The International Journal of Biostatistics

Volume 3, Issue 1

2007

Article 10

An Integrated Analysis of Individual and Aggregated Health Data Using Estimating Equations

Jose Miguel Martinez, *Pompeu Fabra University*

Joan Benach, *Pompeu Fabra University*

Josep Ginebra, *Technical University of Catalonia*

Fernando G. Benavides, *Pompeu Fabra University*

Yutaka Yasui, *University of Alberta (Corresponding
Author)*

Recommended Citation:

Martinez, Jose Miguel; Benach, Joan; Ginebra, Josep; G. Benavides, Fernando; and Yasui, Yutaka (2007) "An Integrated Analysis of Individual and Aggregated Health Data Using Estimating Equations," *The International Journal of Biostatistics*: Vol. 3: Iss. 1, Article 10.

DOI: 10.2202/1557-4679.1060

An Integrated Analysis of Individual and Aggregated Health Data Using Estimating Equations

Jose Miguel Martinez, Joan Benach, Josep Ginebra, Fernando G. Benavides, and Yutaka Yasui

Abstract

Analyses of individual disease-exposure data within a population are useful when exposure of interest varies sufficiently within the population. When the within-population variance of exposure is limited, however, power of the individual-data analysis is reduced. In such situations, aggregated-data analyses of disease data across populations, with a sample of individual exposure data from each population, can be powerful in estimating the exposure effect if between-population variation of exposure is large. In this paper, we consider a new analytical framework that is a combination of the individual- and aggregated-data analyses, based on an estimating equation approach. The proposed analysis utilizes strengths from individual data and aggregated data in the estimation of the exposure effect of interest, depending on which of the exposure variations (within- versus between-population) dominates. Simulation studies under various different scenarios were performed to show the strengths of the proposed approach in the estimation of the exposure effects of interest.

KEYWORDS: aggregated data, cohort, random effects, small areas

Author Notes: This work was partially funded by a grant from the Fondo de Investigaciones Sanitarias (FIS 03/0586) and the BBVA Foundation. The Occupational Health Research Unit was recognized as a standing research group by the Government of Catalonia (2002-2004: SGR/0005; 2005-2008: SGR/00699). YY is supported by the Canada Research Chair Program and Alberta Heritage Foundation for Medical Research.

1. INTRODUCTION

Analyses of individual disease-exposure data within a population are useful when the exposure of interest varies sufficiently within the population. When the within-population variance of exposure is limited, however, power of the individual-data analysis within a population is reduced. In such situations, aggregated-data analyses of disease data across populations, with a sample of individual exposure data from populations, can be powerful in estimating the exposure effect if between-population variation of the exposure is large (Prentice and Sheppard, 1995; Sheppard and Prentice, 1995). Both approaches are useful depending on where the exposure variation exists.

Observational studies usually examine multiple covariates (exposures and confounding variables) that can have different types of variation. Some covariates may have a high within-population variability and others may have a high between-population variability. An individual-data analysis approach may perform poorly on the estimation of the covariables with high between-population variability because, if there is substantial within-group clustering in the risk of the outcome, the effective individual sample size is lower. In other words, if the within-group variability is low compared to the between-group variability, the power of individual data analysis is reduced. On the other hand, an aggregated-data analysis approach may perform poorly on the estimation of the covariates with high within-population variability. In addition, the individual data analysis may perform poorly in estimating the exposure effect even if the exposure is subject to high within-population variability. This may happen if the effect of a confounding variable, which is by definition related to the exposure of interest (Szklo and Nieto, 2000), is poorly adjusted for due to its high between-population variability. Similarly, the aggregated-data analysis approach could perform poorly in estimating exposure effects even if the exposure is subject to high between-population variability, due to the influence of a confounding variable with high within-population variability.

Motivated by these considerations, we consider a new analytical framework that combines the individual- and aggregated-data analyses by extending the estimating equation approach for aggregated-data analyses of Prentice and Sheppard (Prentice and Sheppard; Sheppard and Prentice) to the combination of the individual- and aggregated-data analyses. The proposed analysis utilizes strengths from individual data and aggregated data in estimating the exposure effect of interest, depending on which of the exposure variations (within- versus between-population) dominates. This approach can be useful in epidemiological observational studies where we include exposure and confounding variables that can have different source of within- and between-population variability. For example, in the study of bladder cancer etiology we

can jointly include variables where the within-population variability is higher than the between-population variation, such as smoking status, and variables where the between-population variation can be higher than the within-population, such as chlorinated drinking water (Villanueva et al., 2003).

