Practice of Epidemiology

Overcoming Ecologic Bias using the Two-Phase Study Design

Jon Wakefield[1] and Sebastien J.-P. A. Haneuse[2]

[1] Departments of Statistics and Biostatistics, University of Washington, Seattle, WA.
[2] Center for Health Studies, Group Health Cooperative of Puget Sound, Seattle, WA.

Ecologic (aggregate) data are widely available and widely utilized in epidemiologic studies. However, ecologic bias, which arises because aggregate data cannot characterize within-group variability in exposure and confounder variables, can only be removed by supplementing ecologic data with individual-level data. Here the authors describe the two-phase study design as a framework for achieving this objective. In phase 1, outcomes are stratified by any combination of area, confounders, and error-prone (or discretized) versions of exposures of interest. Phase 2 data, sampled within each phase 1 stratum, provide accurate measures of exposure and possibly of additional confounders. The phase 1 aggregate-level data provide a high level of statistical power and a cross-classification by which individuals may be efficiently sampled in phase 2. The phase 2 individual-level data then provide a control for ecologic bias by characterizing the within-area variability in exposures and confounders. In this paper, the authors illustrate the two-phase study design by estimating the association between infant mortality and birth weight in several regions of North Carolina for 2000–2004, controlling for gender and race. This example shows that the two-phase design removes ecologic bias and produces gains in efficiency over the use of case-control data alone. The authors discuss the advantages and disadvantages of the approach.

bias (epidemiology); case-control studies; confounding factors (epidemiology); data interpretation, statistical; research design; sampling studies

Epidemiologists continue to use ecologic and aggregate data. Despite their known drawbacks, these data, often aggregated across geographic areas, offer the advantages of widespread availability and gains in statistical power from large populations and increased exposure ranges. Data availability and exposure variability often determine the scale of examination and the suitability of a study. Exposures arising from a point or line source offer exposure contrasts on small scales, requiring small-area data; in contrast, dietary variables show little variation across small scales, and consequently international studies are used (1, 2).

In addition to the usual biases that may arise in observational studies, ecologic studies suffer from several biases unique to their design. The primary challenge is that ecologic data alone are generally insufficient to characterize within-area variability in exposures and confounding variables. The collective impact of the various biases that result is often referred to under the umbrella term *ecologic bias*. When ecologic bias causes a mismatch between conclusions concerning individual-level associations drawn from aggregate and individual-level data, this is known as the *ecological fallacy*. Many authors have examined the various aspects of ecologic bias (3–9). The only reliable way to characterize within-area variation in exposures and confounders, and hence control ecologic bias, is to collect and incorporate individual-level data. To help epidemiologists achieve this goal, in this paper we describe the use of the two-phase design in an ecologic setting.

To implement a two-phase design, an initial phase 1 cross-classification by the binary disease outcome and stratification variables is required; in phase 2, samples of individuals are drawn from each of the cross-classification cells, with data on additional variables being drawn from the subsamples of individuals (10, 11). Intuitively, the stratified sampling is

Correspondence to Dr. Jon Wakefield, Box 357232, Department of Biostatistics, University of Washington, Seattle, WA 98195-7232 (e-mail: jonno@u.washington.edu).

TABLE 1. Numbers of deaths, numbers of births, and probability of infant mortality according to race, gender, and birth weight category, North Carolina, 2000–2004

Birth weight and gender	Whites			Non-Whites		
	No. of deaths	No. of births	Probability of infant mortality (×100)	No. of deaths	No. of births	Probability of infant mortality (×100)
Normal birth weight						
Female	425	226,103	0.19	229	82,713	0.28
Male	609	240,581	0.25	302	87,313	0.35
Low birth weight						
Female	892	19,047	4.7	995	13,748	7.2
Male	1,175	17,360	6.8	1,227	12,170	10.1

focused on informative cells, and estimation methods use both phases of data for efficiency and to acknowledge the outcome-dependent sampling. In the simplest ecologic setting, the cross-classification is by outcome and area only, and if area is a surrogate for important risk factors this design will be efficient. We are particularly interested in situations where an initial classification is available by outcome, area, and confounders such as age and gender—this is the case in a semi-ecologic study. Phase 2 may then provide detailed exposure information on a subset of the phase 1 individuals.

To illustrate these methods, we consider infant mortality in the state of North Carolina. For this example, we have access to complete individual-level data, permitting a "gold standard" individual-level analysis. For these data, we construct an ecologic study, implement the two-phase approach, and compare the results with those of the full individual-level analysis.

MATERIALS AND METHODS

Infant mortality data

The North Carolina State Center for Health Statistics provides information on vital statistics for all North Carolina residents. Data are available from the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (http://www.irss.unc.edu/). We considered data from all 100 counties in North Carolina for the period 2000–2004. Over these 5 years, 699,035 infants were born and 5,854 died; across counties, the number of births ranged between 267 and 70,590, and the number of deaths ranged between 2 and 510.

