Aggregation Gain Or Loss? Modelling the Effects of Group Variables with Binary Responses

Li, L.[1,2], **M. Hudson**[1] **and J. Ma**[1]

[1] Department of Statistics, Macquarie University, NSW
[2]PhD scholar, NHMRC Clinical Trials Centre, University of Sydney, NSW
Email: lli@efs.mq.edu.au

Keywords: *aggregation, group-level variable, bias, loss of efficiency*

EXTENDED ABSTRACT

Interest in grouped or aggregated data in epidemiology has grown over the last decade because of increasing emphasis on data confidentiality and also because not all disease determinants can be applied at an individual-level. Often researchers must choose between using the publicly available aggregated data or gathering individual data without prior knowledge of a gain in efficiency (Lang and Gottschalk, 1996).

Individual-level analyses are often case-centered and focused on identifying individual susceptibility, but may fail to identify the underlying cause of incidence, i.e. the ecologic effects (Rose, 1985). Diez Roux's (2004) summaries issues on study of group-level factors in epidemiology and calls for further research to lead to an understanding of how interactions within and between levels affect health.

The efficiency loss from fitting aggregated data to estimate coefficients of group-level variables, rather than using individual data is discussed by Lang and Gottschalk (1996) using ordinary least squares (OLS) linear models. However, the impact of estimating parameters of group-level variables in a logistic regression model remains unclear. It may be unwarranted to apply the results from linear regression directly to logistic regression.

In this paper, we focus on the potential benefits of using group-level variables over individual-level variables in aggregated data analyses. Two logistic regression models were used in our analyses:

1. Individual-level model (binary model)

   Individual observation $y_{ij} \sim Bernoulli(p_{ij})$

   $$log\Big(\frac{p_{ij}}{1-p_{ij}}\Big) = C^{(1)} + \beta^{(1)} x_i + \delta^{(1)} z_{ij}$$

   where p_{ij} is the probability of $y_{ij} = 1$.

2. Aggregated-level model (binomial model)

   Number of events at the aggregated level $y_{i\cdot} \sim Binomial(m_i, \bar{p}_i)$

   $$log\Big(\frac{\bar{p}_i}{1-\bar{p}_i}\Big) = C^{(2)} + \beta^{(2)} x_i + \delta^{(2)} \bar{z}_i$$

   where $\bar{p}_i$ is the average probability of $y_{ij} = 1$ in group i.

We asses the bias and efficiency of $\hat{\beta}^{(2)}$. Using simulation, we explore the effects of aggregation by comparing the results of group-level analyses with those from individual-level analyses.

These simulation results suggest that there are benefits of using aggregated data when our research interest is in group-level variables. In general, the bias from fitting aggregated data is negligible if the effects of individual-level variables on the outcome are moderate. Loss of efficiency resulting from aggregation in estimating group-level effects is also small if the strength of the correlation between explanatory variables is moderate. A large number of observations per group or a large number of groups do not affect the accuracy of estimated parameters of group-level variables and provide little benefit in efficiency either. The loss of efficiency can appear when the number of groups is less than 10.

Under certain circumstances, if we examine the group-level effects, we can choose to use the publicly available aggregated data instead of gathering individual data. In some cases, as pointed out by Lang and Gottschalk (1996), even if the loss of efficiency exists in aggregated data analyses, the gain in efficiency in gathering individual-level data may not warrant trade-offs with time and monetary costs involved.

1 INTRODUCTION

Privacy has become a sensitive issue in recent years (Behlen and Johnson, 1999; Denley and Smith 1999; Gostin and Hadley, 1998). Growing concerns about confidentiality and privacy increase the difficulty of obtaining the information from individuals. In epidemiology, much attention has been drawn to the fact that not all disease determinants can be conceptualized as individual-level attributes (Diez Roux, 2004). Alternatively, aggregated or grouped data, to which individuals belong, are relatively easier and cheaper to obtain than individual-level information. Such data will encourage greater interest and study of aggregation in epidemiologic research in the future.

