STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:2591–2602 (DOI: 10.1002/sim.1524)

Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data

Melissa D. Begg*,† and Michael K. Parides

Department of Biostatistics, Mailman School of Public Health of Columbia University, 722 West 168th Street, New York, New York 10032, U.S.A.

SUMMARY

The focus of this paper is regression analysis of clustered data. Although the presence of intracluster correlation (the tendency for items within a cluster to respond alike) is typically viewed as an obstacle to good inference, the complex structure of clustered data offers significant analytic advantages over independent data. One key advantage is the ability to separate effects at the individual (or item-specific) level and the group (or cluster-specific) level. We review different approaches for the separation of individual-level and cluster-level effects on response, their appropriate interpretation and give recommendations for model fitting based on the intent of the data analyst. Unlike many earlier papers on this topic, we place particular emphasis on the interpretation of the cluster-level covariate effect. The main ideas of the paper are highlighted in an analysis of the relationship between birth weight and IQ using sibling data from a large birth cohort study. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: clustered data analysis; between-cluster effects; within-cluster effects; model misspecification; covariate selection

1. INTRODUCTION AND STATEMENT OF THE PROBLEM

Correlated data arise in a number of different settings including school-based research, dental research and longitudinal study designs. The key feature of these correlated (or clustered) data is that items under study are bound together in sets (or clusters) that are known to the data analyst. For example, students in a school tend to cluster together (that is, respond alike), as do sites around a tooth in an individual patient, and multiple measures on a single subject over time. For many years, research statisticians have stressed the importance of correcting for correlation among items within a cluster (see, for example, references [1–3]). The complexity of the research design calls for more complex techniques (for example, regression models) to obtain valid analytic results. This fact is now well known and accepted in the broader research community. While this complexity is often seen as a problem, it is important to

*Correspondence to: Melissa D. Begg, Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street (R626B), New York, NY 10032, U.S.A.
†E-mail: mdb3@columbia.edu

Received August 2002
Accepted January 2003

recognize that the complexity of the data can be an advantage, permitting more interesting, insightful analysis. Specifically, in addition to studying how individual-level factors impact individual-level responses, clustered data allow us to study the effect of cluster-level factors on individual-level responses. Fuller, more thoughtful modelling of individual-level and cluster-level effects might enable us to exploit the data more completely, thereby leading us to better understand the sets of relationships that underlie important questions in medicine, public health and other domains.

Two common regression approaches for analysing clustered data are the random effects approach (for example, see references [4–7]) and the generalized estimating equations (GEE) approach [8, 9]. Both techniques are widely used and applicable to both continuous data and discrete data responses. An excellent review of these approaches can be found in Diggle *et al.* [1], Vonesh and Chinchilli [2], Zeger *et al.* [10] or Zeger and Liang [11]. To summarize briefly, the random effects approach seeks to explicitly model the correlation among items within clusters, while the GEE approach treats the correlation as a nuisance and corrects for it apart from the regression model in the estimation of the covariance parameters. Because these methods take different approaches to accounting for intracluster correlation, they lead to different interpretations for the regression coefficients (as described by Neuhaus [12]). Regression coefficients from a random effects model are characterized as 'subject-specific' estimates; that is, the exposure regression coefficient can be interpreted as an expected difference in the response for the same cluster at different levels of the exposure measurement. In contrast, a marginal model yields regression coefficients that are described as population averaged; they represent expected differences among different clusters with different exposure levels. (For linear models, the random effects and marginal estimators share the same interpretation, and so this distinction does not exist.)

Apart from this difference in interpretation of regression parameters, there is little in the literature to offer guidance with respect to model selection with clustered data. This is unfortunate, as the structure of clustered data permit greater flexibility in model fitting than do independent data. In particular, clustered data allow for separation of effects at the individual (or item-specific) level and the group (or cluster-specific) level.

