue) is biased toward the null and the ecologic result (red) is biased away from the null. The average risks (Y_i) in each group are unchanged but the average exposures move closer together. This causes the resulting ecologic regression line to have higher slope.

by aggregation (loss of information) magnifies the individual-level bias due to confounding by group and/or effect modification of the risk difference by group. Other things equal, the magnification factor is maximized if exposure within groups is dichotomous [11]. Thus textbook examples, which typically use two-by-two tables, tend to overstate the amount of bias magnification occurring in many real studies.

Bias magnification provides a useful tool for theoretical considerations of ecologic bias. Does it have any practical use? When designing ecologic studies, one should try to minimize M by increasing between-group differences in exposure while making within-group exposure as homogeneous as possible; see also [7]. Bias magnification also suggests an approach to sensitivity analysis of ecologic bias from confounding by group and/or effect modification of the risk difference by group. Assume an individual-level model as a reference. The ecologic bias has two components: the amount of bias caused by ignoring group on the individual-level (b_c - b_w), and the magnification of this bias caused by aggregation (M). When analyzing ecologic data, we will not know the size of the bias on the individual level, but we can make various assumptions. For example, we may be able to make educated guesses about the direction and possible amount of confounding by

group on the individual level. We may then be able to estimate the magnification factor. For binary exposures, we can compute M from the ecologic data alone (appendix 8). In other situations we may be able to estimate M from samples of the study population or from routinely collected environmental data. For example, we might use air pollution measurements and spatial statistics to estimate the variation in exposure within cities, comparing it to variation in average air pollution between cities. If M is small, we gain confidence that the ecologic bias isn't too different from the bias on the individual level. See [21] for another approach to sensitivity analysis of ecologic bias.

Ecologic studies using exposure variables of the type "fraction exposed" may be particularly problematic. Such ecologic exposure variables are the aggregated form of binary exposures on the individual level. Other things equal, use of such variables tends to maximize the magnification factor (increasing any bias present due to confounding by group or effect measure modification by group), increase bias in non-linear models, and bias results away from the null when non-differential exposure misclassification occurs. Studies with small variation between average exposures (as in Figure 5F) are particularly worrisome.

The approach discussed here examines three important sources of ecologic bias: confounding by group, effect modification of the risk difference by group, and non-differential exposure misclassification. It does not take into account other potential problems, e.g., confounding within groups or effect measure modification within groups (However, confounding within groups alone does not cause ecologic bias when the relationship between exposure and outcome is linear; see appendix 9. The plausibility of the linearity assumption must be carefully considered).

In this paper we have employed a simple, abstract approach based on several assumptions: availability of the same types of information on both individual and group levels, use of average group exposure as the ecologic exposure measure, linear models, weighted least squares regression, estimation of the risk difference. Our purpose was to display, as simply as possible, the underlying mechanisms causing the magnification of individual bias upon aggregation into groups. The loss of information about within-group exposure variance is seen to be the culprit.

The ideas here can be extended, with some modifications of results, to examine additional issues: inclusion of covariates, confounding and effect measure modification within groups, group-level exposure measures other than means, and partially ecologic studies [4,11]. This paper has focused on the risk difference, but the results can be readily applied to the rate difference and studies with continuous outcomes (appendix 3). Generalization to the more commonly used relative risks and rate ratios is underway (see also [21]).

In addition to theoretical investigations like this one, we also need to know more about the amount of ecologic bias encountered in practice [7,22-24]. By helping us focus on how ecologic studies go astray, we hope to move toward the goal of domesticating ecologic bias: treating it as another source of epidemiologic bias that needs to be analyzed and quantified [22].