Utilizing both types of data has been proposed under the fully Bayesian framework by Jackson et al. (Jackson et al., 2006). Our proposal follows the same basic concept, but extends the estimating equation approach that Prentice and Sheppard proposed for aggregated-data analyses.

In Section 2, we explain the study design and data structure of the proposed analytical framework. Following a brief review of the individual- and aggregated-data random-effects models (IRM and ARM), we describe their combination, a “population-based estimating equation” (PBEE) approach, in Section 3. In Section 4 we describe a simulation study that illustrates advantages of the PBEE over individual- and aggregated-data analyses. Finally, we summarize discussions in Section 5.

2. STUDY DESIGN

We consider a study design in which 1) aggregated data on a disease or mortality outcome are available on K populations with population sizes n_k ($k=1,\dots,K$); and 2) individual data for a random sample of m_k individuals ($m_k \leq n_k$) in each of K populations are collected. A diagram of the data structure is given in Figure 1.

Population 1	Population k	Population K
<i>Population-level data</i> $[Y_1, n_1, Z_1]$	<i>Population-level data</i> $[Y_k, n_k, Z_k]$	<i>Population-level data</i> $[Y_K, n_K, Z_K]$
<i>Individual-level data</i> $[(Y_{11}, z_{11}), \dots, (Y_{1m_1}, z_{1m_1})]$	<i>Individual-level data</i> $[(Y_{k1}, z_{k1}), \dots, (Y_{km_k}, z_{km_k})]$	<i>Individual-level data</i> $[(Y_{K1}, z_{K1}), \dots, (Y_{Km_K}, z_{Km_K})]$

Figure 1. Data structure diagram.

We denote the disease outcome variable of the i th individual in the k th population as Y_{ki} and a p times 1 vector of covariates associated as $z_{ki} = (z_{ki}^{(1)}, \dots, z_{ki}^{(p)})^T$. The variable Y_{ki} takes a value of 1 if the outcome that is of interest (disease or death) occurs for the individual within the defined study follow-up period, and 0 otherwise. In each population's aggregate data, we have the total number of disease cases, Y_k , and the total number of individuals at risk, n_k , during the study period, and possibly a j times 1 vector of population-level covariates $Z_k = (Z_k^{(1)}, \dots, Z_k^{(j)})^T$. These aggregated data are often available from and published periodically by governmental agencies.

3. RELATIVE RATE ANALYSIS OF INDIVIDUAL AND AGGREGATED DATA

This section briefly reviews the individual- and aggregated-data random-effects models following Prentice and Sheppard's work for aggregated-data analyses and proposes the use of both individual- and aggregated-data under a single estimating equation.

3.1 The individual data random effects model

Let p_{ki} denote the probability that the i th individual, $i=1, \dots, m_k$, in the k th population, $k=1, \dots, K$, develops a certain disease within a defined follow-up period. We consider a relative rate model:

$$p_{ki} = p_{k0} e^{z_{ki}^T \beta}$$

where p_{k0} is a 'baseline' disease probability for the k th population corresponding to $z_{ki} = 0$ and β is a p times 1 parameter vector to be estimated. The random effects assumption gives the following:

$$p_{k0} = h_k e^{\gamma_0}$$

where e^{γ_0} denotes the expected baseline rate and h_k denotes the residual baseline rate or 'frailty' of the k th population. We consider that the h_k 's are independent random effects with mean 1 and variance σ^2 . Using the random effects assumption, the model can be written as

$$p_{ki} = E(Y_{ki} | h_k) = h_k e^{x_{ki}^T \alpha},$$

where $x_{ki} = (1, z_{ki})^T$ and $\alpha = (\gamma_0, \beta^T)^T$.

3.2 The aggregated data random effects model

An aggregated-data model can be induced from the random effects model for individual data by averaging $h_k e^{x_{ki}^T \alpha}$ over the m_k individuals in the sample within the k th population, and considering the average disease probability $\bar{p}_k$ of the population among the n_k individuals in the population:

$$\bar{p}_k = E(Y_k / n_k | h_k) = h_k \varepsilon_{m_k} \{ e^{x_k^T \alpha} \}$$

where $\varepsilon_{m_k} \{ a_k \} = m_k^{-1} (a_{k1} + \dots + a_{km_k})$ denotes the average of the argument over the m_k individuals in the k th population following the previous notation of Prentice and Sheppard. Note that the left-hand side of the model equation is the average based on the aggregated health data in the entire population (i.e., n_k individuals), while the right-hand side is the average based on the individual covariate data in the sample (i.e., m_k individuals). As Prentice and Sheppard considered in their previous research, this model assumes that the individual covariates are random samples of a sufficient size from the population.