To fix ideas, we consider the following example. Suppose that we wish to investigate the impact of birth weight on childhood intelligence (or IQ) as measured by the Wechsler Intelligence Scale for Children [13]. Numerous studies have appeared in the literature to address this question (for example, those by Richards *et al.* [14] and Breslau [15]), with most demonstrating that heavier babies tend to have higher IQs (as both children and adults). A major criticism of most of these studies is inadequate adjustment for confounding by family socio-economic status (SES). Birth weight is well known to be associated with family SES, with families of higher status having larger babies. SES of a family is also related to the IQ measurements of children in that family. Thus, SES can act as a potent confounder of the relationship between birth weight and IQ. Theoretically one could attempt to adjust for such confounding bias by including a measure of SES in regression models; however, SES is notoriously difficult to measure accurately. While the addition of an imperfect measure of SES to the regression model may succeed in partially removing confounding bias, residual bias may continue to obscure the true relationship between birth weight and IQ. Alternatively, if one had measures of birth weight and IQ for multiple siblings within the same family, one could use these data to make 'within-family' comparisons that would be tightly controlled for SES (as well as a host of other important family-level factors). In addition, it would be

possible not only to evaluate individual-level birth weight as a predictor of individual IQ, but to evaluate the effect of 'family averaged' birth weight on IQ. This leads to a separation, or partitioning, of the effect of birth weight that conforms to interesting research hypotheses. Using this approach we could even study the potential for interaction between individual-level and family-level effects. In summary, by breaking down the birth weight effect into individual-level and family-level components, we are able to obtain better estimates of these effects and draw more accurate conclusions.

Several authors have considered the question of separating, or partitioning, item-level and cluster-level effects. Because these discussions have appeared in diverse fields of research, there are various different (but identically defined) terms for the same problem. The social sciences literature discusses the notion of estimating individual versus 'contextual' or 'compositional' effects (that is, cluster-level effects) and the usefulness of 'centring' individual-level effects around a group (or context) mean (for example, see reference [16], chapter 5, and reference [17]). In the statistical literature, Neuhaus and Kalbfleisch [18] introduce the concept of differential 'between-cluster' and 'within-cluster' covariate effects, and the implications of ignoring this distinction. Their idea of separating between- and within-cluster effects parallels that of separating cross-sectional and longitudinal effects in the analysis of repeated measures [1, 19]. Berlin *et al.* [20] present the idea of 'confounding by cluster effects' in the analysis of data from a multi-centre study. Finally, regression estimation of 'between-cluster' and 'within-cluster' effects is reviewed by Mancl *et al.* [21] in the setting of dental research.

In this paper we review earlier findings on the best ways to estimate within-cluster (or item-level) effects on response. We go beyond the problem of within-cluster effects, however, to focus specifically on cluster-level effects. Like Raudenbush and Bryk [16] and Kreft *et al.* [17], we discuss the proper formulation and interpretation of between-cluster (cluster-level or cluster-averaged) effects on outcome. Our goal is to provide a thorough accounting of the appropriate assessment of cluster-level effects and its implications for the data analyst.

In Section 2 we introduce some notation and specify a series of models that seeks to partition item-specific and cluster-specific covariate effects on the outcome measure. An example illustrating these ideas is described in Section 3. Section 4 describes the rationale for separating individual-level and group-level effects, and compares currently existing recommendations. Section 5 addresses the definition of the between-cluster effect. Finally, Section 6 gives a summary discussion and recommendations for model fitting.

2. NOTATION AND MODEL STATEMENTS

2.1. Notation

Let the individual response measurements be denoted by Y_{ij}, where $i = 1, 2, 3, \ldots, K$ indexes the cluster or group, and $j = 1, 2, 3, \ldots, n_i$ indexes the item within cluster i. For simplicity, suppose that there is only one exposure (or treatment) variable of interest, denoted by X_{ij}, which varies from item to item within a cluster. (Note that for study designs in which the exposure variable X_{ij} is fixed for all items in a cluster, the issues discussed in this paper do not apply.) The mean exposure measurement for the group can be computed as $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$. Naturally, just as the exposure measurement varies from item to item in a cluster, the cluster-specific exposure means can vary from cluster to cluster.

Table I. Models that relate the conditional mean of Y_{ij} to X_{ij}.