The primary scientific goal of this illustrative study is to estimate the association between infant mortality and birth weight, controlling for gender and race. Of particular interest is potential effect modification of the infant mortality–birth weight association by race. Table 1 provides a cross-tabulation of the data, collapsed across counties, by outcome, gender, race, and low birth weight status (<2,500 g).

Individual-level analysis

Consider the following logistic regression model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_g G_i + \beta_r R_i + \beta_w W_i + \beta_{rw} W_i R_i, \quad (1)$$

where p_i is the probability of infant mortality for child i in birth weight category W_i (0 = normal, 1 = low) with gender G_i (0 = female, 1 = male) and race R_i (0 = White, 1 = non-White). The parameters of interest are the odds ratio

TABLE 2. Estimated relative risk of infant mortality from various study designs, North Carolina, 2000–2004

	Individual-level discrete*		Individual-level continuous†		Ecologic study‡	
	RR§	95% CI§	RR	95% CI	RR	95% CI
Gender	1.42	1.34, 1.49	1.43	1.35, 1.52	1.65	0.78, 3.49
Race	1.41	1.27, 1.57	1.12	1.05, 1.19	0.88	0.75, 1.03
Weight	27.5	25.5, 29.7	1.71	1.70, 1.73	3.56	1.77, 7.13
Weight × race	1.11	0.98, 1.25	1.04	1.03, 1.06	1.12	0.95, 1.33

* Weight was a binary indicator of low birth weight (<2,500 g).

† Weight was a continuous measure, standardized so that a 1-unit difference corresponded to a decrease of 250 g.

‡ Analysis based on the ecologic log-linear model, model 2. Coefficients are for a 10% increase in the corresponding ecologic percentages.

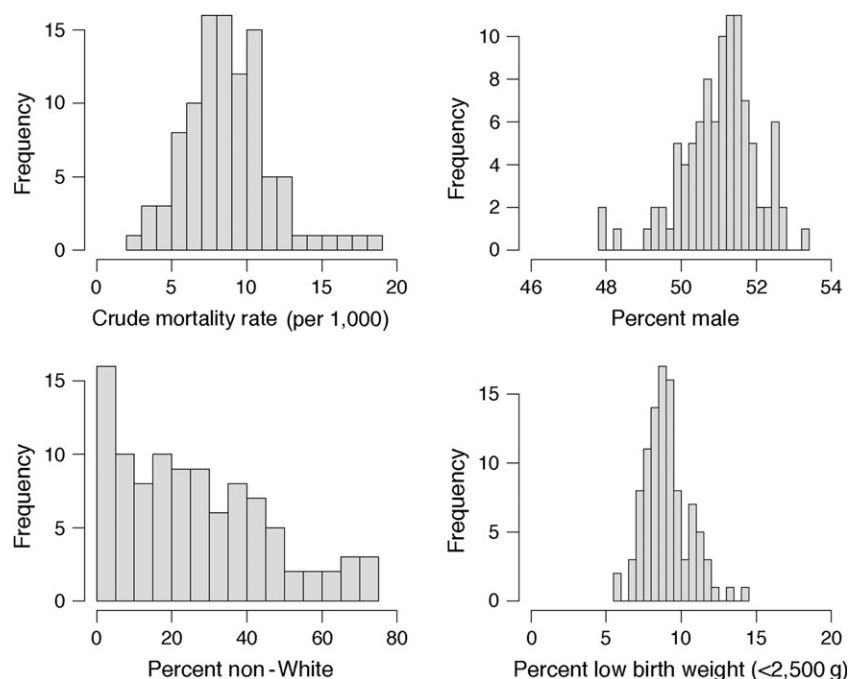§ RR, relative risk; CI, confidence interval.

FIGURE 1.   Histograms of the crude infant mortality rate (per 1,000 livebirths) and the percentages of infants born male, non-White, and of low birth weight (<2,500 g), North Carolina, 2000–2004.

(hereafter referred to as the relative risk because of the rarity of infant mortality) associated with low birth weight for White babies, $\exp(\beta_w)$, and the additional multiplicative change in relative risk associated with low birth weight for non-White babies, $\exp(\beta_{rw})$.

Table 2 provides estimates from model 1, based on the complete individual-level data. The main effect indicates a strong association between infant mortality and low birth weight among the White babies. Based on the estimate for the interaction term, there is modest evidence of an additional 11 percent increase in risk for non-White low birth weight babies.

A simulated ecologic study

Because of the unavailability of individual-level data, ecologic data may be resorted to and may come in a variety of forms. For example, a purely ecologic study would consist of marginal death counts across the 100 counties as well as the marginal proportions of babies who were born male, non-White, and of low birth weight. Alternatively, a semi-ecologic study might consist of the numbers of births and infant deaths cross-classified by gender and race, along with the proportion of babies that were of low birth weight, in each county.

For the North Carolina infant mortality data, we collapsed the counts within the 100 counties to mimic a purely ecologic study. Figures 1 and 2 provide histograms and maps, respectively, of the ecologic data. As expected, the proportion male is tightly clustered around 0.5 across counties,

while the proportion non-White varies between 0.002 and 0.380 and the proportion with low birth weight varies between 0.06 and 0.14.