Studies limited to characteristics of aggregates (groups) of individuals are commonly termed *ecologic studies* (Langbein and Lightman, 1978; Morgenstern, 1998). Ecologic study benefits research with focus on ecologic effects, i.e. effects on groups rather than individuals. Individual-level analyses are often case-centered and focused on identifying individual susceptibility, but may fail to identify the underlying cause of incidence, i.e. the ecologic effects (Rose, 1985). Ecologic effects are particularly relevant when evaluating the impacts of social processes such as programs, policies, or legislation (Morgenstern, 1998). Group-level variables, such as alcohol consumption, health performance or life expectancies of different countries or regions, are research interests. Diez Roux (2004) summarizes issues on study of group-level factors in epidemiology and called for further research to understand how interactions within and between levels affect health.

Ecologic analysis in linear regression is well understood (Jargowsky, 2005; Langbein and Lightman, 1978). If the aggregate regression model is correctly specified, the analysis will provide equally unbiased estimates as those from individual level models. In fact, there is no guarantee that the individual regressions are better than aggregate ones (Jargowsky, 2005). Aggregation sometimes does not produce an aggregation loss, but may instead produce an aggregation gain (Grunfeld and Griliches, 1960).

Binary outcome response is a common type of outcome in epidemiology. Logistic regression is perhaps the most commonly used model (Greenland, 1998). Recent studies have explored the implications of covariate aggregation in logistic regression (Johnston *et al*., 2002). Models were constructed at individual level with an aggregated individual-level variable.

Previous ecologic research has mainly focused on the aggregation impacts on estimating parameters of individual-level variables in the presence of both individual-level and aggregated-level variables. Few studies demonstrate the effects of aggregation on estimating parameters of group-level variables. Lang and Gottschalk (1996) discuss, in linear models, the efficiency loss from fitting aggregated data to estimate coefficients of group-level variables, rather than ordinary least squares (OLS) estimates from individual data. However, the impact of estimating parameters of group-level variables in a logistic regression model remains unclear. It may also be unwarranted to apply the results from linear regression directly to logistic regression.

Often researchers must choose between using the publicly available aggregated data or gathering individual data without prior knowledge of a gain in efficiency from gathering individual data at considerable expense (Lang and Gottschalk, 1996). In this paper, we assess the bias and the efficiency loss of estimated parameters of group-level variables using aggregated data in logistic regression models. This knowledge has an immediate relevance in the design of epidemiological studies.

2 AGGREGATING EFFECTS IN ORDINARY LEAST SQUARES LINEAR REGRESSION

Before the simulation study of logistic regression, we briefly examine findings on the properties of estimated parameters of group-level variables in individual-level OLS models versus those estimated in corresponding models using aggregated data.

2.1 Individual-level model

Consider the general two-level linear regression model as follows:

$$y_{ij} = \sum_{k=1}^{a} \beta_k x_{ik} + \sum_{p=1}^{b} \delta_p z_{ijp} + \varepsilon_{ij} \qquad (1)$$

where $i = 1, \ldots n$ and $j = 1, \ldots, m_i$. Here y_{ij} as response variable, x_{ik} are group-level explanatory variables ($k = 1, \ldots, a$) and z_{ijp} are individual-level explanatory variables ($p = 1, \ldots, b$). Set n as the number of groups or clusters and m_i as the number of the observations per group. The total number of observation, denoted as N, equals $\sum_{i=1}^{n} m_i$.