3.3 Relative rate inference with population based estimating equations

To utilize the entire data for parameter estimation under the Figure 1 study design, we propose to combine estimating equations for the individual- and aggregated-data analyses into one equation. Using $Y_k^I = (Y_{k1}, \dots, Y_{km_k})^T$ and $\mu_k^I = (\mu_{k1}, \dots, \mu_{km_k})^T$ with $\mu_{ki} = E(Y_{ki}) = E(E(Y_{ki} | h_k)) = e^{x_{ki}^T \alpha}$, the individual-data estimating equation is

$$\sum_{k=1}^K (D_k^I)^T (V_k^I)^{-1} (Y_k^I - \mu_k^I) = 0,$$

where $D_k^I = \partial \mu_k^I / \partial \alpha^T$, $V_k^I = \Delta_k + \sigma^2 \mu_k^I (\mu_k^I)^T$ with $\Delta_k = \text{diag}[\mu_{ki} \{1 - (1 + \sigma^2) \mu_{ki}\}]$. The inverse of the variance-covariance matrix $(V_k^I)^{-1}$ can be computed (Sheppard and Prentice) by

$$(V_k^I)^{-1} = (\Delta_k)^{-1} - \sigma^2 (\Delta_k)^{-1} \mu_k^I (\mu_k^I)^T (\Delta_k)^{-1} \{1 + \sigma^2 (\mu_k^I)^T (\Delta_k)^{-1} \mu_k^I\}^{-1}.$$

By excluding the individual data from the aggregated data in each population, that is, considering $\bar{Y}_k^A = (n_k - m_k)^{-1} (Y_k - \sum_{i=1}^{m_k} Y_{ki})$ and $\hat{\mu}_k^A = E(\bar{Y}_k^A) = E(E(\bar{Y}_k^A | h_k)) = \varepsilon_{m_k} \{e^{x_k^T \alpha}\}$, where we assume that the m_k -individual samples are random samples from the population and that, consequently, they are representative of the $(n_k - m_k)$ individuals in the same population, the aggregated-data estimating equation is

$$\sum_{k=1}^K (\hat{D}_k^A)^T (\hat{V}_k^A)^{-1} (\bar{Y}_k^A - \hat{\mu}_k^A) = 0,$$

where

$$\hat{D}_k^A = \partial \hat{\mu}_k^A / \partial \alpha^T$$

and

$$\hat{V}_k^A = \sigma^2 \{(\hat{\mu}_k^A)^2 - \hat{\phi}_k (n_k - m_k)^{-1}\} + (\hat{\mu}_k^A - \hat{\phi}_k)(n_k - m_k)^{-1}$$

with $\hat{\phi}_k = \varepsilon_{m_k} \{e^{2x_k^T \alpha}\}$ (Prentice and Sheppard; Sheppard and Prentice). Note that we do not observe the individual covariates of every individual in the population, except those in the m_k -individual sample. Therefore, we estimate the average values $\mu_k^A = \varepsilon_{n_k - m_k} \{e^{x_k^T \alpha}\}$ and $\phi_k = \varepsilon_{n_k - m_k} \{e^{2x_k^T \alpha}\}$ with the average values in the sample, $\hat{\mu}_k^A$ and $\hat{\phi}_k$, respectively.

The two estimating equations above can be combined to utilize both the individual and aggregate components of the entire data:

$$\sum_{k=1}^K \begin{pmatrix} D_k^I \\ \hat{D}_k^A \end{pmatrix}^T \begin{pmatrix} V_k^I & 0 \\ 0 & \hat{V}_k^A \end{pmatrix}^{-1} \begin{pmatrix} Y_k^I - \mu_k^I \\ \bar{Y}_k^A - \hat{\mu}_k^A \end{pmatrix} = 0.$$

Note that we propose a simple addition of the two estimating equations. The combined estimating equation deviates slightly from the optimal linear estimating function (Godambe and Thompson, 1989) of the form,

