

To GEE or Not to GEE

Comparing Population Average and Mixed Models for Estimating the Associations Between Neighborhood Risk Factors and Health

Alan E. Hubbard,^a Jennifer Ahern,^b Nancy L. Fleischer,^b Mark Van der Laan,^a Sheri A. Lippman,^b Nicholas Jewell,^a Tim Bruckner,^c and William A. Satariano^{b,d}

Abstract: Two modeling approaches are commonly used to estimate the associations between neighborhood characteristics and individual-level health outcomes in multilevel studies (subjects within neighborhoods). Random effects models (or mixed models) use maximum likelihood estimation. Population average models typically use a generalized estimating equation (GEE) approach. These methods are used in place of basic regression approaches because the health of residents in the same neighborhood may be correlated, thus violating independence assumptions made by traditional regression procedures. This violation is particularly relevant to estimates of the variability of estimates. Though the literature appears to favor the mixed-model approach, little theoretical guidance has been offered to justify this choice. In this paper, we review the assumptions behind the estimates and inference provided by these 2 approaches. We propose a perspective that treats regression models for what they are in most circumstances: reasonable approximations of some true underlying relationship. We argue in general that mixed models involve unverifiable assumptions on the data-generating distribution, which lead to potentially misleading estimates and biased inference. We conclude that the estimation-equation approach of population average models provides a more useful approximation of the truth.

(*Epidemiology* 2010;21: 467–474)

A growing body of research has examined neighborhood-level characteristics in association with patterns of health, functioning, and survival in populations.¹ Many of these analyses employ a multilevel approach, examining neighborhood-level exposures in association with individual-

level outcomes, while adjusting for individual- and neighborhood-level confounders. One objective of these studies has been to determine the associations of community characteristics (eg, crime statistics and environmental exposures) with health outcomes after adjusting for individual characteristics of residents.²

Two modeling approaches are commonly used to estimate the associations between neighborhood characteristics and health outcomes in multilevel studies. One is the random effects or mixed model, which uses maximum likelihood estimation,³ and the other is the population average model, which typically uses a generalized estimating equations (GEE).⁴ These methods are used in place of basic regression because the health status of residents in the same neighborhood may be correlated, thus violating the independence assumptions of traditional regression models. In the neighborhood-effects literature to date, the mixed model has been favored, perhaps because it involves explicit modeling and partitioning of the covariance structure of the outcomes within and between neighborhoods. Partitioning the variance allows the calculation of the proportion of variance in the outcome due to neighborhood-to-neighborhood variation against that due to the variance among individuals within a neighborhood, as well as changes in these variance components after adjustment for exposures and confounders at both the neighborhood and individual-levels. The ability to partition variance within and between neighborhoods is an alluring feature of mixed models, but it is only one of a larger set of issues that should be considered when selecting an analytic approach. Our paper aims to contrast regression methods for studying associations between neighborhood-level exposures and individual-level outcomes; other issues regarding causal inference in neighborhood studies have been addressed elsewhere.⁵

There are many overviews of mixed models and population average models.^{6–9} The purpose here is to review the assumptions of each approach as relevant to studies of neighborhood effects. Our overarching perspective that it is generally most realistic to think of regression models as approximations of the truth, with results from mixed models possibly biased given the reliance on untestable assumptions

Submitted 22 May 2008; accepted 7 July 2009; posted 9 March 2010.

From the Divisions of ^aBiostatistics and ^bEpidemiology, Berkeley School of Public Health, University of California, Berkeley, CA; ^cIrvine Program in Public Health, College of Health Sciences, University of California, Irvine, CA; and ^dDivision of Community Health & Human Development, Berkeley School of Public Health, University of California, Berkeley, CA.

Editors' note: A commentary on this article appears on page 475.

Correspondence: Alan E. Hubbard, Division of Biostatistics, University of California, Berkeley School of Public Health, 140 Warren Hall 7360, Berkeley, CA 94720. E-mail: hubbard@stat.berkeley.edu.

Copyright © 2010 by Lippincott Williams & Wilkins

ISSN: 1044-3983/10/2104-0467

DOI: 10.1097/EDE.0b013e3181caeb90

on the data-generating distribution. It is typically most realistic to think of a high-dimensional (eg, many covariates) regression model as an approximation of some true underlying and unknowable model. A mixed model requires correct specification of the regression models for the so-called fixed effects coefficients as well as distributional assumptions and regression models for the random effects. If the model is misspecified (ie, an incorrect model form is used to estimate the mean of outcome Y_{ij} given covariates X_{ij} for person i within neighborhood j) one can still define the parameter of interest as the regression model that would be estimated if the entire target population was used as the sample. However, it is unclear how one interprets coefficients in a misspecified mixed-effects model. As we discuss below, there is no generally interpretable parameter that can be defined as the projection of a misspecified mixed model onto the true underlying model, whereas with the population average model, there is hope to explicitly define this projection (approximation). In fact, even when the fixed effects part of a model is correctly specified, depending on how the true random-effects model relates to the estimating model, the estimated fixed effects can be very misleading. Since the latent, random-effects distribution is nonidentifiable, larger sample sizes do not help.