The above model for observations in group i can be written in matrix form as:

$$y_i = \mathbf{X}_i\beta + \mathbf{Z}_i\delta + \varepsilon_i \qquad (2)$$

for $i = 1, 2, \cdots, n$, where $E(\varepsilon_i) = 0$; $Var(\varepsilon_i) = \mathbf{V_i} = \sigma^2\mathbf{I_i}$, and $\varepsilon_1, \cdots, \varepsilon_n$ are independent. Here y_i and ε_i are $m_i \times 1$ vectors; $\mathbf{X}_i$ is an $m_i \times$

a design matrix for the number of a group-level variables in group i; and $\mathbf{Z}_i$ is an $m_i \times b$ design matrix for the number of b individual-level variables in group i. We have $\mathbf{X_i} = \mathbf{1_i}\mathbf{x_i}^T$, in which $\mathbf{x_i}^T = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ia} \end{bmatrix}$ and $\mathbf{1_i^T} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$ of length m_i.

2.2 Aggregated-level model

When we aggregate the data (1), observations are the means of all variables:

$$\bar{y_i} = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}, \quad \bar{x}_{ik} = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ik} = x_{ik},$$

$$\bar{z}_{ip} = \frac{1}{m_i}\sum_{j=1}^{m_i} z_{ijp}, \quad \bar{\varepsilon}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} \varepsilon_{ij}$$

so that the relationship is

$$\bar{y}_i = \sum_{k=1}^{a} \beta_k x_{ik} + \sum_{p=1}^{b} \delta_p \bar{z}_{ip} + \bar{\varepsilon}_i \tag{3}$$

The above model can be written in the matrix form corresponding to equation (2) as:

$$\bar{y}_i = \mathbf{x}_i^T\beta + \bar{\mathbf{z}}_i^T\delta + \bar{\varepsilon}_i \tag{4}$$

where $E(\bar{\varepsilon}_i) = 0$; $\bar{v}_i = Var(\bar{\varepsilon}_i) = \mathbf{1_i^T V_i 1_i}/m_i^2 = \sigma^2/m_i$ and $\bar{v}_i$ is a scalar. Here $\bar{\mathbf{z}}_\mathbf{i}^T = \begin{bmatrix} \bar{z}_{i1} & \bar{z}_{i2} & \cdots & \bar{z}_{ib} \end{bmatrix} = (m_i)^{-1}\mathbf{1_i^T Z_i}$.

2.3 Bias and efficiency

The method of OLS provides $\hat{\beta}$, the Best Linear Unbiased Estimate (BLUE). Similarly, weighted least squares (WLS) provides the BLUE estimator $\hat{\beta}^{agg}$, of β for aggregated data in equation (4). $\hat{\beta}^{agg}$ is, by definition, unbiased for β.

We compare the variances of estimated coefficients from an individual-level model and aggregated model to examine loss of efficiency using aggregated data. Since we are interested in the group-level variables,our interest is restricted to the variances of estimated coefficients of group-level variables, denoted as $\mathbf{Var}(\hat{\beta})$ and $\mathbf{Var}(\hat{\beta}^{\mathbf{agg}})$. We shall not take the covariance into account, rather we compare the corresponding diagonal elements of these matrices.

For the individual-level model, we have

$$\mathbf{Var}(\hat{\beta}) = \sigma^2[\textstyle\sum_\mathbf{i} m_i \mathbf{x_i x_i^T} - (\sum_\mathbf{i} \mathbf{x_i 1_i^T Z_i})(\sum_\mathbf{i} \mathbf{Z_i^T Z_i})^{-1}(\sum_\mathbf{i} \mathbf{Z_i^T 1_i x_i^T})]^{-1} \tag{5}$$

For aggregated data, WLS estimation provides:

$$\mathbf{Var}(\hat{\beta}^{\mathbf{agg}}) = \sigma^2[\textstyle\sum_\mathbf{i} m_i \mathbf{x_i x_i^T} - (\sum_\mathbf{i} \mathbf{x_i 1_i^T Z_i})(\sum_\mathbf{i} (m_i)^{-1}\mathbf{Z_i^T Z_i})^{-1}(\sum_\mathbf{i} \mathbf{Z_i^T 1_i x_i^T})]^{-1} \tag{6}$$