Model number m	Model statement
1	$h(E[Y_{ij}\|X_{ij}) = \alpha_1 + \beta_1 X_{ij}$
2	$h(E[Y_{ij}\|X_{ij}, \bar{X}_i) = \alpha_2 + \beta_2 X_{ij} + \gamma_2 \bar{X}_i$
3	$h(E[Y_{ij}\|X_{ij}, \bar{X}_i) = \alpha_3 + \beta_3 (X_{ij} \quad \bar{X}_i) + \gamma_3 \bar{X}_i$
4	$h(E[Y_{ij}\|X_{ij}, \bar{X}_i) = \alpha_4 + \beta_4 (X_{ij} \quad \bar{X}_i)$
5	$h(E[Y_{ij}\|X_{ij}, \bar{X}_i) = \alpha_5 + \gamma_5 \bar{X}_i$

2.2. *Model specification*

One can employ regression analysis to model the relationship of Y_{ij} and X_{ij}. Five possible regression models for the Y–X relationship are described in Table I. To begin, one might fit a simple model that specifies the expected value of Y_{ij} (or some function, h, of its expected value) as a linear function in the covariate X_{ij}, as specified by model 1 in the table. Model 2 builds on model 1 by adding the cluster-level mean exposure. Model 3 is mathematically equivalent to model 2, but substitutes a 'centred' form of the individual measurement ($X_{ij} \quad \bar{X}_i$) for the raw measurement (X_{ij}). Since the individual measurement is replaced by its deviation from the cluster-level mean, this new version of the individual score represents how much larger, or smaller, the individual measurement is compared to other items in its cluster. Model 4 specifies the deviation ($X_{ij} \quad \bar{X}_i$) as the sole predictor of outcome, and model 5 specifies the group mean ($\bar{X}_i$) as the sole predictor variable.

2.3. *Rationale for cluster-level mean adjustment*

There are several reasons for adjusting for the cluster mean. As noted by Berlin *et al.* [19], variability in cluster means is common, and can confound the estimated association between the individual-level exposure measurement and outcome. Adjusting for the cluster mean may remove confounding bias. In the same spirit, papers by Neuhaus and Kalbfleisch [18] and Mancl *et al.* [21] argue that inference on the individual-level exposure effects can be misleading without adjustment for cluster-level means. Kreft *et al.* [17] and Raudenbush and Bryk [16] articulate the need for evaluating cluster-level effects as predictor variables in their own right. Thus, all five models might be useful, depending on the specific research question(s).

2.4. *Rationale for centring*

Centring a regression covariate around its grand mean is a common practice in ordinary regression analysis of independent data. It sometimes permits easier interpretation of the regression coefficients, and can avoid computational problems in parameter estimation. In the analysis of clustered data, one can centre around the grand mean of a predictor variable, or around the cluster-level means (yielding a cluster-specific deviation, defined as $X_{ij} \quad \bar{X}_i$). We refer to the latter as 'cluster-specific centring'. If there is concern about confounding of the relationship

between X_{ij} and Y_{ij} by cluster means (the $\bar{X}_i$'s), then cluster-specific centring serves to reduce that bias.

3. EXAMPLE: RE-EVALUATION OF BIRTH WEIGHT AND IQ IN THE NATIONAL COLLABORATIVE PERINATAL PROJECT

Consider the example of evaluating the relationship between birth weight and childhood IQ. We use data from the National Collaborative Perinatal Project (NCPP), a longitudinal study funded by the National Institutes of Health to study prenatal and perinatal determinants of health and well-being in early childhood. Broman [22] gives a full description of the study design and demographic characteristics of study participants. From 1959 to 1966, tens of thousands of pregnant women at twelve centres nationwide accepted the invitation to join the study. Investigators collected data on conditions during pregnancy, labour and birth characteristics, and early childhood health and development. In this paper we focus on the assessment at birth (which included birth weight) and the assessment made on children at about 7 years of age, since this examination included determination of each child's IQ (Wechsler Intelligence Scale for Children [13]). A recent report demonstrated that higher birth weight was associated with higher IQ within same-sex sibling pairs from the NCPP [23].

We also use data from the NCPP to illustrate certain principles about model fitting when analysing clustered data. Our sample consists of those children from live-born singleton births for whom a valid IQ measure at age 7 years is available in the publicly available archival data [24]. If a subject's IQ was recorded as less than 40, that subject was deleted from the analysis because of the strong likelihood of an incomplete administration of the IQ test.