$(\partial E[Y]/\partial \alpha^T)^T (Var[Y])^{-1} (Y - E[Y])$. Specifically, $\begin{pmatrix} V_k^I & 0 \\ 0 & \hat{V}_k^A \end{pmatrix}$ is not the

variance-covariance matrix of $\begin{pmatrix} Y_k^I - \mu_k^I \\ \bar{Y}_k^A - \hat{\mu}_k^A \end{pmatrix}$; $(Y_k^I - \mu_k^I)$, $(\bar{Y}_k^A - \hat{\mu}_k^A)$ are

correlated, and $\hat{\mu}_k^A$ has a sampling variation that is unaccounted for in $\hat{V}_k^A$. As these second-order assumptions are difficult to verify, we propose keeping the “weights” of the combined estimating function as a correspondence to a simple sum of the two estimating equations, and we use a robust-sandwich variance estimator (Royall, 1986) of $\hat{\alpha}$ that reflects empirical second-order characteristics of $\begin{pmatrix} Y_k^I - \mu_k^I \\ \bar{Y}_k^A - \hat{\mu}_k^A \end{pmatrix}$. This is in the spirit of Prentice and Sheppard and Liang and

Zeger (Liang and Zeger, 1986; Zeger and Liang, 1986) in their use of a robust-sandwich variance estimator of mean parameters.

Statistical inference on α can generally be based on the asymptotic normality of $\hat{\alpha}$, whose variance can be estimated consistently by the following robust-sandwich variance estimator:

$$(I_\alpha)^{-1} \left[\sum_{k=1}^K \begin{pmatrix} D_k^I \\ \hat{D}_k^A \end{pmatrix}^T \begin{pmatrix} V_k^I & 0 \\ 0 & \hat{V}_k^A \end{pmatrix}^{-1} \begin{pmatrix} Y_k^I - \mu_k^I \\ \bar{Y}_k^A - \hat{\mu}_k^A \end{pmatrix} \begin{pmatrix} Y_k^I - \mu_k^I \\ \bar{Y}_k^A - \hat{\mu}_k^A \end{pmatrix}^T \begin{pmatrix} V_k^I & 0 \\ 0 & \hat{V}_k^A \end{pmatrix}^{-1} \begin{pmatrix} D_k^I \\ \hat{D}_k^A \end{pmatrix} \right] (I_\alpha)^{-1},$$

where

$$I_\alpha = \sum_{k=1}^K \begin{pmatrix} D_k^I \\ \hat{D}_k^A \end{pmatrix}^T \begin{pmatrix} V_k^I & 0 \\ 0 & \hat{V}_k^A \end{pmatrix}^{-1} \begin{pmatrix} D_k^I \\ \hat{D}_k^A \end{pmatrix}$$

with all quantities evaluated at $\hat{\alpha}$. As in Prentice and Sheppard, the estimation procedure is completed by inserting a $K^{1/2}$ -consistent estimator for σ^2 . Such estimators are given by moment estimators defined for the individual and aggregated components of the estimating equation, respectively:

$$\begin{aligned}
(\hat{\sigma}^2)_I &= \frac{1}{K} \sum_{k=1}^K \left(\left[\varepsilon_{m_k} \{Y_k\} (\varepsilon_{m_k} \{Y_k\} m_k - 2\hat{\mu}_k^A m_k - 1) + 2\varepsilon_{m_k} \{\mu_k Y_k\} \right] \left[(\hat{\mu}_k^A)^2 m_k - \hat{\phi}_k \right]^{-1} + 1 \right) \\
(\hat{\sigma}^2)_A &= \frac{1}{K} \sum_{k=1}^K \left[(\bar{Y}_k^A - \hat{\mu}_k^A)^2 - (\hat{\mu}_k^A - \hat{\phi}_k)(n_k - m_k)^{-1} \right] \left[(\hat{\mu}_k^A)^2 - \hat{\phi}_k(n_k - m_k)^{-1} \right]^{-1}.
\end{aligned}$$

Recall that $\hat{\mu}_k^A$ is based on the m_k individual samples, even though it includes a superscript ‘‘A.’’ We do not unify the two estimators of σ^2 in line with the idea of combining two estimating equations into one. The estimating equation can be solved through the Newton-Raphson procedures.

4. SIMULATION DESIGN AND EFFICIENCY COMPARISON

A simulation study was conducted to compare the inferential performance of three approaches (IRM, ARM and PBEE) described in Section 3. We considered four sample-size scenarios depending on the number of populations, K , and the sample size in each population, m_k : $(K, m_k) = (100, 100)$, $(100, 50)$, $(50, 100)$, and $(50, 50)$.