To illustrate the various concepts in this paper, we will refer to a theoretical study of neighborhood crime rates and their influence on the probability that neighborhood residents walk more than 2 hours/week. We begin by reviewing the general characteristics of the mixed model and GEE/population average model approaches. Next, we contrast the interpretation of regression coefficients of mixed versus population average models, particularly in the context of neighborhood studies. We then discuss the implications of considering the regression model (the fixed effects part of the model) as an informative approximation of the true model based on a misspecified class of regression models. The paper ends with a discussion comparing the 2 approaches.

MIXED MODELS

A mixed-model approach is predicated on the idea that heterogeneity exists across neighborhoods for some of the regression coefficients, and that the heterogeneity can be represented by a probability distribution. This approach provides a specific model for the conditional distribution of the outcome given covariates and random effects, and the distribution of random effects given covariates, implying a fully specified model for the distribution of outcome, given covariates. Let Y_{ij} be the outcome for subject i within neighborhood j , $\mu(X_{ij} | \beta)$ the average “response” of a person with the same covariates X_{ij} , β a set of fixed coefficients, and $U_{ij}(\alpha_j, X_{ij})$ an error term that is a function of “neighborhood” random effects, α_j , and perhaps is also a function of the covariates. The mean of the i^{th} person in the j^{th} neighborhood in this context can be written as:

$$E(Y_{ij} | X_{ij}, \alpha_j) = g[\mu(X_{ij} | \beta) + U_{ij}(\alpha_j, X_{ij})] \quad (1)$$

g is the link function depends on the regression (eg, linear: $g^{-1}(u) = u$,

log: $g^{-1}(u) = \log(u)$, logistic: $g^{-1}(u) = \log[u/(1-u)]$) and $E\{U_{ij}(\alpha_j, X_{ij}) | X_{ij}\} = 0$. To make this notation more concrete, consider $Y_{ij} = 1$ if subject i in neighborhood j walks more than 2 hours per week, 0 otherwise and X_{ij} is a continuous measure of crime rate for neighborhood j (would be the same for each i within j). Here is a simple random-effects model that relates Y_{ij} and X_{ij} , allowing for both neighborhood-to-neighborhood variability in the underlying probability of walking (at $X_{ij} = 0$, represented by α_{0j}) as well as variability in the slope of the logit of the probability versus crime rate (represented by α_{1j}):

$$\log \left[\frac{P(Y_{ij} = 1 | X_{ij}, \alpha_j)}{1 - P(Y_{ij} = 1 | X_{ij}, \alpha_j)} \right] = \beta_0 + \alpha_{0j} + (\beta_1 + \alpha_{1j})X_{ij}, \alpha_j \sim \text{MVN}(0, \Sigma). \quad (2)$$

In this case, the logit link is used because Y_{ij} is binary, and thus

$E(Y_{ij} | X_{ij}, \alpha_j) = P(Y_{ij} = 1 | X_{ij}, \alpha_j)$, $\mu(X_{ij} | \beta) = \beta_0 + \beta_1 X_{ij}$, where $\beta = (\beta_0, \beta_1)$, $U_{ij}(\alpha_j, X_{ij}) = \alpha_{0j} + \alpha_{1j} X_{ij}$ and $\alpha_j = (\alpha_{0j}, \alpha_{1j})$. Thus, the model for the mean walking variable of an individual within neighborhood j is a function of measured variables (X_{ij}), unknown parameters (β), and unmeasured (latent) variables (α_j). The estimates of the so-called fixed effects (β) and parameters of the distribution of the error terms are then derived via maximum likelihood (or restricted maximum likelihood). Due to the conditional error distribution, the mean model implies a model for Y_{ij} , given X_{ij} and α_j . In addition, if one also asserts a model for the distribution of α_j , given X_{ij} (as we have in the example above) then one can derive the likelihood of the observed data, specifically of Y_{ij} given X_{ij} . This likelihood is derived by integrating the likelihood for theoretical data (as if one observes $U_{ij}(\alpha_j, X_{ij})$) over the proposed distributions of the residual error and α_j . The inference (ie, standard error calculations) for the estimates of the coefficients are typically derived via standard maximum likelihood inference (or restricted maximum likelihood), the accuracy of which relies on both the underlying mean and error distribution models being correctly specified.

