Published in final edited form as:
Epidemiology. 2011 May ; 22(3): 382–389. doi:10.1097/EDE.0b013e3182125cff.

Designs for the combination of group- and individual-level data

Sebastien Haneuse[1] and **Scott Bartell**[2]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA

[2]Department of Epidemiology and Program in Public Health, University of California – Irvine, Irvine, CA

Abstract

Background—Studies of ecologic or aggregate data suffer from a broad range of biases when scientific interest lies with individual-level associations. To overcome these biases, epidemiologists can choose from a range of designs that combine these group-level data with individual-level data. The individual-level data provide information to identify, evaluate, and control bias, while the group-level data are often readily accessible and provide gains in efficiency and power. Within this context, the literature on developing models, particularly multi-level models, is well-established, but little work has been published to help researchers choose among competing designs and plan additional data collection.

Methods—We review recently proposed "combined" group- and individual-level designs and methods that collect and analyze data at two levels of aggregation. These include aggregate data designs, hierarchical related regression, two-phase designs, and hybrid designs for ecologic inference.

Results—The various methods differ in (i) the data elements available at the group and individual levels and (ii) the statistical techniques used to combine the two data sources. Implementing these techniques requires care, and it may often be simpler to ignore the group-level data once the individual-level data are collected. A simulation study, based on birth-weight data from North Carolina, is used to illustrate the benefit of incorporating group-level information.

Conclusions—Our focus is on settings where there are individual-level data to supplement readily accessible group-level data. In this context, no single design is ideal. Choosing which design to adopt depends primarily on the model of interest and the nature of the available group-level data.

In ecologic studies the fundamental unit of investigation is a group of individuals, rather than individuals themselves.[1] Ecologic studies are widely used because group-level (or aggregated) data are easy and inexpensive to obtain, particularly through data depositories such as disease registries and census data. Further, developments in computing (e.g. geographical information systems) let researchers combine information at varying levels of aggregation.[2] Taking advantage of these strengths, ecologic designs continue to be employed in many epidemiologic settings, including studies of environmental risk factors,[3-7] cancer screening,[8,9] investigations of chronic disease[10] and infectious disease.[11,12]

Notwithstanding their continued use, ecologic studies are controversial because they directly assess group-level associations: that is, relationships between group-level outcomes and

Corresponding author: Sebastien Haneuse, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115, Phone: (617) 432-3980, Fax: (617) 432-5619, shaneuse@hsph.harvard.edu.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

group-level exposure measures. Such associations are sometimes of interest,[13] particularly for policymaking.[14] Typically, though, the scientific goal in epidemiology is to assess individual-level associations. With group-level data alone, one generally cannot estimate individual-level associations, although one may be tempted to interpret results from an ecologic study in terms of such associations. Doing so has many pitfalls.[1,15-21] Of particular concern is that misuse of ecologic results may give rise to the ecologic fallacy, in which conclusions based on a group-level analysis differ from those that would have been drawn had an individual-level analysis been performed.

The fundamental difficulty is that ecologic studies cannot characterize within-group joint outcome/exposure/confounder distributions. This makes estimation of individual-level associations extremely difficult and is analogous to the challenge faced when an important confounder is missing. Unfortunately, the problem cannot be overcome solely via post-hoc analytic methods, at least not without making untestable assumptions.[22] The only reliable way to address the problem is to collect and incorporate appropriate individual-level data.

Combining group- and individual-level data has intuitive appeal. The individual-level data permit identifiability of individual-level associations via three mechanisms: (i) evaluation and control of bias, (ii) separation of contextual, within-, and between-group effects, and (iii) the ability to check models. Once identifiability is established, ecologic data may provide gains in power and efficiency, particularly if they represent large sample sizes and if the exposure of interest exhibits large between-group variation.

The past 20 years have seen numerous study designs and methods proposed to combine group- and individual-level data. Despite an extensive literature on developing models, particularly multi-level models,[23,24] little work has been published to help researchers choose among alternative designs and, consequently, to help them plan additional data collection efforts. This paper reviews recently proposed “combined” epidemiologic study designs and describes the statistical frameworks they use to estimate individual-level associations. These ideas are illustrated with a simple, hypothetical study of birth weight, using data from North Carolina. Given individual-level data, the additional complexity of combining two sources of information at the analysis stage may make it appealing to ignore the group-level data. However, a simulation study illustrates the potentially substantial benefits of accommodating group-level data.

MODEL SPECIFICATION

Fundamental to each design or method reviewed here is the premise that scientific interest lies with some underlying individual-level model. Suppose the population of interest can be stratified into K groups of sizes $N_1, \ldots, N_K$; in environmental epidemiology, such groups are often based on geographic location.[20]

Let Y_{ki} be some binary outcome of interest for the i^th^ individual in the k^th^ group and $\boldsymbol{\pi}_{ki} = P(Y_{ki} = 1)$ the corresponding outcome probability. Consider the following general regression model:

$$g(\pi_{ki}) = X_{ki}\beta, \qquad (1)$$

where g() is a link function [e.g. log() for a log-linear model and logit() for a logistic model] and X_{ki} denotes a vector of covariates. The latter may include exposures of interest, confounders, and potential effect modifiers and, further, may be defined at either the individual or group level. When both individual- and group-level covariates are included, such models are often called multi-level models.[23,24]

Regardless of the level at which components of X_{ki} are defined, we call the β parameters in model (1) individual-level associations because they correspond to differences in risk between two individuals (because the outcome is defined at the individual level).