The population size, n_k , in each population was fixed at 2,000. Note that these samples sizes are realistic for surveys of randomly sampled individuals that cover a given geographical area (e.g., a cross-sectional national health survey), which can be analyzed with aggregated data (e.g., census data) using an integrated analysis method. Two covariates, denoted as $z_k^{(1)}$ and $z_k^{(2)}$, were generated, where $z_k^{(1)}$ represents exposure of interest and $z_k^{(2)}$ represents a confounding factor. For the i th person in the k th population, we generate individual covariate values $(z_{ki}^{(1)}, z_{ki}^{(2)})$ from a bivariate normal distribution with population mean

$(M_k^{(1)}, M_k^{(2)})$ and variance-covariance matrix $\begin{pmatrix} \sigma_w^2 & \theta_l \sigma_w \\ \theta_l \sigma_w & 1 \end{pmatrix}$, where the

population mean $(M_k^{(1)}, M_k^{(2)})$ is also a bivariate normal random vector with mean

$(0, 0)$ and variance-covariance $\begin{pmatrix} 1 & \theta_p \\ \theta_p & 1 \end{pmatrix}$. To consider a range of exposure

variance in within- and between-populations, the within-population variance σ_w^2 of exposure were set at 0.25, 0.5, 1, 2, 4 or 16 but the between-population variance of exposure was fixed at 1. This covers a wide range of the within-versus between-variance ratio of exposure from 0.25/1 where the between-population variance dominates the within-population variance, to 16/1 where the within-population variance dominates the between-population variance. We

considered $(\theta_I, \theta_P) = (0,0)$ for a case without confounding, in other words a “no confounding case” (NCC), and $(0.3,0.3)$ for a simple confounding case (SCC) by $z_k^{(2)}$.

We also considered an extended confounding case (ECC) with $(\theta_I, \theta_P) = (0.3,0.3)$, where the within- versus between-population variance ratio for the two covariates, $z_k^{(1)}$ and $z_k^{(2)}$, were above and below 1, respectively, i.e., $z_k^{(1)}$'s within- and between-population variances are 0.85 and 3.4, respectively, and $z_k^{(2)}$'s within- and between-population variances are 4 and 0.25, respectively. In this way, the two covariates have the same total variance, but $z_{ki}^{(1)}$ has the variance ratio of 0.25 indicating that the between-population variance dominates this covariate's variability and that $z_{ki}^{(2)}$ has the variance ratio of 16 indicating that the within-population variance dominates.

With the NCC and SCC, we can study situations with various within- and between-population variabilities of an exposure variable; in addition we can have, respectively, a no confounding and a confounding variable without changing their within- and between-population variances. In this way we can assess the hypothesized patterns of performance: 1) the IRM may perform well when the exposure of interest has a high within-population variability and poorly when it has a high between-population variability; 2) the ARM may perform well when the exposure of interest has a high between-population variability and poorly when it has a high within-population variability; and 3) the PBEE approach may perform relatively well in both situations.

With the ECC, we can also evaluate the hypothesized patterns of performance when the exposure and confounding variables have dissimilar within- and between-population variabilities: 1) the IRM may perform poorly in estimating the exposure effect in spite of the exposure's high within-population variability, due to the influence of a confounding variable may not be estimated well because of its high between-population variability; 2) the ARM may perform poorly in estimating the exposure effect in spite of the exposure's high between-population variability due to the influence of a confounding variable that may not be estimated well because of its high within-population variability; and 3) the PBEE approach may perform relatively well in both situations.

Population-specific frailties, h_k , were generated as independent realized values from a gamma distribution with mean 1 and variance 0.05. The disease event indicator, Y_{ki} , was generated as a Bernoulli random variable with probability $h_k \exp(\gamma_0 + \beta_1 z_{ki}^{(1)} + \beta_2 z_{ki}^{(2)})$ where $\gamma_0 = -3$, $\beta_1 = 0.2$, and $\beta_2 = 0.2$. The value $\gamma_0 = -3$ is equivalent to a baseline disease risk of about 5%, and both

$\beta_1 = 0.2$ and $\beta_2 = 0.2$ are equivalent to relative risk of 1.2. In each of the parameter sets, 1,000 simulation runs were carried out. The simulation was programmed in the free software R (R development core team, 2005), and our codes are available at www.ualberta.ca/~yyasui/homepage.html.

