STATISTICAL AND SUBSTANTIVE INFERENCES IN PUBLIC HEALTH: Issues in the Application of Multilevel Models

Jeffrey B. Bingenheimer¹ and Stephen W. Raudenbush²

¹ School of Public Health, and ²School of Education, Survey Research Center, and Departments of Statistics and Sociology, University of Michigan, Ann Arbor, Michigan 48109; email: bartbing@umich.edu, rauden@umich.edu

Key Words epidemiologic methods, hierarchical linear models, causation, cluster-randomized trials, empirical Bayes

Abstract Multilevel statistical models have become increasingly popular among public health researchers over the past decade. Yet the enthusiasm with which these models are being adopted may obscure rather than solve some problems of statistical and substantive inference. We discuss the three most common applications of multilevel models in public health: (a) cluster-randomized trials, (b) observational studies of the multilevel etiology of health and disease, and (c) assessments of health care provider performance. In each area of investigation, we describe how multilevel models are being applied, comment on the validity of the statistical and substantive inferences being drawn, and suggest ways in which the strengths of multilevel models might be more fully exploited. We conclude with a call for more careful thinking about multilevel causal inference.

INTRODUCTION

Over the past decade, interest in multilevel statistical models has increased dramatically in public health. This is evidenced not only by the proliferation of published articles in which multilevel modeling techniques are used, but also by the growing number of books (52, 62, 66, 78, 84, 87, 107, 118, 141) on the subject, as well as by the appearance of numerous invited commentaries and review articles (1, 18, 20, 21, 22, 26, 33, 70, 76, 111, 114, 123, 124, 126, 140, 142, 149) addressing this topic in public health and social scientific periodicals. As Mason (92) observed several years ago, however, the diffusion of a new statistical methodology within any field of study follows a predictable pattern: Overzealous early adopters tout the method as a panacea, whereas critics charge that it offers nothing new to the field. Ultimately, this process is resolved only when the legitimate advantages and limitations of the novel methodology become widely recognized. The methodology then finds its rightful place within the field’s “armamentarium” (92, p. 221).
Such a process now appears to be well underway with multilevel statistical models in public health.

In this review, we attempt to accelerate Mason’s diffusion process by presenting a critical discussion of the ways in which public health investigators have used multilevel models. For many types of data and a wide range of research questions, multilevel models provide a stronger basis for statistical inference than traditional, single-level models. Like any technology, however, multilevel models have their limitations. Our goal here is to identify both the advantages and the limitations of multilevel models, distinguishing the inferences that are strengthened by their use from those that are not.

After first introducing the terminology and notation to be used throughout the review, we identify three sets of questions to which public health investigators have applied multilevel models. These questions address (a) the effects of group-level interventions, (b) the multilevel etiology of health outcomes, and (c) the relative performance of health service providers. For each set of questions, we show how multilevel models are being applied, comment on the validity of the interpretations and inferences that are being drawn, and provide suggestions about how the strengths of multilevel data analysis might be more fully exploited. We conclude with a call for more sophisticated thinking about multilevel causal inference.

**SOME SIMPLE MULTILEVEL MODELS**

Suppose that we wish to study variations in body mass index (BMI) among young adults, focusing on sex differences and differences related to the presence or absence of fast-food restaurants in individuals’ neighborhoods. We let $Y_{ij}$ denote the BMI of the $i$th person living in the $j$th neighborhood. To begin our analysis, we assume that within each neighborhood BMI follows a normal distribution with a neighborhood-specific mean, $\beta_{0j}$, and a variance, $\sigma^2$. Furthermore, we assume that the neighborhood-specific means themselves vary according to a normal distribution with mean $\gamma_{00}$ and variance $\tau_{00}$. These simple assumptions lead to the following two-level model:

$$Y_{ij} = \beta_{0j} + r_{ij} \quad (1a)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (1b)$$

$$r_{ij} \sim N(0, \sigma^2); \quad u_{0j} \sim N(0, \tau_{00}); \quad \text{cov}(r_{ij}, u_{0j}) = 0. \quad (1c)$$

Equation 1a represents the variability of BMI within each neighborhood and is called the level-1 model, whereas Equation 1b represents the variability between neighborhoods and is called the level-2 model. The expressions in 1c constitute the variance-covariance structure of the model, and the symbol $\sim$ should be read as “are independent and take the following distribution.” We can substitute Equation 1b into Equation 1a to obtain the combined model

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}, \quad (2)$$
with the same variance-covariance structure as given in Equation 1c. This model, which incorporates no covariates, is identical to a one-way random effects analysis of variance. It is often used to partition variation into two levels. Note that the total variance of \( Y_{ij} \) is \( \sigma^2 + \tau_{00} \), whereas the between-neighborhood component of this variance is simply \( \tau_{00} \). This leads to the following expression for the intracluster correlation coefficient, which represents the proportion of variability in BMI occurring between, rather than within, neighborhoods:

\[
\rho = \frac{\tau_{00}}{\sigma^2 + \tau_{00}}. \tag{3}
\]

As a next step in our investigation, we introduce a level-1 covariate, \( X_{ij} \), which takes the value zero if the respondent is female and one if the respondent is male. Now we can rewrite the model given in Equations 1a through 1c as follows:

\[
Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \tag{4a}
\]

\[
\beta_{0j} = \gamma_{00} + u_{0j} \tag{4b}
\]

\[
\beta_{1j} = \gamma_{10} + u_{1j} \tag{4c}
\]

\[
r_{ij} \sim N(0, \sigma^2); \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \right]. \tag{4d}
\]

Note that the level-2 model now comprises two equations, 4b and 4c, and that the variance-covariance structure given in Equation 4d is somewhat more complicated than before. This model implies that, within a given neighborhood, the mean BMI for females is \( \beta_{0j} \), and the mean BMI for males is \( \beta_{0j} + \beta_{1j} \). Thus, the mean difference in BMI between males and females in this neighborhood is \( \beta_{1j} \). Across neighborhoods, the mean BMI is \( \gamma_{00} \) for females and \( \gamma_{00} + \gamma_{10} \) for males, and the mean difference is thus \( \gamma_{10} \). The female-male difference is not constant across neighborhoods, but varies according to a normal distribution with mean \( \gamma_{10} \) and variance \( \tau_{11} \). Setting \( \tau_{11} = 0 \) simplifies the model by implying that the female-male difference is the same in all neighborhoods. Again, substitution may be used to combine Equations 4a, 4b, and 4c into a single combined model.