DESIGN OPTIONS

The focus of this paper is on designs that supplement readily available group-level data with a sample of individuals for whom outcome and covariate information are observed. The following describes four general classes of such designs and their statistical methods.

Aggregate data methods

Suppose one has access to the number of cases in each group, denoted N_{1k}. Given group-specific population totals, one can calculate the observed proportion of cases for the k^{th} group as $\overline{Y}_k = N_{1k}/N_k$. Consider the induced model for the group-level outcome, $\boldsymbol{\pi}_k = E[\overline{Y}_k]$, obtained by averaging the individual-level model (1) over the N_k individuals in the group:

$$\pi_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\pi_{ki} = \frac{1}{N_k}\sum_{i=1}^{N_k} g^{-1}(X_{ki}\beta). \qquad (2)$$

The right-hand side of (2) shows that the induced model for $\overline{Y}_k$ is a function of the underlying individual-level β parameters. Further, it demonstrates that evaluating the model requires only individual-level information on the components of X within the k^{th} group $\{X_{ki}:i = 1,\ldots,N_k\}$. Note that these data constitute the (observed) group-specific (marginal) covariate distribution.

Exploiting these two features of expression (2), the aggregate data design supplements group-level outcome data with individual-level information on the covariate distribution.[25] Such information could be obtained by surveying individuals for information on exposures, confounders, and effect modifiers. As there is no requirement to link this information to individual-level outcomes, it may be possible to take advantage of existing surveys to obtain these data. When the survey represents a complete enumeration of each group (i.e. all N_k individuals), the combined design is called the full-survey aggregate data design. When a complete enumeration is not available or feasible, the survey sub-sample aggregate data design collects individual-level covariate information on a random sub-sample within each group. Assuming a log link for the individual-level model (1), estimates of β under both designs are obtained as the solution to an estimating equation.[25]

Both the full-survey and survey sub-sample aggregate data design are useful when aggregated outcome counts are available and administering a survey solely for covariates is most practical. In some settings, one may be able to administer the survey to collect individual-level information on outcomes and covariates jointly. Combining these data with group-level outcome information, Martinez et al.[26] proposed the integrated aggregate data design and developed an estimating-equations framework for estimation and inference for β. They showed that combining the two sources of data can correspond to improvements of analyses that use only survey-based individual-level outcome/covariate information.[27]

Hierarchical related regression

Each of the full-survey, survey sub-sample, and integrated aggregate data designs employ semiparametric estimating equations for their analyses. The estimating-equations framework is appealing because it does not rely on assumptions regarding the within-group covariate distributions. However, if one is willing to make distributional assumptions, one can take advantage of a fully parametric statistical framework.[15,28] Modifying our notation slightly, let $\boldsymbol{\pi}_{ki}(x) = P(Y_{ki} = 1| X = x)$ denote the outcome probability for the i^{th} individual in the k^{th} group, given a covariate vector value of X=x. As with expression (1), $\boldsymbol{\pi}_{ki}(x)$ is taken to be specified in terms of individual-level associations of interest. Let $f_k(x)$ denote the joint covariate distribution for the k^{th} group. The induced group-level model is obtained by integrating the individual-level model over $f_k(x)$:

$$\pi_k = \int \pi_{ki}(x) f_k(x) dx. \qquad (3)$$

Specifying $f_k(x)$ depends on the components of X. For example, Jackson et al.[29] consider two covariates, X_1 and X_2. X_1 is binary and follows a Bernoulli distribution; X_2 is continuous and assumed to be normally distributed, conditional on X_1. While specification in general settings does require care, adopting a specific distributional form for $f_k(x)$ can improve small-sample bias and efficiency and power if the assumptions are correct.

A key advantage of the parametric approach is that one can imbed (3) into a fully Bayesian analysis.[2,29] Recently, Jackson et al[30] introduced hierarchical related regression as a flexible Bayesian framework for combining group- and individual-level data. Hierarchical related regression extends previous work developments based on expression (3) in that it permits the use of within-group joint outcome/covariate information. Viewed as a parametric analogue of the integrated aggregate data design, the Bayesian formulation of hierarchical related regression is appealing because it provides flexibility for incorporating prior information and accommodating challenging data features such as spatial structure, measurement error, and missingness. The framework also facilitates data synthesis across various data sources leading to improved power to distinguish individual-level and contextual effects.[31]

Two-phase designs

Recently, the two-phase design was proposed as a convenient framework for overcoming ecologic bias.[32] Briefly, two-phase studies were proposed as an extension of the case-control design for settings where the exposure of interest is rare.[33,34] At phase I, the population is cross-classified according to the outcome and some stratification variable, S. The latter takes on a finite number of levels and is observed on all individuals of the population. The phase I stratification provides an efficient sampling frame from which additional individual-level information is collected on a sub-sample at phase II.[35] In this respect, the design resembles a stratified case-control study with the added advantages of being able to (i) estimate coefficients corresponding to the stratification variable (i.e. S) and (ii) obtain general efficiency gains by incorporating stratified outcome totals for the population.