Many families in the NCPP database contributed more than one child to the study. Because our emphasis is on the separation of within-cluster and between-cluster effects, our illustration includes only those families for whom multiple children (at least two) appear in the data set ($N=12\,616$ children from $K=5657$ families). (Generally speaking, single-child families should be retained in such an analysis, as they provide valuable information on the between-cluster effect.) Family size (defined as the number of children from one family included in the analytic sample) varies from 2 to 6 children. Individual IQ in the sample ranges from 40 to 150, with a mean (standard deviation) of 95.4 (14.3) points. Individual birth weight averages 3.21 kg (SD = 0.53), and ranges from 0.79 to 5.56 kg. The mean family-averaged birth weight is 3.21 kilograms (SD = 0.46) across families. Within-family birth weight deviations ranged from −1.95 to 1.77 kg (SD = 0.27), averaging around 0.21 kg in absolute value.

Our first goal in the analysis of these data is to accurately assess the effect of individual birth weight on individual IQ, free of confounding by family-level influences. Our second goal is to evaluate whether family-averaged birth weight has an independent impact on individual IQ, beyond that imparted by individual birth weight. While it is difficult to imagine that the birth weights of one's siblings has a direct effect on an individual's IQ, it is easy to see that family-averaged birth weight might serve as a proxy for relevant family-level characteristics, including both genetic and environmental factors.

We have analysed these data using linear mixed models via the 'xtreg' procedure in Stata 7 [25]. We fit the five models defined in Table I, replacing the fixed intercept α_0 with α_i, a random intercept for each family. The regression coefficient estimates corresponding to each of the five models are presented in Table II. (Note that we also used the GEE approach

Table II. Estimated regression coefficients (and 95 per cent confidence intervals) for the five models posed in Table I, based on a random effects linear regression analysis of the NCPP birth weight/IQ data set. Outcome is IQ (original measure on continuous scale) and exposure is birth weight (in kg). The regression coefficients are interpretable as mean IQ differences.

Model number m	$\hat{\beta}_m$	$\hat{\gamma}_m$
1	3.795 (3.335, 4.255)	—
2	2.554 (1.937, 3.170)	2.785 (1.861, 3.709)
3	2.554 (1.937, 3.170)	5.339 (4.651, 6.026)
4	2.554 (1.937, 3.170)	—
5	—	5.338 (4.651, 6.026)

Table III. Exponentiated regression coefficient estimates (and 95 per cent confidence intervals) for the five models posed in Table I, based on a random effects logistic regression analysis of the NCPP birth weight/IQ data set. Outcome is IQ> 85 (yes/no) and exposure is birth weight (in kg). The parameter estimates presented are interpretable as odds ratios.

Model number m	$\exp(\hat{\beta}_m)$	$\exp(\hat{\gamma}_m)$
1	1.98 (1.76, 2.23)	—
2	1.41 (1.17, 1.69)	1.78 (1.40, 2.26)
3	1.41 (1.17, 1.69)	2.51 (2.14, 2.94)
4	1.42 (1.18, 1.70)	—
5	—	2.50 (2.14, 2.93)

with an exchangeable correlation structure to fit these models, and obtained virtually identical results, as expected.) We see that the estimated regression coefficients for the individual-level covariate (child's birth weight in kilograms) and cluster-level covariate (family averaged birth weight in kilograms) vary, depending on the model specified. For example, the average IQ difference for a 1kg increase in birth weight is given as 3.8 points in model 1 and 2.6 points in models 2, 3 and 4. For the effect of family-averaged birth weight on IQ we obtain an estimate of 2.8 IQ points in model 2 and 5.3 points in models 3 and 5. Clearly these estimates differ substantially, so that choice of model greatly affects inference about the effect sizes for each predictor variable. A crucial question, then, is how to choose the model and corresponding parameter estimates that best address the primary research hypotheses.

In a further analysis, we dichotomized the primary outcome variable in order to generate, for comparison, a set of parameter estimates from the logistic model. The outcome for these analyses was defined as IQ> 85, corresponding to the minimum IQ for a child to be classified as 'learning disabled'. For convenience, we refer to this outcome as IQ 'in the normal range'. Approximately 78 per cent of the children in the sample had IQs above 85; 36 per cent of families had at least one child whose IQ was below the normal range.