Alternatively, we might choose to add a level-2 covariate to the model given by Equations 1a through 1c. Suppose, for instance, that \( W_j \) is an indicator variable taking the value one if the \( j \)th neighborhood contains a fast-food restaurant, and zero if it does not. Because \( W_j \) characterizes neighborhoods rather than individuals, we include it in the level-2 model

\[
Y_{ij} = \beta_{0j} + r_{ij} \tag{5a}
\]

\[
\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \tag{5b}
\]

with the same variance-covariance structure given in Equation 1c. The level-1 model in Equation 5a implies that within each neighborhood BMI follows a normal distribution with neighborhood-specific mean \( \beta_{0j} \) and variance \( \sigma^2 \). The level-2
model characterizes the distribution of these neighborhood-specific means. For neighborhoods without a fast-food restaurant, these means vary around \( \gamma_{00} \); but for neighborhoods with a fast-food restaurant, they vary around \( \gamma_{00} + \gamma_{01} \). If neighborhoods were randomly assigned to \( W_j \), then we would have a cluster-randomized trial with the experimental condition being the presence of a fast-food restaurant, and \( \gamma_{01} \) would be interpreted as the average treatment effect.

Finally, consider a model that includes covariates at level 1 and level 2. Again, we let \( X_{ij} = 1 \) if the individual is male and \( W_j = 1 \) if the neighborhood contains a fast-food restaurant. Then we write the following model:

\[
Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}
\]

\[
\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}
\]

\[
\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}
\]

with the variance-covariance structure given in Equation 4d. In this case, substitution of Equations 6b and 6c into 6a gives the following combined model:

\[
Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + \gamma_{11} X_{ij} W_j + u_{0j} + u_{1j} X_{ij} + r_{ij}.
\]

The combined formulation of the model given in Equation 7 highlights an important feature of this model, namely the presence of a cross-level interaction represented by the term \( \gamma_{11} X_{ij} W_j \). This interaction can be interpreted in two ways. First, the average difference between males and females depends upon whether or not a fast-food restaurant is present in the neighborhood (\( \gamma_{10} \) in neighborhoods without a fast-food restaurant; \( \gamma_{10} + \gamma_{11} \) in those with a fast-food restaurant). Alternatively, the difference between neighborhoods with and without fast food restaurants depends upon the sex of the individual. For females the average difference is \( \gamma_{01} \), whereas for males the average difference is \( \gamma_{01} + \gamma_{11} \). The cross-level interaction may be omitted from this model by dropping \( \gamma_{11} W_j \) from Equation 6c. The model may be further simplified by omitting \( u_{0j} \) to obtain a constant intercepts model or by omitting \( u_{ij} \) as discussed above. Naturally, decisions to omit or include covariates, cross-level interactions, and components of variation should be based upon theoretical and empirical considerations, as well as the purpose of the research.

Models of the type presented in Equations 1 through 7 may be generalized in a variety of ways. Instead of having individuals and neighborhoods respectively as the level-1 and level-2 units, for example, we might have repeated measures (level 1) over time on several individuals (level 2). Designs of this sort can be used to estimate individual growth curves and can also incorporate time-varying covariates (79, 85). Alternatively, different assumptions about the distribution of \( Y_{ij} \) could be incorporated into the level-1 model. If \( Y_{ij} \) represented whether or not a given respondent is obese, for example, then its distribution would be Bernoulli instead of normal and a logistic link function could be used at level 1. Indeed, multilevel models can be specified for normal, binary, categorical, ordinal, count,
rate, and time-to-event outcome variables, incorporating appropriate distributional assumptions and link functions. Moreover, the model can be extended to three or more levels to account for a variety of study designs (e.g., repeated measures within individuals within neighborhoods; individuals within classrooms within schools). Readers interested in further details are encouraged to consult any of the books that are now available on multilevel modeling (52, 62, 66, 78, 84, 87, 107, 118, 141).

We now have all of the notation required for our review of multilevel modeling in public health. We shall refer to the level-1 coefficients $\beta_0j$ and $\beta_1j$ as microparameters; to the level-2 coefficients $\gamma_{00}$, $\gamma_{01}$, $\gamma_{10}$, and $\gamma_{11}$ as macroparameters; to $u_0j$ and $u_1j$ as random effects; and to the terms $\sigma^2$, $\tau_{00}$, $\tau_{11}$, and $\tau_{01} = \tau_{10}$ as variance components (93).

**QUESTIONS ABOUT THE EFFECTS OF GROUP-LEVEL INTERVENTIONS**

Perhaps the most straightforward application of multilevel models in public health is to data arising from cluster-randomized trials. These trials are experiments in which randomization is implemented at the group level, but outcome variables are measured at the individual level. Numerous cluster-randomized trials have been conducted in public health, with randomization of cities (12, 13, 45, 47, 56, 74), housing developments (139), schools (63, 89, 157), classrooms (136), and worksites (54, 64, 69). The natural question that arises in such studies is, Did the intervention make a difference? A simple multilevel model for addressing this question is given in Equations 5a and 5b, and the research question involves the value of the macroparameter, $\gamma_{01}$. The model can be used to obtain an estimate of this macroparameter, to test statistical hypotheses about it, and to construct a confidence interval for it.