In the ecologic context, group-level data can be used as the basis for the phase I stratification. A simple strategy is to cross-classify the population by case status and group membership. A drawback, however, is that if the number of groups is large, the phase I stratification will have many strata and potentially small cell sizes. This may lead to a breakdown in the analysis methodology. An alternative strategy is to base the phase I stratification on observed group-level covariate measures. We illustrate this approach in our simulation study below.

Estimation and inference of individual-level association parameters using data from a two-phase design follows using standard weighting or likelihood-based methods; Wakefield and Haneuse give a detailed summary.[32]

Hybrid designs for ecologic inference

Proposed to address ecologic bias directly, the hybrid design for ecologic inference supplements an ecologic study with case-control data drawn from the same underlying population.[36] Specifically, the design assumes that group-level outcome and covariate data are available and that individual-level covariate data (stratified by outcome status) are collected from each group.

Assuming an individual-level logistic model, estimation/inference proceeds via the induced hybrid likelihood, derived by averaging the individual-level likelihood over all the possible configurations of the unobserved complete individual-level data. This differs from the various aggregate data designs and the hierarchical related regression, which consider the induced group-level model derived by averaging the individual-level model over the unobserved individual-level data (see expressions (2) and (3)). Estimation and inference based on the hybrid likelihood can proceed via either maximum likelihood or within the Bayesian framework.[37]

Like the two-phase design, the hybrid design may be viewed as a stratified case-control design. Indeed, the hybrid design that collects case-control samples from each group is equivalent to the two-phase design where the phase I stratification is based on group membership. A key distinction, however, is that under the hybrid design one can choose not to collect individual-level data or to collect case-control data only from certain areas. This provides flexibility at the design stage, where logistical or financial constraints may preclude or limit individual-level data collection for some groups. In contrast, current analysis techniques for the aggregate data design and two-phase design exclude groups for which no or case-only individual-level data are available. Depending on modeling and distributional assumptions, hierarchical related regression also has flexibility to incorporate information from groups with no or case-only individual-level data.

EXAMPLE: LOW BIRTH WEIGHT DATA

To illustrate these approaches, we introduce a simple study of low birth weight (LBW; <2,500 g) and consider the task of estimating the impact of infant race and sex using data compiled by the North Carolina State Center for Health Statistics (http://www.irss.unc.edu/). Restricting to 2003 and 2004, North Carolina had 237,978 births, of which 21,493 were LBW. Across the K = 100 counties, the LBW rate varied from 6.0% to 15.9%; the percent non-white from 0.0% to 76.4%; and the percent male from 45.0% to 56.8% (Figure 1).

Let Y_{ki} be a binary indicator of LBW for the ith infant born in the kth county, and π_{ki} the corresponding probability of LBW. Consider the following individual-level model

$$g(\pi_{ki})=\beta_0+\beta_X X_{ki}+\beta_Z Z_{ki}, \qquad (4)$$

where X_{ki} indicates race (0/1 = white/non-white), Z_{ki} indicates sex (0/1 = female/male), and g() is a link function. In model (4), β_X and β_Z are the individual-level associations of interest.

Notational framework

To make explicit differences in observed data structures between the reviewed designs/methods, Tables 1 and 2 present a notational framework for combining group- and individual-level data, based on the North Carolina LBW example. For ease of exposition, a county-specific subscript is omitted but should be taken as implicit throughout.

Consider a generic county with a population size of N. Let N_{0xz} and N_{1xz} denote the number of LBW non-cases and cases with race/sex pattern [X = x/Z = z], respectively (see Table 1a). Summed across the levels of race and sex, the marginal LBW non-case and case totals are N_0 and N_1. Summing across the levels of LBW, M_{xz} denotes the number of individuals with race/sex pattern [X = x/Z = z]. Table 1a shows the M_{xz} as the marginal totals for N_{yxz}. Table 1b provides the M_{xz} as the joint race/sex distribution directly, together with notation for the marginal race and sex distributions: counts M_{x+}, x=0/1, and M_{+z}, z=0/1, respectively.

Tables 1a and 1b provide upper-case notation representing all individuals in the county; Table 1c provides analogous, lower-case notation for a sub-sample of size n. For example, n_{1xz} denotes the number of LBW cases with race/sex pattern [X = x/Z = z] observed in the sub-sample. Following our review, individual-level data may be observed only on covariates (full-survey and survey sub-sample aggregate data designs) or jointly on outcomes and covariates (integrated aggregate data design, hierarchical related regression, two-phase and hybrid designs).

Data structures

Using the notation of Table 1, Table 2 summarizes observed data structures across various study designs. In an individual-level study, for example, one would observe either the N_{yxz} totals of Table 1a or the n_{yxz} totals of Table 1c, depending on whether data were obtained on all individuals or a sub-sample. Taken across the levels of Y/X/Z, the totals are denoted $\mathbf{N_{yxz}}$ and $\mathbf{n_{yxz}}$, respectively. In contrast, an ecologic study design would observe only county-specific marginal LBW, race, and sex totals: $\{\mathbf{N_y},\mathbf{M_{x+}},\mathbf{M_{+z}}\}$, where $\mathbf{N_y} = \{N_0,N_1\}$, $\mathbf{M_{x+}} = \{M_{0+},M_{1+}\}$, and $\mathbf{M_{+z}} = \{M_{+0},M_{+1}\}$. Tables 1a and 1b make this explicit by presenting the N_{yxz} and M_{xz} counts within square brackets.