We analysed the binary data example by fitting a series of random effects logistic models in Stata [25], using the 'xtlogit' procedure. The estimated regression coefficients corresponding to each of the five models specified in Table I are presented in Table III. (We note that when we fit these same models using GEE logistic regression with exchangeable correlation

structure, the estimated regression coefficients were attenuated relative to the random effects estimates.) As observed for the linear model, the five models give different estimates for the individual-level and cluster-level effects. The odds ratios for the within-cluster effect vary from 1.4 to 2.0, while the odds ratios for the cluster-level effect range from 1.8 to 2.5. Again we must choose which of these estimators to report, based on which best capture the effects of interest according to the study hypotheses.

4. EVALUATION OF THE INDIVIDUAL-LEVEL OR WITHIN-CLUSTER EFFECT

Typically, one of the primary aims of a study of clustered data is to estimate the effect of an individual predictor value on an individual response. In this sense, clustering in the data might seem a nuisance, in that the dependence among groups of responses complicates the analysis. As discussed previously, however, clustering among observations should not always be viewed as an impediment; it sometimes becomes an advantage when evaluating a hypothesis. The availability of additional information on fellow cluster members (who are more homogeneous than independent subjects) makes a more thorough evaluation of the research question(s) possible. This type of distinction is often made when highlighting the difference between random effects approaches and marginal approaches (for example, GEE) for modelling clustered data. The random effects approach is typically characterized as modelling a subject-specific, or within-cluster, effect, whereas GEE is described as providing a population-averaged effect. We think that this type of description can be misleading, however; both techniques can be used to obtain meaningful estimates of individual-level (within-cluster) and cluster-level (between-cluster) exposure effects. Unless care is taken to fit an appropriate model that adjusts for cluster-level effects, both approaches can yield biased estimates of the individual exposure effect (as noted by Berlin *et al.* [19]). On the other hand, when the regression model includes adjustment for cluster-level effects, both approaches can yield more accurate individual-level effect estimates.

Berlin *et al.* [20], Neuhaus and Kalbfleisch [18], Mancl *et al.* [21], Ten Have *et al.* [26, 27] and Raudenbush and Bryk [16] all emphasize the importance of removing bias due to confounding by cluster-level mean exposure. Furthermore, Neuhaus and Kalbfleisch [18] have shown that the estimate of the individual-level effect that is adjusted for the cluster-level mean is equivalent to the estimate obtained from a conditional likelihood analysis. Thus, it is now recommended that individual-level effects should be reported after adjustment for cluster-level effect. This can be accomplished via modelling in two ways: (i) by including the cluster-level mean exposure (as in model 2); or (ii) by substituting the cluster-specific deviation for the individual exposure measure with or without inclusion of the cluster-level mean in the regression model (as in models 3 and 4). Hence, for our continuous data example (see Table II), the appropriate estimate of the effect of individual birth weight on IQ is 2.6 IQ points for every 1 kg higher birth weight, given the same family-averaged birth weight. The larger estimate, 3.8 IQ points for every kilogram, is distorted due to confounding by family-averaged birth weight, and, therefore, should not be reported. For the binary data results in Table III, the appropriate estimate of the individual-level effect is 1.4, indicating that the odds of a normal range IQ are 40 per cent higher for every 1 kg increase in birth weight, adjusted for the family-averaged birth weight. The higher odds ratio estimate (2.0) is biased due to failure to adjust for the family-level effect.

5. EVALUATION OF THE CLUSTER-LEVEL OR BETWEEN-CLUSTER EFFECT

We have discussed inclusion of a cluster-level average exposure effect as an adjustment variable in regression in order to remove bias in the estimate of the individual-level exposure effect. There may be circumstances, however, in which an estimate of the cluster-level mean effect may itself be of primary interest. In the sociology literature, this would be referred to as a contextual effect (that is, the effect that an individual item's milieu, or context, exerts on an individual item's response). While proper estimation of the individual exposure effect has been discussed to a great extent in the literature, considerably less attention has been given to the estimation of the cluster-level effect. As a consequence, there are few published recommendations to guide model fitting for this purpose.

One reference does appear in Chapter 5 of the comprehensive text on hierarchical modelling by Raudenbush and Bryk [16]. In a subsection entitled 'Disentangling person-level and compositional effects', the authors propose an adjusted model and describe how to use it to estimate the contextual effect. We will return to their recommendation after discussing the estimation of the cluster-level effect from the birth weight/IQ data.