Multilevel models offer several advantages over other data analytic strategies for cluster-randomized trials. Public health investigators have realized for several decades that individual-level analyses of data from cluster-randomized trials produce excessive Type I errors. Such analyses ignore the dependence of individuals in the same cluster, causing the precision of the estimate of $\gamma_{01}$ to be overstated. Indeed, in 1978, Cornfield described the practice of applying individual-level analytic techniques to data from cluster-randomized trials as “an exercise in self-deception” (14, p. 102). The primary advantage of multilevel models in this setting, therefore, is that they provide an accurate representation of the sources of variability in the data, and thus lead to more honest test statistics, $p$-values, and confidence intervals.

An alternative data analytic strategy for cluster-randomized trials is to conduct a single-level analysis at the level of the group. The investigator first computes a collection of cluster-specific means, then treats these means as the dependent variable in a single-level model. Individual-level covariate adjustment can be incorporated in the first stage by applying some specified covariate distribution to each cluster, and then treating the adjusted means as outcomes. Multilevel models,
in contrast, implement both stages simultaneously, allowing for direct adjustment for individual-level covariates through their inclusion in the level-1 model, as in Equations 6a through 6c. They also allow the investigator to test hypotheses about the variability of treatment effects across subgroups defined by individual-level covariates, through the inclusion of cross-level interaction terms. Thus, compared with the single-level analysis of (adjusted) cluster-specific means, multilevel models offer advantages of convenience and flexibility. In most cases they also provide greater statistical power.

In addition to these advantages, insights from multilevel modeling have been used to aid in the design of cluster-randomized trials. Recognizing that statistical power in such studies depends upon the number of clusters in each treatment condition, the number of individuals measured in each cluster, and the intracluster correlation coefficient, several authors (e.g., 27, 77, 116) have developed methods for determining the sample size requirements for these trials. It should be noted that the extreme case of only one cluster per treatment condition (e.g., 9, 96) provides no basis at all for statistical inference. Other investigators (e.g., 60, 109, 110, 138) have provided estimates of intracluster correlation coefficients to help facilitate sample size calculations.

Because including additional clusters in such trials is typically quite expensive, adjustment for preintervention measures of outcome variables is often used as a cost-effective way of increasing power. This can be done through either of two designs: (a) the longitudinal cohort design, in which measurements are taken on the same individuals before and after the intervention, and (b) the repeated crosssections design, in which samples are drawn independently before and after the intervention. Diehr and colleagues (19) have considered the relative merits of these strategies. Multilevel models can be adapted easily to either of these design options. Several references are available concerning the design and analysis of cluster-randomized trials (28, 29, 76, 106, 107, 140, 142).

Several limitations should be noted in the interpretation of estimated treatment effects. The first stems from the possibility that, in theory, the effect of any treatment could vary across clusters. Although a given intervention may decrease smoking rates in one worksite, for example, it may have no effect on smoking rates in another. The study designs discussed above provide estimates of the average treatment effect, but provide no information about the magnitude of the treatment effect within a specific cluster, nor the degree of variability of the treatment effect between clusters. Thus, although cluster randomized trials can help answer the question, Did this cluster-level intervention make a difference on average?, they cannot answer the question, Did the intervention make a difference in this specific cluster? Testing interaction terms between cluster-level treatment indicators and other cluster- or individual-level variables can, however, provide some insight into the possible systematic variation in the effect of the treatment.

The second limitation concerns the generalizability of the average treatment effect. Typically, the clusters in cluster-randomized trials are selected on the basis of convenience and do not contain a random sample from any well-defined population.
Any inference that the estimated treatment effect will apply to other clusters, therefore, cannot be made on the basis of probability theory but must rely instead upon careful judgments about the similarity of possible future intervention sites to those that were studied in the trial. In the absence of random selection of clusters, therefore, cluster-randomized trials cannot answer the question, What would be the effect of implementing this intervention in other clusters?

We wish to emphasize the importance of randomization to the interpretation of the treatment effect as a causal quantity. On many occasions, major group-level intervention studies have allocated clusters to treatment conditions on the basis of convenience or the preferences of local policymakers or administrators (9, 36, 41–44, 46, 88, 90, 96, 108, 112). Nonrandom assignment threatens the validity of the causal interpretation of the treatment effect because treatment and control clusters may then differ systematically in terms of (a) the preintervention levels of dependent variables, (b) the distribution of level-1 covariates that interact with the treatment, or (c) the potential effect of the treatment. This third factor is especially worrisome because clusters may self-select into treatment conditions based on the magnitude of their expected benefit. School administrators, for instance, might be more likely to opt for an intervention condition, given the choice, if they believe (possibly for good reason) that their students will benefit from that program. Moreover, this type of self-selection process may not be related to observable covariates, and thus may not be amenable to statistical adjustment to reduce the resulting bias.

Finally, for a variety of reasons, statistical inference in cluster-randomized trials depends strongly upon the number of clusters in each treatment condition. As discussed above, the number of clusters per condition is an important determinant of statistical power (27, 116). Furthermore, when the number of clusters in each condition is greater than one but still relatively small, key results from asymptotic theory may not apply. In particular, most hypothesis testing and confidence interval construction for cluster-randomized trials rely on the assumption that estimates of the treatment effect are distributed normally around their true value. When the number of clusters is large, the central limit theorem guarantees that this assumption will be approximately true. When the number of clusters in each condition is small, however, the distribution of treatment effect estimates may be dramatically nonnormal. Moreover, the normality assumption is especially difficult to test in these circumstances because there are too few residuals for assessing the shape of their distribution. Relatedly, when the number of clusters is relatively small and the within-cluster sample sizes vary widely, the standard error of the treatment effect will generally be too small. Although a Bayesian analysis can mitigate this problem to some degree [see (118), pp. 410–12], we recommend striving for equal within-cluster sample sizes in cluster-randomized trials with a small number of clusters in each experimental condition.

