STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:5745–5763
Published online 11 August 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3384

Combining data from primary and ancillary surveys to assess the association between neighborhood-level characteristics and health outcomes: the Multi-Ethnic Study of Artherosclerosis

B. N. Sánchez[1,*,†], T. E. Raghunathan[1], A. V. Diez Roux[2], Y. Zhu[3] and O. Lee[1]

[1]*Department of Biostatistics, University of Michigan, School of Public Health, Ann Arbor, MI 48109, U.S.A.*
[2]*Department of Epidemiology, University of Michigan, School of Public Health, Ann Arbor, MI 48109, U.S.A.*
[3]*Department of Biostatistics, Boston University, School of Public Health, Boston, MA 02118, U.S.A.*

SUMMARY

There is increasing interest in understanding the role of neighborhood-level factors on the health of individuals. Many large-scale epidemiological studies that accurately measure health status of individuals and individual risk factors exist. Sometimes these studies are linked to area-level databases (e.g. census) to assess the association between crude area-level characteristics and health. However, information from such databases may not measure the neighborhood-level constructs of interest. More recently, large-scale epidemiological studies have begun collecting data to measure specific features of neighborhoods using ancillary surveys. The ancillary surveys are composed of a separate, typically larger, set of individuals. The challenge is then to combine information from these two surveys to assess the role of neighborhood-level factors. We propose a method for combining information from the two data sources using a likelihood-based framework. We compare it with currently used *ad hoc* approaches via a simulation study. The simulation study shows that the proposed approach yields estimates with better sampling properties (less bias and better coverage probabilities) compared with the other approaches. However, there are cases where some *ad hoc* approaches may provide adequate estimates. We also compare the methods by applying them to the Multi-Ethnic Study of Atherosclerosis and its Neighborhood Ancillary Survey. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: maximum likelihood; ancillary survey; social epidemiology

*Correspondence to: B. N. Sánchez, Department of Biostatistics, University of Michigan, School of Public Health, Ann Arbor, MI 48109, U.S.A.
†E-mail: brisa@umich.edu

Contract/grant sponsor: Publishing Arts Research Council; contract/grant number: 98-1846389
Contract/grant sponsor: National Heart, Lung, and Blood Institute; contract/grant numbers: N01-HC-95159, N01-HC-95166, R01 HL071759

Received 24 October 2007
Accepted 10 June 2008

1. INTRODUCTION

Several studies have shown that people living in disadvantaged neighborhoods have adverse health conditions [1–3]. Most of these studies use area-level data available at small geographical levels, such as census data at the census tract level, to develop neighborhood-level disadvantage measures. However, these measures of neighborhood disadvantage do not directly measure the specific exposures (such as the ability to walk or the availability of healthy foods) that may be related to health outcomes (such as cardiovascular disease and its risk factors). The absence of direct measures of the specific features hypothesized to be related to health limits the ability to draw inferences regarding neighborhood health effects. Identifying the specific features that are relevant is also fundamental from the point of view of developing neighborhood-level interventions that could subsequently be tested in randomized trials. For these reasons the development of measures of neighborhood-level attributes is an important need in epidemiology.

There are several methodological challenges in moving beyond simple census-based measures of neighborhood characteristics. For instance, the data sources that directly measure some of the relevant neighborhood-level characteristics have to be identified, and the analytical methods to integrate them with data that include individual-level outcomes have to be developed. Although administrative databases (such as data on the location of parks and recreational facilities) may be available to construct some neighborhood-level measures, they are limited in scope. In addition, some neighborhood-level constructs, such as social cohesion, cannot be assessed using administrative databases. Thus, to address these limitations, sociological and epidemiologic studies have begun implementing ancillary surveys or equivalent data collection from informants or raters to develop specific measures of neighborhood conditions [4–7]. These collect data on neighborhood features of interest from local informants, separate from the people on whom health outcomes are assessed.

One such large-scale epidemiological study is the Neighborhood Ancillary Study to the Multi-Ethnic Study of Atherosclerosis (MESA). MESA is a 10-year longitudinal study of 6814 men and women without clinical cardiovascular disease at baseline from four racial/ethnic groups (Non-Hispanic White, African-American, Chinese and Hispanic). The participation rate for those screened and eligible was 59.8 per cent. Participants reside in six study sites (Baltimore, MD; Chicago, IL; Forsyth County, NC; Los Angeles, CA; Northern New York, NY; St. Paul, MN). In each study site, participants were identified using lists of dwellings and telephone numbers; full details have been documented elsewhere [8]. These data provide primary individual-level outcome variables.

The Neighborhood Ancillary Study to MESA conducted a survey of other residents from the same neighborhoods as the MESA participants in three (Baltimore, Forsyth County and New York) of the six MESA sites. The ancillary survey was restricted to these three sites due to funding restrictions. The purpose of this ancillary study was to assess a variety of neighborhood characteristics using a series of scales. The collection of data from a sample of residents who are not MESA participants has two important advantages: (a) it allows a larger sample size per neighborhood potentially increasing the validity and reliability of the measure and (b) it avoids reporting bias which may occur when neighborhood and behavioral information is obtained from the same subject (e.g. persons with worse diets may be less likely to report healthy food availability, regardless of actual availability in their neighborhood) [4]. The survey sample consisted of 5988 individuals that resided in the same census tracts as the MESA participants. It was conducted through random digit dialing, with one person sampled per household. The response rate was

DOI: 10.1002/sim

46.5 per cent and the sample was approximately representative of the geographic areas from which it was drawn [9]. The study participants were interviewed over the phone to measure four neighborhood features: the walking environment (seven items), the availability of healthy foods (three items), the neighborhood safety (three items) and the neighborhood social cohesion (four items). These domains were conceptualized as being related to cardiovascular disease and associated risk factors.

The objective of this paper is to evaluate methods that combine information from these two types of data sources. To illustrate these methods, we will assess the relationship between the availability of healthy foods (exposure) in the neighborhood, captured by the ancillary study, and hypertension (outcome) at baseline, as reported by the MESA participants. We adjust for potential individual-level covariates. In Section 2, we specify models relating the exposure, outcome and covariates, and in Section 3 we discuss four potential approaches for estimating the parameters in the models. In Section 4, we report on a simulation study evaluating these methods. In Section 5, we discuss the application of these methods to the actual MESA study that motivated this research. Finally, Section 6 concludes with a discussion of the implications of this work.

2. DATA STRUCTURE AND MODEL

For each neighborhood j, $j=1,\ldots,J$, let n_j be the number of participants from the original health study (e.g. MESA), and let m_j be the number of participants from the ancillary neighborhood study. Let $Y_j=(Y_{j1},\ldots,Y_{jn_j})$ be the health outcomes and $X_j=(X_{j1},\ldots,X_{jn_j})$ be covariates for the individuals in neighborhood j. For simplicity of notation we assume that X_{ji}, $i=1,\ldots,n_j$, is univariate, but this can be easily extended. Suppose that U_{jk}, $k=1,\ldots,m_j$, is a univariate response to a neighborhood measure from subject k in the ancillary survey, and suppose that θ_j is the population average of the responses $U_j=(U_{j1},\ldots,U_{jm_j})$ in neighborhood j. We assume that

$$U_{jk}|\theta_j \sim \text{iid } \mathrm{N}(\theta_j,\sigma^2) \tag{1}$$

The substantive question of interest is the association between θ_j and Y_{ji} after adjustment for X_{ji}.