Tables 1 and 2 present the bias and coverage of the 95% confidence interval for the 1,000 simulation runs for NCC and SCC. These results were presented in all four (K, m_k) combinations for the three approaches, IRM, ARM and PBEE. In the case of NCC (Table 1), when the between-population variance is larger than the within-population variance, the ARM model generally presents lower bias than the IRM model, and when the within-population variance is larger, the IRM generally presents lower bias than the ARM model. The PBEE approach presents the lowest or close to the lowest value of bias in the three approaches, regardless of the within- and between-variances. The results for the confidence coverage interval are similar for the three models when the variance ratio is not large: they are all slightly lower than the 95% coverage. However, when the within-variance dominates, the ARM's coverage probability becomes low due to the large bias in estimating the parameter: the PBEE approach is affected by the same problem but to a much lesser degree. Similar results and patterns are observed for SCC (Table 2).

Figures 2 and 3 show mean squared errors of parameter estimation for NCC and SCC, respectively. In all four (K, m_k) combinations for the NCC and SCC, the mean squared error of the IRM decreases, and that of ARM increases as the ratio of within- to between-population exposure variance increases. However, the PBEE approach consistently provides the smallest (or close to the smallest) mean squared error among the three methods in all the scenarios considered.

Table 3 compares the estimation performance for ECC. The bias is small in both β_1 and β_2 estimation by the three approaches except for ARM's β_2 estimation, subject to the large within- versus between-variation of z_{k2} . In terms of the mean squared error, the ARM results in smaller errors than the IRM for the covariate with 0.25 variance ratio, while the IRM provides smaller errors than the ARM for the covariate with 16 variance ratio. The PBEE approach, however, performs well for both of the covariates variance ratios evaluated, providing the best (or close to the best) results in all the scenarios we considered.

Table 1. Bias and 95% confidence interval coverage in β_1 estimation by the individual random-effects model (IRM), aggregated random-effects model (ARM) and population-based estimating equation approach (PBEE) for the no confounding case (NCC), in differing within- versus between-population variance ratios for the four scenarios $(K, m_k) = (100, 100), (100, 50), (50, 100),$ and $(50, 50)$.

Number of Populations		Survey sample size in kth population											
		100						50					
		% Bias			95% Interval coverage			% Bias			95% Interval coverage		
	Variance ratio	IRM	ARM	PBEE	IRM	ARM	PBEE	IRM	ARM	PBEE	IRM	ARM	PBEE
100	0.25	0.20	0.00	0.07	92	93	93	1.99	-0.50	0.10	92	93	93
	0.5	-0.33	-0.28	-0.40	94	95	94	1.51	-1.58	-1.04	92	93	93
	1	0.74	-0.56	-0.08	92	94	93	1.86	-2.66	-1.51	92	92	92
	2	0.90	-2.22	-0.56	94	93	93	0.84	-4.84	-2.68	92	90	91
	4	0.99	-4.63	-0.59	94	91	93	1.18	-9.12	-3.26	92	87	90
	8	0.39	-8.98	-1.01	92	87	91	0.80	-17.25	-3.55	93	67	90
	16	-2.79	-19.16	-3.47	90	56	87	-2.45	-30.90	-5.17	91	13	83
50	0.25	0.85	0.32	0.60	91	93	92	-1.94	-0.09	-0.39	91	94	93
	0.5	-1.49	-1.13	-1.24	93	92	92	2.86	-0.89	-0.11	92	93	93
	1	-0.55	-1.56	-1.14	93	92	93	2.52	-1.92	-0.67	92	93	93
	2	0.62	-2.50	-0.92	93	92	92	2.54	-4.04	-1.56	93	94	92
	4	1.25	-4.71	-0.60	92	91	93	2.83	-8.22	-2.38	91	91	91
	8	1.14	-9.44	-0.45	92	88	92	2.21	-16.83	-2.63	92	81	90
	16	-2.32	-19.20	-3.04	91	74	89	-1.51	-29.21	-4.29	90	46	89

Table 2. Bias and 95% confidence interval coverage in β_1 estimation by the individual random-effects model (IRM), aggregated random-effects model (ARM) and population-based estimating equation approach (PBEE) for the simple confounding case (SCC), in differing within- versus between-population variance ratios for the four scenarios $(K, m_k) = (100, 100), (100, 50), (50, 100),$ and $(50, 50)$.