Virtually all of the limitations discussed above concern the design of cluster-randomized trials. Studies with too few clusters provide little or no basis for statistical inference; those that fail to randomize clusters to treatment conditions undermine the basis for causal inference; and even well-designed and well-executed
cluster-randomized trials provide limited information about the variability and generalizability of treatment effects. These limitations are inherent in the data. No statistical model, multilevel or otherwise, can substitute for careful study design.

QUESTIONS ABOUT MULTILEVEL ETIOLOGY

Although the analysis of cluster-randomized trials may represent the most straightforward application of multilevel models in public health, the most widespread use of such models to date may be found in observational studies of multilevel health etiology. Researchers working in this area pose the general question, Do neighborhoods matter for health? Social epidemiologists have reported stark differences between neighborhoods in age-adjusted mortality rates, with the residents of impoverished areas being much more likely to die young (51, 59, 98). Several ecological studies likewise revealed higher mortality rates and worse health in geographic aggregates characterized by widespread deprivation or concentrated disadvantage (37, 97, 101, 105, 115, 122). These findings raise the possibility that characteristics of shared physical or social environments may make important contributions to individual health. Yet the difficulties in interpreting ecological associations are well known to epidemiologists (58, 113, 144, 146), and the need for multilevel data, and multilevel models, is therefore clear.

Initial Partitioning of Variance

A preliminary question that arises in investigations of the multilevel etiology of health and disease is, How much do neighborhoods (or other geographical aggregates) vary on key health measures? This question concerns the magnitude of the variance component, \( \tau_{00} \), in Equation 1. Fitting this simple model provides an estimate of \( \tau_{00} \) as well as a test of the null hypothesis that \( \tau_{00} = 0 \) (i.e., that there is no variation between neighborhoods). Moreover, if the variable in question is continuous and an identity link function is chosen, this model also provides an estimate of \( \sigma^2 \), which can then be used along with the estimate of \( \tau_{00} \) to compute the intracluster correlation coefficient, \( \rho \), using Equation 3. Several investigations of the multilevel etiology of disease begin with precisely this approach (7, 30, 31, 32, 48, 55, 61).

Although the resulting statistics provide an interesting descriptive summary of the data, they should not be overinterpreted. Three points should be borne in mind. First, the statistical test of the null hypothesis \( \tau_{00} = 0 \) may be relatively weak, leading to potential type II errors. Therefore, failure to reject this null hypothesis does not necessarily imply that there is no variation between neighborhoods. Second, when the outcome variable is not continuous and a link function other than identity (e.g., log or logit) is used at level 1, the level-1 variation cannot be easily summarized by a single term like \( \sigma^2 \), and therefore \( \rho \) cannot be computed. A variety of approaches have been proposed for dealing with this problem (e.g., 118, p. 298; 141, p. 224), but none is wholly satisfactory.
Third, and most important, there is no direct correspondence between the amount of variation falling at a given level and the extent to which explanatory variables may be found at that level. Indeed, a relatively small value of $\rho$ may correspond to relatively large standardized mean differences (35). It is well known and often noted that variation between neighborhoods may be due to the individual-level characteristics of residents. Conversely, neighborhood-level covariates may play an important etiological role even when there is virtually no between-neighborhood variation. This somewhat counterintuitive point reflects the complex ways in which etiologically important covariates at multiple levels can interact to produce different patterns of variation in an outcome measure. Investigators should not be discouraged from exploring the possible contributions of neighborhood-level covariates, even when $\tau_{00}$ is small.

**Context and Composition**

As a natural follow-up to the unconditional partitioning of variation described above, investigators often pose the question, To what extent are observed variations between neighborhoods due to characteristics of the individuals residing in them? This question is typically framed using the language of “context versus composition” (20, 33). For example, divergent mortality rates in different neighborhoods could be explained by the concentration of retirement communities and nursing homes in some areas, and of young families in others. In such a situation, one might observe considerable between-neighborhood variation in mortality rates, but this variation could be wholly explained by the ages of individual residents.

To address this problem, investigators (7, 30, 31, 32, 48, 55, 61) often introduce several individual-level covariates, $X_{1ij}, \ldots, X_{kij}$, to the model at level 1, as in Equation 4, usually holding their coefficients fixed (i.e., $\beta_{ij} = \gamma_{10}, \ldots, \beta_{kj} = \gamma_{00}$; or equivalently, $\tau_{11} = \ldots = \tau_{kk} = 0$). This yields a new estimate of the level-2 variance component, $\tau^*_{00}$, with variation due to the individual-level covariates removed. The investigator may again conduct a statistical test of the null hypothesis that $\tau^*_{00} = 0$ and may compare this new estimate directly with the estimate of $\tau_{00}$ from the unconditional model. If $\tau^*_{00}$ remains nonzero, the investigator may conclude that some proportion of the between-neighborhood variation is due to characteristics of the neighborhoods rather than to their composition. Indeed, the ratio of $\tau^*_{00}$ to $\tau_{00}$ is sometimes interpreted as the proportion of between-neighborhood variation that is caused by neighborhood-level factors.

Although this data analytic strategy has intuitive appeal, the inferential problems involved are actually rather thorny. One problem is that etiologically important individual-level covariates may be omitted from the level-1 model, and those that are included may be measured with error. As a result, the model may be vulnerable to criticism that compositional variation has not been completely removed, in which case $\tau^*_{00}$ might still be considered an overestimate of variation between neighborhoods that is not accounted for by the characteristics of residents. A
The second problem is that the omission of level-2 covariates from this model may cause the contribution of level-1 covariates to be overstated. This will be the case whenever there exists some neighborhood-level covariate, $W_j$, that is (a) omitted from the model, (b) correlated with the dependent variable, and (c) correlated with one or more of the individual-level compositional variables in the level-1 model. In this situation, $\tau^*_00$ underestimates the variation between neighborhoods because some of the variance related to $W_j$ has been inadvertently removed.