To illustrate a typical ecologic analysis, let Y_k and N_k denote the observed numbers of deaths and births in county k ($k = 1, \ldots, 100$). Further, let Q_{Wk} denote the proportion of low birth weight babies in county k, and let Q_{Gk} and Q_{Rk} denote the corresponding proportions of babies that are male and non-White. A typical ecologic analysis might fit the log-linear model:

$$\log E(Y_k) = \log N_k + \beta_0^* + \beta_g^* Q_{Gk} + \beta_r^* Q_{Rk} + \beta_w^* Q_{Wk} \\ + \beta_{rw}^* Q_{Rk} Q_{Wk}, \qquad (2)$$

where $\log N_k$ is a county-specific offset that accounts for the denominator sizes. Although model 2 bears some resemblance to the individual-level model, model 1, the interpretation of its components is quite different. For example, interpreted literally, $\exp(\beta_w^*)$ is the relative risk associated with an all-White area whose newborns are all low birth weight as compared with an all-White area containing no low birth weight infants, with both areas having the same proportion of male births. Hence, the interpretation of $\exp(\beta_w^*)$ resembles more closely that of a contextual effect and is therefore not comparable to the individual-level effect $\exp(\beta_w)$.

The outcome, Y_k, is a count, and to allow for extra-Poisson variability, we fit model 2 using quasi-likelihood (12). Table 2 shows results from a fit in which each of the proportions has been multiplied by 10. That is, each relative risk estimate compares two areas that differ in the
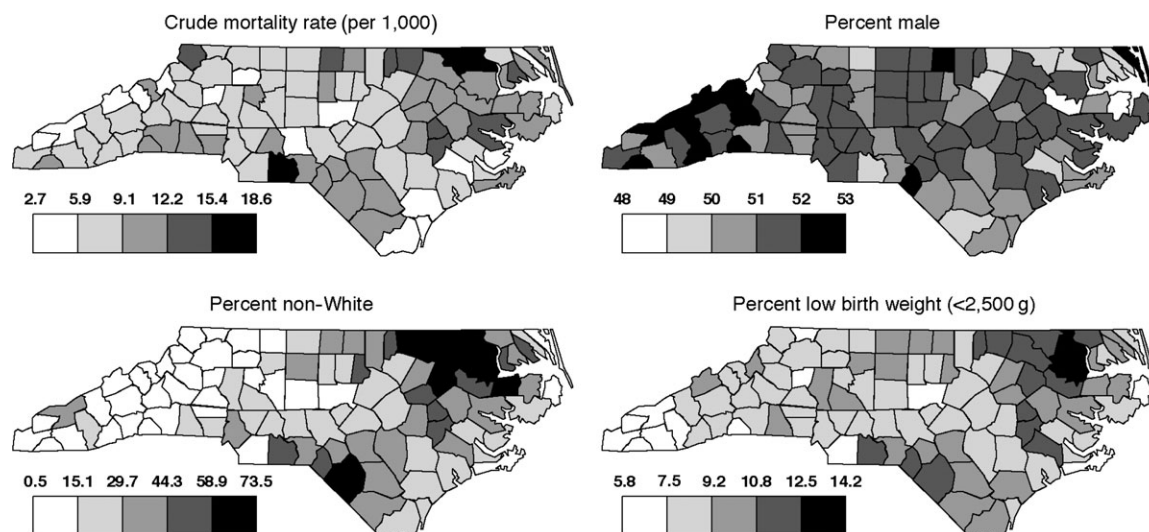
FIGURE 2. Maps of the crude infant mortality rate (per 1,000 livebirths) and the percentages of infants born male, non-White, and of low birth weight (<2,500 g), North Carolina, 2000–2004.

corresponding proportion by 10 percent. We see that the ecologic relative risk associated with low birth weight is completely incomparable with the individual-level coefficient. Furthermore, non-White race now appears protective, rather than detrimental as the individual-level analysis suggests. This spurious result provides an example of the ecological fallacy, in which conclusions (here regarding race) based on ecologic data are opposite of those drawn on the basis of individual-level data.

The inherent difficulty in estimating individual-level associations from ecologic data can be illustrated by examining the induced aggregate model. Let N_{kgrw} denote the number of children in county k and in gender, race, and birth weight categories g, r, and w, respectively. For a rare outcome, model 1 may be approximated by a log-linear model, and aggregation within area k yields

$$(N_{k000}e^{\beta_0} + N_{k100}e^{\beta_0+\beta_g} + N_{k010}e^{\beta_0+\beta_r} + N_{k001}e^{\beta_0+\beta_w}$$
$$+ N_{k110}e^{\beta_0+\beta_g+\beta_r} + N_{k101}e^{\beta_0+\beta_g+\beta_w} + N_{k011}e^{\beta_0+\beta_r+\beta_w}$$
$$+ N_{k111}e^{\beta_0+\beta_g+\beta_r+\beta_w+\beta_{wr}})/N_k, \qquad (3)$$

where, again, N_k is the total number of births in region k. Intuitively, model 3 provides the average risk associated with area k taken as the sum of the numbers in each of the eight gender/race/low birth weight strata, each multiplied by their specific mortality probabilities. Estimation of this model requires information on the N_{kgrw}, which is generally not available in an ecologic study. The differences between models 2 and 3, both in terms of the predictor variables and in terms of parameter interpretation, explain the ecological fallacy for the North Carolina data. In the case of a continuous exposure, it may be possible to predict the direction of the bias (9), since the individual and ecologic models are both log-linear. However, with a discrete exposure like the

one we have here, such an endeavor is not possible because of the differing functional forms.