Number of Populations		Survey sample size in kth population											
		100						50					
		% Bias			95% Interval coverage			% Bias			95% Interval coverage		
	Variance ratio	IRM	ARM	PBEE	IRM	ARM	PBEE	IRM	ARM	PBEE	IRM	ARM	PBEE
100	0.25	0.59	0.26	0.29	92	93	92	1.05	0.05	0.24	91	93	93
	0.5	-0.30	-0.84	-0.48	93	93	92	0.92	-1.40	-0.96	92	93	93
	1	-0.33	-1.43	-0.83	92	94	92	1.88	-2.51	-1.18	93	93	92
	2	-0.08	-2.87	-1.12	93	93	93	0.72	-5.16	-2.57	93	91	92
	4	0.68	-5.55	-0.77	92	90	94	1.04	-10.34	-3.30	92	85	90
	8	0.29	-11.44	-1.01	94	83	93	0.92	-19.90	-3.42	92	61	89
	16	-5.01	-22.89	-5.34	84	45	81	-4.03	-35.04	-6.42	86	8	78
50	0.25	0.71	0.28	0.48	91	93	92	-1.81	0.77	0.13	92	93	91
	0.5	1.34	-0.13	0.11	90	92	91	2.59	-0.85	0.20	90	93	92
	1	1.74	-0.83	0.19	90	92	91	1.60	-2.19	-0.75	91	92	91
	2	1.11	-2.38	-0.14	92	92	92	3.67	-4.48	-0.64	92	92	92
	4	1.51	-5.08	-0.09	92	91	92	3.10	-9.78	-1.76	92	88	91
	8	0.76	-10.91	-0.75	92	87	92	1.66	-19.56	-2.74	91	74	89
	16	-4.55	-22.54	-5.03	86	68	85	-3.37	-34.86	-5.67	87	31	82

Figure 2. Mean squared error of β_1 estimates from the individual random-effects model (IRM), aggregated random-effects model (ARM) and population-based estimating equation approach (PBEE) for the no confounding case (NCC), in the differing within- versus between-population variance ratios for the four scenarios $(K, m_k) = (100, 100), (100, 50), (50, 100), \text{ and } (50, 50)$.

Figure 3. Mean squared error of β_1 estimates from the individual random effects model (IRM), aggregated random effects model (ARM) and population-based estimating equation approach (PBEE) for the simple confounding case (SCC), in the differing within- versus between-population variance ratios for the four scenarios $(K, m_k) = (100, 100), (100, 50), (50, 100),$ and $(50, 50)$.

Table 3. Bias and mean squared error in β_1 and β_2 estimation by the individual random-effects model (IRM), aggregated effects model (ARM) and population-based estimating equation approach (PBEE) for the extended confounding case (ECC), in differing within- versus between-population variance ratios for the four scenarios $(K, m_k) = (100, 100), (100, 50), (50, 100),$ and $(50, 50)$.

			Survey sample size in kth population											
			100						50					
			% Bias			Mse x 10 ⁻³			% Bias			Mse x 10 ⁻³		
			IRM	ARM	PBEE	IRM	ARM	PBEE	IRM	ARM	PBEE	IRM	ARM	PBEE
100	$z_{ki}^{(1)}$	0.25	0.03	0.86	-0.09	1.00	0.21	0.34	-0.15	1.50	0.16	1.56	0.22	0.36
	$z_{ki}^{(2)}$	16	0.27	-18.04	-1.51	0.65	3.45	0.56	0.87	-31.85	-4.86	1.20	5.90	1.05
50	$z_{ki}^{(1)}$	0.25	-0.04	0.83	0.12	2.01	0.43	0.72	-0.38	1.62	0.21	3.46	0.43	0.76
	$z_{ki}^{(2)}$	16	0.72	-17.59	-1.43	1.29	5.82	1.10	2.15	-33.09	-4.22	2.75	8.14	2.14

5. DISCUSSION

This article considers a new integrated data design and analysis, valuable for epidemiological investigations, that combines all the available information in the health outcomes and exposure variables at the various levels of data organization (Figure 1). Our integrated design differs from the aggregated data design with the ARM proposed by Prentice and Sheppard, because the ARM uses only aggregated health information on K groups and individual explanatory covariate data from a sample within each of K groups. In addition to the aggregated health and individual explanatory covariates data, the integrated design uses individual health information in the random sample within the K populations. This individual health information links with the individual explanatory covariates for each subject as in cohort studies. Therefore, the integrated design with the PBEE approach can be considered an extension of both the ARM proposed by Prentice and Sheppard and the IRM. Note that, while our simulation experiment did not include any population level covariate $Z_k = (Z_k^{(1)}, \dots, Z_k^{(j)})^T$ discussed in Section 2, the PBEE's ability to assess the population-level covariates' effects is critical in social and environmental epidemiology research in studying contextual and environmental factors that may influence health.