Thus, investigators wishing to separate context from composition using this approach are faced with a vexing specification problem. On the one hand, omitting important individual-level covariates can cause $\tau^*_00$ to be overestimated; on the other hand, including individual-level covariates that correlate with etiologically significant neighborhood-level covariates can cause $\tau^*_00$ to be underestimated. Moreover, in most applications it is likely that both of these problems will apply, making $\tau^*_00$ virtually uninterpretable within the context versus composition framework. The problem, in substantive terms, is that context and composition are deeply confounded. Selection processes operate to place individuals with certain characteristics into certain types of neighborhoods. Within neighborhoods, these and other individual characteristics combine interactively to shape individual and collective outcomes. Meanwhile, the characteristics of neighborhoods may affect health outcomes in part by modifying individual-level covariates. The complexity of these multilevel selection, interaction, and mediation processes makes it difficult to separate context from composition using observational data.

One strategy that provides a partial remedy to these difficulties is a two-stage analysis (120). In the first stage, one fits a model similar to that given in Equations 4a through 4c, but holding the slopes fixed (i.e., $u_{ij} = 0$) and group-mean centering the individual-level covariates:

\[
\begin{align*}
Y_{ij} &= \beta_0 + \beta_{1j}(X_{ij} - \bar{X}_j) + r_{ij} \quad 8a. \\
\beta_{0j} &= \gamma_{00} + u_{0j} \quad 8b. \\
\beta_{ij} &= \gamma_{10} \quad 8c.
\end{align*}
\]

with the same variance-covariance structure given in Equation 1c. Here $\bar{X}_j$ is the neighborhood-specific mean of the covariate $X_{ij}$. The group-mean centered covariate has the useful property of being independent of all neighborhood-level covariates, $W_j$, and the estimate $\hat{\gamma}_{10}$ therefore represents the average within-neighborhood effect of $X_{ij}$. The next stage is to generate an adjusted dependent variable, $Y^*_{ij} = Y_{ij} - \hat{\gamma}_{00} - \hat{\gamma}_{10}X_{ij}$. Finally, one estimates the model given in Equations 1a through 1c, with $Y^*_{ij}$ as the dependent variable. This model produces a new $\tau^*_00$ that is adjusted for all of the individual-level covariates included in the first stage, but still contains all of the variation attributable to neighborhood-level covariates. Because some important $X_{ij}$ may still be omitted, and those that are included may be measured with error, this new $\tau^*_00$ should be considered an upper bound on the true variability between neighborhoods that cannot be attributed to composition.
Effects of Cluster-Level Variables

An alternative approach for studying the multilevel etiology of health focuses on the effects of specific cluster-level covariates. This research strategy seeks to answer questions of the type, Does this cluster-level variable affect people’s health? Investigators using this approach generally fit models of the form given in Equations 6a through 6c to cross-sectional data; their primary interest is in the value of the macroparameter, \( \gamma_{01} \), associated with the cluster-level covariate, \( W_j \) (10, 23–25, 34, 38, 40, 49, 67, 70, 71, 75, 95, 102, 103, 121, 125, 131, 132, 137, 143, 156). The use of multilevel models in this context produces efficient estimates and realistic standard errors for these parameters, and thereby provides a sound basis for hypothesis testing and confidence interval construction. Yet several uncertainties arise regarding the proper specification of the model, and these uncertainties may undermine the interpretability of \( \gamma_{01} \) as a scientifically meaningful quantity.

One problem involves the treatment of individual-level covariates in the level-1 model. In studies of the relationship between area socioeconomic characteristics and individual health, for instance, some investigators include an extensive list of individual-level variables in their models (e.g., 10, 23, 49, 67, 70, 71, 95, 121, 125, 131, 132, 156). Others, in contrast, include only a small number of individual-level demographic covariates such as age, sex, and a crude measure of socioeconomic position (e.g., 24, 25, 75, 103, 143). Whether their approach to level-1 covariates is inclusive, restrictive, or somewhere in between these two extremes, investigators seldom provide an explicit justification for their overall model specification approach or for their decisions regarding the inclusion or omission of specific individual-level covariates.

The striking absence of critical discussions of level-1 model specification in this area of research may reflect investigators’ reluctance to confront the nettlesome substantive issues that are involved. The two major considerations are mediation and confounding. In this context, an individual-level variable may be regarded as a potential confounder if it contributes to the processes that sort individuals into neighborhoods and affects the local health outcome independently of neighborhood characteristics. For example, individual poverty could limit one’s access to quality medical care and also limit one’s residential options to neighborhoods characterized by concentrated poverty; it might therefore create a spurious association between neighborhood poverty and individual health. Investigators concerned about confounding typically include potential confounders in the level-1 model in order to reduce their biasing impact on the estimate of \( \gamma_{01} \), leading to an inclusive approach to level-1 model specification.

Alternatively, individual-level covariates could act as mediators, accounting mechanistically for the association between a neighborhood-level variable and an individual-level health outcome. For example, living in a neighborhood with a high concentration of bars and liquor stores could lead to high levels of individual alcohol consumption, with the resulting impacts on cardiovascular and liver function. Including an individual-level measure of alcohol consumption as a covariate in a
model linking neighborhood-level liquor store concentration to individual health yields an interpretation of $\gamma_{01}$ as the direct (i.e., unmediated) effect of liquor store concentration (4). Yet the investigator may desire an estimate of the total, not just the direct, effect. In such cases, inclusion of the potential mediator in the level-1 model could be characterized as overadjustment. This line of reasoning leads to the adoption of a restrictive approach to level-1 model specification.