Collecting individual-level data

It is generally well recognized that in order for ecologic data to provide reliable inferences, they need to be supplemented with individual-level data. With a rare outcome, the aggregate data design (2) collects supplemental individual-level survey information on exposures and confounders; intuitively, these provide estimates of the N_{kgrw} in model 3. An approach that also uses similar data but assumes a parametric form for the within-area distributions and then fits the implied disease risk model has also been suggested (13–15). While these approaches can overcome ecologic bias, they are still ecologic in nature, since there is no linkage between outcome and exposures or confounders at the level of the individual (16).

In the setting of a nonrare outcome, a scheme for combining ecologic data with a series of 2 × 2 tables with simple random samples has been outlined (17). More recently, the parametric aggregate data design has been extended to incorporate prospectively collected information on individuals (18). In this paper, we focus on studies of rare outcomes, and we therefore consider outcome-dependent sampling. Previously we considered case-control sampling within areas (19, 20); here we consider the use of two-phase sampling to obtain individual-level data jointly on both outcomes and exposures/confounders.

Ecologic two-phase studies

Two-phase study designs are a generalization of matched case-control designs in which, initially, the entire sampling population is cross-classified according to case/control status and some stratification variable, S. The latter depends on

TABLE 3.   Discrete low birth weight example*

Phase 1 stratification†	% bias				Relative efficiency			
	θ_g	θ_r	θ_w	θ_{wr}	θ_g	θ_r	θ_w	θ_{wr}
Y alone	2.2	3.0	12.6	8.4	100.0	100.0	100.0	100.0
$Y \times A$	2.4	3.0	10.6	11.2	99.0	105.2	92.0	106.9
$Y \times A \times G$	1.1	3.5	11.3	10.9	67.1	106.3	91.4	107.2
$Y \times A \times R$	2.2	0.5	14.0	6.4	99.3	74.1	120.0	89.4
$Y \times A \times W$	1.4	3.3	0.5	1.9	77.1	93.5	28.2	46.8
$Y \times A \times G \times R$	1.1	0.3	14.9	6.3	66.7	74.0	120.6	89.6
$Y \times A \times G \times W$	0.1	1.4	0.9	1.8	18.8	94.0	28.2	46.5
$Y \times A \times R \times W$	0.3	0.2	0.2	0.2	80.0	26.1	14.7	15.2

* Percent bias and relative efficiencies of relative risk estimators (in comparison with a case-control design) for various two-phase schemes, based on 10,000 simulations. Each repetition samples 500 cases and 500 controls.

† Y = outcome, A = region, G = gender, R = race, W = binary low birth weight (<2,500 g) indicator.

covariates observed in all individuals and may include exposures of interest, proxy exposure measures, or potential confounders. In settings like those we are considering, such as environmental epidemiology, S may also depend on geographic area, which can act as a surrogate for the totality of confounders associated with each area, as well as provide a well-defined sampling frame for the controls. For example, Flick et al. (21) recently reported results from a case-control study of the association between nonsteroidal antiinflammatory drugs and non-Hodgkin's lymphoma in which the data were matched by county.

Let us assume that S takes on J levels. After the initial cross-classification, the phase 1 data consist of $2J$ counts, N_{ij}, with $i = 0/1$ (corresponding to noncase/case status) and with j indexing stratum ($j = 1, \ldots, J$). In phase 2, samples of size n_{ij} are taken within each of the phase 1 strata, and individual-level measurements, x_{ijk}, are taken on these individuals ($k = 1, \ldots, n_{ij}; i = 0, 1; j = 1, \ldots, J$). Such individual-level data may include additional covariates not readily available on all subjects and/or accurate measurements for proxy exposures available on all individuals in phase 1 but subject to measurement error or misclassification. We note that the traditional case-control design corresponds to an initial classification based solely on case/control status (and so does not involve S), while a matched case-control design classifies additionally on confounders. Whereas case-control designs ignore the phase 1 data, in a two-phase approach these data are exploited to provide efficiency gains and to enable the estimation of intercepts and interactions, including those involving phase 1 stratification variables (11).

The outcome-dependent nature of the phase 2 sampling must be accounted for when analyzing two-phase data; a number of approaches have been developed (22–27). Software for implementing the methods in an ecologic context (along with the North Carolina data) is available from the first author (http://faculty.washington.edu/jonno/cv.html). Unless otherwise stated, all of the analyses presented here implement full maximum likelihood estimation, which, under correct model specification, provides the most efficient

estimates (25). We emphasize that both the phase 1 and phase 2 data are exploited for estimation; further details are provided in the Appendix.