Under the full-survey aggregate data design, group-level outcome totals are supplemented with a survey collecting individual-level data on the covariate distribution. For the LBW example, these correspond to the marginal LBW and joint race/sex counts: $\{\mathbf{N_y},\mathbf{M_{xz}}\}$, where $\mathbf{M_{xz}} = \{M_{00},M_{01},M_{10},M_{11}\}$. When a full survey is unavailable or unfeasible, the survey sub-sample aggregate data design supplements the outcome totals with race/sex information on a random sub-sample of n individuals: $\{\mathbf{N_y},\mathbf{m_{xz}}\}$, where $\mathbf{m_{xz}}=\{m_{00},m_{01},m_{10},m_{11}\}$. If one can survey joint individual-level LBW/race/sex information further on a random sub-sample, the integrated aggregate data design combines these data with the group-level outcome totals: $\{\mathbf{N_y},\mathbf{n_{yxz}}\}$. The hierarchical related regression framework, which can be seen as a parametric analogue of the integrated aggregate data design, uses these data structures and any additional covariate information: hence, the observed data may consist of $\{\mathbf{N_y},\mathbf{n_{yxz}}\}$, $\{\mathbf{N_y},\mathbf{M_{x+}},\mathbf{M_{+z}},\mathbf{n_{yxz}}\}$ or $\{\mathbf{N_y},\mathbf{M_{xz}},\mathbf{n_{yxz}}\}$. As noted above, the flexibility of hierarchical related regression also permits contributions from counties where individual-level data are either unavailable (i.e. $\{\mathbf{N_y},\mathbf{M_{x+}},\mathbf{M_{+z}}\}$) or case-only (i.e. $\{\mathbf{N_y},\mathbf{M_{x+}},\mathbf{M_{+z}},\mathbf{n_{1xz}}\}$).

The simplest two-phase study stratifies the entire population by outcome status and county membership. That is, the phase I strata are determined by the $\mathbf{N_y}$ across the K = 100 counties. Within each county, a sub-sample of n_0 non-cases and n_1 LBW cases are sampled and their race/sex status retrospectively determined. Thus the observed data structures are $\{\mathbf{N_y},\mathbf{n_{0xz}},\mathbf{n_{1xz}}\}$. An alternative is to use group-level exposure information to stratify the population. For example, Figure 1b shows county-specific percent non-white rates using five strata; Table 3a provides the corresponding phase I stratification. From each of these 10 strata, one could retrospectively sample individuals and observe their race/sex status. Under this design, the observed data structures are $\{\mathbf{N_y},\mathbf{M_{x+}},\mathbf{n_{0xz}},\mathbf{n_{1xz}}\}$.

Finally, the hybrid design supplements an ecologic study with individual-level case-control data: hence, the available data structures are $\{\mathbf{N_y},\mathbf{M_{x+}},\mathbf{M_{+z}},\mathbf{n_{0xz}},\mathbf{n_{1xz}}\}$. As with HRR, the hybrid design permits contributions from some counties from which either no individual-level data or case-only data are observed: $\{\mathbf{N_y,M_{x+},M_{+z}}\}$ and $\{\mathbf{N_y,M_{x+},M_{+z},n_{1xz}}\}$, respectively.

Simulation study

To further illustrate methods for combining group- and individual-level data, we present a short simulation study based on the North Carolina LBW data. To estimate components of model (4), we considered combined eight designs: (i) full-survey aggregate data design; (ii) survey sub-sample aggregate data design with n = 200 sampled from each county; (iii) integrated aggregate data design supplementing the survey sub-sample aggregate data design with n = 500 more random samples from each of the four largest counties, for which joint outcome/covariate data are surveyed; (iv) two-phase design with phase I stratification based on county membership and n = 2000; (v) two-phase design with phase I stratification based on county-specific non-white prevalence rates (Table 3a) and n = 2000; (vi) two-phase design with phase I stratification based on county-specific sex prevalence rates (Table 3b) and n = 2000; (vii) hybrid design with 250 cases and 250 controls from each of the four largest areas; and (viii) hybrid design with 250 cases from each of the four largest areas. For the two-phase designs, phase II sample sizes were balanced across the phase I strata and estimation based on maximum likelihood.[35] For simplicity, we present only frequentist methods in our simulation study and, in particular, present no results for the hierarchical related regression approach. An online eAppendix (http://links.lww.com) provides the data and code for the simulation study.

For each design, we simulated 10,000 combined group-/individual-level datasets. Throughout, the total number of births and within-county race/sex distributions were held at those in the observed data. Outcome data were generated based on model (4); a log link was used for each aggregate data design; a logit link was used for the two-phase and hybrid designs. Coefficient values for the “true” models were obtained from a fit of the complete individual-level data: (-2.52, 0.59, -0.17) for the log-linear model; (-2.44, 0.66, -0.18) for the logistic model.

Table 4 presents small-sample percent bias, relative efficiency and mean squared error. As analysis techniques for each design/method have been shown to be consistent (asymptotically unbiased), reported bias is due to small samples and, more specifically, not ecologic bias. Further, we note that relative efficiency is defined here as the ratio of the standard error under each design to the standard error for an analysis using individual-level outcome/exposure data. This ratio may be interpreted as how much tighter confidence intervals could be, on average, when one combines the two sources of information compared with using the individual-level data only.