Given the class of mixed models usually considered (Equation 1), the regression coefficients in the linear and log-linear mixed models can be interpreted as either coefficients of the conditional mean of Y_{ij} and X_{ij} (conditional on α_j) or as coefficients of the population average association of Y_{ij} and X_{ij} : $E(Y_{ij} | X_{ij})$. We discuss in more detail later the specific instance of logistic regression models where, unlike

the linear and log-linear case, the coefficients are not equivalent in the random-effects and population-average models. Because of the overlap of interpretation of the coefficients in the linear/log-linear case, one can obtain “robust” inference for the coefficients, (relying only on correct model for $E(Y_{ij} | X_{ij})$), either by appropriately bootstrapping¹⁰ or using a sandwich-type estimator.¹¹ However, the typical inference provided by mixed models algorithms is based on the assumption that the model for the underlying random effects distribution is correctly specified.

The likelihood of the observed data with respect to the distribution of both the observed and unobserved latent variables (density of Y_{ij} , given X_{ij}) in a mixed model is as follows:

$$L(Y_{ij} | X_{ij}) = \int_{\alpha_j} f(Y_{ij} | X_{ij}, \alpha_j) h(\alpha_j | X_{ij}) d\alpha_j, \quad (3)$$

which can be interpreted as the average probability density function of Y_{ij} given (X_{ij}, α_j) averaged over the probability density of α_j , $h(\alpha_j | X_{ij})$. For instance, in our example above (Equation 2):

$$L(Y_{ij} | X_{ij}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P(Y_{ij} | X_{ij}, \alpha_{0j}, \alpha_{1j}) h(\alpha_{0j}, \alpha_{1j}) d\alpha_{0j} d\alpha_{1j},$$

where h is the multivariate normal probability density function with mean vector $(0, 0)$ and covariance matrix, Σ . The mixed model in which data for individuals i , within the same neighborhood j , are generated from common random variables α_j , will imply correlation of the observations within the same neighborhood. This often motivates the use of such models. However, an infinite variety of combinations of densities f and h could provide the same marginal distribution of Y_{ij} , given X_{ij} . This means that the random-effects model for the density is nonparametrically nonidentifiable, because only the distribution of the observed data (the Y_{ij}, X_{ij}) can provide information about the fit of competing models. There are an infinite number of combinations of $f(Y_{ij} | X_{ij}, \alpha_j)$ and $h(\alpha_j | X_{ij})$, which result in the same $L(Y_{ij} | X_{ij})$. When these models are used, the hope is that the misspecification of the joint distribution of the error terms and random effects does not make the estimates and inference provided by this procedure unduly misleading.

POPULATION AVERAGE MODELS

The coefficient estimates returned by the generalized estimating equations (GEE) typically used to estimate population average models (sometimes called marginal models) describe changes in the population mean given changes in covariates, while accounting for within-neighborhood non-independence of observations when deriving the variability

estimates of these coefficients. One can also use maximum-likelihood-based methods for estimation of this parameter (for instance, see Heagerty and Zeger⁸), so we do not want to confuse the estimation method (GEE) with the parameter (population-average models). The GEE approach does not require distributional assumptions because estimation of the population-average model depends only on correctly specifying a few aspects of the observed data-generating distribution (ie, the mean of the outcome given the covariates), not on the entire joint distribution of observed data and random effects. The GEE approach requires only (1) proposing a parameter of interest (in our case, the coefficients in the model of $E[Y_{ij} | X_{ij}]$ or the probability of walking among individuals that live in neighborhoods with crime rate X_{ij}) and (2) finding an estimating function that has mean 0 if the true parameters are entered into that estimating function. We provide an associated technical report¹⁵ with a general form of the estimating function on which the estimating equations are based.

For linear models, the estimating-equation approach sometimes provides practically the same estimator of the parameters (for instance, a type of weighted least squares) to a specific mixed model (eg, the estimator that is based on maximum likelihood and a mixed model). For instance, a simple random intercept linear model implies equal variances for all observations and equal covariances of all possible paired observations within the statistical unit (neighborhood) and as always no correlation of observations made on different units. This will yield the same estimates as the exchangeable working correlation model in GEE. Thus, in specific circumstances, the estimates provided by a GEE approach and a mixed-model approach (where both result in the same estimated variance-covariance model of observations on the same unit) are equal. However, the 2 approaches depart in this case when deriving the inference for what might be equivalent parameter estimates of β . With the estimating-equation approach, no likelihood has been specified, so maximum likelihood inference is not available for these estimators. Instead, robust or sandwich inference is typically provided.⁴ With the more technical detail available in the associated technical report,¹⁵ one can show that these estimators are asymptotically linear (ie, they can be written asymptotically as a sum of independent and identically distributed random variables, called the influence curve). If there is a closed-form representation of the influence curve (which will be the same dimension as the number of coefficients, β) one can derive robust inference of $\hat{\beta}$ and the resulting standard errors by estimating the sample variance-covariance of these random variables (influence curve components), making no assumptions about the underlying distribution of the data. To gain some intuition, we note that an equivalently valid method of inference in this case would be the nonparametric bootstrap, where one (1) randomly resamples neigh-