The relative efficiency of $\hat{\beta}_k^{agg}$ for kth group-level variable is given as

$$RE_k = \frac{Var(\hat{\beta}_k)}{Var(\hat{\beta}_k^{agg})} \qquad k = 1, \cdots, a. \tag{7}$$

We expect the ratio to be equal to or close to one if there is no or little efficiency loss by using aggregated data. From equation (5) and (6), we conclude:

- If there is no correlation or weak correlation between group-level explanatory variables and individual-level explanatory variables, RE_k is equal to or close to 1. There is no loss of efficiency of $\hat{\beta}_k^{agg}$. On the other hand, if the correlation is very strong, then the RE_k is expected to be as small as zero.
- $(\sum_\mathbf{i}(m_i)^{-1}\mathbf{Z_i^T Z_i})^{-1}$ is the only item that is different in equation (5) and (6). As group size increases, $Var(\hat{\beta}_k^{agg})$ increases as well. This means RE_k decreases towards zero. Loss of efficiency of $\hat{\beta}_k^{agg}$ is getting bigger.

Corresponding to equation (7), Lang and Gottschalk (1996) derived the simplification of equation (7) for relative efficiency of the kth group-level variable when numbers of observations in each group are equal.

$$RE_k = \frac{1 - R_{agg}^2}{1 - R_{ind}^2}, \tag{8}$$

where R_{ind}^2 is the uncentered R-square from the auxiliary individual-level regression of x_k, a group-level variable, on the all the other remaining explanatory variables. Similarly, R_{agg}^2 is the uncentered R-square from the aggregated-level auxiliary regression of x_k on all other aggregated variables. Equation (8) indicates that the loss of efficiency from estimating coefficients of group-level variables with aggregated data depends only on the two uncentered R-squares.

Lang and Gottschalk (1996) concluded that the loss in efficiency hinges on the correlations between group-level variable of interest x_k and remaining explanatory variables. When they are orthogonal, there is no loss in efficiency from using aggregated data, which is also what we have concluded. When they are not orthogonal, however, relative efficiency depends on whether the within-group variation in individual-level variables is large relative to the between-group variation. If all the values within each group are similar, the loss of efficiency will be modest, otherwise it will be larger. As long as the two R-squares are small or are roughly of equal size, there is little loss from using aggregated data.

3 LOGISTIC REGRESSION SIMULATION

Using simulation, we generated data sets with two hierarchical levels (individual and group). Individuals are nested in the different groups. Following the notation in section 2, the outcome variable y_{ij} is an independent Bernoulli random variable (1 for event, 0 for non-event) at the individual-level. There are two explanatory variables in the individual-level study, one individual level variable z_{ij} and one group-level variable: x_i. The individual-level variable z_{ij} represents each individual's profile in the study. The group-level variable x_i is the study variable of interest, which might correspond to a group-level performance or policy variable.

The outcomes of group-level data are sums aggregated from those outcomes in individual-level data. After aggregation, the number of event in each group equals $\sum_{i=1}^{n} y_{ij}$, denoted by $y_{i\cdot}$, and the total number of observations (events and non-events) is N.

In this section, we firstly specify the problems to be studied by introducing relevant models, and then explain the simulation procedure to be applies for data analysis. All simulations and calculations were performed with R (version 2.4.1).

3.1 Models and study problem

Two logistic regression models were used in the analysis:

1. Individual-level model (binary model)

$$log\Big(\frac{p_{ij}}{1-p_{ij}}\Big) = C^{(1)} + \beta^{(1)}x_i + \delta^{(1)}z_{ij} \quad (9)$$

where $y_{ij} \sim Bernoulli(p_{ij})$ and p_{ij} is the probability of $y_{ij} = 1$.

2. Aggregated-level model (binomial model)

$$log\Big(\frac{\bar{p}_i}{1-\bar{p}_i}\Big) = C^{(2)} + \beta^{(2)}x_i + \delta^{(2)}\bar{z}_i \quad (10)$$

where $y_{i\cdot} \sim Binomial(m_i, \bar{p}_i)$ and $\bar{p}_i$ is the average probability of $y_{ij} = 1$ in group i.

