Improving power with repeated measures: diet and serum lipids\textsuperscript{1–3}

Julie A Marshall, Sharon Scarbro, Susan M Shetterly, and Richard H Jones

**ABSTRACT** The inability to detect associations between diet and serum cholesterol in cross-sectional population studies has been attributed to measurement error in diet assessments and between-subject variability in lipid concentrations. Current statistical methods can reduce the effects of measurement error and allow within-subject comparisons when replicate measures on individuals are available, even if the time between replicates is as long as 4 y and replicate data are not available for all subjects. Data from 928 nondiabetic participants of the San Luis Valley Diabetes Study with measures of 24-h dietary intake and fasting lipid concentrations at baseline, at a 4-y follow-up visit, or both were analyzed in a random-effects model that allowed for an unbalanced design. Sex was included as a non-time-varying covariate and age, body mass index, and energy intake were included as time-varying covariates. The findings when LDL cholesterol (mmol/L) was regressed on saturated fat intake (20 g/d) with all observations in a random-effects model (\( \beta = 0.14, P = 0.0016 \)) were compared with results with observations restricted to the first visit only (\( \beta = 0.05, P = 0.52 \)), a balanced design using averages across visits (\( \beta = -0.12, P = 0.28 \)), and a balanced design with random effects obtained by excluding subjects without two observations (\( \beta = 0.12, P = 0.0092 \)). Study power was greatest in the random-effects model using all observations and time-varying covariates. These findings highlight the importance of even a single replicate observation on a subsample of subjects. We recommend analyzing all data rather than averaging measures across visits or omitting observations to create a balanced design. *Am J Clin Nutr* 1998;67:934–9.

**KEY WORDS** Longitudinal data analysis, repeated measures, serum lipids, diet, epidemiologic methods, measurement error, random-effects model, statistical analysis

**INTRODUCTION** Experimental studies have shown consistently that higher dietary cholesterol and saturated fatty acid intakes are associated with increased serum total and LDL-cholesterol concentrations. Population-based cross-sectional studies, however, have generally not shown these associations. These negative findings have resulted from measurement error in diet and serum lipid assessment, small differences in intake among subjects, and large variations in cholesterol concentrations among individuals, usually attributed to genetic differences (1, 2). Genetic typing of subjects in the San Luis Valley Diabetes Study according to polymorphisms of candidate genes for heart disease, combined with the availability of replicate measures of serum lipid concentrations and dietary intake, have put us in a unique position to investigate associations of dietary intake and serum lipids with less measurement error and the ability to account for potential genetic modifiers in a free-living population (3).

Longitudinal studies are defined broadly as studies in which data on individuals are collected at two or more points in time. Multiple measures taken in the same individual may introduce correlation that is not accounted for in standard regression. In addition, longitudinal studies typically have unbalanced designs in which the number of observations and observation intervals are not the same for all subjects. Missing data and time-varying covariates make standard multivariate procedures inapplicable. In the past, investigators sometimes eliminated observations in their data sets to create a balanced design or averaged the observations collected over time to make the data conform to the requirements of standard software. Averaging data with unequal numbers of observations per subject produces unequal variances, which violates the assumptions of simple regression. In this situation, unequal variances can be accommodated by using generalized least squares in the analysis. However, averaging may also introduce bias by removing informative within-subject variability that would otherwise contribute information to parameter estimation. Methods (4–6) and software (7) that account for these unbalanced characteristics of longitudinal data are now readily available. The analyses presented here show how the magnitude and statistical significance of the association between saturated fat intake and serum LDL concentrations depends on the method used for analysis.

**SUBJECTS AND METHODS**

**Study population**

A geographically based sample of Hispanic and non-Hispanic white men and women aged 20–74 y and living in the San Luis Valley.
Valley of southern Colorado (Table 1) was examined initially between 1984 and 1988 (3, 8). Subjects included in the present study were determined to have normal glucose tolerance (9) at baseline and 74% were rescreened after an average of 4.3 y and determined to be nondiabetic. On the basis of a medication inventory, visits at which subjects were taking thiazides, β-blockers, estrogens, progestosterone, or lipid-lowering drugs were excluded. We attempted to include two visits for each subject, but because of nonresponse at follow-up or use of lipid-altering medications at either the first or second visit the final data set contained 318 subjects with only a baseline visit, 552 subjects with two visits, and 58 subjects with only a follow-up visit.