Given the true neighborhood characteristic, θ_j, and fixed covariates, X_{ji}, assume that the health outcome follows a generalized linear model [10]. That is, for subject i in neighborhood j, $i=1,\ldots,n_j$, $j=1,\ldots,J$, the expected response $\mu_{ji}=E(Y_{ji}|\theta_j,X_{ji})$ can be modeled as

$$g(\mu_{ji})=\beta_0+\beta_1\theta_j+\beta_2X_{ji} \tag{2}$$

where β_1 represents the association between the neighborhood-level covariate and the outcome and is of primary interest, and $g(\cdot)$ is a known link function, for example, the logistic transformation.

This model could also be extended to include interactions between θ_j and X_{ji}. For example, the association between the availability of healthy foods in the neighborhood and hypertension may be different for men and women. The model of interest would instead be

$$g(\mu_{ji})=\beta_0+\beta_1\theta_j+\beta_2X_{ji}+\beta_3X_{ji}\theta_j \tag{3}$$

where β_3 is the difference in association between the neighborhood-level characteristic and outcome at different levels of the individual-level covariate.

DOI: 10.1002/sim

3. ESTIMATION

Several estimation procedures can be used to estimate the health effects of neighborhood characteristics derived from ancillary surveys. Some procedures consist of substituting an estimate of θ_j in model (2) or (3) [6, 7]. For example, the sample mean, $\overline{u}_j = \sum_{k=1}^{n_j} U_{jk}/n_j$, may be used as an estimate of θ_j. However, the number of individuals in the ancillary study, m_j, may be small in many neighborhoods, making $\overline{u}_j$ an unreliable estimate of θ_j. Furthermore, the sample means $\overline{u}_j$ are subject to sampling variability, and discarding such sampling variability may lead to invalid estimates of β_1 and β_3. Parameter estimates obtained by plugging in $\overline{u}_j$ in (3) will be labeled 'Plug-in Mean'.

A more sophisticated approach is to substitute an Empirical Bayes estimate of θ_j, $\widehat{\theta}_j$, after adding the assumption that the neighborhood characteristics vary in the population according to

$$\theta_j \sim \mathrm{N}(Z_j^t \gamma, \tau^2) \tag{4}$$

In (4), Z_j is a vector of population-level characteristics of the neighborhood obtained, for example, from census data. Imposing the additional structure (4) provides a framework to incorporate other neighborhood characteristics and borrow strength across neighborhoods to obtain better estimates of θ_j. Such estimates can be obtained using, for example, SAS Proc MIXED [11]. This approach, however, also discards the sampling variability of $\widehat{\theta}_j$ and may lead to invalid estimates of β_1 and β_3. Estimates obtained from this approach will be labeled 'EB'.

An alternative approach consists of a slight modification as follows. Noting that $\theta_j = \widehat{\theta}_j + \xi_j$, where ξ_j is an error of estimation, we can rewrite the model as

$$\begin{aligned} g(\mu_{ji}) &= \beta_0 + \beta_1(\widehat{\theta}_j + \xi_j) + \beta_2 X_{ji} \\ &= \beta_0 + \beta_1 \widehat{\theta}_j + \beta_2 X_{ji} + \xi_j^* \end{aligned} \tag{5}$$

where $\xi_j^* = \beta_1 \xi_j$ is a random quantity shared by all responses μ_{ji} in neighborhood j. In other words, the sampling variability in the estimated $\widehat{\theta}_j$ induces correlation among the health outcomes. To account for such correlation, one could fit a mixed model with a random intercept for each neighborhood. However, because the number of individuals in the ancillary study, m_j, may vary by neighborhood, the variance of the random effect will also vary by neighborhood, i.e. $\xi_j^* \sim \mathrm{N}(0, \delta_j)$.

Implementation of this alternative may not be straightforward with standard software for hierarchical models such as SAS [11] or HLM [12], although fully Bayesian implementations may be possible in software such as WinBUGS [13]. However, by making the simplifying assumption that $\xi_j^* \sim \mathrm{N}(0, \delta)$, (5) becomes a generalized linear mixed model [14]. Thus, an approximation to this approach is easily implementable in both SAS [11] and HLM [12]. However, this approximation disregards the fact that the magnitude of the correlation among the health outcomes induced by the sampling variability of $\widehat{\theta}_j$ depends on the magnitude of the true association β_1. Furthermore, the errors ξ_j^* may not be independent across neighborhoods because the estimates $\widehat{\theta}_j$ are simultaneously obtained from fitting a model to the data from all neighborhoods. This induces non-independence of neighborhoods, which could result in biases in the standard errors of regression coefficients. Lastly, simplifying $\xi_j^* \sim \mathrm{N}(0, \delta_j)$ to $\xi_j^* \sim \text{iid } \mathrm{N}(0, \delta)$ assumes that the precision of the neighborhood-level measure is the same for all neighborhoods, which may decrease the efficiency of $\widehat{\beta}_1$. Although with some limitations, this approach, which we will call 'EB+RE', may yield some advantages over the two previous approaches.

DOI: 10.1002/sim

A preferred approach is to simultaneously estimate the parameters in (1)–(4) via maximum likelihood estimation. From each neighborhood, the contribution to the joint likelihood based on the observed data conditional on θ_j's is

$$L_j(\beta, \sigma^2 | \theta_j, Y_j, X_j, U_j) = \prod_{i=1}^{n_j} f(Y_{ji} | \theta_j, X_{ji}, \beta) \prod_{k=1}^{m_j} f(U_{jk} | \theta_j, \sigma^2) \tag{6}$$

Since the θ_j are iid, the likelihood based on the observed data is obtained by integrating (6) with respect to θ_j and taking the product across neighborhoods

$$L(\beta, \sigma^2, \gamma, \tau^2 | Y, X, U) = \prod_{j=1}^{J} \int L_j(\beta, \sigma^2 | \theta_j, Y_j, X_j, U_j) f(\theta_j | \gamma, \tau^2, U_j) \, \mathrm{d}\theta_j \tag{7}$$

where $Y = (Y_1, \ldots, Y_J)$, $X = (X_1, \ldots, X_J)$ and $U = (U_1, \ldots, U_J)$. We implemented maximization of the joint likelihood, (7), as a function of $\beta, \gamma, \sigma^2, \tau^2$ in Proc NLMIXED in SAS [11]. Although other software can be used for this purpose, we chose SAS because of its widespread use in epidemiological applications. Appendix A includes prototype programming statements as well as a graphical description of how the data sets from both surveys should be merged. In what follows, this fourth approach will be called 'MMLE'.

4. SIMULATIONS

We conducted a simulation study to compare the sampling properties of parameter estimates obtained using the four estimation procedures described in the previous section. We chose the number of neighborhoods and the number of respondents per neighborhood to be the same as in the MESA study and the ancillary survey described in the introduction.

4.1. Main effects model

To generate simulated data for a model with only a main effect for the neighborhood characteristic we used the

$$\text{main effects model:} \quad g(\mu_{ji}) = \beta_0 + \beta_1 \theta_j$$

where $g(\cdot)$ was assumed to be the logit link. Responses from the ancillary study were simulated using

$$U_{jk} | \theta_j \sim \text{iid } \mathrm{N}(\theta_j, \sigma^2)$$

and

$$\theta_j \sim \text{iid } \mathrm{N}(\gamma_0, \tau^2)$$

The simulation was conducted in a 5^3 factorial design, where the true values for the parameters were chosen to cover a range of situations. The values for β_1 were $0, 0.5, 1, 2$ and 3, and the values for the between-cluster variance, τ^2, were $0.25, 0.5, 1, 5$ and 10. The values for the within-cluster variance, σ^2, were chosen such that the within-to-between variance ratios, σ^2/τ^2, were $0.25, 0.5, 1, 2$ and 4. For a given set of parameters, we generated 500 simulated data sets. For each

DOI: 10.1002/sim

data set, parameter estimates were obtained through the four methods described in the previous section. The maximum standard deviation of the Monte Carlo error for $\widehat{\beta}_1$ was 0.01 over all parameter scenarios.