We asses the bias and efficiency of $\hat{\beta}^{(2)}$. The bias of group-level exposure effect is defined as the difference between $\hat{\beta}^{(2)}$ and its true value used in simulation to generate data. If $\hat{\beta}^{(2)}$ is unbiased, we would expect that $\beta^{(2)} = \beta$. Relative efficiency (RE) is defined as ratio between variances of $\hat{\beta}^{(1)}$ and of $\hat{\beta}^{(2)}$. RE is used to measure the loss of efficiency of $\hat{\beta}^{(2)}$. If there is no loss of efficiency of group-level model estimates, we would expect that relative efficiency is equal to or close to 1. Next, simulation was employed to test these claims.

3.2 Simulation and data analysis

To facilitate the simulation, several assumptions were required. We generated two continuous explanatory variables. At group-level, these two variables were independent, identically distributed variables with a bivariate normal distribution.

To simplify the simulation and explanation, we set the mean of x_i and μ_i, the expected value of z_{ij} in group i, to zero ($\mu_x = \mu_\mu = 0$) and $\sigma_x = \sigma_\mu = 1$. The correlation coefficient between x_i and μ_i is denoted as ρ. We set the number of observations in each group to be equal. First we generated the group-level data with two explanatory variables using the bivariate normal distribution with parameters as above and obtained 50 group-level values for each variable.

Then, we expanded this matrix to individual-level data by duplicating group-level variable x_i to 100 individual-level values in each group; and by generating z_{ij} using $z_{ij} = \mu_i + \varepsilon'_{ij}$ where $\varepsilon'_{ij} \sim N(\mu_i, 1)$. The linear predictor (LP) at the individual level was:

$$LP = C + \beta x_i + \delta z_{ij} \quad (11)$$

where C was a constant, equal to -2.2. C was chosen to make the probability of outcome equal to 10% when all the variables are equal to their mean value, i.e. zero. 10% was chosen because the rate of incidence is generally low, such as the death rate of heart attack. Next we generated the outcome variable y_{ij} from a Bernoulli distribution with probability calculated from a linear predictor.

1000 replicates of 5000 individual observations in each group were generated. There were two sets of data in each replicate, i.e. individual-level and aggregated-level data. We fitted individual and aggregate level models respectively for the two different levels of data to obtain $\hat{\beta}$ in each. For each model, we then calculated the mean of 1000 $\hat{\beta}$s. The limits of the Wald 95% confidence interval for each simulation were calculated to test the inclusion of the true underlying group-level study effect, i.e. β. The percentage of coverage, i.e. the proportion of simulations in which the calculated confidence interval included the true value, was presented to examine the behavior of confidence intervals. We also calculated the sample variances of $\hat{\beta}$ from both models and relative efficiency was calculated.

4 RESULTS

We varied a number of variables to determine the reliability and limitations of the aggregated-level analysis on the effects of group-level study variable. We considered the effects of correlation between two explanatory variables (ρ), the different impact of

individual risk variables on the outcome (δ), numbers of observations in each group (m), and numbers of groups (n). The simulation outputs include: the mean of $\hat{\beta}$ from 1000 simulations, denoted as $\bar{\hat{\beta}}$; bias ($\bar{\hat{\beta}}^{(2)} - \beta$); the percentage of coverage of the true β in 1000 replicates; and the average standard errors of $\hat{\beta}$ and the average relative efficiency. We were interested in whether we would obtain unbiased and efficient $\hat{\beta}^{(2)}$ under different situations compared with results from individual-level analyses. We also calculated the RE calculate from OLS equation (8).

We used the same value of β in all the replicates since changes of β were observed to have very little impact on bias and loss of efficiency of group-level variable estimates from aggregated data. The choice of β is not a critical parameter to the findings presented below. In all the simulations, we modeled a small effect for the group-level variable ($\beta = 0.5$).