The appropriate specification of the level-1 model becomes problematic in the absence of a strong basis for classifying individual-level covariates as either confounders or mediators. For many individual-level covariates, existing theoretical and empirical considerations simply do not permit an unambiguous classification, especially when the available data are cross-sectional, as is common in studies of this type. Epidemiologists may therefore disagree on whether or not an individual-level covariate should be included in the level-1 model. Investigators may then consider specifying two versions of the model, one including and the other omitting the questionable covariate. Convergent results would reduce uncertainty, while divergent results would increase it.

In many cases, however, a single individual-level variable may play both roles, confounder and mediator, simultaneously. The possible confounding role of individual poverty was considered above, but poverty may also mediate the association between socioeconomic characteristics of neighborhoods and individual health. For instance, residing in a neighborhood with high levels of unemployment may deprive one of contact with social networks through which access to employment opportunities may be obtained; this could lead to ongoing individual poverty, which in turn could have consequences for health. In situations where a single individual-level covariate plays both confounding and mediating roles, neither level-1 model specification (including or omitting the covariate in question) is entirely satisfactory. Including the covariate leads to an overadjusted estimate of $\gamma_{01}$, whereas omitting the covariate leads to an estimate that is biased by confounding. In such cases, there may be no level-1 model specification that yields a scientifically interpretable estimate of $\gamma_{01}$.

Another source of uncertainty in the interpretation of $\gamma_{01}$ involves the selection of neighborhood-level variables for inclusion in the model. Some investigators (e.g., 24, 25, 121, 125) include a single neighborhood-level variable in their models, usually derived from census data aggregated to the tract or zip code level. Examples include median income, median house value, percent living below the poverty line, and proportion of female-headed households. Aggregated census variables such as these, however, tend to be highly correlated with one another, making their unique effects on health outcomes difficult to distinguish (50). For example, if $Y_{ij}$ is a measure of respiratory function and $W_j$ is the proportion of neighborhood residents living in overcrowded housing, then $\gamma_{01}$ might be interpreted as the effect of widespread overcrowding in a neighborhood on respiratory health. But because neighborhoods characterized by overcrowding may also have lower median incomes, higher unemployment rates, and higher proportions of people working in manual labor, any effects that these variables have on respiratory
health could make $\gamma_{01}$ a biased measure of the effect of overcrowding. On the other hand, including these potential confounders in the model could greatly increase the variance of the estimate of $\gamma_{01}$, making even strong effects difficult to detect.

For this reason many investigators have turned to combining aggregate census variables into indices. Such indices are derived by standardizing and summing these census variables, or by means of principle components or factor analysis, and are often given names like deprivation (34, 38, 70, 71, 75, 143), social environment (23, 156), or concentrated disadvantage (8, 132). This approach has the advantage of avoiding, to some degree, the erroneous ascription to one variable (e.g., overcrowding) of effects that are actually attributable to other variables (e.g., neighborhood poverty and unemployment). The downside of this solution is the vague meaning of $\gamma_{01}$. If $W_j$ were median income, then we could interpret $\gamma_{01}$ as the expected change in $Y_{ij}$ that would result if we raised the median income in a neighborhood by one unit. When $W_j$ is a summary index comprised of many census variables, however, the practical interpretation of $\gamma_{01}$ is no longer obvious. The alternatives, then, could be characterized as dishonest specificity and honest ambiguity. Investigators must rely upon their judgment to determine which, in a particular research setting, is the lesser evil.

Not all research on the multilevel etiology of health relies exclusively on census data for the characterization of neighborhoods. Several investigators, instead, have undertaken the direct quantification of the features of neighborhood physical and social environments that are thought to play important etiological roles in health. Examples of such measures include the presence or absence of supermarkets, fast-food restaurants, and other commercial food establishments in a neighborhood (103), neighborhood social support (8), collective efficacy (135), and physical and social disorder (119). To the extent that census variables are used as proxies for these environmental variables, direct measurement is clearly to be preferred. Direct measures can be costly and time-consuming to obtain, however, and share with census variables the problem of possible confounding by other measured and unmeasured individual- and neighborhood-level covariates.

**Effects of Individual-Level Variables in a Multilevel Setting**

Although seldom employed in public health [but see (8)], a third approach to studying the multilevel etiology of health is to place well-established group differences into a multilevel context. Consider, for example, the average difference between African American and White adults in systolic blood pressure (151). One potential explanation for this difference is that, in the context of striking residential segregation (94), African American and White neighborhoods may differ systematically in terms of factors that influence blood pressure. An obvious strategy for handling this possibility would be to compare the blood pressures of African Americans and Whites who reside in the same neighborhood. This is precisely what economists attempt to do by adding a fixed effect (an indicator variable) for every neighborhood studied. These fixed effects should capture all of the effects
on blood pressure of heterogeneity between neighborhoods. Any remaining difference between African Americans and Whites cannot, therefore, be explained by their residential environments. An analogous but more parsimonious approach is possible within a multilevel modeling framework. The model to be specified is identical to that given in Equations 8a through 8c, with $X_{ij}$ being an indicator variable for race/ethnicity and $\bar{X}_j$ being its mean value within the $j$th neighborhood. Following this approach, the macroparameter $\gamma_{10}$ is interpreted as average within-neighborhood difference in blood pressure between African Americans and Whites. It may be substantially different from the overall race/ethnic disparity in blood pressure.

Several extensions of the group-mean centered model are possible. The first is to include a random effect, $u_{1j}$, in Equation 8c. A nonzero variance, $\tau_{11}$, of this random effect would imply that the disparity in blood pressure varies across neighborhoods. One might then attempt to model variations in the disparity by introducing measured neighborhood-level variables ($W_j$) into Equation 8c.