Figure 1 shows the per cent of bias in $\widehat{\beta}_1$ for the main effects model from each estimation method. When $\beta_1 = 0$, the estimates from all methods were unbiased. When $\beta_1 \neq 0$, the estimates obtained from the methods that substitute an estimate of θ_j in the outcome model were biased toward the null. In general, bias increased with increasing value of β_1; the within-to-between neighborhood variance ratio in the responses of the ancillary study, σ^2/τ^2; and the values of τ^2. Using the Empirical Bayes estimate for θ_j and including a random effect (EB+RE), as in (5), yielded estimates that were only slightly biased in comparison with the marginal maximum likelihood estimates (MMLEs) obtained from maximizing the joint likelihood (7). This bias was more evident for larger effect sizes (e.g. $\beta_1 = 2$). In contrast, the bias in $\widehat{\beta}_1$ from substituting $\widehat{\theta}_j$ or $\overline{u}_j$ for θ without including a random effect was much larger.

Figure 2 shows the empirical coverage probabilities of 95 per cent nominal level confidence intervals for β_1 in the main effects model. When $\beta_1 = 0$, the coverage from all estimates remained at approximately 95 per cent. When $\beta_1 \neq 0$, the coverage probabilities from the EB+RE method were in general close to those from the MMLE. However, for large effect sizes, the coverage probabilities from the EB+RE were slightly lower. The coverage probability from methods that substitute $\overline{u}_j$ or $\widehat{\theta}_j$ without including a random effect was very poor (e.g. <80 per cent) for most scenarios.

Table I gives a numerical summary of empirical bias of $\widehat{\beta}_1$, as well as the impact of simulation parameters on the bias in the main effects model. The overall mean and standard deviation of the empirical bias are given in the first row. The second row shows the sum of the ranks of the absolute bias for each method. For each parameter setting, the methods were ranked from 1 to 4, with 1 meaning the lowest bias. The ranks for each method over all parameter scenarios were then summed. Clearly, the MMLE approach suffers from less bias overall, whereas the Plug-in Mean has the largest average bias, as well as the most variable bias. Lastly, the table gives the estimated coefficients from a linear regression model where $100 * \text{Bias}(\widehat{\beta}_1)$ is the dependent variable. For each method, the R^2 represents the per cent of variability in the empirical bias explained by the simulation parameters. This table complements Figures 1 and 2 as it quantifies the impact of a given parameter, or combination of parameters, on the bias of $\widehat{\beta}_1$ for each method. For example, the bias in the EB+RE method is significantly predicted by the value of the product $\beta_1 \times \tau^2$, with larger values incurring larger attenuation. The bias from the EB method is more strongly impacted by this same product (higher coefficient), but even more strongly by the combination of the within-to-between variance ratio and the exposure coefficient: $\sigma^2/\tau^2 \times \beta_1$. Lastly, this table may be used to approximate the bias in the regression parameter $\widehat{\beta}_1$ from a given set of estimated parameters $(\widehat{\beta}_1, \widehat{\sigma}^2, \widehat{\tau}^2)$ using the fitted regression model.

Including covariates X_{ji} in the health outcome model (2) does not make a difference in making inferences for β_1 when θ_j and X_{ji} are uncorrelated. When θ_j and the covariates are correlated and X_{ji} is predictive of the health outcome, estimates of β_1 would be affected due to confounding. If confounders attenuate the observed association, then the differences between the methods will be less apparent. As seen from the simulation results, the methods yield more similar results when associations are small.

In a practical situation, including neighborhood-level covariates in (4) will partially explain the variance of the neighborhood-level characteristic (i.e. decrease the magnitude of the estimated τ^2). Since larger τ^2 increases the bias in the EB and EB+RE approaches, including covariates in the

DOI: 10.1002/sim

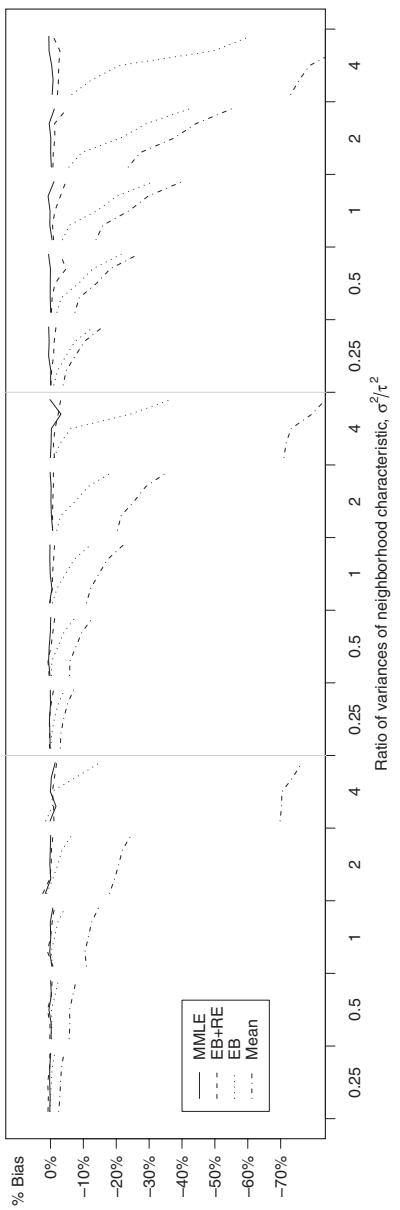

Figure 1. Per cent bias in $\widehat{\beta}_1$ from main effects model, as a function of simulation parameters. Within each value of σ^2/τ^2, the values of τ^2 are, from left to right, 0.25, 0.5, 1, 5 and 10. The true effect is (a) $\beta_1 = 0.5$, (b) $\beta_1 = 1$ and (c) $\beta_1 = 2$.

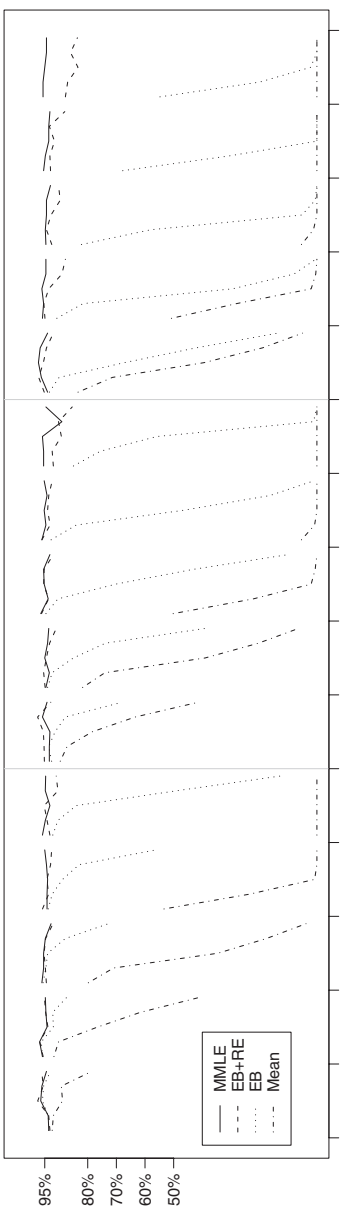

Figure 2. Empirical 95 per cent coverage probabilities for β from main effects model, as a function of simulation parameters. Within each value of σ^2/τ^2, the values of τ^2 are, from left to right, 0.25, 0.5, 1, 5 and 10. The true effect is (a) $\beta_1 = 0.5$, (b) $\beta_1 = 1$ and (c) $\beta_1 = 2$.