We first examined the impact on group-level variable estimation from aggregated data when the strength of correlation between two explanatory variables varies. Table 1 shows that there is very little bias of $\hat{\beta}^{(2)}$ from aggregated data and its coverage of true β is around 95% as the absolute value of ρ increases. Relative efficiency is close to 1 when ρ is small or moderate. This means loss of efficiency is small. However, the loss of efficiency grows bigger as the strength of correlation increases.

Table 2 indicates that the association between outcome and individual-level risk variable has a strong impact on $\hat{\beta}^{(2)}$. As the effect of individual-level risk variable on outcome is getting stronger, bias of $\hat{\beta}^{(2)}$ is increasing and coverage of the true β drops dramatically. As a result of this, relative efficiency of $\hat{\beta}^{(2)}$ increases.

In Table 3, we can see that the changing number of observations in each group has no impact on estimates of β and there is no bias from aggregated-level models. As the group size increases, estimates of s.e. of $\hat{\beta}$ decrease in both individual-level and group-level analyses. We see relative efficiency indicates a slight drop, which is negligible.

There is no indication of bias in Table 4 when we vary the number of groups. Loss of efficiency is modest even when we have as few as 10 groups. But the relative efficiency is small when the number of groups is less than 10. When the number of groups is 5, the average RE is 0.78 as shown in the table. There is a much larger variation (lower quartile 0.63 and upper quartile 0.98 from 2000 simulations) than those under other situations (for example, when the number of groups is 10, lower quartile is 0.83 and upper quartile is 0.99 from 2000 simulations). Loss of efficiency can be large when the number of groups is less than 10.

RE based on equation (8) for linear model outcomes is shown in the last row of all four tables. In general, RE (linear) is slightly lower than RE for binary outcomes calculated from simulations except when δ is large. In Table 2, as $|\delta|$ increases, RE from simulations increases while RE (linear) remains constant. Therefore, RE (linear) could be used as a rough indicator of relative efficiency in logistic regression when the effect of individual risk variables on the outcome is moderate.

5 DISCUSSION

In this paper, our main focus is group-level variables using aggregated data. This focus distinguishes our research from previous studies, which mainly examined individual-level variables. Our simulation results suggest that there are benefits of using aggregated data when our research interest is on group-level variables. In general, bias from fitting aggregated data is negligible if the model is correctly specified and the effects of individual-level variable on the outcome are moderate. Loss of efficiency resulting from aggregation in estimating group-level effects is also small if the strength of the correlation between explanatory variables is moderate. A large number of observations per group or a large number of groups do not affect the accuracy of estimated parameters of group-level variables and provide little benefit in efficiency either. The loss of efficiency exists when the number of groups is less than 10.

Under certain circumstances, if we examine the effect of group-level effects, researchers can choose to use the publicly available aggregated data instead of gathering individual data. In some cases, even when the loss of efficiency exists in aggregated data analyses, the gain in efficiency may not warrant trade-offs with time and monetary costs involved in gathering individual-level data (Lang and Gottschalk, 1996).

We examined the group-level effects for the linear model in section 2 based on the assumption of independent and identically distributed disturbance terms. We have made the same assumption in our simulation study of logistic regression. However, this assumption is not valid any more if disturbance terms are correlated and not identically distributed. In many applications, we are confronted with the specification of grouped data or hierarchical data structure. For example, asthma patients may be included from different cities with different levels of air pollution. We need to take the variance-covariance structure of error term into consideration when we research aggregation effects on group-level variables. The logistic regression model is no longer an appropriate model to use. Few options are available: generalized nonlinear least-squares