The integrated design performed with the proposed PBEE method presents a powerful analytical framework that takes into account both within- and between-population exposure variation and that combines the strengths of both individual- and aggregated-data. The design is applicable without knowledge of which of the exposure variations (within- versus between-population) dominates. In addition, although we may have knowledge of which variations dominate on each variable, two or more exposure variables of interest do not necessarily have the same type of variation. As we have shown by ECC simulations, the PBEE approach will be more advantageous particularly in terms of bias and mean squared errors over the IRM and the ARM in such cases.

Utilization of individual and aggregated health data has been proposed under the fully Bayesian framework by Jackson et al. to improve the ecological inference (Jackson et al.). Our proposal follows the same basic concept of Jackson et al., i.e., to use individual and aggregated health data. In contrast to the specifications of likelihood and prior distributions required in the fully Bayesian approach, however, we apply the integrated-analysis concept under an estimating equation framework requiring specification of only the first two moments of the outcome distribution. Jackson et al. multiplied the likelihood of individual data and that of aggregated data, while we removed the samples in the individual data from the aggregated data. If the sample sizes of individuals m_k 's are appreciable, this difference may be important. Jackson et al.'s method could also be applied, without any additional difficulty, with the individuals in the individual data

omitted from the aggregate data. In addition, the estimating equation approach makes the mean-parameter inference robust against misspecification of the second-order characteristics of disease outcomes. As a future research topic, a more detailed study can evaluate the advantages and disadvantages of the PBEE approach versus the fully Bayesian approach.

Recently, Haneuse and Wakefield (Haneuse and Wakefield, 2007) proposed a hierarchical Bayesian inference combining ecological and case-control data. They considered a hybrid scheme where ecological data on marginal distributions of deaths and covariates are supplemented with individual data from a sample of case-control data. While our integrated design with the PBEE approach is for cohort rather than case-control studies, it is of interest consider an extension of the PBEE approach to the case-control setting and compare with the Bayesian approach of Haneuse and Wakefield.

The individual covariate data are used in both the individual and aggregated parts of PBEE. Although covariate measurement errors are of concern, particularly in certain applications such as the diet-covariate considered by Prentice and Sheppard, nondifferential measurement errors did not cause bias in the examples in either Prentice and Sheppard or Jackson et al. In future research, we plan to consider how the integrated design with the PBEE approach behaves with smaller samples, for example, 5 to 10 individuals. In addition, consider the applicability of the adjustment method for small-sample bias proposed by Prentice and Sheppard to the estimating equations of PBEE and when such an adjustment is beneficial.

The study design/data structure considered here is currently not a study design used in epidemiology, but it offers certain advantages discussed in this article. The gain in the parameters' estimates can be achieved easily using aggregated mortality or disease data available from governmental agencies. We recommend the integrated design/data structure with the PBEE approach as a new analytical framework that can be considered in future epidemiological studies.

REFERENCES

- Godambe, V.P., Thompson, M.E. "An extension of quasi-likelihood estimation" *Journal of Statistical Planning and Inference*. 1989;22:137-52.
- Haneuse S., Wakefield J. "Hierarchical Models for Combining Ecological and Case-Control Data" *Biometrics*. 2007;63:128-36.
- Jackson, C. et al. "Improving ecological inference using individual-level data" *Statistics in Medicine*. 2006;25:2136-59.
- Liang, K.Y., Zeger, S. "Longitudinal data analysis using generalized linear models" *Biometrika*. 1986;73:13-22.

- Prentice, R.L., Sheppard, L. "Aggregate data studies of disease risk factors" *Biometrika*. 1995;82:113-125.
- R Development Core Team. "R: A language and environment for statistical computing" R Foundation for Statistical Computing, Vienna, Austria. 2005 <http://www.R-project.org>.
- Royall, R.M. "Model robust confidence intervals using maximum likelihood estimators." *International Statistical Review* 1986. 54, 221–226.
- Sheppard, L., Prentice, R.L. "On the reliability and precision of within- and between-population estimates of relative rate parameters" *Biometrics*. 1995;51:853–63.
- Szklo, M., Nieto, F.J. *Epidemiology. Beyond the basics*. 2000. Gaithersburg: Aspen Publication.
- Villanueva, C.M et al. "Meta-analysis of studies on individual consumption of chlorinated drinking water and bladder cancer" *Journal of Epidemiology and Community Health*. 2003;57:166–73.
- Zeger, S., Liang, K.Y. "Longitudinal data analysis for discrete and continuous outcomes" *Biometrics*. 1986;42:121–30.