DOI: 10.1002/sim

Table I. Summary of the empirical bias of $\widehat{\beta}_1$*100, and linear regression coefficients to predict bias($\widehat{\beta}_1$)*100 from main effects model.

	MMLE	EB+RE	EB	Mean
Average bias (Std)	−0.3 (2.6)	−1.4 (4.4)	−26.0 (45.8)	−47.9 (66.9)
Sum of ranks	245	290	431	554
Intercept	−0.19	0.94	0.13	1.17
β_1	0.77**	−1.87*	−0.20	4.73
τ	0.00	−0.19	1.83**	2.18*
σ^2/τ^2	−0.05	−0.53	2.66	0.90
$\beta_1 \times \tau^2$	−0.09	0.01	−3.78*	−4.91*
$\sigma^2/\tau^2 \times \beta_1$	−0.32	1.27*	−5.84*	−20.23*
$\sigma^2/\tau^2 \times \tau^2$	0.09	0.25*	0.04	−0.25
$\sigma^2/\tau^2 \times \beta_1 \times \tau^2$	−0.05	−0.30*	−0.74*	0.43
Residual variance	5.7	7.5	255.4	256.5
R-square	0.20	0.60	0.90	0.95

*Denotes p-value <0.05.
**Denotes p-value <0.1.

equation of θ_j would ultimately reduce their bias. As neighborhood covariates are not used to derive the sample mean for the Plug-in Mean method, the regression coefficient corresponding to this approach would not be changed. The MMLE approach would remain approximately unbiased because, as seen from Table I, for fixed values of σ^2 and β_1, changes in τ^2 lead to non-significant changes in the bias of $\widehat{\beta}_1$.

4.2. Model with cross-level interactions

We also conducted simulations with a model that includes an interaction between the neighborhood-level feature and individual's characteristics

$$\text{interaction model:} \quad g(\mu_{ji}) = \beta_0 + \beta_1\theta_j + \beta_2 X_{ji} + \beta_3 X_{ji} * \theta_j$$

where $g(\cdot)$ was assumed to be the logit link. Responses from the ancillary study were again simulated with

$$U_{jk}|\theta_j \sim \text{iid } \mathrm{N}(\theta_j, \sigma^2)$$

and

$$\theta_j \sim \text{iid } \mathrm{N}(\gamma_0, \tau^2)$$

In the simulations, we chose X_{ji} to be gender. The parameter values used for the simulations were as follows: $\beta_1 = 1$ and 2; $\beta_2 = 1$ and 2; $\beta_3 = -1, -0.5, 0, 0.5$ and 1; and $\tau^2 = 5$ and 10. The values for σ^2 were chosen such that the within-to-between variance ratios, σ^2/τ^2, were 0.5, 1 and 1.5. For a given set of parameters, we generated 500 simulated data sets. For each data set, parameter estimates were obtained through the four methods described in the previous section. The maximum standard deviation of the Monte Carlo error for $\widehat{\beta}_3$ was 0.018 over all parameter scenarios.

DOI: 10.1002/sim

Figures 3 and 4 show the simulation results for the model with interactions. In terms of their dependence on simulation parameters, the bias and coverage followed a pattern similar to those in the main effects model. Again, for non-zero interaction effects, $\beta_3 \neq 0$, all methods that substituted an estimate of θ_j were biased toward the null, including the case when a random effect was included, as in (5).

Additionally, the magnitude of the bias depended on the sign of the true interaction effect. When β_3 is negative, the magnitude of the bias is slightly lower than the bias when β_3 is positive (e.g. Figure 3(b) vs (d)). This finding might seem counterintuitive at first. However, this difference in bias might arise because the magnitude of the correlation among health outcomes depends on the value of β_3. To see the differences in the magnitude of the correlations, suppose that $g(\cdot)$ is the identity link. Then the correlation between two observations within neighborhood j would be $var(\xi_i)\{\beta_1^2+\beta_3^2 X_{ji} X_{jk}+\beta_1\beta_3(X_{ji}+X_{jk})\}$. For positive β_1 and non-negative covariate values (as in our simulations), a negative β_3 reduces the degree of correlation among health outcomes within the neighborhood, whereas positive β_3 increases it. For the EB approach, ignoring larger correlations among health outcomes increases the bias in regression parameters. The EB+RE approach ignores the fact that the correlation is heterogenous within neighborhoods (e.g. differs by gender). A possible alternative might be to include a second random effect in the outcome model, namely a random slope for X_{ji}.

Finally, the magnitude of the main effects β_1 and β_2 did not have a great impact on the bias of $\widehat{\beta}_3$ or on the coverage probabilities. Therefore, the results shown in the figures are the bias and coverage values obtained for the case when $\beta_1 = 1$. When $\beta_1 = 2$, the magnitude of the bias was only slightly larger and the coverage was slightly lower than in the case when $\beta_1 = 1$. The bias comparing the values of the association with gender β_2 ($\beta_2 = 1$ vs 2) was very similar; thus the figure presents the bias averaged over both values.

5. EXAMPLE

We applied all four methods of estimation to the MESA study and its neighborhood ancillary study to assess the association of healthy food availability on hypertension. Other neighborhood characteristics may also have an association with hypertension, but here we use healthy food availability as an example. Neighborhoods were defined using census tracts from the 2000 Census. To be included in this analysis, each neighborhood had to contain at least one participant from both the primary and ancillary studies. A total of 436 neighborhoods were included, with a total of 2676 MESA participants and 5074 ancillary survey participants. For the three MESA sites included in the MESA Neighborhood study, the average number of participants per tract was 6.5 (median 4, IQR: 2–8). For the neighborhood ancillary survey, the average number of participants per tract was 11.5 (median 9, IQR: 5–15).

For participants in the MESA study, hypertension was defined according to JNC VI (1997) criteria [15]. Participants were deemed hypertensive if one of the following was true: their systolic or diastolic blood pressure was greater than or equal to 140 or 90 mmHg, respectively, or they reported being hypertensive *and* were taking anti-hypertension medication. We included the following individual-level covariates in the analysis: age, gender, education, income and race/ethnicity (Non-Hispanic White, Hispanic, African-American); complete covariate information was required to be included in this analysis. Table II gives a summary of the demographic characteristics of the sample. The ancillary study participants tend to be younger, more educated and more likely to be female.