Table 1. Group-level variable effect estimates varying $|\rho|$

$\|\rho\|$		0	0.3	0.5	0.7	0.9
$\bar{\hat{\beta}}$	Individual Model	0.50	0.50	0.50	0.50	0.50
	Aggregated Model	0.49	0.49	0.49	0.49	0.50
Bias$(\bar{\hat{\beta}}^{(2)} - \beta)$		-0.01	-0.01	-0.01	-0.01	0
Coverage (%)	Individual Model	95.0	95.3	95.2	95.5	95.4
	Aggregated Model	94.1	95.5	96.2	95.9	96.0
S.E. of $\hat{\beta}$ $(\times 10^{-3})$	Individual Model	48	49	51	54	60
	Aggregated Model	48	50	54	65	103
RE		1.01	0.96	0.87	0.68	0.33
RE (OLS)		0.99	0.94	0.85	0.67	0.33

The correlation between explanatory variables (ρ) varies in each model. The true value of β is equal to 0.5. Estimates are averages of 1000 replicates with 50 groups and 100 observations in each group. The association between the outcome and the individual-level risk variable is fixed ($\delta = 0.4$).

Table 2. Group-level variable effect estimates varying $|\delta|$

$\|\delta\|$		0	0.5	1.0	1.5	2.0
$\bar{\hat{\beta}}$	Individual Model	0.50	0.50	0.50	0.50	0.50
	Aggregated Model	0.50	0.49	0.44	0.38	0.32
Bias$(\bar{\hat{\beta}}^{(2)} - \beta)$		0	-0.01	-0.06	-0.12	-0.18
Coverage (%)	Individual Model	96.5	95.4	96.4	96.0	96.2
	Aggregated Model	96.0	95.7	75.0	19.3	1.1
S.E. of $\hat{\beta}$ $(\times 10^{-3})$	Individual Model	51	48	47	48	50
	Aggregated Model	52	49	45	43	41
RE		0.94	0.97	1.10	1.28	1.49
RE (OLS)		0.94	0.94	0.94	0.94	0.94

The association between the outcome and the individual-level risk variable (δ) changes. The true value of β is equal to 0.5. Estimates are averages of 1000 replicates with 50 groups and 100 observations in each group. The correlation between explanatory variables (ρ) is fixed in each model ($\rho = 0.3$).

Table 3. Group-level variable effect estimates varying group size (m)

Group Size		5	10	20	50	100	150	500
$\bar{\hat{\beta}}$	Individual Model	0.51	0.52	0.51	0.50	0.50	0.50	0.50
	Aggregated Model	0.50	0.51	0.50	0.49	0.49	0.50	0.48
Bias$(\bar{\hat{\beta}}^{(2)} - \beta)$		0	0.01	0	-0.01	-0.01	0	-0.02
Coverage (%)	Individual Model	95.0	95.8	95.2	96.2	94.8	96.9	97.3
	Aggregated Model	95.7	95.8	95.6	96.9	94.5	96.2	92.2
S.E. of $\hat{\beta}$ $(\times 10^{-3})$	Individual Model	224	157	111	70	49	41	22
	Aggregated Model	227	159	113	71	50	42	22
RE		0.98	0.97	0.97	0.97	0.96	0.95	0.98
RE (OLS)		0.96	0.95	0.95	0.95	0.94	0.94	0.94

The number of observations in each group varies. Estimates are averages of 1000 replicates from 50 groups with different number of observations in each group. The true value of β is equal to 0.5. The correlation between explanatory variables (ρ) is fixed in each model ($\rho = 0.3$). The association between the outcome and the individual-level risk variable is fixed ($\delta = 0.4$).

(GNLS) model, generalized estimating equations (GEE) and latent variable modelling. We will conduct further research on this topic using such approches.