Mathematical appendix
1. Linear risk functions and aggregation
Assume risk to individuals is a linear function of exposure x

$$r_{ij} = q_i + b_i x_{ij} \tag{A1}$$

We index groups with i and subjects with j, and allow the background risk q_i and risk difference b_i to vary between groups (Alternatively, one can think in terms of an unmeasured group-level covariate Z_i: $q_i = q + \gamma Z_i$, $b_i = b + \eta Z_i$). Conceiving of risks as probabilities of developing

disease, the expected value of the binary outcome for individual y_{ij} is equal to their risk, so that

$$y_{ij} = q_i + b_i x_{ij} + e_{ij} \tag{A2}$$

where e_{ij} is an error term. Averaging within groups (of size n_i) produces the aggregate equation

$$\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n_i}\sum_{j=1}^{n_i}(q_i + b_i x_{ij} + e_{ij}) = q_i + b_i\left(\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}\right) + \frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij} \tag{A3}$$

$$Y_i = q_i + b_i X_i + e_i \tag{A4}$$

X_i and Y_i are the average exposure and outcome per group (See [14,21] for an in-depth discussion of statistical models in ecologic studies). The e_{ij} and e_i vanish under expectation.

Alternatively, we can think of the risks as proportions. Under the proportion model, the e_{ij} and e_i are residuals (e_i then vanishes in equation A4 because the sum of residuals within groups equals zero). Both the probability and proportion interpretations can be applied in this paper: the proportion model may help in thinking about ecologic bias in particular data sets. For example, the individual-level model(s) assumed for an analysis need not be correct; instead it is a reference against which we measure ecologic bias.

We assumed a linear risk model (A2) at the individual level. Although very simple, it still yields insight into many aspects of ecologic bias (Non-linear functions, described below, add some additional features). Linear risk models can also be easily analyzed using ordinary least squares (OLS). The risk difference (b_i) is the natural effect measure to use in this situation. For example, applying OLS to the individual-level data in a two-by-two table yields q_i and b_i (OLS picks the line that runs through the mean values of the outcomes at the two exposure levels). Although OLS is not commonly used for analyzing binary outcome data, it simplifies the understanding of ecologic bias. OLS, and other methods described here, are directly applicable to many studies of continuous outcomes.

While we chose to use a linear risk model, the relationship between the average outcome in a group (Y_i) and the average exposure (X_i) in two-by-two tables must be linear [10,11]; using the notation in Table 6, the average risk in the group is given by:

$$Y_i = \frac{p_i m_{i1} + q_i m_{i0}}{n_i} = p_i X_i + q_i(1 - X_i) = q_i + (p_i - q_i)X_i$$
$$= q_i + b_i X_i \tag{A5}$$

Table 5: Effect of non-differential exposure misclassification on individual and ecologic studies

	Group 0			Group 1			Crude		
Correct									
	expose	unexpose	sum	expose	unexpose	sum	expose	unexpose	sum
cases	160	80	240	720	10	730	880	90	970
noncases	40	720	760	180	90	270	220	810	1030
total	200	800	1000	900	100	1000	1100	900	2000
risk	0.8	0.1		0.8	0.1		0.8	0.1	
RD	0.7			0.7			0.7		
X_i	0.20			0.90					
Y_i	0.24			0.73					
Misclassify									
	expose	unexpose	sum	expose	unexpose	sum	expose	unexpose	sum
cases	144	96	240	578	152	730	722	248	970
noncases	176	584	760	162	108	270	338	692	1030
total	320	680	1000	740	260	1000	1060	940	2000
risk	0.45	0.14		0.78	0.58		0.68	0.26	
RD	0.31			0.20			0.42		
X_i	0.32			0.74					
Y_i	0.24			0.73					

The data are misclassified assuming the same sensitivity and specificity in each group, both equal to 0.8. Since there are no other sources of bias in this example, the crude individual and ecologic estimates of the RD are identical (0.7) for the correct data. For the misclassified data, the crude individual level estimate of the RD (0.42) is biased toward the null while the ecologic estimate (1.17) is biased away from the null. RD = risk difference; X_i and Y_i are the average exposure and average risk in group *i*.