neighborhoods with replacement to create a new pseudo-population of the same size (with regard to number of independent units) as the original data, (2) performs the same estimation procedure as conducted on the original data, and (3) repeats this procedure many times. For each run of the bootstrap, b , let $\hat{\beta}^b$, $b = 1, \dots, B$ be the estimated vector of coefficients, $\hat{\beta}^b = (\hat{\beta}_0^b, \hat{\beta}_1^b, \dots, \hat{\beta}_p^b)$. Now, the variance-covariance of the estimated coefficients among these bootstrap samples is provided by simple sample variance-covariance estimates of

$$\hat{\beta}^b, \text{ eg, } \text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \frac{1}{B} \sum_{b=1}^B (\hat{\beta}_i^b - \hat{E}\hat{\beta}_i)(\hat{\beta}_j^b - \hat{E}\hat{\beta}_j), \text{ where}$$

$$\hat{E}\hat{\beta}_i = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_i^b.$$

The one caveat to an otherwise straightforward approach is that the number of neighborhoods has to be sufficiently large; the inference is asymptotically correct but not necessarily accurate in smaller samples. Thus, one should be skeptical of the robust standard error estimates based on the influence curve in GEE in cases in which the number of neighborhoods is relatively small. In the case of a few neighborhoods with large numbers of observations, it might be preferable to base inference on certain assumptions (say exchangeability) and thus capitalize on the large number of subunits (people) available for estimating the few components of the variance-covariance matrix; this would rely on the so-called “naive” inference returned by some GEE procedures.

PARAMETER INTERPRETATION IN MIXED VERSUS POPULATION AVERAGE MODELS

To illustrate the differences in parameter interpretation between mixed models and population-average models, we continue with the example relating crime rates within neighborhoods to the mean walking level of neighborhood residents. We now use a simpler random intercept version of Equation 2:

$$\text{logit}[P(Y_{ij} = 1 | X_{ij}, \alpha_{0j})] = \beta_0 + \alpha_{0j} + \beta_1 X_{ij}, \alpha_{0j} \sim N(0, \sigma). \quad (4)$$

Within this model, the interpretation of the coefficient β_1 relates changes in the mean of the outcome (proportion walking) via changes in the crime rate within the higher units (neighborhoods). In general, for multivariable random-intercept logistic-regression models, the nonintercept fixed-effect parameters are interpreted as the log [odds ratio (OR)] for a change in the associated explanatory variable, holding the neighborhood fixed. In our example, β_1 is the log(OR) of walking for a unit increase in crime rate holding the neighborhood fixed.

Now, assume a logistic regression population average model, or $\text{logit}[P(Y_{ij} = 1 | X_{ij})] = \beta_0^* + \beta_1^* X_{ij}$ (note that the

random intercept model above does not imply this form as the correct population average model), where * indicates that these coefficients are not the same parameters as in the random intercept model. Specifically, β_1^* is a measure of association relative to changes in explanatory variables across neighborhoods or the log(OR), comparing the probability of walking for individuals who live in neighborhoods that differ by 1 unit of crime rate.

Although they estimate theoretically different parameters in this logistic case, for linear and log-linear models, the coefficients from the typical specification of mixed-effects models will be numerically equivalent to coefficients from the population-average models, given that both models imply the same model for $E(Y_{ij} | X_{ij})$. In the linear case, this can be seen by noting that population-average model is derived by averaging the neighborhood-specific model across neighborhoods, or

$$\begin{aligned} E(Y_{ij} | X_{ij}) &= E_{\alpha_j}[E(Y_{ij} | X_{ij}, \alpha_j)] \\ &= \mu(X_{ij} | \beta) + E_{\alpha_j}[U_{ij}(\alpha_j, X_{ij})] = \mu(X_{ij} | \beta). \end{aligned}$$

Thus, for linear and log-linear models, the distinction between estimating a within-neighborhood effect using maximum likelihood estimation and a population-average effect using GEE is less critical; the coefficient estimates returned from the mixed-model estimating procedures are estimates of both such effects.

In contrast, much has been made of the differences in the actual numerical values of coefficients from a logistic mixed model versus the population-average model. For instance, one can show that the population average OR is always closer to the null value of 1 than the corresponding mixed effects OR when the true model is a simple random intercept model.¹² To be more concrete, consider a situation with X_{ij} is binary (crime rate 1 = high, 0 = low), then the slope coefficients in a random-intercept logistic-regression model is the average within neighborhood log(OR) of walking when the neighborhood is at high versus low crime, whereas the population-average logistic regression produces an estimate of the log(OR) that compares the average probabilities of walking (averaged across all neighborhoods) in high-crime versus low-crime neighborhoods.