DOI: 10.1002/sim

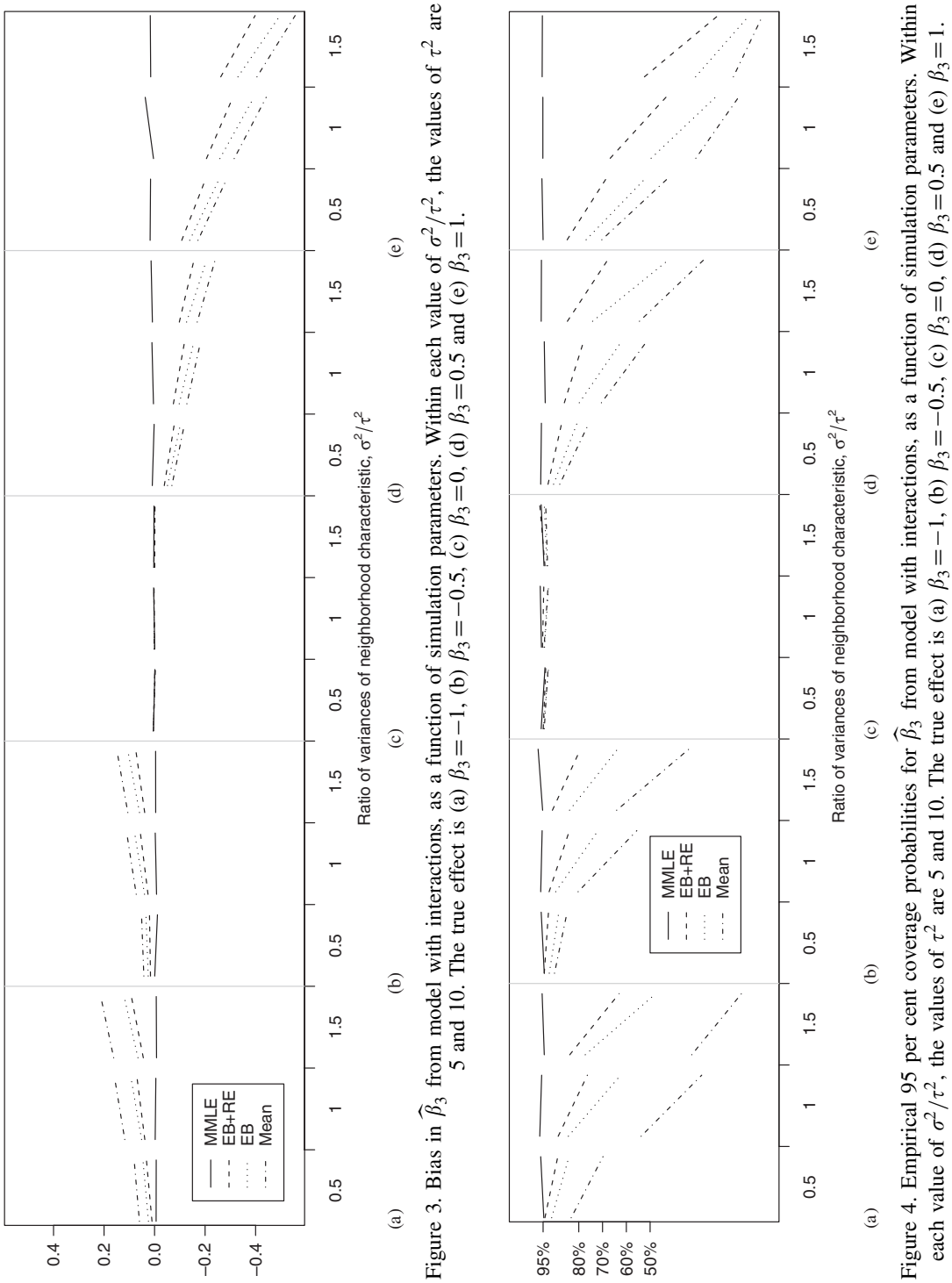

Figure 3. Bias in $\widehat{\beta}_3$ from model with interactions, as a function of simulation parameters. Within each value of σ^2/τ^2, the values of τ^2 are 5 and 10. The true effect is (a) $\beta_3=-1$, (b) $\beta_3=-0.5$, (c) $\beta_3=0$, (d) $\beta_3=0.5$ and (e) $\beta_3=1$.

Figure 4. Empirical 95 per cent coverage probabilities for $\widehat{\beta}_3$ from model with interactions, as a function of simulation parameters. Within each value of σ^2/τ^2, the values of τ^2 are 5 and 10. The true effect is (a) $\beta_3=-1$, (b) $\beta_3=-0.5$, (c) $\beta_3=0$, (d) $\beta_3=0.5$ and (e) $\beta_3=1$.

DOI: 10.1002/sim

Table II. Demographic characteristics for MESA study and Neighborhood Ancillary Study to MESA participants, 2000–2004.

	MESA	Ancillary survey
Hypertension, per cent	50.8	n/a
Gender, per cent male	45.3	35.3
Age, mean (st. dev.)	62.3 (9.98)	46.9 (17.1)
Education category,* per cent		
< H.S. Diploma	14.5	12.3
H.S. Graduate/GED	34.8	17.1
Some college/technical or associate degree	24.7	25.2
College graduate	17.7	23.3
Master degree or beyond	17.8	21.8
Income,† per cent		
$\leqslant$24 999	27.5	25.2
25 000 − $\leqslant$49 999	31.7	24.4
$\geqslant$50 000	40.7	38.2
Race/ethinicity, per cent		
Non-hispanic white	41.6	52.4
African–American	42.3	28.6
Hispanic	16.1	13.2
Other	n/a	5.8

*In the outcome model, education was disaggregated into eight categories ranging from no schooling to graduate or professional school and treated as a continuous variable.

†In the outcome model, income was disaggregated into 13 categories ranging from $\leqslant$\$5000 to $>$\$100 000 and treated as a continuous variable.

Demographic differences between MESA and ancillary survey participants are due to the fact that ancillary survey participants are approximately representative of the neighborhoods in which MESA participants live, whereas MESA had inclusion criteria based on age and race/ethnicity.

Healthy food availability was measured in the ancillary survey by three items scored in a five-point Likert scale ranging from 'strongly disagree' (scored at 1) to 'strongly agree' (scored at 5). The items were (1) a large selection of fresh fruits and vegetables is available in my neighborhood (2) the fresh fruits and vegetables in my neighborhood are of high quality and (3) a large selection of low-fat foods is available in my neighborhood. The responses from each survey participant were averaged to summarize the individual's impression of the availability of healthy foods in his/her neighborhood (i.e. U_{jk} is the average of the three items). Thus, given the coding for the items, a high value of θ_j describes a neighborhood with good availability of healthy foods. Using a mixed model, like (4), the overall average of neighborhood healthy food availability across all sites was 3.4. From the same model, the estimated between-neighborhood variability was $\widehat{\tau}^2 = 0.19$, and the within-neighborhood variability was $\widehat{\sigma}^2 = 0.83$ ($\widehat{\sigma}^2/\widehat{\tau}^2 = 4.34$, intra-cluster correlation 0.19).

Table III gives the log-odds ratios for the age- and gender-adjusted association between neighborhood healthy food availability and hypertension, as estimated by the methods discussed in Section 3. The direction of the magnitude of the coefficients is consistent with the simulation findings: using the mean value of the responses within a neighborhood produces an estimate closest to zero, whereas the MMLE approach produces the largest point estimate. Predictably, given the

DOI: 10.1002/sim

Table III. Association* of neighborhood healthy food availability and hypertension (log-odds scale).

Method	Overall model $\widehat{\beta}_1$ (s.e.)	Model with interaction by gender	
		Male, $\widehat{\beta}_1$ (s.e.)	Female, $\widehat{\beta}_1+\widehat{\beta}_3$ (s.e.)
Plug-in Mean	−0.34 (0.08)	−0.33 (0.13)	−0.36 (0.10)
Plug-in EB	−0.62 (0.12)	−0.54 (0.19)	−0.69 (0.14)
Plug-in EB+RE	−0.63 (0.13)	−0.55 (0.17)	−0.71 (0.15)
MMLE	−0.65 (0.12)	−0.58 (0.17)	−0.72 (0.15)

Note: MESA Study and Neighborhood Ancillary Study to MESA, 2000–2004.
*Adjusted for gender and age only; all coefficients significantly different from zero at $p<0.01$.

Table IV. Association† of neighborhood healthy food availability and hypertension (log-odds scale).