Equation A5 is identical to A4 (except for e_i).

2. Extensions; non-linear risk functions and pure specification bias

Equations 1–3 (in the main text) and Figure 1 were constructed to describe two-by-two tables and risks, but the approach is readily extended to rates, continuous outcomes and continuous exposures. Instead of modeling risks in a closed cohort, we can also examine incident cases in person-time [11]. In this type of individual-level study, we would know the interior of the table (the number of exposed and unexposed cases); we could therefore compute the rates for the exposed and unexposed as well as the rate difference. In an ecologic study of this type, we would know only the marginal rate and the marginal exposure distribution. Rate diagrams are very similar to risk diagrams except that the *y* axis has no upper bound. For continuous outcomes y_{ij} with normally distributed errors e_{ij}, OLS is a conventional model of analysis; *b* is then the change in outcome per unit change in exposure. Instead of restricting x_{ij} to zero or one as in a binary exposure, we can also let x_{ij} be a continuous measure of exposure. For equations 1–3 (and Figure 1) to hold, the risk

function would have to be linear but non-linear functions are also possible.

Instead of the linear risk function (A2), suppose we assume the following log-linear model:

$$y_{ij} = \exp[q + bx_{ij}] + e_{ij} \qquad (A6)$$

Aggregating yields (ignoring error terms)

$$Y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \exp[q + bx_{ij}] \qquad (A7)$$

Equation A7 is generally not equal to $\exp[q+bX_i]$, i.e., for non-linear models the individual-level and aggregate models do not have the same functional form. This discrepancy is the source of pure specification bias [7,9]. If x_{ij} is normally distributed within groups, then applying expectations and cumulants to (A6) yields:

$$Y_i = \exp\left[q + bX_i + \frac{b^2}{2}\sigma_i^2\right] \qquad (A8)$$

where σ_i^2 is the exposure variance within group i [9]. More generally, we can approximate (A8) by applying Taylor series and aggregation to (A6) (see also [25-27]).

Assuming that counts of cases fit a Poisson model, one can fit ecologic data using (A8). One can, however, gain some insight into pure specification bias using an approximation. Linearly regressing X_i against $\log[Y_i]$ yields:

$$b_e \approx b + \frac{b^2}{2}\frac{\mathrm{cov}[X_i,\sigma_i^2]}{\mathrm{var}[X_i]} \qquad (A9)$$

The exponential of b_e now estimates the relative risk. As shown by (A9), in the absence of other sources of ecologic bias, log-linear ecologic regression is subject to pure specification bias, approximated by the second term in (A9). For such models, the within-group exposure variances can often be expected to covary with average exposures. There is little or no bias if b is small (low curvature), the within group exposure variance (σ_i^2) does not depend on X_i, or exposure is uniform within groups ($\sigma_i^2 = 0$), results consistent with those found by Richardson *et al.* [9] for the normal distribution case. Risk diagrams drawn using log-transformed risks turn exponential risk functions into straight lines; the line describing the upper bound of the error in Figure 2 becomes curved (bowing downward).

3. Bias magnification equation

Assume the individual-level model of equation A2. The crude individual-level and population-weighted ecologic estimates of the risk difference are, respectively:

$$b_c = \frac{\mathrm{cov}[y_{ij},x_{ij}]}{\mathrm{var}[x_{ij}]} = \frac{C_T}{V_T} \qquad (A10)$$

$$b_e = \frac{\mathrm{cov}_B[Y_i,X_i]}{\mathrm{var}_B[X_i]} = \frac{C_B}{V_B} \qquad (A11)$$

V_T and V_B are the total and between-group exposure variances (This derivation assumes V_T and V_B are non-zero). C_T and C_B are the total and between-group covariances of outcome and exposure. C_B and V_B are weighted by group size n_i:

$$C_B = \mathrm{cov}_B[X_i,Y_i] = \frac{1}{n}\sum_{i=0}^{N-1} n_i X_i Y_i - (\bar{X})(\bar{Y})$$
$$(A12)$$