Across all designs, small-sample bias for the race effect is low (at most -2.8%). For the sex effect, bias under the integrated aggregate data design, two-phase, and hybrid designs is low. For the log-linear model, the full-survey and survey sub-sample aggregate data design estimators exhibit substantial small-sample biases of 16.7% and -91.0%, respectively. This contrasts with the two estimators that use individual-level outcome/exposure data (1.9% and -2.4%). The contrasting performance is due to the reliance of the full-survey and survey sub-sample aggregate data designs on between-county exposure variation as their source of information, together with the low variation in the percent male across the 100 counties (Figure 1c). As the percent non-white exhibits substantial between-county variation, the full-survey and survey sub-sample aggregate data designs perform relatively well for the race parameter.

Overall, designs that use group-level data have improved efficiency for estimating the race effect compared with those that use individual-level data only. For the sex effect, the two aggregate data designs that do not access individual-level outcome data suffer from substantially reduced efficiency; in contrast, the integrated aggregate data designs retains much of the benefit for the race effect (48.7% relative efficiency) with no tradeoff in the sex effect (95.4% relative efficiency). Each of the two-phase and hybrid designs outperforms or does no worse than a case-control design. Not surprisingly, the two-phase design that stratifies on group-level race measures has greater efficiency gains than the design that stratifies on group-level sex measures (48.8% reduction vs. 8.4%). In addition to substantial gains for the race effect (standard errors reduced by approximately 62%) and despite low between-county variation in the proportion male, the hybrid likelihood exploits this information to provide moderate efficiency gains of approximately 20% for the sex effect. Further, comparing the two hybrid designs indicates that, at least in this context, a case-only hybrid design may be a reasonable approach. The results for mean squared error reflect those of relative efficiency.

DISCUSSION

When scientific interest lies in individual-level associations, either alone or jointly with group-level associations, the only reliable solution to the ecologic inference problem is to collect individual-level data. Epidemiologists have at their disposal a range of designs that facilitate this; we have sought to provide a comprehensive overview of “combined” designs and associated analysis techniques. A short simulation study highlights potentially substantial efficiency gains associated with combining the two types of information in the analyses.

In practice, the specific choice of design will depend on the individual-level model of interest, the nature of available information and assumptions regarding the data. Currently, the integrated aggregate data design, two-phase, and hierarchical related regression may provide the most convenient and powerful designs for researchers. In addition to the general benefits of the Bayesian framework, hierarchical related regression has the unique advantage of permitting arbitrary link functions [i.e., both log() and logit()] to consider both non-rare and rare outcomes. However, hierarchical related regression requires additional input from researchers in distributional and modeling assumptions. While sensitivity analyses are an option, the semi-parametric analyses of the integrated aggregate data design and two-phase design reduce the need for assumptions and, hence, may be appealing. Under the integrated aggregate data design, the individual-level data are obtained via simple random sampling so that the design will be most useful for non-rare outcomes. Further, Martinez et al.[26,27] developed their analytic framework assuming a log-linear model. The two-phase design, in contrast, is a stratified case-control study with analysis approaches having been developed assuming a logistic model for the outcome. Hence, it would likely be most appealing for rare outcomes.

Our simulation suggests the hybrid design experiences the greatest efficiency gain from the inclusion of group-level data. This is likely due to the induced likelihood’s direct use of group-level covariate data when characterizing possible configurations of the unobserved joint outcome/covariate data. The aggregate data design does not exploit such information; the two-phase design may use between-group covariate information but only indirectly as part of the phase I stratification. The hierarchical related regression approach of Jackson et al[30] also uses group-level covariate data to help inform and estimate within-group covariate distributions. A drawback of the hybrid likelihood, however, is that it is computationally expensive and the statistical development has so far been limited to a few categorical covariates. None of the other reviewed designs are limited in this respect. While the simulation study did not examine hierarchical related regression, we anticipate its performance being similar to the integrated aggregate data design and hybrid designs. A comprehensive statistical evaluation of each of the designs and methods is beyond the scope of this paper but could be useful for researchers considering these designs.

Beyond statistical considerations, when choosing between combined designs, researchers need to weigh numerous practical and epidemiologic issues. For example, logistical and financial constraints may preclude the collection of individual-level data from each group or area. In other settings, researchers may look to supplement readily available individual-level data with appropriate group-level data.[38] From an epidemiologic perspective, model specification and interpretation can be challenging in multi-level settings. Specific issues include distinguishing between- from within-group effects; appropriately using between- and within-group exposure variation; characterizing and identifying between- and within-group confounding; identifying potential contextual effects; and ensuring compatibility of differing data sources. These issues are crucial to the design process in that they determine the data elements that require collection.[23,24,31]

We emphasize that no single design is ideal, and researchers have flexibility to tailor their choice to their specific setting. Indeed, the sequential nature of the designs (that is, collecting individual-level data given group-level data) lends itself to considering design issues that may improve efficiency, with group-level characteristics potentially being incorporated into decision-making. To date, little work has focused on study design in this context.[32,39,40] Further work on these competing strategies of sampling and analyses would give researchers practical guidance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sources of financial support: This study was funded, in part, by NCI R-01 grant #CA125081.