RESULTS

Simulation study

To implement a two-phase design, one must initially specify the variables upon which the phase 1 stratification is based and, in particular, define S. In the North Carolina example, we could stratify on area, gender, race, low birth weight, or some combination. Caution is required, though, as too fine a stratification may leave some cells empty, leading to a breakdown of the analysis method. In phase 2, one must decide how to allocate the individual-level samples across these strata (i.e., the n_{ij} sample sizes). One recommended choice is that of a balanced scheme, where equal numbers of individuals are sampled across phase 1 strata (23). An alternative is to consider optimal sampling strategies for two-phase sampling, typically tailored to the specific setting (28).

Below we report on a simulation study for which the results were based on 10,000 simulated data sets. In each case, we generated complete individual-level data using the parameter estimates from the full data (table 2). This provided the basis for the phase 1 stratification, from which cases and controls were sampled in phase 2.

Infant mortality data

Discrete birth weight status. We implemented schemes with eight different phase 1 stratifications, as outlined in table 3. With 100 counties, it is not possible to use county (with 100 levels) as a phase 1 stratification variable, since some counties will contain very few cases; further cross-classification will result in zero entries, leading to estimation difficulties. Hence, we constructed 10 regions based on contiguous counties, shown in figure 3, and matched on this new variable.
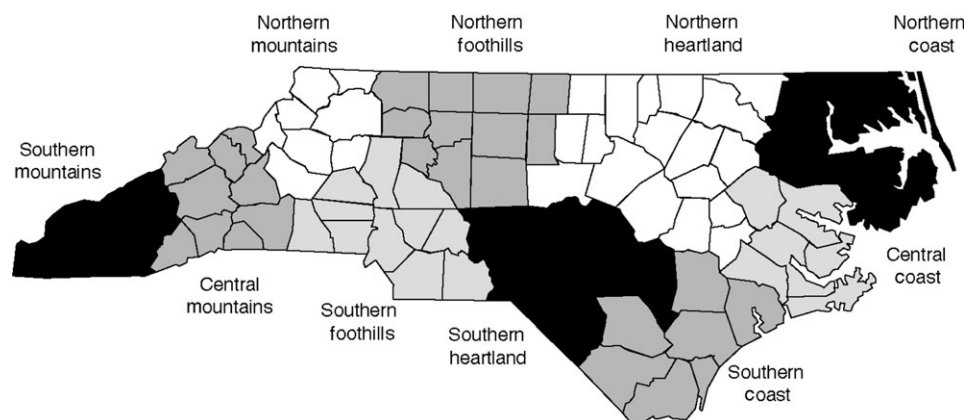
FIGURE 3.   Grouping of North Carolina's 100 counties into 10 regions for a study of the relation between infant mortality and birth weight.

Under each scheme, we sampled 500 cases and 500 controls. We adopted a balanced design in which equally sized samples were taken, where possible, across the J strata; when sufficient numbers of cases were not available, noncases were sampled instead. Table 3 gives the percent bias across simulations as compared with the fitting of the individual-level model to the totality of data, as well as the relative efficiency, taken as the ratio of the variance of the two-phase relative risk estimators (across simulations) relative to that of the case-control design. The only difference between conventional case-control and two-phase sampling with phase 1 stratified by outcome only is in the estimation of the intercept, which may be estimated under the two-phase approach; inference for the relative risks is identical. The results in table 3 are presented in terms of relative risk; thus, for example, $\theta_r = \exp(\beta_r)$ is the relative risk corresponding to race.

It is apparent from table 3 that since region is only weakly associated with outcome, little is gained by stratifying on region alone. When we stratify results by gender or race or low birth weight, estimation of the corresponding relative risks (including the interaction) improves correspondingly. When we stratify on low birth weight in phase 1, efficiency improves markedly over case-control sampling. For example, the standard error of the relative risk for the interaction, θ_{wr}, is 0.53 under case-control sampling and 0.08 under two-phase sampling; analysis of the individual data gives a standard error of 0.07, so the two-phase analysis is almost as efficient as the analysis using the full data, even though it is based on only 500 cases and 500 controls. When we use low birth weight in phase 1, the parameter estimates are unbiased; in particular, the ecologic bias evident in table 2 is eliminated. There is some finite sample bias when we do not use low birth weight in phase 1.