Table 6: Notation for a general two-by-two table

	exposed	unexposed	sum
cases	$p_i m_{i1}$	$q_i m_{i0}$	$p_i m_{i1}+q_i m_{i0}$
noncases	$(1-p_i)m_{i1}$	$(1-q_i)m_{i0}$	$(1-p_i)m_{i1} + (1-q_i)m_{i0}$
total	m_{i1}	m_{i0}	$n_i = m_{i1}+m_{i0}$
risks	p_i	q_i	
RD	p_i-q_i		
X_i			m_{i1}/n_i
Y_i			$(p_i m_{i1}+q_i m_{i0})/n_i$

Let p_i and q_i be the risks of the exposed and unexposed in group i. Let m_{i1} and m_{i0} be the numbers of exposed and unexposed in group i. RD = risk difference; X_i and Y_i are the average exposure and average risk in group i; n_i = size of group i.

$$V_B = \mathrm{var}_B[X_i] = \frac{1}{n}\sum_{i=0}^{N-1} n_i X_i^2 - (\bar{X})^2 \qquad (A13)$$

where N is the number of groups, n is the total population, and $\bar{X}$ and $\bar{Y}$ are the overall means. Expand equation A10, partitioning the total covariance and variance into within-group and between-group pieces [18,19]:

$$\begin{aligned} b_c &= \frac{C_T}{V_T} = \frac{C_W + C_B}{V_T} \\ &= \frac{C_W}{V_W}\frac{V_W}{V_T} + \frac{C_B}{V_B}\frac{V_B}{V_T} \qquad (A14) \\ &= b_w\frac{V_W}{V_T} + b_e\frac{V_B}{V_T} \end{aligned}$$

C_W and V_W are the within-group covariance and variance. Equation A14 shows that the crude individual estimate of the risk difference b_c is a weighted average of the within-group and ecologic estimates: $b_w \le b_c \le b_e$ since $V_W + V_B = V_T$ and variances are nonnegative. b_w, equal to C_W/V_W, is the within-group individual-level estimate of the risk difference. It is a weighted average of the b_i with weights equal to $n_i\,\mathrm{var}_i(x_{ij})$, where $\mathrm{var}_i(x_{ij})$ is the exposure variance within group i. b_w is also the result of ordinary least squares regression of the individual-level data, adjusted for group using indicator variables [11,16]. After substituting $V_W = V_T - V_B$, a little algebra yields the bias magnification equation:

$$(b_e - b_w) = (b_c - b_w)\frac{V_T}{V_B} = (b_c - b_w)M \qquad (A15)$$

Subtracting (b_c - b_w) from both sides of equation A15 yields:

$$(b_e - b_c) = (b_c - b_w)(M - 1) = (b_c - b_w)\,F \qquad (A16)$$

where $F = M - 1$ is called the *inflation factor* [10,16]. Equation A16 measures the amount of bias added by aggregation (Figure 9).

The simplified equation describing bias magnification of confounding by group (equation 10 in the main text) can be derived in another way: insert $y_{ij} = q_i + bx_{ij} + e_{ij}$ into equation A10 and expand; insert $Y_i = q_i + bX_i + e_i$ into equation A11 and expand:

$$(b_c - b) = \frac{\text{cov}[q_i, x_{ij}]}{\text{var}[x_{ij}]} \qquad (A17)$$

$$(b_e - b) = \frac{\text{cov}_B[q_i, X_i]}{\text{var}_B[X_i]} \qquad (A18)$$

Dividing equation A18 by equation A17 yields:

$$\frac{(b_e - b)}{(b_c - b)} = \frac{\text{cov}_B[q_i, X_i]}{\text{cov}[q_i, x_{ij}]} \frac{\text{var}[x_{ij}]}{\text{var}_B[X_i]} = M \qquad (A19)$$

Note that $\text{cov}[q_i, x_{ij}] = \text{cov}_B[q_i, X_i]$, a consequence of using population weighting. Confounding by group is absent if $\text{cov}_B[q_i, X_i] = 0$, i.e., the background risks and average exposures are uncorrelated (e.g., $q_0 = q_1$ for two 2 × 2 tables).