References

1. Morgenstern, H. Ecologic studies. In: Rothman, KJ.; Greenland, S.; Lash, T., editors. Modern Epidemiology. Third. Philadelphia: Lippincott Williams & Wilkins; 2008. p. 511-531.
2. Best NG, 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 Stat Soc Ser A Stat Soc. 2001; 164(1):155–174.
3. Wilkinson P, Thakrar B, Walls P, et al. Lymphohaematopoietic malignancy around all industrial complexes that include major oil refineries in Great Britain. Occup Environ Med. 1999; 56(9):577–80. [PubMed: 10615289]
4. Whitley, E.; Darby, S. Quantifying the risks from residential radon. In: Barnett, V.; Stein, A.; Turkman, K., editors. Statistics for the Environment 4: Statistical Aspects of Health and the Environment. Chichester: John Wiley & Sons; 1999. p. 71-89.
5. Maheswaran 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(4):455–60. [PubMed: 10490560]
6. Hu S, Ma F, Collado-Mesa F, Kirsner RS. Ultraviolet radiation and incidence of non-Hodgkin's lymphoma among Hispanics in the United States. Cancer Epidemiol Biomarkers Prev. 2004; 13(1): 59–64. [PubMed: 14744734]

7. Reynolds P, Hurley SE, Gunier RB, et al. Residential proximity to agricultural pesticide use and incidence of breast cancer in California, 1988-1997. Environ Health Perspect. 2005; 113(8):993–1000. [PubMed: 16079069]

8. Shaw PA, Etzioni R, Zeliadt SB, et al. An ecologic study of prostate-specific antigen screening and prostate cancer mortality in nine geographic areas of the United States. Am J Epidemiol. 2004; 160(11):1059–69. [PubMed: 15561985]

9. Das B, Feuer EJ, Mariotto A. Geographic association between mammography use and mortality reduction in the US. Cancer Causes Control. 2005; 16(6):691–9. [PubMed: 16049808]

10. Grant W. Ecological study of dietary and smoking links to lymphoma. Altern Med Rev. 2000; 5(6):563–72. [PubMed: 11134979]

11. Pepin J. From the Old World to the New World: an ecologic study of population susceptibility to HIV infection. Trop Med Int Health. 2005; 10(7):627–39. [PubMed: 15960701]

12. Simonsen L, Reichert TA, Viboud C, et al. Impact of influenza vaccination on seasonal mortality in the US elderly population. Arch Intern Med. 2005; 165(3):265–72. [PubMed: 15710788]

13. Goldhagen J, Remo R, Bryant T 3rd, et al. The health status of southern children: a neglected regional disparity. Pediatrics. 2005; 116(6):e746–53. [PubMed: 16263972]

14. Michael Y, Yen I. Built environment and obesity among older adults--can neighborhood-level policy interventions make a difference? [Commentary]. Am J Epidemiol. 2009; 169(4):409–412. [PubMed: 19153213]

15. Richardson S, Stucker I, Hemon D. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. Int J Epidemiol. 1987; 16(1):111–20. [PubMed: 3570609]

16. Piantadosi S, Byar D, Green S. The ecological fallacy. Am J Epidemiol. 1988; 127(5):893–904. [PubMed: 3282433]

17. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol. 1989; 18(1):269–74. [PubMed: 2656561]

18. Greenland S. Divergent biases in ecologic and individual-level studies. Stat Med. 1992; 11(9): 1209–1223. [PubMed: 1509221]

19. Greenland S, Robins J. Ecologic studies--biases, misconceptions, and counterexamples [Commentary]. Am J Epidemiol. 1994; 139(8):747–60. [PubMed: 8178788]

20. Richardson, S.; Monfort, C. Spatial epidemiology: methods and applications. Oxford: Oxford University Press; 2000. Ecological correlation studies.

21. Wakefield J. Ecological inference for 2×2 tables. J R Stat Soc Ser A Stat Soc. 2004; 167(3):385–445.

22. Wakefield J. Sensitivity analyses for ecological regression. Biometrics. 2003; 59(1):9–17. [PubMed: 12762436]

23. Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. Am J Public Health. 1998; 88(2):216–22. [PubMed: 9491010]

24. Diez-Roux AV. The study of group-level factors in epidemiology: rethinking variables, study designs, and analytical approaches. Epidemiol Rev. 2004; 26:104–11. [PubMed: 15234951]

25. Prentice RL, Sheppard L. Aggregate data studies of disease risk factors. Biometrika. 1995; 82(1): 113–125.

26. Martinez JM, Benach J, Ginebra J. An Integrated Analysis of Individual and Aggregated Health Data Using Estimating Equations. Int J Biostat. 2007; 3(1):10.

27. Martinez JM, Benach J, Benavides FG, et al. Improving multilevel analyses: the integrated epidemiologic design. Epidemiology. 2009; 20(4):525–32. [PubMed: 19436212]

28. Lasserre V, Guihenneuc-Jouyaux C, Richardson S. Biases in ecological studies: utility of including within-area distribution of confounders. Stat Med. 2000; 19(1):45–59. [PubMed: 10623912]

29. Jackson CH, Best NG, Richardson S. Improving ecological inference using individual-level data. Stat Med. 2006; 25(12):2136–59. [PubMed: 16217847]

30. Jackson CH, Best NG, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. J R Stat Soc Ser A Stat Soc. 2008; 171(1):159–178.