An important point about this approach is that the estimated effects of specific neighborhood-level covariates on health are not the focus. In the fixed effects approach discussed previously, the research questions involved the value of the macroparameter $\gamma_{01}$ associated with the cluster-level covariate $W_j$, and the interpretation of this parameter was problematized by uncertainty about the appropriate specification of the level-1 and level-2 models. The group-mean centering approach described in this section is free of such problems. Because no cluster-level variables are included in the model, we need not worry that the estimate of their effects could be biased owing to confounding or overadjustment. Instead, by situating a well-known individual-level disparity within a multilevel framework, we remove the entire contribution of this disparity to cluster-level heterogeneity. The group-mean centering approach is not appropriate, however, for studying the effects of neighborhood-level covariates when adjustment for individual-level covariates is desired. In this context, grand-mean centering of the individual-level covariates leads to a more meaningful interpretation of $\gamma_{01}$ [see (118), p. 142].

**QUESTIONS ABOUT THE VALUES OF MICROPARAMETERS**

Our discussion thus far has focused on applications of multilevel modeling in which the parameters of interest are either the variance components ($\sigma^2$, $\tau_{00}$, $\tau_{11}$, and $\tau_{01} = \tau_{10}$) or the macroparameters ($\gamma_{00}$, $\gamma_{01}$, $\gamma_{10}$, and $\gamma_{11}$). Sometimes, however, investigators are most interested in the values of the microparameters ($\beta_0$ and $\beta_1$) for each of the $j$ clusters studied. In these applications, the theory of empirical Bayes estimation comes into play. Consider the multilevel model given in Equations 1a through 1c. For the $j$th cluster, we have two possible estimates of $\beta_{0j}$: the cluster-specific sample mean, $\bar{Y}_j$, and the grand-mean, $\bar{Y}_{00}$. It turns out that an optimal estimator is actually a weighted average of these two:

$$\beta_{0j}^* = \lambda_j \bar{Y}_j + (1 - \lambda_j) \bar{Y}_{00}. \tag{9}$$
where $\lambda_j$ is called the reliability and represents the ratio of true score to total score variance in the cluster-specific sample mean. The larger the size of the cluster-specific sample is, the more reliable this mean will be and the more heavily weighted $\beta^*_0j$ will be toward $\bar{Y}_j$. When the cluster-specific sample size is small, however, $\beta^*_0j$ “borrows” information for the other clusters in order to compensate for the statistical fluctuations associated with small samples. Because $\lambda_j$ must be estimated from the data, $\beta^*_0j$ is called an empirical Bayes estimator.

Empirical Bayes estimators have diverse applications in public health. Perhaps the most common use has been the estimation of rates of unusual health outcomes in small populations. In this setting, the observed rate, $\bar{Y}_j$, may not be a good indicator of the true underlying rate, but applying the population average rate to all clusters may obscure important clues to disease etiology. Empirical Bayes estimators, or adaptations thereof, have been used to characterize toxoplasmosis rates in cities in El Salvador (39, 104); to obtain age- and sex-specific stomach, bladder, and lung cancer rates in Missouri counties (147, 148); to derive age-standardized lip cancer rates for areas in Scotland (11); to map county-level, sex-specific, age-standardized cancer and fire- and burn-related mortality rates for the United States (17, 18, 91); and to map breast cancer and Hodgkin’s lymphoma rates in health districts in Sardinia (3). Unfortunately, virtually all of these applications have appeared in the statistical and biostatistical literature, and the use of empirical Bayes methods for estimating rates of health outcomes has not been widespread among other public health investigations. Further applications of empirical Bayes methods, such as the simultaneous estimation of effects of multiple exposures on health (57, 154) and modeling the progression of HIV infection (16, 80), have likewise been confined to the statistical literature.

One area in which the use of empirical Bayes methods has moved beyond the technical literature is the relative performance of different health care providers. The publication in England, Scotland, and the United States of such performance indicators as physician- or hospital-level mortality rates for a given surgical procedure has generated intense debate about the meaningfulness and utility of such indicators (68, 99, 100). Two issues often arise in this debate, both of which have relevance for the use of empirical Bayes methods: reliability and risk adjustment. The issue of reliability is founded on the idea that the observed average level of some outcome for a given health care provider reflects both systematic and chance components. The systematic components are the desired basis for comparison, but the chance components may be so large that little can be inferred from observed data about the systematic components. Investigations have shown, for example, that the physician-specific average level of glycemic control in diabetics is not stable enough to support meaningful comparisons of physician performance (65).

The second problem, risk adjustment, relates to the possibility that health care providers may have very different patient mixes. Whereas one surgeon handles routine procedures that involve little risk of complication, another may consistently be sought out by patients for the most complicated procedures. In such a case, the latter physician may do better work yet the former may achieve a lower
postoperative mortality rate. The need for some kind of risk adjustment in comparing providers’ performance is therefore widely recognized, and the methods for accomplishing this adjustment have received considerable attention (5, 6, 134).

Empirical Bayes procedures may be useful for addressing both of these problems. As discussed above, empirical Bayes estimators have often been used to achieve greater stability in situations in which standard estimators may be unreliable. Moreover, some degree of risk adjustment may be obtained by including patient-level measures of risk or severity in a multilevel model. Several investigators, therefore, have used multilevel models of the type given in Equations 4a through 4c, with patients at level 1 and providers at level 2, to obtain stabilized, risk-adjusted indices of provider performance (15, 72, 81–83). In these applications, $X_{ij}$ represents some measure of risk for the $i$th patient being treated by the $j$th health care provider. The effects of these measures of risk are usually treated as fixed (i.e., $\tau_{11} = \ldots = \tau_{kk} = 0$) and $\beta^*_{0j}$ is interpreted as the risk-adjusted performance indicator for the $j$th provider. The desirability of including variables measuring provider characteristics in the level-2 model have been discussed in the literature on evaluation of educational institutions (120) and may be applicable to health care evaluation.