When M equals one – i.e., exposure within groups is homogeneous – equation A15 shows that the ecologic and individual-level results are equal. This result, desirable for the nominally ecologic study, is a consequence of using population weighting; unweighted ecologic regression will generally produce different results from the individual-level analysis. For example, in (A19) the weighted and unweighted covariances of X_i and q_i will usually not be equal [11].

4. M ≥ 1 when X_i is the ecologic exposure measure
Suppose we use the mean exposure in each group (X_i) as our ecologic measure of exposure. Partitioning total exposure variance within and between groups shows that M is always at least one:

$$M = \frac{\text{var}[x_{ij}]}{\text{var}_B[X_i]} = \frac{V_T}{V_B} = \frac{V_W + V_B}{V_B} \geq 1 \qquad (A20)$$

5. Effect measure modification and b_c
Using the notation in Table 6, the crude risk difference is

$$b_c = \frac{\sum_i p_i m_{i1}}{\sum_i m_{i1}} - \frac{\sum_i q_i m_{i0}}{\sum_i m_{i0}} \qquad (A21)$$

where p_i is the risk in the exposed, q_i is the risk in the unexposed and we sum over groups i.

To consider effect modification of the risk difference, assume $q_i = q$ and $p_i = q + b_i$. Substituting into (A21), we obtain

$$b_c = \frac{\sum_i b_i m_{i1}}{\sum_i m_{i1}} = \sum_i b_i w_i \qquad (A22)$$

where the w_i are non-negative weights. Thus b_c must be between the minimum and maximum values of b_i. For additional discussion of when b_c and b_e are bounded, see [11].

6. Computation of b_w for Table 3, Figure 6
b_w is a weighted average of the b_i using weights equal to group size times the exposure variance within the group:

$$w_0 = n_0 \, \text{var}_0(x_{0j}) = 200(X_0)(1-X_0) = 200(0.5)(0.5) = 50 \qquad (A23)$$
$$w_1 = n_1 \, \text{var}_1(x_{1j}) = 200(X_1)(1-X_1) = 200(0.4)(0.6) = 48 \qquad (A24)$$

$$b_w = \frac{w_0 b_0 + w_1 b_1}{w_0 + w_1}$$
$$= \frac{50(0.1) + 48(0.5)}{50 + 48} \approx 0.296 \qquad (A25)$$

7. Non-differential exposure misclassification
From Table 4, the average exposure in the misclassified two by two table is

$$
\begin{aligned}
U_i &= \frac{s(a_i + c_i) + (1-t)(b_i + d_i)}{n_i} \\
&= s\frac{(a_i + c_i)}{n_i} + (1-t)\frac{(n_i - a_i - c_i)}{n_i} \\
&= sX_i + (1-t)(1 - X_i) \\
&= (s + t - 1)X_i + (1-t) \\
&= \lambda X_i + (1-t)
\end{aligned}
\qquad (A26)
$$

where $\lambda = s+t-1$. Assuming s and t are the same in all groups, the ecologic estimate of the RD for the misclassified data is

$$
\begin{aligned}
b'_e &= \frac{\mathrm{cov}_B[Y_i, U_i]}{\mathrm{var}_B[U_i]} \\
&= \frac{\mathrm{cov}_B[Y_i, \lambda X_i + (1-t)]}{\mathrm{var}_B[\lambda X_i + (1-t)]} \\
&= \frac{\lambda \, \mathrm{cov}_B[Y_i, X_i]}{\lambda^2 \, \mathrm{var}_B[X_i]} \\
&= \frac{b_e}{\lambda}
\end{aligned} \qquad \text{(A27)}
$$