31. Jackson CH, Richardson S, Best NG. Studying place effects on health by synthesising individual and area-level outcomes. Soc Sci Med. 2008; 67(12):1995–2006. [PubMed: 18950921]

32. Wakefield J, Haneuse S. Overcoming ecologic bias using the two-phase study design. Am J Epidemiol. 2008; 167(8):908–16. [PubMed: 18270370]

33. White E. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol. 1982; 115(1):119–28. [PubMed: 7055123]

34. Weinberg CR, Wacholder S. The design and analysis of case-control studies with biased sampling. Biometrics. 1990; 46(4):963–75. [PubMed: 2085641]

35. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. J R Stat Soc Ser C Appl Stat. 1999; 48(4):457–68.

36. Haneuse S, Wakefield J. Geographic-based ecological correlation studies using supplemental case-control data. Stat Med. 2008; 27(6):864–87. [PubMed: 17624917]

37. Haneuse S, Wakefield J. Hierarchical Models for Combining Ecological and Case Control Data. Biometrics. 2007; 63(1):128–136. [PubMed: 17447937]

38. Stromberg U, Bjork J. Incorporating group-level exposure information in case-control studies with missing data on dichotomous exposures. Epidemiology. 2004; 15(4):494–503. [PubMed: 15232411]

39. Plummer M, Clayton D. Estimation of population exposure in ecological studies. J R Stat Soc Series B Stat Methodol. 1996; 58(1):113–26.

40. Sheppard L, Prentice RL, Rossing MA. Design considerations for estimation of exposure effects on disease risk, using aggregate data studies. Stat Med. 1996; 15(17-18):1849–58. [PubMed: 8888477]

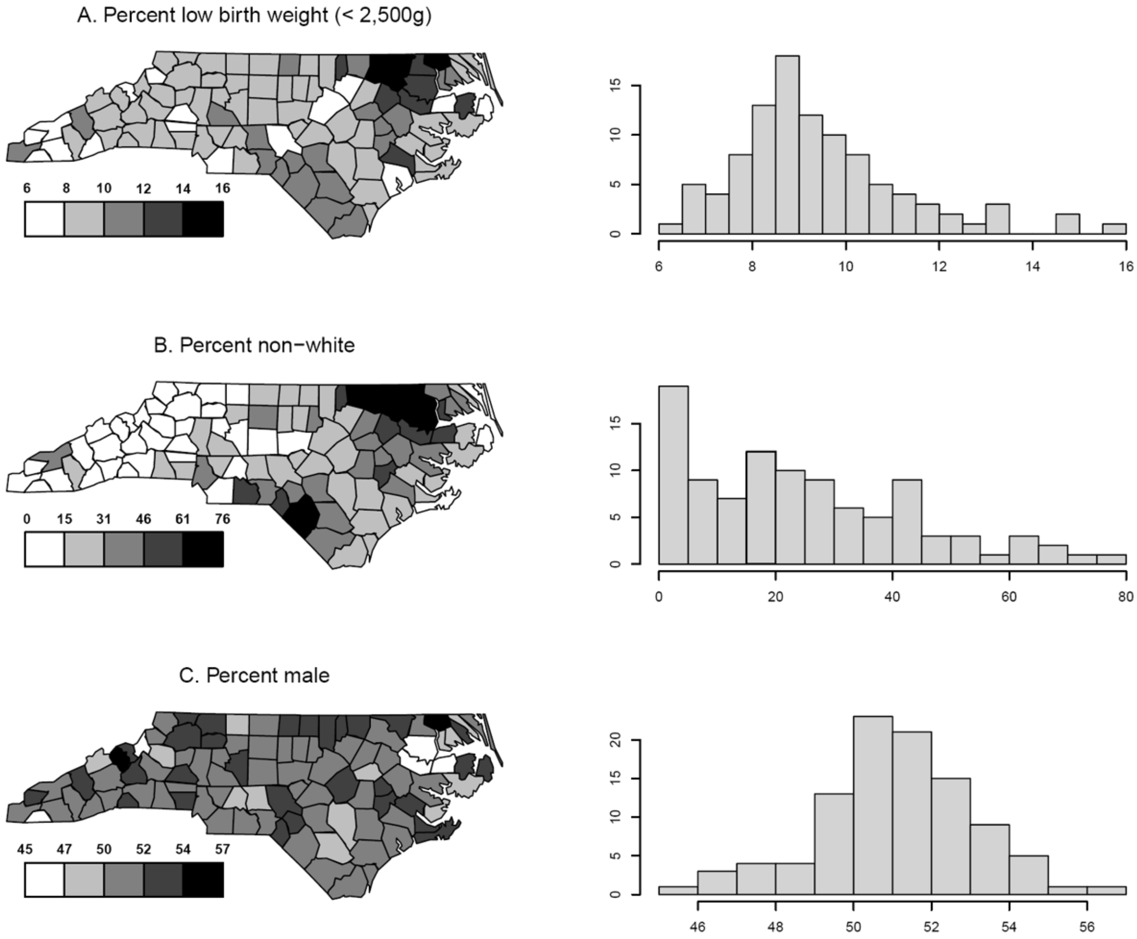

Figure 1.