Method	Overall model $\widehat{\beta}_1$ (s.e.)	Model with interaction by gender	
		Male, $\widehat{\beta}_1$ (s.e.)	Female, $\widehat{\beta}_1+\widehat{\beta}_3$ (s.e.)
Plug-in Mean	−0.11 (0.09)	−0.08 (0.11)	−0.13 (0.11)
Plug-in EB	−0.25 (0.13)*	−0.17 (0.19)	−0.33 (0.17)*
Plug-in EB+RE	−0.26 (0.12)*	−0.17 (0.17)	−0.33 (0.16)*
MMLE	−0.26 (0.12)*	−0.17 (0.17)	−0.35 (0.16)*

Note: MESA study and Neighborhood Ancillary Study to MESA, 2000–2004.
*$p<0.1$.
†Adjusted for gender, age, education, income and race/ethnicity.

magnitude of the variance components ($\widehat{\tau}^2=0.19$, $\widehat{\sigma}^2=0.83$) and a small effect size, small differences were found between the EB approaches and the MMLE, with the Plug-in Mean being the most attenuated. The bias of the Plug-in Mean method can be predicted from the linear regression coefficients given in Table I by substituting $\widehat{\tau}^2=0.19$, $\widehat{\sigma}^2=0.83$ and $\widehat{\beta}_1=-0.34$ in the regression model. The predicted bias is 0.34; in other words, a corrected regression coefficient would be −0.68, much closer to the MMLE estimate.

Table IV gives the log-odds ratios estimated after including additional covariates in the model. The estimated association between healthy food availability and hypertension is attenuated for all methods, reflecting confounding by the additional covariates. As seen from the simulation results, the methods differ less when the effect sizes are smaller. Thus, the differences between the EB+RE and the MMLE approaches are less apparent when additional covariates are included.

Several additional analyses were conducted in the example data set to assess the impact of potential differences in the mean and variance of healthy food availability across sites. Indicator variables denoting site were included in (4), and the variance of the neighborhood level measure, τ^2, was allowed to differ by site. There was significant evidence that healthy food availability differed by site ($p=0.001$), with the mean for New York, Maryland and North Carolina being 3.44, 3.28 and 3.24, respectively. There was marginal evidence that the variability differed by site. Instead of an average $\widehat{\tau}^2=0.19$, the site-specific variabilities were 0.24, 0.15 and 0.10 for NY, MD and NC, respectively. The estimated $\widehat{\beta}_1$ coefficients were slightly attenuated after considering site differences (less than 10 per cent attenuation); however, they still showed the same ordering as given in Table III. For example, the magnitude of the difference between the MMLE and the EB+RE

DOI: 10.1002/sim

remained approximately 0.02. Adjusting for mean and variance differences in the neighborhood characteristic did not change the conclusions about the method comparison, but lead to slightly attenuated substantive interpretation of the results. Studies with multiple sites should thoroughly explore how site differences in the mean and variance of neighborhood characteristics may impact their results. The MMLE and EB approaches permit the investigator to incorporate such differences in the modeling strategy, whereas the Plug-in Mean does not.

6. CONCLUSIONS

We evaluated four methods that combine information from two surveys to assess the association of neighborhood-level characteristics with health outcomes. These methods apply to studies in which a primary survey collects data on individual's health and risk factors, while an ancillary study collects neighborhood information from a separate sample of informants. This data collection approach has become more and more common in newly designed studies as a way to link neighborhood-level information to existing health studies.

We compared frequently used 'plug-in' approaches with a novel approach using maximum likelihood estimation. The plug-in approaches consist of summarizing the ancillary study information in a way that an estimate of the neighborhood characteristic of interest can be obtained. For example, a simple mean of the ancillary study responses for each neighborhood or an Empirical Bayes estimate is typically calculated. The estimated characteristic is then substituted in a model for the disease outcome to estimate exposure parameters. In contrast, the maximum likelihood approach simultaneously combines information from both surveys without the need for the intermediate step of summarizing ancillary survey information.

We demonstrated through simulation studies that among the plug-in approaches, the regression coefficient obtained from substituting the simple mean of the responses had the poorest sampling properties. In contrast, the method that plugs in an Empirical Bayes estimate and includes a random effect for neighborhood in a mixed model (EB+RE) provides adequate estimates in certain cases. The maximum likelihood estimate was superior to all plug-in methods. The EB+RE approach may be further improved by removing the simplifying assumption that the random effect is homoschedastic across neighborhoods, i.e. $\xi_j^* \sim \mathrm{N}(0, \delta)$. SAS code to relax such an assumption is provided in Appendix B. However, when we conducted analysis relaxing the homoschedasticity assumption in the example, no changes in the estimated coefficient were observed in comparison with the constant variance EB+RE approach.

In light of the results presented here, we encourage researchers who might still prefer to use the EB+RE method to report the ecometric properties of their measures of neighborhood characteristics (e.g. [4, 5, 9]). Ecometric properties of a scale refer to the validity and reliability of the scale in assessing neighborhood-level properties (as opposed to individual-level properties). Adequate reporting of the ecometric properties of estimated neighborhood characteristics provides a way of evaluating the extent of downward bias in the estimated exposure effects, as in Table I. The bias of the EB+RE method increases when the within-to-between variability ratio in the ancillary responses increases. This bias also increases when the between-neighborhood variability increases for a fixed value of the within-to-between variability ratio. Furthermore, it is important to emphasize that, for reasons detailed more fully below, the variance parameter estimated in the EB+RE method is not an unbiased estimate of residual neighborhood clustering of the disease outcomes after accounting for predictors of interest as is often cited (e.g. [16, 17]).

DOI: 10.1002/sim

For ease of exposition and given limited space, we have focused our discussion to the case where health outcomes and ancillary study responses are independent within and across neighborhood, conditional on neighborhood characteristics. That is, the true outcome model in (2) is a generalized linear model [10], not a generalized linear mixed model [14, 18]. The case when residual correlation between outcomes within neighborhoods is present can be accommodated in the MMLE approach, as detailed in Appendix A. In this case, the MMLE approach can unbiasedly estimate such clustering effect, assuming that the additional clustering is independent of the neighborhood characteristic θ_j. In contrast, the EB + RE method cannot distinguish between the variance component in the outcome that is due to neighborhood clustering in the outcome and the component of correlation among the outcomes that is due to measurement error of the estimated neighborhood characteristic (as in (5)). Expanding the fitting procedures to enable estimation of a longitudinal outcome model and time changing covariates may not be straightforward, particularly in software such as SAS Proc NLMIXED, but may be feasible in other software. Similar to the outcome model, the model for the ancillary responses (1) is a linear model without further levels of nesting. Extending this part of the model would be easily feasible by including further random effects into model (1). However, depending on the number and complexity of the levels of nesting, fitting such a model may no longer be feasible in a package such as SAS Proc NLMIXED. Other SAS procedures, such as the macro NLINMIX, or other software such as WINBUGS may more easily accommodate such extensions.

The discussion was also focused to the case of cross-sectional, univariate ancillary responses U_{jk}. In the example, U_{jk} was the average of three items from the ancillary survey; all items scored 1 to 5. In essence, the ancillary response data were collected as a multivariate response, which we reduced to an average. Averaging the items inherently assumes that they measure the construct in the same scale. That is, that a 1 unit increase in the response to one of the items is approximately the same as a 1 unit increase in all other items. This assumption may not always be tenable. In such a case, multivariate extensions of (1) would be required, for example, factor analysis. Extensions of (1) would also be necessary for cases when the ancillary survey is conducted at more than one point in time. Such extensions would allow modeling of changes in the neighborhood characteristic with time.