Since $0 \le \lambda \le 1$, the variance of U_i is smaller than the variance of the X_i, i.e., the U_i are closer together. If we think of equation A26 as an iteration equation, then NDEM moves the average exposure one step closer to the stationary point given by solving $X_c = \lambda X_c + (1\text{-}t)$:

$$
X_c = \frac{1-t}{1-\lambda} = \frac{1-t}{(1-s)+(1-t)} \qquad \text{(A28)}
$$

For $s = t$, $X_c = 0.5$.

In the absence of any other sources of bias except NDEM, one can show that the individual-level estimate of the RD is given by

$$
\frac{b_c}{b} = \lambda \frac{\mathrm{var}[x_{ij}]}{\mathrm{var}[u_{ij}]} \le 1 \qquad \text{(A29)}
$$

where u_{ij} is the misclassified exposure on the individual level (If we allow values of s and t below 0.5, the absolute
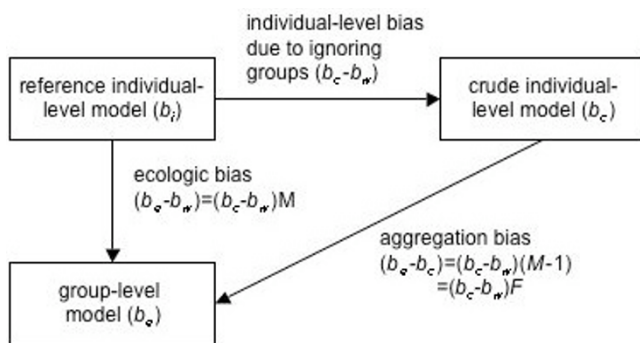
Figure 9
Bias magnification and inflation. The ecologic bias (b_e - b_w) equals the individual-level bias (b_c - b_w) – due to confounding by group and effect modification of the risk difference by group – multiplied by the magnification factor M, assuming no other sources of ecologic bias. The ecologic bias also equals the sum of the individual-level bias due to ignoring groups and the bias caused by aggregation. The latter is measured by F, the inflation factor, equal to M - 1.

value of (A29) is less than or equal to unity). Using (A14), we can expand (A29) into within-group (left term) and between-group (right term) portions:

$$
\left(\frac{b_c}{b} \right) = \left(\lambda \frac{\mathrm{var}_W[x_{ij}]}{\mathrm{var}_W[u_{ij}]} \right)\left(\frac{\mathrm{var}_W[u_{ij}]}{\mathrm{var}[u_{ij}]} \right) + \left(\frac{1}{\lambda} \right)\left(\frac{1}{M_U} \right) \qquad \text{(A30)}
$$

where M_U is the magnification factor for the misclassified exposure. All expressions in parentheses in equation A30 are less than or equal to one except $(1/\lambda)$. Thus, while the ecologic estimate of the RD is biased away from the null by $1/\lambda$, this tendency is counterbalanced in individual level studies by the inverse of the magnification factor M_U. $\lambda M_U \ge M \ge 1$, as can be derived from (A29):

$$
\begin{aligned}
\frac{b_c}{b} &= \lambda \frac{\mathrm{var}[x_{ij}]}{\mathrm{var}[u_{ij}]} \frac{\mathrm{var}_B[U_i]}{\mathrm{var}_B[U_i]} \frac{\mathrm{var}_B[X_i]}{\mathrm{var}_B[X_i]} \\
&= \lambda \frac{\mathrm{var}[x_{ij}]}{\mathrm{var}_B[X_i]} \frac{\mathrm{var}_B[U_i]}{\mathrm{var}[u_{ij}]} \frac{\mathrm{var}_B[X_i]}{\mathrm{var}_B[U_i]} \\
&= \lambda M \frac{1}{M_U} \frac{\mathrm{var}_B[X_i]}{\lambda^2 \, \mathrm{var}_B[X_i]} \\
&= \frac{M}{\lambda M_U} \le 1
\end{aligned} \qquad \text{(A31)}
$$