County-specific outcome and exposure data for the North Carolina low birth weight data.

Table 1

Notational framework for combining group- and individual-level data from a generic county, for the North Carolina low-birth-weight example. Upper-case letters denote information representing all individuals in the county; lower-case letters denote information on a sub-sample. Quantities in square brackets are not observed in an ecological study.

(a) Group-level outcome totals				**(b) Group-level covariate totals**				**(c) Individual-level data**			
Race/ Sex	**LBW = 0**	**LBW = 1**			**Race = 0**	**Race = 1**		**Race/ Sex**	**LBW = 0**	**LBW = 1**	
0/0	[N_{000}]	[N_{100}]	[M_{00}]	Sex = 0	[M_{00}]	[M_{10}]	M_{+0}	0/0	n_{000}	n_{100}	m_{00}
1/0	[N_{010}]	[N_{110}]	[M_{10}]	Sex = 1	[M_{01}]	[M_{11}]	M_{+1}	1/0	n_{010}	n_{110}	m_{10}
0/1	[N_{001}]	[N_{101}]	[M_{01}]		M_{0+}	M_{1+}	N	0/1	n_{001}	n_{101}	m_{01}
1/1	[N_{011}]	[N_{111}]	[M_{11}]					1/1	n_{011}	n_{111}	m_{11}
	N_0	N_1	N						n_0	n_1	n

LBW, low birth weight 0/1 = no/yes; Race 0/1 = white/non-white; Sex 0/1 = female/male.

Table 2

Data structures under various designs that consider group- and/or individual-level data. Upper-case letters denote information on the entire population and lower-case letters denote information on a sub-sample.[a]

	Group-level data	**Individual-level data**
Individual-level study		
Complete		$\mathbf{N_{yxz}}$
Sub-sample		$\mathbf{n_{yxz}}$
Ecological study	$\mathbf{N_y, M_{x+}, M_{+z}}$	
Aggregate data designs		
Full survey	$\mathbf{N_y}$	$\mathbf{M_{xz}}$
Survey sub-sample	$\mathbf{N_y}$	$\mathbf{m_{xz}}$
Integrated	$\mathbf{N_y}$	$\mathbf{n_{yxz}}$
Hierarchical related regression	$\mathbf{N_y, M_{x+}, M_{+z}}$	$\mathbf{M_{xz}, n_{yxz}}$
Two-phase study	$\mathbf{N_y, M_{x+}, M_{+z}}$	$\mathbf{n_{0xz}, n_{1xz}}$
Hybrid design for ecological inference		
Case-control hybrid	$\mathbf{N_y, M_{x+}, M_{+z}}$	$\mathbf{n_{0xz}, n_{1xz}}$
Case-only hybrid	$\mathbf{N_y, M_{x+}, M_{+z}}$	$\mathbf{n_{1xz}}$

[a] Y=LBW, low birth weight 0/1=no/yes; X=Race 0/1=white/non-white; Z=Sex 0/1=female/male.

Table 3

Phase I stratification for the North Carolina low birth weight data, with S stratifying births according to county-specific percentage of non-white births and of male births.

	County-specific percentage non-white births				
	0.0%–15.0%	**15.0%–31.0%**	**31.1%–46.0%**	**46.1%–61.0%**	**61.1%–76.0%**
Y = 0	49,130	89,341	69,248	6406	6491
Y = 1	4131	8455	7153	817	937
	County-specific percentage male births				
	45.0%–47.0%	47.1%–50.0%	50.1%–52.0%	52.1%–54.0%	54.1%–57.0%
Y = 0	1440	20,290	155,846	33,918	4991
Y = 1	175	2069	15,395	3321	533

Table 4

Operating characteristics of estimators for the parameters from an individual-level model of low birth weight, under various designs. Values are based on 10,000 simulated datasets.

	Percent Bias		Relative Efficiency[e]		Mean Squared Error	
	Race	**Sex**	**Race**	**Sex**	**Race**	**Sex**
Log-linear model[a]						
Individual-level data only[b]	0.0	1.9	100.0	100.0	19.0	19.7
Full ADD	0.0	16.7	27.3	491.7	1.4	495.5
Survey sub-sample ADD	-2.8	-91.0	27.9	137.5	2.2	863.9
Integrated ADD	-1.7	-2.4	48.7	95.4	4.8	18.2
Logistic model[a]						
Case-control only	0.2	0.7	100.0	100.0	9.7	8.0
Two-phase design: race[c]	0.1	-0.6	51.2	102.7	2.5	8.4
Two-phase design: sex[d]	0.2	0.7	91.6	103.3	8.2	8.6
Hybrid design: case-control	0.6	1.0	37.2	79.1	1.4	5.1
Hybrid design: case-only	0.8	1.4	37.3	79.8	1.4	5.3

[a] $(\beta_0, \beta_X, \beta_Z) = (-2.52, 0.59, -0.17)$ for the log-linear model and $(-2.44, 0.66, -0.18)$ for the logistic model

[b] n=500 individual-level samples from each of the four largest counties

[c] Phase I stratification based on outcome and county-specific proportion of births that are non-white (Table 3)

[d] Phase I stratification based on outcome and county-specific proportion of births that are male

[e] Ratio of standard error estimates, with denominator taken to be standard error for estimator based solely on individual-level data.