TABLE 4.   Continuous weight example

Phase 1 stratification†	% bias				Relative efficiency			
	θ_g	θ_r	θ_w	θ_{wr}	θ_g	θ_r	θ_w	θ_{wr}
Y alone	2.8	1.1	−0.6	−0.2	100.0	100.0	100.0	100.0
$Y \times A$	2.9	2.0	−0.5	−0.3	98.4	106.5	97.3	105.4
$Y \times A \times G$	1.6	1.8	−0.5	−0.3	75.4	106.2	97.6	104.7
$Y \times A \times R$	2.7	−0.6	−0.7	0.1	100.4	76.8	114.4	93.8
$Y \times A \times W$	1.5	2.0	−0.1	0.0	75.1	86.1	43.9	55.1
$Y \times A \times$ misclassified W	2.1	2.2	−0.2	−0.1	82.8	94.1	60.8	69.9
$Y \times A \times G \times R$	1.8	−0.7	−0.8	0.2	75.5	75.7	115.3	94.1
$Y \times A \times G \times W$	0.6	2.2	−0.1	0.0	42.2	87.3	44.0	55.9
$Y \times A \times G \times$ misclassified W	0.9	2.0	−0.2	−0.1	54.9	95.1	62.0	70.7
$Y \times A \times R \times W$	1.8	0.2	−0.1	0.0	77.3	42.1	41.2	36.9
$Y \times A \times R \times$ misclassified W	1.9	−0.1	−0.2	0.0	84.8	54.1	65.1	55.2

* Percent bias and relative efficiencies of relative risk estimators (in comparison with a case-control design) for various two-phase schemes, based on 10,000 simulations. Each repetition samples 500 cases and 500 controls.
† Y = outcome, A = region, G = gender, R = race, W = binary low birth weight (<2,500 g) indicator.

TABLE 5. Comparison of strategies for phase 2 sample collection*

Phase 1 stratification†	Scheme A‡				Scheme B§			
	θ_g	θ_r	θ_w	θ_{wr}	θ_g	θ_r	θ_w	θ_{wr}
Y alone	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$Y \times A$	98.5	107.1	94.7	104.0	99.6	108.0	96.5	105.4
$Y \times A \times G$	71.5	105.9	94.7	102.6	74.9	108.7	95.6	103.9
$Y \times A \times R$	103.0	68.1	112.4	88.1	101.5	78.0	112.1	95.3
$Y \times A \times W$	69.2	82.4	38.6	48.3	73.3	86.6	42.6	53.8
$Y \times A \times$ misclassified W	79.3	89.5	56.2	65.2	83.5	92.5	59.5	71.1
$Y \times A \times G \times R$	71.2	68.5	112.2	87.6	77.0	79.1	110.5	94.1
$Y \times A \times G \times W$	34.3	88.0	38.7	49.8	41.6	89.0	43.4	55.2
$Y \times A \times G \times$ misclassified W	49.1	95.5	56.9	66.0	54.9	96.1	59.6	70.6
$Y \times A \times R \times W$	72.7	33.8	36.5	31.1	77.3	42.8	40.6	36.8
$Y \times A \times R \times$ misclassified W	84.3	46.5	61.1	49.2	86.1	55.6	64.6	55.5

* Relative efficiencies of relative risk estimators (in comparison with a case-control design) for various two-phase schemes, based on 10,000 simulations.

† Y = outcome, A = region, G = gender, R = race, W = binary low birth weight (<2,500 g) indicator.

‡ Scheme A used 200 cases and 200 controls, equally allocated across phase 1 strata.

§ Scheme B used 500 cases and 500 controls, allocated in proportion to phase 1 strata.

Continuous birth weight.   The two-phase ecologic design is particularly useful for investigating the association between health outcomes and environmental exposures. In this context, it may be possible to obtain an estimate of the proportion exposed to high concentrations within each area, perhaps by confounder strata such as age, gender, race, and socioeconomic indices. This may be achieved by first modeling a concentration surface (29) and then using finer geographic information within each area (e.g., census blocks) to estimate the fractions of different demographic groups who are above or below a concentration threshold. In phase 2, one may then sample individuals to obtain more accurate exposure measures at residential addresses. An important aspect is that the proportions exposed (in the phase 1 data) can be error-prone; the benefits of two-phase sampling when a surrogate exposure is available have been demonstrated in other contexts (30). To summarize the approach, we assume the existence of a discrete exposure and stratify by this variable in phase 1, before measuring a continuous version in phase 2.

We consider two situations: the first in which the phase 1 binary exposure is accurate and the second in which it is subject to measurement error. For the latter, we consider a hypothetical situation in which measurement error is added to the low birth weight classification that is used in the phase 1 classification. In particular, we let

$$\Pr(W=0|X=0,j) = p_j \text{ and}$$
$$\Pr(W=1|X=1,j) = q_j,$$

where X is the binary low birth weight and W is the binary surrogate, so that p_j is the specificity and q_j the sensitivity in stratum j ($j = 1, \ldots, J$); we choose $p_j = q_j = 0.9$ across all strata.

Returning to the North Carolina example, we assume that individual-level associations are again given by model 1, but with W_i now a continuous measure of birth weight. Table 2 provides estimates for this model based on the complete individual-level data.

Again using 500 cases and 500 controls in phase 2, we summarize the percent bias and efficiency over various phase 1 stratifications (with equal samples across the phase 1 stratification) and in situations where both accurate and error-prone measures of low birth weight are available. Results are shown in table 4. The benefits of the two-phase design are again evident; the use of the binary low birth weight information clearly allows efficient estimation by sampling of informative individuals. There is some loss of efficiency when the error-prone phase 1 classification is used, as compared with the error-free classification. However, in this setting no bias is introduced, and it is still clearly worthwhile to use the error-prone version.