8. Computation of M *when exposure is binary*
When exposure x_{ij} is binary, M can be computed from ecologic exposure data (X_i) alone. Calculate the total individual-level exposure variance using the standard equation:

$$
\mathrm{var}[x_{ij}] = \bar{X}(1 - \bar{X}) \qquad \text{(A32)}
$$

where $\bar{X}$ is the population-weighted mean of the X_i. Use equation A13 for $\mathrm{var}_B[X_i]$.

9. Confounding within groups
Assume a reference individual-level model with a linear relationship between outcome and exposure. Assume that outcome is also related to an individual-level covariate z_{ij} via a possibly nonlinear function $h()$:

$$
y_{ij} = q + bx_{ij} + h(z_{ij}) + e_{ij} \qquad \text{(A33)}
$$

Aggregation produces

$$
Y_i = q + bX_i + H_i(z_{ij}) + e_i \qquad \text{(A34)}
$$

$$
H_i(z_{ij}) = \frac{1}{n_i} \sum_{j=1}^{n_i} h(z_{ij}) \qquad \text{(A35)}
$$

where $H_i(z_{ij})$ is the average value of $h(z_{ij})$ within group i. The crude individual-level estimate b_c is derived by inserting equation A33 into equation A10 and expanding:

$$b_c = b + \frac{\text{cov}[x_{ij}, h(z_{ij})]}{\text{var}[x_{ij}]} \qquad (A36)$$

Partitioning the covariance within and between groups yields

$$b_c = b + \frac{\text{cov}_W[x_{ij}, h(z_{ij})]}{\text{var}[x_{ij}]} + \frac{\text{cov}_B[X_i, H_i(z_{ij})]}{\text{var}[x_{ij}]} \qquad (A37)$$

The ecologic estimate b_e is derived by inserting equation A34 into equation A11 and expanding:

$$b_e = b + \frac{\text{cov}_B[X_i, H_i(z_{ij})]}{\text{var}_B[X_i]} \qquad (A38)$$

The individual-level estimate, equation A37, is biased by two terms: confounding within groups and confounding between groups. The ecologic estimate, equation A38, is biased only by confounding between groups (These results remain true if $h()$ is linear, e.g., $h(z_{ij}) = \gamma z_{ij}$ and $H(z_{ij}) = \gamma Z_i$). Note that an individual-level variable (z_{ij}) can cause confounding between groups even if it doesn't cause confounding within groups. For models like equation A33, confounding within groups alone does not bias the ecologic estimate. For models that are nonlinear in both exposure and covariates, confounding within groups remains important. For further discussion of confounding within groups, see [11,21].

The approach of assuming an individual-level model, aggregating to obtain an ecologic model, and then comparing biases on the individual and group level is very powerful [11].

Abbreviations
NDEM, non-differential exposure misclassification

OLS, ordinary least squares

RD, risk difference

Competing interests
The author(s) declare that they have no competing interests.

Authors' contributions
TW is the sole author of this paper.