Three important difficulties characterize this approach. The first concerns the specification of the level-1 model. The goal should be to include a set of measures of patient variables that will completely remove differences in risk that are beyond the control of the provider, without removing differences that may be attributable to the quality of the care received from that provider. This ideal may be difficult to achieve. Indeed, much of the literature on risk adjustment in performance evaluation focuses on the proper specification of the risk-adjustment model (e.g., 6, 99, 100). A second difficulty is that, even with the increased reliability of empirical Bayes estimators, the resulting indicators of provider performance may still be too unstable to permit a useful comparison of individual providers (53, 65, 145).

A third difficulty that arises in using empirical Bayes estimators as performance indicators relates to bias. In most settings, statistical theory guarantees that $\bar{Y}_j$ is an unbiased estimator of the mean in the $j$th cluster. Yet $\beta^*_{0j}$ is pulled away from this unbiased estimator by the term $(1 - \lambda_j)\hat{\gamma}_{00}$ in Equation 8. This bias can be very small when $\lambda_j$ is large (in fact, it approaches zero as the sample size within the $j$th cluster increases to infinity), but it may be substantial in practical settings. Moreover, the degree of bias in $\beta^*_{0j}$ will vary across clusters when the cluster-specific sample sizes differ. In the context of health care provider ratings, if all providers had an identical number of patients whose outcomes could be observed, the bias affecting the empirical Bayes estimator would be inconsequential. It would shift all provider-specific outcome estimates toward a common mean, $\hat{\gamma}_{00}$, but would not alter the rank order of providers. With unequal sample sizes, however, the degree of bias could vary substantially, leading to a reordering of ranks as small, high-performing providers are unfairly “punished” and small, poor-performing providers are inordinately “rewarded.” In any possible application of empirical Bayes estimators, in fact, the investigator must consider carefully the implied trade-off between reliability and bias.
CONCLUSION

Public health investigators have found a variety of uses for multilevel statistical models. The diversity of the applications discussed above is a testimony to the flexibility of these models for answering a wide range of seemingly disparate research questions. Nevertheless, the full potential of multilevel models has yet to be realized in public health. The inferences that public health investigators wish to make, in fact, could be strengthened by more carefully framing research questions, and by more fully exploiting the capacity of multilevel models.

A major difficulty in public health, as in many other fields, is the threat to the internal validity of inferences posed by confounding in observational studies (73). These difficulties do not characterize experiments, since the randomization of units to treatments guarantees that, on average, there will be no confounding. Yet in many instances, ethical or logistical considerations render experimental randomization impossible. Statistical thinking about these inferential problems draws upon a counterfactual framework sometimes known as the Rubin causal model (86, 129, 133). From this framework, several methods for obtaining measures of treatment effects from observational data have been developed, including propensity score matching, the use of instrumental variables, and selection modeling (2, 130, 152, 153).

Public health applications of multilevel models, particularly those using non-experimental data, could be substantially improved by integrating counterfactual thinking into the framing of research questions. This can often be accomplished by considering what randomized experiment a given nonexperimental design is intended to approximate. In multilevel studies of the etiology of disease, for example, it is usually unclear which of at least two possible experiments the research aims to emulate. The first experiment is one in which neighborhoods or other clusters are assigned at random to two or more treatments, such as median income or levels of some socioeconomic index. This is precisely analogous to a cluster-randomized trial. If it could be done, this type of experiment would be useful for answering questions about what would happen to the residents of a given neighborhood if the characteristics of that neighborhood could be changed. Yet because experiments in which the socioeconomic characteristics of entire neighborhoods are systematically altered (without altering the socioeconomic characteristics of individual residents) may never be technically feasible, the question becomes, How can the standard observational study of this type be made to resemble as closely as possible a cluster randomized trial?

A second possible experiment involves the random assignment of individuals or families to different neighborhoods. Such a design was employed in the Moving to Opportunity study, in which families living in publicly assisted housing in high-poverty neighborhoods were assigned to one of three experimental conditions: (a) housing assistance and mobility counseling, with a requirement to move to a low-poverty neighborhood, (b) Section 8 housing vouchers, with the ability to move anywhere, and (c) no housing assistance. Preliminary results of this study (128) suggest that families in the first condition experienced significant
improvement, relative to the other two groups, in a variety of variables related to well-being. Clearly this experiment would answer quite different questions from those addressed by the hypothetical cluster-randomized trial described above. Yet questions about the effects of moving to particular types of neighborhoods may have great relevance to public health policy. How can observational studies be conducted to resemble as closely as possible the experimental design exemplified by the Moving to Opportunity study?

It seems clear that if the usefulness of observational data is to be maximized, those data will often need to be longitudinal. Fortunately, multilevel models are very well suited to longitudinal data. Indeed, some of the earliest proposed multilevel models (150, 155) were intended for fitting a set of individual linear growth curves to longitudinal blood pressure data. Since then, growth curves estimated via multilevel models have only rarely appeared in the public health literature but have been common in educational and psychological research (117, 127). Multilevel modeling approaches to the study of individual change have many advantages, including (a) individuals need not be observed at the same times or on the same number of occasions, (b) time-varying covariates can be incorporated into the model, and (c) individual change parameters can be situated in larger contexts such as neighborhoods and workplaces. Growth curve models represent the most underexploited use of multilevel models in public health. Combining these models with longitudinal data and a careful posing of research questions within the counterfactual framework may go a long way toward advancing scientific knowledge about the public’s health and how best to improve it.

ACKNOWLEDGMENTS

We gratefully acknowledge the John D. and Catherine T. MacArthur Foundation, the National Institute of Justice, and the National Institute of Mental Health for their support of the Project on Human Development in Chicago Neighborhoods, from which many of the ideas presented here emerged.

The Annual Review of Public Health is online at http://publhealth.annualreviews.org

LITERATURE CITED

49. Gee GC. 2002. A multilevel analysis of the relationship between institutional and individual racial discrimination and
of chronic illness within a multilevel modelling framework. *Health & Place* 1:27–40


MULTILEVEL INFERENCE


146. Thorndike EL. 1939. On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. *Am. J. Psychol.* 52:122–24