In this presentation of the MMLE approach, we do not employ survey weights. MESA participants were sampled at each site using a variety of approaches including random digit dialing and community lists [8]. Hence, there are no sampling weights for the MESA study because the sample is not specifically representative of any well-defined finite population. The weights for the ancillary neighborhood survey arise from post-stratification to the collection of all census tracts covered by the MESA community. As such, using post-stratification weights would smoothen the healthy food availability scale over the collection of census tracts and would have biased the neighborhood-specific estimates toward the common mean. Instead, the correct weights to be used in this analysis would be based on the 'within-neighborhood' selection probabilities. These within-neighborhood probabilities are roughly constant, however, such that weighted and unweighted analyses give similar results. Furthermore, the weights are not correlated with the healthy food availability scale, and therefore there is not much gain in using the weights as covariates either [19].

The results reported here can be compared and contrasted with the measurement error literature [20]. For example, the sampling variability in the Empirical Bayes estimate can be viewed as measuring the neighborhood characteristic with error. However, an important difference is that in the measurement error literature, it is typically assumed that the error of estimation has a much smaller variance when compared with the population variance of the true quantity; in

DOI: 10.1002/sim

other words, $(\sigma^2/m_j)/\tau^2$ is small. This scenario of variance components may not generally be true in epidemiological investigations combining data from two surveys since the heterogeneity of responses within neighborhoods is often much larger than the variance across neighborhoods. The within-neighborhood error variance, $(\sigma^2/m_j)/\tau^2$, can be reduced by increasing the number of respondents within a neighborhood; however this may incur in higher study costs.

Careful design of ancillary surveys is warranted to maximize the amount of data available when combining surveys. For example, the number of neighborhoods with MESA participants contributing health outcome data was 496, of which 436 had overlap with the ancillary survey. Thus, there were 60 excluded census tracts with primary study participants due to missing healthy food availability information. Assuming that these tracts do not systematically differ from the census tracts included, then the consistency of the presented results is not affected. Although issues such as non-response cannot always be avoided, sampling efforts for the ancillary survey should be focused on obtaining data in all neighborhoods represented in the primary survey.

The informativeness of the ancillary survey and its utility in validly estimating neighborhood-level properties also rely its design. An important point is that if the purpose of the ancillary survey is to objectively assess neighborhood characteristics, then the ancillary sample's demographics need not be similar to (or representative of) the sample to which it will be linked and for which health outcomes will be assessed. This is because the point of the ancillary study would be to characterize neighborhood characteristics as objectively as possible rather than to obtain an estimate of what the health study sample would have reported had they been surveyed.

Although the advantages of the MMLE approach vs the plug-in approaches are many, it is not without limitations. One drawback of this approach is its potential susceptibility to convergence problems given a larger number of parameters estimated jointly. These problems are due to the numerical routines used to implement the method. In our simulations we found that convergence issues were more likely to emerge when the scales of the parameters were disparate (e.g. small β relative to σ). One solution was to standardize the ancillary survey responses by a fixed number, estimating the parameters, and then transforming to the original scale. Another potential difficulty is the susceptibility of the methods to distributional assumptions (e.g. violation of normality of the neighborhood characteristic). However, the EB approaches would be similarly affected, as they too rely on the normality assumptions of the random effects in estimating the EB estimate of the neighborhood characteristic. To shed light in this issue, we conducted a few simulations where the neighborhood characteristic had a chi-squared distribution with three degrees of freedom. When β_1 was zero, all methods remained unbiased. However, for non-zero β_1, we found that all methods yielded attenuated regression coefficients, and the coverage probabilities were diminished. However, the MMLE approach performed equally well to, or slightly better than, its closest competitor. For example, with parameters $\beta_1=1$, $\tau^2=1$, $\sigma^2/\tau^2=2$, the biases were $-0.027, -0.031, -0.045, -0.42$, and the empirical 95 per cent coverage probabilities were $0.92, 0.92, 0.65$ and 0 for the MMLE, EB+RE, EB and Plug-in Mean approaches, respectively. Similarly, with parameters $\beta_1=1$, $\tau^2=5$, $\sigma^2/\tau^2=4$, the biases were, respectively, $-0.033, -0.041, -0.24, -0.78$, and the empirical 95 per cent coverage probabilities were $0.96, 0.84, 0.76$ and 0. Furthermore, upon diagnosing the lack of normality in the neighborhood characteristic, for example, with a Q–Q plot of the EB estimates obtained from Proc MIXED, the MMLE approach could be easily extended to include a different distribution for the random effects.

We contrasted available methodologies used to combine information from two surveys for studies that aim at estimating the health effects of neighborhood characteristics. Additionally, we provided a new tool that uses maximum likelihood estimation to combine the information from both surveys

DOI: 10.1002/sim

and improves upon available methodologies to unbiasedly estimate such health effects. Further, we provided prototype code for the implementation of the maximum likelihood approach. Our methods are applicable not only to studies of health effects but also to other situations where a construct is assessed by combining the reports of multiple informants or raters in order to investigate the effects of that characteristic on individual-level outcomes.

APPENDIX A: BASIC DATA SETUP AND PROGRAMMING STATEMENTS FOR MMLE APPROACH

We briefly describe how the data for the primary and ancillary studies should be merged, and subsequently sketch the programming statements in SAS Proc NLMIXED. The data sets should be staked, containing one row per study participant, as shown in Figure A1. The data from the primary study are listed first as marked by the primary study indicator variable (`indic` is 1 for the primary study and 0 for the auxiliary study). The participant's neighborhood identifier is labeled `id`. Health outcomes and covariates for the primary study participants are stored in the columns `Yji`, `Xji`. Note that these columns are zero for the ancillary study participants. The column `Ujk` contains the responses from the ancillary study participants. In the example of Section 5, this column would contain the average of the three items for food availability. Again note that the values of this variable are set to zero for the primary study participants. Lastly, the column `Zj` contains other neighborhood-level covariates.

The SAS Proc NLMIXED programming statements below assume the model outlined in equations (1)–(4), assuming a logistic link for the outcome model. These statements make use of starting values, which could be obtained by first fitting the model with the EB+RE method.

`id` Neigh-Level identifier	`indic` Primary study indicator	`Yji` Health Outcome, Y_{ji}	`Xji` Indiv-level covariates, X_{ji}	`Ujk` Ancillary Study Responses, U_{jk}	`Wjk` Ancillary Study covariates, W_{jk}	`Zj` Neigh-level Covariates, Z_j
1	1	y_{11}	x_{11}	0	0	Z_1
1	1	y_{12}	x_{12}	0	0	Z_1
⋮	⋮	⋮	⋮		⋮	⋮
1	1	y_{1n_1}	x_{1n_1}	0	0	Z_1
⋮	⋮	⋮	⋮	⋮	⋮	
J	1	y_{J1}	x_{J1}	0	0	Z_J
J	1	y_{J2}	x_{J2}	0	0	Z_J
⋮	⋮	⋮	⋮	⋮	⋮	⋮
J	1	y_{Jn_J}	x_{Jn_J}	0	0	Z_J
1	0	0	0	u_{11}	w_{11}	Z_1
1	0	0	0	u_{12}	w_{12}	Z_1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	0	0	u_{1m_1}	w_{1m_1}	Z_1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
J	0	0	0	u_{J1}	w_{J1}	Z_J
J	0	0	0	u_{J2}	w_{J2}	Z_J
⋮	⋮	⋮	⋮	⋮	⋮	⋮
J	0	0	0	u_{Jm_J}	w_{Jm_J}	Z_J

Figure A1. Data setup for MMLE approach.