For typical cross-sectional neighborhood-level data (the X_{ij} is constant for all i within a j), the estimate of the association within the random effects model has to come from comparisons across neighborhoods, because no within-neighborhood log(OR)s are estimable directly. Thus, the estimate of the slope coefficient (the β_1 in Equation 4) depends on assumptions about the distribution of the random effect: by assuming, for example, normality of an unmeasured random effect, only then is the within-neighborhood effect estimable from the data. Thus, papers such as Larsen and Merlo,¹³ which provide statistics for characterizing the

amount of between-neighborhood variability returned from these random effects models, are useful only insofar as the assumptions behind these random-effects models are not strongly violated. Note that Carlin et al¹⁴ discuss another interesting latent-variable model based on discrete mixtures, and provide an example in which the typical normal assumption for the random effect is questionable. However, one should be cautious, when interpreting estimates that are sensitive to distributional assumptions on unmeasured variables.

REGRESSION MODELS AS PROJECTIONS

A appropriate analytic approach is to: (1) propose the scientific question of interest, (2) determine the relevant parameter of interest that addresses the question, (3) propose an estimator that might require empirically nonidentifiable assumptions, and (4) report the estimate with appropriate caveats.⁸ Latent variable models, including mixed models (see Equation 2 for instance), always require assumptions untestable by the data. These models might uniquely address the specific scientific question of interest, for instance by evaluating the association of latent variables, but the information they provide in the form of the estimates of the parameters of interest (for instance, the within-unit associations of neighborhood-level variables) can be misleading if the latent variable model is misspecified. To derive reliable likelihood-based inference, the model should not represent an approximation, but rather the true form of the mean model given the covariates and latent variables. Lack of compelling theory for a particular model choice, and an insufficient sample size to estimate the mean model flexibly, results in estimated regressions models that are, at best, reasonable approximations to the conditional mean. If the latent variable part is not modeled correctly, it is unclear how to interpret the fixed-effects estimates even if that part of the model is

correctly specified. The situation becomes even more complicated with the misspecification of the fixed-effects portion.

In contrast, one could reasonably justify estimating an informative approximation of the true population-average model (and in theory this parameter is estimable with few assumptions in large sample sizes). With this approach, one could avoid any model specification (ie, the model is non-parametric) and the parameter of interest would be some approximation of $E[Y_{ij} | X_{ij}]$ within the proposed class of estimating models (eg, linear with only main effects). That is, one can define explicitly the parameter of interest as a function of the distribution of the observed data (X, Y). (Note: the approximation can be different depending on what working correlation structure one uses in GEE—see an associated technical report for details.¹⁵) In essence, one can think of the parameter of interest as the coefficients one would derive if the proposed model, $\mu(X_{ij} | \beta)$, using a particular algorithm, was fit to not just a sample of people in a sample of neighborhoods, but instead the entire target population of interest using the proposed estimation algorithm (eg, weighted least squares using the weight matrix implied by the working correlation model specified in the GEE models).

For example, consider a simple nonrepeated-measures data structure and the true model $Y = b_0 + b_1X + b_2X^2 + e$ with $e \sim N(0, \sigma^2)$. Assume one fits a linear model of Y on X to data from the underlying population generated by this quadratic model. Also assume that the parameters of interest are the c_0 and c_1 one would obtain if one fit a linear model $Y = c_0 + c_1X + \epsilon$ to the entire population. Figure 1 shows (1) the underlying true mean model, (2) the projection of a linear model on the truth, (3) actual data from the underlying true model, and (4) the fit of a linear model to that data. This caricature illustrates what generally happens when regression models are fit for explanatory purposes. If one thinks of the

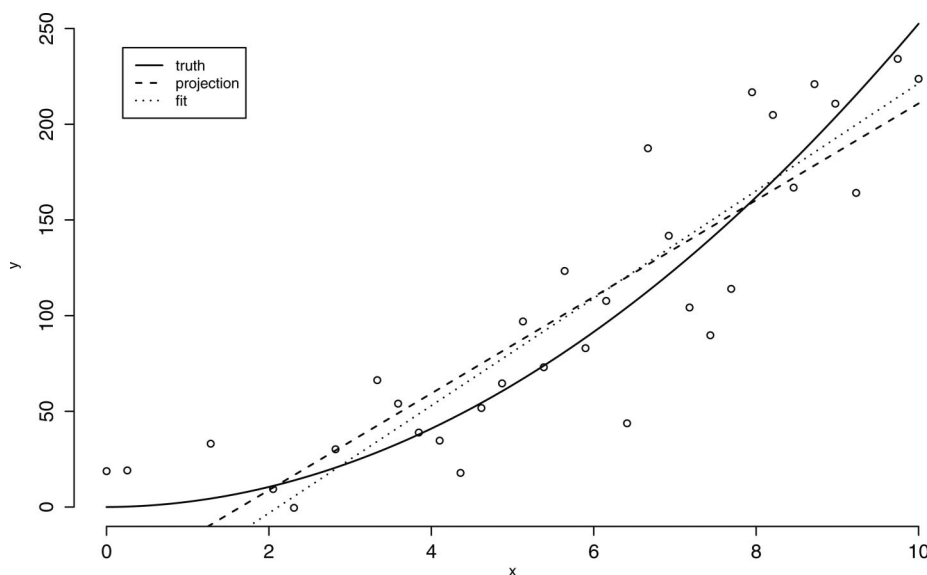


FIGURE 1. Example of truth, projection, data, and fit.