Further design considerations

While the results of tables 3 and 4 focus on alternative schemes for defining the phase 1 stratification, a variety of other design considerations can be investigated. Table 5 considers two extensions. Scheme A examines how reducing the phase 2 samples from 500 cases and controls to 200 cases and controls affects efficiency. The reductions are not substantial, which suggests that the two-phase design has benefits even when resources for obtaining individual-level data are limited. In scheme B, we return to case/control sizes of 500, but we now sample individuals in proportion to the phase 1 stratum sizes, rather than taking equal numbers (the standard errors for the "Y only" stratification are equal in table 4 and table 5, scheme B). The efficiencies are very

similar to those in table 4, suggesting that, in this setting, efficiency is driven primarily by the choice of the variables used to define S, rather than by the specific allocation of samples in phase 2.

DISCUSSION

In this paper, we have described the two-phase study design as a means of avoiding the numerous and often severe pitfalls associated with the analysis of ecologic and/or aggregate data. The results of our simulation studies point to the benefits associated with combining the two sources of data, in terms of both bias and efficiency. Rather than supplement an ecologic study with individual-level data, it may be of interest to combine existing individual-level data with external group-level data. Strategies that combine both types of data have been shown to alleviate participation bias and improve efficiency in case-control studies with missing data (31).

When designing a two-phase study, a variety of choices must be made, including the variables which form the basis of the phase 1 stratification, the total numbers of cases and controls sampled in phase 2, and the way in which resources are allocated across phase 1 strata. It is clear that important variables should be used as a basis for the phase 1 stratification. Typically, however, one will not know the appropriate individual-level model and an educated guess will be required. While choosing a nonoptimal set of stratification variables reduces efficiency, the ability of the two-phase design to help overcome ecologic bias is not affected.

An example of an ecologic study for which we believe two-phase sampling could be particularly useful is a study of the association between death from myocardial infarction and magnesium in domestic water in northwestern England (32). In this study carried out by Maheswaren et al. (32), ecologic-level magnesium concentrations were measured in the domestic water supply, with an average of six measurements being taken per water zone (containing up to 50,000 people). The study did not provide evidence to support the protective hypothesis. The main ecologic-bias difficulties arising here were due to the within-zone variability in magnesium levels and confounding factors, particularly socioeconomic status and the water constituents fluoride, calcium, and lead, and the inability to characterize the within-area distribution of magnesium levels across all age, gender, and socioeconomic status strata. In general, the relative risks due to environmental exposures will be in the range of 1.2–1.5 (33), making control for confounding particularly important. For the magnesium example, a two-phase study would sample individual cases and noncases with the potential strata water zone, gender, age, socioeconomic status, and a categorical version of exposure based on measurements taken initially (or on historic data). Magnesium concentrations could be sampled at selected case/noncase residences and be augmented with information on confounding water constituents and individual-level confounders such as smoking. Within the two-phase framework, information on multiple exposures could be collected in phase 2 and incorporated into a single disease model. To fully characterize the joint distribution of exposures and con-

founders, larger phase 2 sample sizes will be required, particularly if the exposures are highly correlated.

A number of methods have been proposed for combining ecologic- and individual-level data, and our method builds on these approaches. The aggregate data method (2) does not stratify in phase 1, either by outcome or by stratum, although the latter would be possible via the inclusion of stratum-specific intercepts. In a semi-ecologic study, an ecologic exposure is combined with individual-level outcomes and confounders. A two-phase approach is particularly useful for such a study, with the phase 2 data corresponding to stratified sampling of individual exposures.

We have presented the two-phase approach from the perspective of supplementing available ecologic data with individual-level data, but it is also feasible to start with population-based matched case-control data and then add ecologic data, perhaps from the Census Bureau and a disease registry. Thinking of the design in this way emphasizes that the phase 1 and phase 2 data must be comparable; this is straightforward to think about statistically, but in any application it will be complex and require great care. Clearly a population-based case-control study is more amenable to the two-phase design than is a hospital-based study, since the geographic catchment area of the latter will be difficult to determine. An existing cohort provides an alternative sampling frame. It is now common practice to embed case-control studies within a larger cohort. For example, the multicountry European Prospective Investigation into Cancer and Nutrition (1) has provided a population from which numerous case-control studies have been constructed. With two-phase methodology, it is possible to use the data from the complete cohort to inform confounder relations, particularly confounding by geographic area.

ACKNOWLEDGMENTS

REFERENCES

1. Riboli E. Nutrition and cancer: background and rationale of the European Perspective Investigation into Cancer and Nutrition (EPIC). Ann Oncol 1992;3:783–91.
2. Prentice RL, Sheppard L. Aggregate data studies of disease risk factors. Biometrika 1995;82:113–25.
3. Morgenstern H. Ecologic study. In: Armitage P, Colton T, eds. Encyclopedia of biostatistics. Vol 2. New York, NY: John Wiley and Sons, Inc, 1998:1255–76.
4. Piantadosi S, Byar DP, Green SB. The ecological fallacy. Am J Epidemiol 1988;127:893–904.
5. Greenland S, Morgenstern H. Ecological bias, confounding and effect modification. Int J Epidemiol 1989;18:269–74.
6. Greenland S. Divergent biases in ecologic and individual level studies. Stat Med 1992;11:1209–23.