We begin by examining the estimated regression coefficients for the family-averaged birth weight covariate in models 2, 3 and 5 from Table II. Note first that models 2 and 3 are mathematically equivalent (as noted by Kreft and deLeeuw [17]). The estimates of the contextual effect of family-averaged birth weight (represented by γ_m), however, are strikingly different in the two models (while the estimated effects of individual birth weight on IQ are identical). Model 2 estimates an increase in IQ of about 2.8 points for every kilogram increase in the family-averaged birth weight, while models 3 and 5 estimate the effect to be about 5.3 points, almost twice as large. In choosing which to report, one must consider the interpretation of each estimator.

The relevant parameter from model 2, γ_2, can be interpreted as the effect of a 1 kg increase in family-averaged birth weight on IQ, holding fixed the individual birth weight. To see this, consider the expected IQ for a model that specifies as covariates $X_{ij} = r$ and $\bar{X}_i = s$ versus a model that specifies as covariates $X_{ij} = r$ and $\bar{X}_i = (s + 1)$. The expected IQ is given by $\alpha_2 + \beta_2 r + \gamma_2 s$ for the first covariate set and $\alpha_2 + \beta_2 r + \gamma_2(s + 1)$ for the second. Subtracting the first expected value from the second, we obtain an estimated difference in expected IQ of γ_2 points. Thus, given two subjects with the same individual birth weight, the subject whose family has a 1 kg higher average birth weight can be expected to have an IQ about 2.8 points higher than the other subject, on average.

The interpretation of the estimate from model 3, γ_3, is more complicated. It can be interpreted as the effect of a 1 kg increase in family-averaged birth weight, holding fixed the individual birth weight *deviation*, but where the individual birth weight changes necessarily. To see this, let us write out an expression for the expected IQ under model 3 when the family-averaged birth weight is equal to s, and when the family-averaged birth weight is equal to $(s + 1)$, holding individual birth weight fixed at $X_{ij} = r$. The expected IQ for the first covariate set is $\alpha_3 + \beta_3(r \quad s) + \gamma_3 s$, while for the second set of covariates, the expected IQ is given by $\alpha_3 + \beta_3(r \quad s \quad 1) + \gamma_3(s + 1)$. Taking the difference between these two expressions, we find that the expected change in IQ for a 1-unit increase in family-averaged birth weight is equal to $\gamma_3 \quad \beta_3$ (not γ_3) when the individual exposure measurement (not the deviation) is held constant. The only way to derive γ_3 as the effect of interest is to consider comparing expected IQ from a model with covariates $X_{ij} = r + 1$ and $\bar{X}_i = s + 1$ to a

model with covariates $X_{ij} = r$ and $\bar{X}_i = s$. Thus, γ_3 represents the mean expected difference in IQ associated with a 1 kg increase in family-averaged birth weight simultaneous with a 1 kg increase in individual birth weight. This is due to the fact that when the cluster-level mean changes, the deviation changes by definition. To keep the deviation constant when the cluster-level mean changes, the individual exposure measure must also change by the same amount.

This observation leads to another fact about the γ_3 effect: the γ coefficient in model 3 (5.339) is identical to the sum of the β (2.554) and γ (2.785) coefficients from model 2 ($\gamma_3 = \gamma_2 + \beta_2$). This is a direct consequence of specifying the two model covariates as U and ($W \quad U$) (versus specifying the covariates as U and W in model 2). Furthermore, the γ estimate from model 5 (which makes no adjustment for individual exposure, either as raw measurement or deviation) is identical, apart from rounding error, to the γ estimate from model 3 (which is *apparently* adjusted for the individual birth weight via inclusion of the birth weight deviation). Coefficient γ_5 represents the effect of *average* family birth weight on *average* family IQ. (Diggle *et al.* [1] refer to this as the 'cross-sectional effect', while Neuhaus and Kalbfleisch [18] and Berlin *et al.* [20], use the term 'between-cluster' or 'among-cluster' effect.) While this is an interpretable quantity, it is very different in both magnitude and meaning from the (individual-specific) γ estimate in model 2. Therefore, although model 3 appears to offer a direct estimate of the cluster-level mean exposure that is adjusted for the individual exposure, this is not the case. We feel that this point is largely unrecognized by practitioners.