parameter of interest as the true conditional mean of Y , given the covariates of interest, then one is almost always wrong. However, if one views the parameter of interest as a projection of the “truth” onto some smaller set of models (eg, linear), this is often a reasonable parameter of interest. (Of course, one can easily come up with exceptions, for example a linear model would fail to capture a U-shaped relationship.) For instance, one might want to know the average trend. In that case, a linear fit to data generated from an underlying nonlinear model is a legitimate parameter. Other examples could include fitting a linear model when the true model is log-linear, etc.

Consider the case of a mixed model in which the true data generating distribution has the following form (binary outcome) of:

$$\log \left[\frac{P(Y_{ij} = 1 | X_{ij}, b_{0i})}{1 - P(Y_{ij} = 1 | X_{ij}, b_{0i})} \right] = b_0 + b_1 X_{ij} + b_{0i} \exp(b_2 X_{ij}), b_{0i} \sim N(0, 1)$$

This is a random effects model, but one in which the OR for a unit change in X_{ij} , by unit i , depends on the level of the random effect, b_{0i} . Now, we examine different fits to data generated from this data-generating distribution (note that X_{ij} is uniformly distributed over the integers 0 to 10, $b_0 = -5$, $b_1 = \log(2)$, $b_2 = 0.5$, 100 units and 100 subunits/unit). The results of this analysis are presented in Figure 2, which depicts (1) the true marginal probability of the outcome, by X_{ij} , $P(Y_{ij} = 1 | X_{ij})$, (2) the best approximation based on a simple logit-linear model and independence working correlation model, (3) the corresponding

GEE estimate of this approximation, (4) the estimate using the same data as that for the GEE model of a simple random effects (random intercept) logistic regression model (the plot is done at the random effect, $b_{0i} = 0$), and (5) the population-average mean estimate derived from the misspecified random effects model by marginalization over the estimated random effects distribution.

The results emphasize that the projection of the population average is something one can hope to estimate, and that it bears a rigorously definable relation to the true underlying association. In contrast, the misspecified random-effects model estimate bears little relation to the true underlying marginal association (of course, it is not an estimate of this parameter). It is unclear how to interpret the results of this model because the distribution of individual curves (ie, neighborhood-level curves) can differ widely depending on the form of the random-effects model and the distribution of the random effects.

Finally, the population-average mean estimate from a misspecified random-effects model is of course biased and unnecessarily so, as one need not specify the correct random-effects part to estimate this parameter. If the random effects part is misspecified, both the resulting neighborhood-specific effects (eg, coefficients in a random effects model) and the implied estimated population average model can be unpredictably biased relative to the quantities they are estimating. Heagerty and Zeger⁸ make a very similar point, that the regression parameters in conditionally-specified models (the fixed effects in random effects models) are much more sensitive to random-effects assumptions than are their counterparts in the population-average model.

Given the objective of providing a reasonable approximation of the population average within covariate groups,