7. Greenland S, Robins J. Invited commentary: ecologic studies—biases, misconceptions and counterexamples. Am J Epidemiol 1994;139:747–60.
8. Richardson S, Montfort C. Ecological correlation studies. In: Elliott P, Wakefield JC, Best NG, et al, eds. Spatial epidemiology: methods and applications. New York, NY: Oxford University Press, 2000:205–20.
9. Wakefield JC. Sensitivity analyses for ecological regression. Biometrics 2003;59:9–17.
10. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 1982;115:119–28.
11. Weinberg CR, Wacholder S. The design and analysis of case-control studies with biased sampling. Biometrics 1990;46:963–75.
12. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London, United Kingdom: Chapman and Hall Ltd, 1989.
13. Richardson S, Stucker I, Hémon D. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. Int J Epidemiol 1987;16:111–20.
14. Wakefield JC, Salway RE. A statistical framework for ecological and aggregate studies. J R Stat Soc Ser A 2001;164:119–37.
15. Best N, Cockings S, Bennett J, et al. Ecological regression analysis of environmental benzene exposure and childhood leukemia: sensitivity to data inaccuracies, geographical scale and ecological bias. J R Stat Soc Ser A 2001;164:155–74.
16. Sheppard L. Insights on bias and information in group-level studies. Biostatistics 2003;4:265–78.
17. Wakefield J. Ecological inference for 2 × 2 tables (with discussion). J R Stat Soc A 2004;167:385–445.
18. Jackson S, Best N, Richardson S. Improving ecological inference using individual-level data. Stat Med 2006;25:2136–59.
19. Haneuse SJ, Wakefield J. Hierarchical models for combining ecological and case-control data. Biometrics 2007;63:128–36.
20. Haneuse SJ, Wakefield J. The combination of ecological and case-control data. J R Stat Soc B 2007;70:73–93.
21. Flick ED, Chan KA, Bracci PM, et al. Use of nonsteroidal antiinflammatory drugs and non-Hodgkin lymphoma: a population-based case-control study. Am J Epidemiol 2006;164:497–504.
22. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. Am J Epidemiol 1988;128:1198–206.
23. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika 1988;75:11–20.
24. Flanders W, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. Stat Med 1991;10:739–47.
25. Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. J R Stat Soc Ser B 1997;59:447–61.
26. Breslow NE, Holubkov R. Weighted likelihood, pseudo likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. Stat Med 1997;16:103–16.
27. Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika 1997;51:54–71.
28. Reilly M. Optimal sampling strategies for two-stage studies. Am J Epidemiol 1996;143:92–100.
29. Jerrett M, Afrain A, Kanaroglou P, et al. A review and evaluation of intraurban air pollution exposure. J Expo Anal Environ Epidemiol 2005;15:185–204.
30. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. Appl Stat 1999;48:457–68.
31. Stromberg U, Bjork J. Incorporating group-level exposure information in case-control studies with missing data on dichotomous exposures. Epidemiology 2004;15:494–503.
32. Maheswaren R, Morris S, Falconer S, et al. Magnesium in drinking water supplies and mortality from acute myocardial infarction in north west England. Heart 1999;82:455–60.
33. Pekkanen J, Pearce N. Environmental epidemiology: challenges and opportunities. Environ Health Perspect 2001;109:1–5.

APPENDIX

The likelihood for the two-phase design consists of two components for the phase 1 and phase 2 data, respectively. Following the notation of Breslow and Holubkov (25), the likelihood may be written as

$$\prod_{i=1}^{1} \Pr(\{N_{ij}\}, \{\mathbf{x}_{ijk}\}) \propto \prod_{i=1}^{1} \prod_{j=1}^{J} \Pr(S=j|Y=i)^{N_{ij}}$$

$$\times \prod_{i=1}^{1} \prod_{j=1}^{J} \prod_{k=1}^{n_{ij}} \Pr(\mathbf{X} = \mathbf{x}_{ijk}|Y=i, S=j).$$

Given a logistic disease model (1), each contribution may be reparameterized into a prospective component for the outcome and a marginal component for the distribution of the explanatory variables (S for the phase 1 data and **X** for the phase 2 data). Such a reparameterization results in a series of constraints imposed on the parameter space. Under the traditional case-control design, the marginal distribution of the explanatory variables may be ignored, since an ordinary (prospective) logistic analysis yields a solution that satisfies the constraints. In the two-phase setting, this is not the case, and full maximum likelihood requires consideration of the constraints (25). Alternative analyses for the two-phase design, such as weighted likelihood (24) and pseudolikelihood (23), generally ignore the phase 1 and/or phase 2 constraints, yielding inefficient, though still consistent, estimates.