Model 2, then, is the model of choice for most purposes. The regression estimates (β and γ) obtained directly from model 2 correspond to the influence of each predictor (individual exposure and cluster-level mean exposure) on the individual-level response. We note that the same effect estimates can be obtained from model 3 by defining the 'true' cluster-level mean effect, γ_3, as the difference between two model parameters

$$\gamma_3 = \gamma_3 \quad \beta_3$$

This is precisely how Raudenbush and Bryk define the contextual effect in equation 5.42 of their text (reference [16], p. 139). Berlin *et al.* [20] also characterize this difference parameter as representing the association between individual outcome and the cluster-level mean exposure, giving equal weight to the two possible model formulations (corresponding to models 2 and 3 from Table I). In contrast, we stress that model 3 is undesirable, in that it is certain to lead to misinterpretation of the γ_3 parameter in practice.

Results for the logistic random effects models are completely analogous to those for the linear models, with one exception: perfect equality of corresponding effects from the different models is not achieved in certain cases due to the non-linearity of the logit-based model. In our example, we see that the estimated cluster-level effect from model 2 differs substantially from that of models 3 and 5. Model 2 estimates that a 1 kg increase in family-averaged birth weight increases the odds of IQ in the normal range by 80 per cent, adjusted for the influence of individual birth weight. Models 3 and 5, in contrast, point to a 150 per cent increase in the odds of a normal-range IQ for every 1 kg increase in family-averaged birth weight. As we have pointed out earlier, however, the larger effect estimate is not adjusted for the individual birth weight effect. For the same reasons outlined above, we believe it is preferable to report the model 2 cluster-level effect, as it is directly interpretable.

6. DISCUSSION

The ideas presented in this paper underscore the importance of proper model specification and careful parameter interpretation in regression analysis of clustered data. Misspecifying the regression model by omitting the cluster-level mean effect may seriously bias the estimate of the individual-level exposure effect. Equally important, failure to exercise caution in the definition and interpretation of regression parameters can lead to incorrect evaluation of the cluster-level effect estimate. Both problems pose serious threats to valid inference.

With regard to estimation of the individual-level exposure effect, omission of one or more key confounding variables may cause one to report an estimated effect that is dramatically overstated, understated, or completely opposite from the true relationship. Furthermore, it is clear in practice that some confounders are more influential than others. When dealing with clustered data for which exposure information varies from item to item within a cluster, the cluster-level mean exposure is often the most potent confounder of the association between individual-level exposure and response.

Just as the cluster-level mean exposure can confound the relationship between individual-level exposure and response, so, too, can the individual-level exposure influence the estimated relation between response and the cluster-level mean exposure. We must be careful when defining the quantities that represent the individual-level and cluster-level mean exposures, because the definitions of these quantities dictate the interpretation of their regression coefficients. In our example, both models 2 and 3 include an adjustment for the effect of cluster, and, therefore, appear to yield adjusted effect estimates for the individual-level and cluster-level mean effects. Upon further inspection, however, we discover that one parameterization leads to regression coefficients that are directly interpretable (model 2), while the other leads to a directly interpretable coefficient for the individual-level exposure effect, but not for the cluster-level mean exposure (model 3). Although an appropriate estimate of the cluster effect may be derived from model 3 by taking the difference in two parameter estimates ($\hat{\gamma}_3 \quad \hat{\beta}_3$), this step is frequently overlooked in practice. Furthermore, testing the cluster-level effect is straightforward in model 2, involving the evaluation of one parameter (that is, $H_0{:}\gamma_2 = 0$ versus $H_1{:}\gamma_2 \Subset 0$). The equivalent test from model 3 involves evaluating a difference in parameters along with their respective variance and covariance terms.

Another important aspect of this discussion is the influence of cluster size. When clusters are relatively 'large' (as in school-based or dental research), the cluster mean can be estimated much more precisely than in data sets where cluster sizes are small (as in family studies). Smaller clusters result in greater error in the estimation of the cluster-level mean effect. Data analysts should recognize this limitation, and interpret their results cautiously. Nevertheless, the recommendations made in this paper are pertinent for any cluster size. Our goal is to minimize error from a multiplicity of sources – including error due to bias as well as error due to imprecision. If adjustment for the cluster-level mean removes bias in the estimated individual-level exposure effect, then one should make the adjustment. Similarly, if one wants an unbiased estimate of the cluster-level effect, one should adjust for the influence of the individual-level exposure. Concerns about measurement error should not take precedence over concerns about bias.