DOI: 10.1002/sim

```
proc nlmixed data = combined-dataset  options;
   parms starting-values;
   eta = B0 +  B1 * thetaj +  B2 * Xji + other covariates;
   prob = exp(eta) / (1+exp(eta));
   loglikMain = Yij*prob + (1 - Yij)*(1 - prob);
   loglikAncillary = (-1/2) * ((Ujk - thetaj)/sigma)**2 - log(sigma);
   model Yij ~ general( indic * loglikMain + ( 1 - indic ) * loglikAncillary );
   zgamma = C0 + C1*Zj ;
   random thetaj ~ normal( zgamma , tau**2 ) subject= id ;
run;
```

The model could be easily extended to a model with additional within-neighborhood clustering in the outcomes by including an additional random intercept for neighborhood (e.g. $a_j \sim \mathrm{N}(0, \omega^2)$. This would entail replacing the `eta` and `random` statements:

```
eta = B0 +  B1 * thetaj +  B2 * Xji + other covariates + aj;
random thetaj aj ~ normal([zgamma, 0] , [tau**2, 0, omega**2] ) subject= id ;
```

One may further wish to extend the model to adjust the ancillary study responses by characteristics of the respondents, i.e. $U_{jk}|\theta_j, W_{jk} \sim$ iid $\mathrm{N}(\theta_j + \lambda W_{jk}, \sigma^2)$. This would be achieved with an additional statement and a modification to the `loglikAncillary` statement:

```
   meanUjk = thetaj + L1 * Wjk;
   loglikAncillary = (-1/2) * ((Ujk - meanUjk)/sigma)**2 - log(sigma);
```

Finally, the model can be extended to allow for a small number of site-specific variances of the neighborhood characteristic. In the MESA example, three sites could be modeled as

```
random thetaj ~ normal( zgamma ,
                        site1*tau1**2 + site2*tau3**2 + site3*tau3**2 ) subject= id ;
```

where the `site` variables are 0, 1 indicators for site.

APPENDIX B: PROGRAMMING STATEMENTS FOR EB+RE APPROACH

The SAS Proc NLMIXED programming statements below can be used to estimate the EB+RE approach, assuming a logistic link for the outcome model. These statements make use of starting values, which could be obtained by first fitting the model with the EB method.

```
proc nlmixed data = combined-dataset(where=(indic=1))  options;
   parms starting-values;
   eta = B0 +  B1 * EBestimate +  B2 * Xji + other covariates;
   prob = exp(eta) / (1+exp(eta));
   model Yij ~ binomial(1,prob);
   random boj ~ normal(0,deltaJ) subject=id;
run;
```

In this code, the variable `EBestimate` is the Empirical Bayes estimate for each neighborhood characteristic, $\widehat{\theta}_j$, obtained from, for example, the random effects solution in SAS Proc MIXED.

The EB+RE approach can be extended to model neighborhood-specific variances for the random effect. The neighborhood-specific variances would be used to incorporate varying degrees of measurement error in the Empirical Bayes estimate for each neighborhood characteristic due to

DOI: 10.1002/sim

differing ancillary study sample size in each neighborhood. The MMLE automatically incorporates such differences in sample sizes such that extending the MMLE to this scenario is not necessary. For this extension of the EB+RE method, the following would be added to the above code:

```
deltaJ = (delta1* StdErrPred)**2 ;
```

where `StdErrPred` is the standard error of prediction for the Empirical Bayes estimate for each neighborhood characteristic obtained from, for example, the random effects solution in SAS Proc MIXED.

Similar to the MMLE extension in Appendix A, which allows us to model site-specific variances, the EB+RE approach could also be extended by modifying the definition of `deltaJ` to

```
deltaJ = site1*(delta1* StdErrPred)**2 + site2*(delta2* StdErrPred)**2
          + site3*(delta3* StdErrPred)**2;
```

where the `site` variables are 0, 1 indicators for site.

ACKNOWLEDGEMENTS

MESA is supported by contracts N01-HC-95159 through N01-HC-95166 from the National Heart, Lung, and Blood Institute (NHLBI). This work was also supported in part by grant R01 HL071759 (Diez Roux, PI) from NHLBI. We thank the other investigators, staff and participants of MESA for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.

REFERENCES

1. Sundquist K, Winkleby M, Ahlen H, Johansson SE. Neighborhood socioeconomic environment and incidence of coronary heart disease: a follow-up study of 25,319 women and men in Sweden. *American Journal of Epidemiology* 2004; **159**:655–662. DOI: 10.1093/aje/kwh096.
2. Diez Roux AV, Merkin SS, Arnett D, Chambless L, Massing M, Nieto FJ, Sorlie P, Szklo M, Tyroler HA, Watson RL. Neighborhood of residence and incidence of coronary heart disease. *New England Journal of Medicine* 2001; **345**:99–106.
3. Chaix B, Rosvall M, Merlo J. Neighborhood socioeconomic deprivation and residential instability: effects on incidence of ischemic heart disease and survival after myocardial infarction. *Epidemiology* 2007; **18**(1):104–111.
4. Raudenbush SW, Sampson RJ. Ecometrics: toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology* 1999; **29**:1–41.
5. Gauvin L, Richard L, Craig CL, Spivock M, Riva M, Forster M, Laforest S, Laberge S, Fournel MC, Gagnon H, Gagné S, Potvin L. From walkability to active living potential: an 'ecometric' validation study. *American Journal of Preventative Medicine* 2005; **28**:126–133.
6. Gauvin L, Riva M, Barnett T, Richard L, Craig CL, Spivock M, Laforest S, Laberge S, Fournel MC, Gagnon H, Gagné S. Association between neighborhood active living potential and walking. *American Journal of Epidemiology* 2008; **167**(8):944–953. DOI: 10.1093/aje/kwm391.
7. Cohen DA, Finch BK, Bower A, Sastry N. Collective efficacy and obesity: the potential influence of social factors on health. *Social Science and Medicine* 2006; **62**(3):769–778.
8. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacob Jr DR, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology* 2002; **156**:871–881.
9. Mujahid MS, Diez Roux AV, Morenoff JD, Raghunathan T. Assessing the measurement properties of neighborhood scales: from psychometrics to ecometrics. *American Journal of Epidemiology* 2007; **165**:858–867.
10. McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman & Hall: New York, 1989.
11. SAS Institute Inc. *Base SAS 9.1.3 Procedures Guide* (2nd edn). SAS Institute Inc.: Cary, NC, 2006.
12. Raudenbush SW, Bryk A, Cheong YF, Congdon RT. *HLM 6*: *Hierarchical Linear and Nonlinear Modeling*. Scientific Software International: Chicago, IL, 2004.

13. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
14. McCulloch CE, Searle SR. *Generalized, Linear, and Mixed Models*. Wiley: New York, 2001.
15. JNC-VI. The sixth report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. *Archives of Internal Medicine* 1997; **157**(21):2413–2446.
16. Merlo J, Lynch JW, Yang M, Lindström M, Ostergren PO, Rasmusen NK, Rästam L. Effect of neighborhood social participation on individual use of hormone replacement therapy and antihypertensive medication: a multilevel analysis. *American Journal of Epidemiology* 2003; **157**:774–783.
17. McKay CA, Bell-Ellison BA, Wallace K, Ferron JM. A multilevel study of the associations between economic and social context, stage of adolescence, and physical activity and body mass index. *Pediatrics* 2007; **119**:S84–S91.
18. Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. *Biometrics* 1990; **46**: 673–687.
19. Little RJ, Vartivarian S. On weighting the rates in non-response weights. *Statistics in Medicine* 2003; **22**: 1589–1599.
20. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C. *Measurement Error in Nonlinear Models*: *A Modern Perspective* (2nd edn). CRC Press: New York, 2006.

DOI: 10.1002/sim