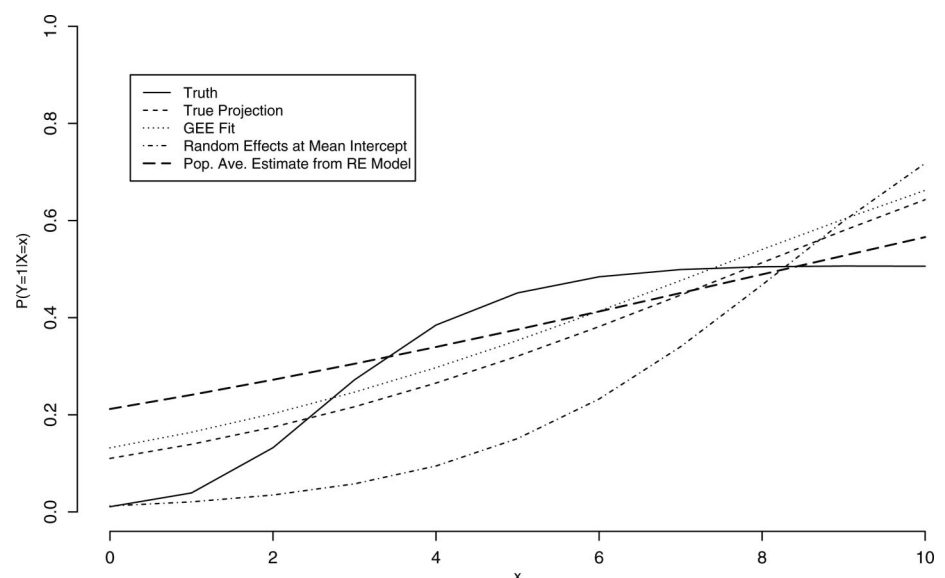


FIGURE 2. Example of truth population average model, true projection onto a logit-linear regression model, a GEE fit assuming a simple logit-linear model as estimated from random sample of data, a corresponding estimate of the fixed effects coefficients in a logit-linear random effects (RE) model with a normally distributed random intercept by unit, and the population average model derived from the estimate of this random effects model.

then inference can be derived from the robust sandwich estimators of variability used by the GEE approach. Freedman¹⁶ discusses how the robust estimators of standard errors are correct even if the mean model is misspecified. However, he argued that one wants a confidence interval for the true parameter, not its projection. In contrast, we take the perspective that these projections can yield useful information about the relationship of neighborhood factors and health outcomes and thus, for many purposes, the important issue is how inferences are derived for this nonparametric model. In Figure 1, the above case of an interpretation of linear projection is relatively straightforward, but further research is needed to interpret the more general case (ie, when the coefficients of a mixed model cannot be interpreted as coefficients in the corresponding population average model, such as in the logistic case). There are some cases where the estimated parameter can be quite misleading relative to the parameter of interest (see for example van der Laan et al¹⁷). Neugebauer and van der Laan¹⁸ describe the parameter of interest as some nonparametric projection of regression models onto the true underlying (causal) regression form. Their development of the theory is modified in our technical report.¹⁵ Essentially, given that the projection of the mean model is interpretable and useful, one can always gain robust inference using the type of robust inference provided by GEE or by using the appropriate bootstrap.

DISCUSSION

The contrast of estimation of the population-average model and corresponding mixed-effects models is part of larger issues in statistical estimation and identifiability. A simple list of the elements of data analysis provides a basis for understanding the intersection of the scientific question, data and statistical estimation: (1) the observed data, (2) the model (set of possible data-generating distributions given background knowledge), (3) parameter of interest of data-generating distribution that addresses scientific question, (4)

estimation of this parameter, and (5) additional nontestable assumptions on data-generating distributions. The data can of course identify (without further assumptions) only the joint distribution of the observed data, say $O_j = (Y_j, X_j)$, where Y_j and X_j are the vector of outcomes and corresponding matrix of explanatory variables measured on neighborhood j . Mixed models, however, estimate parameters of a combined latent variable/observed variable distribution, say $O_j^* = (Y_j, X_j, \alpha_j)$, where again α_j are the neighborhood-level unobserved random variables. Interpretations of parameters of the distribution of O_j^* should require justifications (outside the data) as to why this theoretical data structure is warranted, much as causal inferences from associations in observational data require untestable assertions such as unmeasured confounding. Raudenbush²⁴ sensibly argues that, if the scientific question of interest suggests parameter estimates that involve assumptions on latent variable distributions, then inference based on such assumptions is unavoidable. However, one must acknowledge that the evidentiary weight of such inference is not on par with the inferences made regarding parameters based on assumptions only on the observed data distribution.

We have already discussed reasons why the data can not identify the true distribution of O_j^* (assuming the true distribution can even be represented in that way). However, the data can be used to examine the relative merits (fit) of competing models for O_j^* . In fact, as opposed to the GEE approach, a very attractive approach is to fit various models (random effects and others) that result in different models for the distribution of O_j , and use likelihood-based procedures (such as likelihood-based cross-validation) to select those models. However, inference in the end should, without further untestable assumptions, be only with regard to the resulting distribution of the observed data, O_j .

The Table provides a summary of critical issues to consider when selecting a modeling approach. If the analyst

TABLE. Summary of Approaches for Mixed Models and GEE

| | Mixed Models | GEE |
|-----------------------------------|--|---|
| Focus of interest | Variance components and regression coefficients | Regression coefficients |
| Parameter interpretation | Neighborhood specific | Population average |
| Linear (estimates equivalent) | Change in the mean outcome for a unit change in the associated neighborhood exposure, keeping the random effect (neighborhood) fixed | Change in the mean outcome for a unit change in the associated neighborhood exposure across all of the neighborhoods observed |
| Binary (estimates NOT equivalent) | The log(OR) of an outcome for a unit change in the associated neighborhood exposure, keeping the neighborhood fixed. Not identifiable in cross-sectional studies of neighborhoods without additional assumptions on random effects distribution. | The log(OR) of an outcome for a unit change in the in the associated neighborhood exposure across all of the neighborhoods observed |
| Assumptions | Correctly specified error distribution | No. neighborhoods sufficiently large for robust estimation of standard errors |
| Pitfalls | SE not robust to model misspecification (can use regression diagnostics) | With small no. neighborhoods, SE biased (although not substantially so in simulations with small numbers of units) |