References
1. Last JM: *A Dictionary of Epidemiology* Third edition. New York, NY: Oxford University Press; 1995.
2. Diez-Roux AV: **Bringing context back into epidemiology: variables and fallacies in multilevel analysis.** *Am J Public Health* 1998, **88:**216-222.
3. Greenland S: **Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects.** *Intern J Epidemiol* 2001, **30:**1343-1350.
4. Webster T: **Cross-level bias in partially ecologic studies.** *Spatial Epidemiology Conference, London 2006: Conference Proceedings. 23–25 May* 2006:127-132 [http://www.spatepiconf.org/default.htm]. London. Small Area Health Statistics Unit, Imperial College
5. Jackson C, Best N, Richardson S: **Improving ecological inference using individual-level data.** *Statist Med* 2006, **25:**2136-2159.
6. Künzli N, Tager IB: **The semi-individual study in air pollution epidemiology: a valid design as compared to ecologic studies.** *Environ Health Perspect* 1997, **105:**1078-1083.
7. Greenland S: **Divergent biases in ecologic and individual-level studies.** *Stat Med* 1992, **11:**1209-1223.
8. Morgenstern H: **Ecologic studies.** In *Modern Epidemiology* Second edition. Edited by: Rothman KJ, Greenland S. Philadelphia, PA: Lippincott-Raven; 1998:459-480.
9. Richardson S, Stücker I, Hémon D: **Comparison of relative risks obtained in ecological and individual studies: some methodological considerations.** *Int J Epidemiol* 1987, **16:**111-120.
10. King G: *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data* Princeton, NJ: Princeton University Press; 1997.
11. Webster T: **Bias in Ecologic and Semi-individual Studies.** In *DSc thesis* Boston University School of Public Health, Department of Environmental Health; 2000.
12. Susser M: **The logic in ecological: I. The logic of analysis.** *Am J Public Health* 1994, **84:**825-829.
13. Goodman LA: **Ecological regressions and the behavior of individuals.** *Am Sociol Rev* 1953, **18:**663-664.
14. Wakefield J: **Ecological inference for 2 × 2 tables.** *J R Statist Soc A* 2004, **167:**385-445.
15. Greenland S, Morgenstern H: **Ecological bias, confounding, and effect modification.** *Int J Epidemiol* 1989, **18:**269-274.
16. Palmquist BL: **Ecological Inference, Aggregate Data Analysis of United States Elections, and the Socialist Party of America.** In *PhD thesis* University of California Berkeley, Political Science Department; 1993.
17. Robinson WS: **Ecological correlation and the behavior of individuals.** *Am Sociol Rev* 1950, **15:**351-357.
18. Duncan OD, Cuzzort RP, Duncan D: *Statistical Geography: Problems in Analyzing Areal Data* Glencoe, IL: Free Press; 1961.
19. Piantadosi S, Byar DP, Green SB: **The ecological fallacy.** *Am J Epidemiol* 1988, **127:**893-904.
20. Brenner H, Savitz DA, Jöckel K-H, Greenland S: **Effects of nondifferential exposure misclassification in ecologic studies.** *Am J Epidemiol* 1992, **135:**85-95.
21. Wakefield J: **Sensitivity analysis for ecological regression.** *Biometrics* 2003, **59:**9-17.
22. Webster T: **Commentary: Does the spectre of ecologic bias haunt epidemiology?** *Int J Epidemiol* 2002, **31:**161-162.
23. Lagarde F, Pershagen G: **Parallel analysis of individual and ecologic data on residential radon, cofactors, and lung cancer in Sweden.** *Am J Epidemiol* 1999, **149:**268-274.
24. Björk J, Strömberg U: **Effects of systematic exposure assessment errors in partially ecologic case-control studies.** *Int J Epidemiol* 2002, **31:**154-160.
25. Plummer M, Clayton D: **Estimation of population exposure in ecological studies.** *J R Statist Soc B* 1996, **58:**113-126.

26.   Best N, Cockings S, Bennett J, Wakefield J, Elliott P: **Ecological Regression Analysis of Environmental Benzene Exposure and Childhood Leukaemia: Sensitivity to Data Inaccuracies, Geographical Scale and Ecological Bias.** *J R Statist Soc A* 2001, **164:**155-174.

27.   Richardson S, Montfort C: **Ecological correlation studies.** In *Spatial Epidemiology: Methods and Applications* Edited by: Elliott P, Wakefield JC, Best NG, Briggs DJ. Oxford: Oxford University Press; 2000:205-220.