Estimation is often needed in situations where the model has been incompletely specified due to the omission of important covariates. The omission may be due to either an incorrect understanding of the phenomenon under study or an inability to collect data on all the relevant factors related to the outcome under study (Neuhaus, 1993). For example, unmeasured factors or unspecified variables may confound the baseline disease risk for groups or the effect of the risk factor under study. Confounders in ecological studies may be unmeasured group-level variables or factors which vary between individuals (Jackson *et al.*, 2002). Previous studies have been done to examine the effects of aggregating the individual-level factors. It is claimed that group-level analysis appears to be more reliable than standard individual-

Table 4. Group-level variable effect estimates varying group number (n)

Group Number		5	10	20	50	100	150
$\bar{\hat{\beta}}$	Individual Model	0.52	0.50	0.51	0.50	0.50	0.50
	Aggregated Model	0.51	0.49	0.50	0.49	0.49	0.49
Bias($\bar{\hat{\beta}}^{(2)} - \beta$)		0.01	-0.01	0	-0.01	-0.01	-0.01
Coverage (%)	Individual Model	95.8	96.6	95.9	95.1	95.8	95.7
	Aggregated Model	96.1	97.2	95.7	95.4	95.9	95.7
S.E. of $\hat{\beta}$ ($\times 10^3$)	Individual Model	232	129	83	49	34	27
	Aggregated Model	287	138	85	50	34	28
RE		0.78	0.90	0.95	0.96	0.97	0.98
RE (OLS)		0.78	0.88	0.93	0.94	0.95	0.95

The number of groups varies. Estimates are averages of 1000 replicates from a range of number of groups with 100 observations in each group. The true value of β is equal to 0.5. The correlation between explanatory variables (ρ) is fixed in each model ($\rho = 0.3$). The association between the outcome and the individual-level risk variable is fixed ($\delta = 0.4$).

level analysis under certain situations (Johnston *et al.*, 2002) and aggregating exposure data to the group-level can help absorb measurement error (Richardson and Monfort, 2000). We wish to consider the variable omission problem in further our research.

6 REFERENCES

Denley, I. and S.W. Smith (1999), Privacy in clinical information systems in secondary care, *British Medical Journal*, **318**, 1328-1331

Behlen, F.M. and S. Johnson (1999) Multicenter patient records research: security policies and tools, *Journal of the American Medical Information Association*, **6**, 435-443

Gostin, L. and J. Hadley (1998), Health services research: public benefits, personal privacy, and proprietary interests, *Annala of Interal Medecine*, **129**, 833-835

Diez Roux, A.V. (2004), The study of group-level factors in epidemiology: rethinking variables, study designs and analytical approaches, *Epidemiology Reviews*, **26**, 104-111

Greenland, S. (1998) Introduction to regression models, Modern Epidemiology 2nd Ed., 359-399, Lippincott-Raven Publishers

Grunfeld, Y. and Z. Griliches (1960), Is aggregation necessarily bad?, *Review of Economics and Statistics*, **42**, 1-13

Jackson, C., N. Best and S. Richardson (2006) Improving ecological inference using individual-level data, *Statistics in Medicine*, **25**, 2136-2159

Jargowsky, P.A. (2005), The Ecological Fallacy, The Encyclopedia of Social Measurement, vol. 1, 715-722, San Diego, California: Academic Press

Johnston, S.C., T. Henneman, C.E. McCulloch, and M. van der Laan (2002), Modeling treatment effects on binary outcomes with grouped-treatment variables and individual covariates, *American Journal of Epidemiology*, **156(8)**, 753-756

Lang, K.M. and P. Gottschalk (1996), The loss in efficiency from using grouped data to estimate coefficients of group level variables, *Computational Economics*, **9(4)**, 355-361

Langbein, L.I. and A.J. Lightman (1978), Ecological Inference, Sage University Paper series on Quantitative Application in the Social Science, series no. 07-001, Bevery Hills and London: Sage Publications

Morgenstern, H. (1998), Ecologic Studies, Modern Epidemiology, 2nd Ed., 459-480, Lippincott-Raven Publishers

Neuhaus, J.M. (1993), A geometric approach to assess bias due to omitted covariates in generalized linear models, *Biometrika*, **80(4)**, 807-815

Richardson, S. and C. Monfort (2000), Ecological correlation studies, Spatial Epidemiology, Oxford University Press

Rose, G. (1985) Sick individuals and sick populations, *International Journal of Epidemiology*, **14(1)**, 32-38.