aims to understand the error structure of the data-generating distribution, or wants to derive the estimate of within-unit OR from cross-sectional data, then there is no choice but to use modeling procedures that make explicit untestable assumptions of the data-generating distribution, such as those within mixed models. One should be cautious, given how sensitive the conclusions can be to the assumptions of these models. However, if the focus of the analysis is the estimation of mean effects as well as the estimation of the inference of the coefficients in the model (eg, the association of walking and neighborhood crime rate), then estimating the population-average model via GEE provides a compelling alternative. GEE allows robust inference even if the correlation model is misspecified, or the parameter of interest is not the true mean model but a projection onto a class of approximating models (eg, linear model), as in Figure 1 above. We assert that this is usually the case in statistical models of observational data with many explanatory variables.

We have compared 2 methods for estimating neighborhood-level effects. We contend that researchers should make explicit the assumptions of each method and also consider situations where one method might be preferred over the other. Previous discussions have provided interesting points regarding each method, and philosophical viewpoints on the relative merits.^{19–21} However, the relative merits are not a question of philosophy—once the data are presented, the parameter of interest stated, and the assumptions made explicit and accepted, then the choice should be straightforward. In addition, this choice of estimating population-average models versus mixed models goes beyond a choice between GEE and mixed models. Similar questions are being actively debated in the literature on causal inference. There are issues regarding competing likelihood-based latent variable methods (such as structural equation models) versus estimating function-based methods²² and related targeted maximum likelihood²³ methods that avoid latent-variable formulations. The underlying issues and user recommendations are similar. Does one rely on correct specification of untestable aspects of the data-distribution (the latent variable approach) or on the more narrow assumptions available from the other causal inference methods? The answer will depend on the goals of the analyses, but should also depend on information available to the researchers. Knowing the assumptions of each method and how these assumptions affect the inferences from the analysis will enable researchers to determine the best approach to analyzing their data.

REFERENCES

1. Kawachi I, Berkman LF. *Neighborhoods and Health*. Oxford, New York: Oxford University Press; 2003.
2. Macintyre S, Maciver S, Sooman A. Area, class and health; Should we be focusing on places or people? *J Soc Policy*. 1993;22:213–234.
3. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38:963–974.
4. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13–22.
5. Oakes JM. The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology (with discussion). *Soc Sci Med*. 2004;58:1929–1952.
6. Diggle P, Heagerty P, Liang KY, Zeger S. *Analysis of longitudinal data*. *Oxford Statistical Science Series*, 25. 2nd ed. Oxford, New York: Oxford University Press; 2002.
7. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley-Interscience; 2004.
8. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference (with discussion). *Stat Sci*. 2000;15:1–26.
9. Gardiner JC, Luo ZH, Roman LA. Fixed effects, random effects and GEE: What are the differences? *Stat Med*. 2009;28:221–239.
10. Efron B, Tibshirani R. *An Introduction to the bootstrap*. New York: Chapman & Hall; 1993.
11. Huber PJ. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press; 1967.
12. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev*. 1991;59:25–35.
13. Larsen K, Merlo J. Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *Am J Epidemiol*. 2005;161:81–88.
14. Carlin JB, Wolfe R, Brown CH, Gelman A. A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics*. 2001;2:397–416.
15. Hubbard AE, van der Laan MJ. Nonparametric population average models: deriving the form of approximate population average models estimated using generalized estimating equations. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Berkeley: University of California; 2009. Working Paper 251.
16. Freedman DA. On the so-called “Huber Sandwich Estimator” and “Robust Standard Errors.” *Am Stat*. 2006;60:299–302.
17. van der Laan MJ, Hubbard A, Jewell NP. Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome. *J R Stat Soc Series B Stat Methodol*. 2007;69:463–482.
18. Neugebauer R, van der Laan MJ. Nonparametric causal effects based on marginal structural models. *J Stat Plan Inference*. 2007;137:419–434.
19. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Stat Med*. 2002;21:3291–3315.
20. Merlo J. Multilevel analytical approaches in social epidemiology: measures of health variation compared with traditional measures of association. *J Epidemiol Community Health*. 2003;57(8):550–2.
21. Petronis KR, Anthony JC. Social epidemiology, intra-neighbourhood correlation, and generalized estimating equations. *J Epidemiol Community Health*. 2003;57:914; author reply 914.
22. van der Laan MJ, Robins J. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer; 2002.
23. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;Article 11.
24. Raudenbush SW. Comment on the article marginalized multilevel models and likelihood inference. *Stat Sci*. 2000;15:22–24.