As a final note, we return to the question of choice of modelling approach for analysing correlated data. Earlier papers have observed that random effects and marginal models tend to be used to address different types of research questions [12, 28]. The structure of the

random effects model facilitates assessment of within-cluster effects, while the structure of the marginal model facilitates assessment of cluster-level effects [12]. We want to emphasize that separation of individual-level and cluster-level exposure effects is achievable via appropriate model specification using either the random effects or the marginal modelling approach. It may be important in some cases, therefore, to emphasize careful covariate specification over choice of regression technique (for example, random effects versus GEE) when advising on regression analysis of clustered data.

ACKNOWLEDGEMENTS

We are very grateful to Drs Michaeline Bresnahan, Thomas Matte and Ezra Susser for helpful discussions.

REFERENCES

1. Diggle P, Liang K-Y, Zeger SL. *The Analysis of Longitudinal Data*. Oxford University Press: New York, 1994.
2. Vonesh EF, Chinchilli VM. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker: New York, 1997.
3. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials*. Arnold: London, 2000.
4. Breslow NE, Clayton DG. Approximate inference in generalized linear models. *Journal of the American Statistical Association* 1993; **88**:9–25.
5. Wolfinger R, O'Connell M. Generalized linear mixed models: a pseudo-likelihood approach. *Communications in Statistics – Simulation and Computation* 1993; **22**:1079–1106.
6. Stiratelli R, Laird N, Ware JH. Random effects models for serial observations with binary responses. *Biometrics* 1984; **40**:961–971.
7. Lee Y, Nelder JA. Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B* 1996; **58**:619–656.
8. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
9. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**:121–130.
10. Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
11. Zeger SL, Liang K-Y. An overview of methods for the analysis of longitudinal data. *Statistics in Medicine* 1992; **11**:1825–1839.
12. Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* 1992; **1**:249–273.
13. Wechsler D. *Wechsler Intelligence Scale for Children; Manual*. Psychological Corp: New York, 1949.
14. Richards M, Hardy R, Kuh D, Wadsworth MEJ. Birth weight and cognitive function in the British 1946 birth cohort: longitudinal population based study. *British Medical Journal* 2001; **322**:199–203.
15. Breslau N. Psychiatric sequelae of low birth weight. *Epidemiologic Reviews* 1995; **17**:96–106.
16. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edn. Sage Publications: Newbury Park, 2002.
17. Kreft IGG, de Leeuw J, Aiken LS. The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research* 1995; **30**:1–21.
18. Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998; **54**:638–645.
19. Louis TA. General methods for analyzing repeated measures. *Statistics in Medicine* 1988; **7**:29–45.
20. Berlin JA, Kimmel SE, Ten Have TR, Sammel MD. An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplasty. *Biometrics* 1999; **55**:470–476.
21. Mancl LA, Leroux BG, DeRouen TA. Between-subject and within-subject statistical information in dental research. *Journal of Dental Research* 2000; **79**:1778–1781.
22. Broman S. The collaborative perinatal project: An overview. In *Handbook of Longitudinal Research*, Mednick SA, Finello KM (eds). Praeger: New York, 1984.
23. Matte TD, Bresnahan MA, Begg MD, Susser ES. Influence of variation in birth weight within normal range and within sibships on IQ at age 7 years: cohort study. *British Medical Journal* 2001; **323**:310–314.

24. National Archives and Records Administration, Center for Electronic Records. Record Group 443: Records of the National Institutes of Health – National Collaborative Perinatal Project, 1959–1974.
25. StataCorp. *Stata Statistical Software*: *Release 7.0*. Stata Corporation: College Station, Texas, 2001.
26. Ten Have TR, Landis JR, Weaver SL. Association models for periodontal disease progression: a comparison of methods for clustered binary data. *Statistics in Medicine* 1995; **14**:413–429.
27. Ten Have TR, Landis JR, Weaver SL. Association models for periodontal disease progression: a comparison of methods for clustered binary data [letter]. *Statistics in Medicine* 1996; **15**:1227–1229.
28. Omar RZ, Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Statistics in Medicine* 2000; **19**:2675–2688.