**Data collection and laboratory procedures**

Procedures followed were approved by and in accordance with the University of Colorado Health Sciences Center human subjects guidelines. Serum lipid concentrations were measured in fasting blood samples drawn after subjects had fasted overnight. LDL cholesterol was calculated by subtraction: first, the serum triacylglycerol concentration (10) divided by five was subtracted from the total serum cholesterol value (11); second, from this result, the total serum HDL concentration was subtracted (12). This calculation was performed only for subjects with serum triacylglycerol concentrations < 40 mg/L (13 subjects with triacylglycerol concentrations > 40 mg/L were excluded).

For the diet assessments, subjects were administered a 24-h diet recall by bilingual interviewers trained and certified by the Nutrition Coordinating Center at the University of Minnesota (13). More detailed information on dietary assessment and the intake of these subjects is provided elsewhere (3).

Hispanic ethnicity was defined by self-report to the 1980 US Census question, “Are you of Spanish/Hispanic origin or descent?” (14). Body mass index (BMI) was calculated as current measured weight in kilograms divided by height in meters squared.

**Statistical analysis**

The statistical procedure SAS PROC MIXED (version 6.11) was used for the regression analyses described here (7). This procedure can be used for longitudinal regression analysis in which observations are collected on the same subject or to simplify analyses when the value of covariates having to eliminate observations to produce a balanced design or avoid correlation resulting from multiple observations in a subject. Observations do not need to be balanced by subject, ie, there can be a variable number of observations per subject and the interval between observations can vary by subject. Fixed effects variables can include both time-varying (eg, BMI) and non-time-varying (eg, sex) variables. Models used here account for the lack of independence between observations on the same subject by specifying a random subject effect producing an error structure often referred to as compound symmetry.

The SAS code used for the analyses was as follows:

```
PROC MIXED METHOD = ML;
CLASS subject id variable & classification variables such as sex;
MODEL <dependent variable> = <fixed effects variables> /S CL;
REPEATED/TYPE = CS SUBJECT = <subject id variable>;
```

where METHOD = ML specifies the maximum likelihood method for solving the regression, S requests a solution for the fixed effects estimates (β), CL requests confidence limits around β, REPEATED specifies the error covariance matrix (REPEATED was used instead of RANDOM to accommodate the large sample size), and TYPE = CS specifies the error covariance matrix to have compound symmetry (ie, any two observations on the same subject have a constant correlation regardless of time between replications; this correlation is called the intra-class correlation).

Other error structures are possible, such as an autoregressive within-subject error structure in which the correlation between two observations decreases as the time interval increases. With larger numbers of observations per subject, it is possible to have both a random subject effect and autoregressive within-subject errors (15). In the analyses presented here, the use of an autoregressive error structure was not as good a fit to the data as compound symmetry on the basis of −2 log likelihood. The data set for the model with a random subject effect and unequal number of records per subject contained a record for each visit or observation (Table 2).

Four methods of analyzing the data were compared (Table 3). Use of standard regression methods may result in investigators having to eliminate observations to produce a balanced design or avoid correlation resulting from multiple observations on the same subject or to simplify analyses when the value of covariates (eg, BMI) changes over time. This can be done by restricting the data set to one observation per subject (model A includes 1 baseline visit for 870 subjects) or by restricting the data set to an equal number of observations per subject and using average values for variables that vary over time (model B includes 552 subjects with two visits, ie, 552 records).

**Table 1**

Subject characteristics according to type of visit and visit order, San Luis Valley, CO, 1984–1992

<table>
<thead>
<tr>
<th>Subjects with only a baseline visit (n = 318)</th>
<th>Baseline visit in subjects with two visits (n = 552)</th>
<th>Second visit in subjects with two visits (n = 552)</th>
<th>Subjects with only a second visit (n = 58)</th>
<th>All visits (n = 1480)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic (%)</td>
<td>44.3</td>
<td>40.0</td>
<td>40.0</td>
<td>43.1</td>
</tr>
<tr>
<td>Female (%)</td>
<td>56.9</td>
<td>45.7</td>
<td>45.7</td>
<td>48.3</td>
</tr>
<tr>
<td>Age (y)</td>
<td>49.7 ± 13.3</td>
<td>51.3 ± 11.4</td>
<td>55.5 ± 11.3</td>
<td>62.9 ± 9.8</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>25.1 ± 4.2</td>
<td>25.4 ± 3.9</td>
<td>25.7 ± 4.0</td>
<td>26.6 ± 5.1</td>
</tr>
<tr>
<td>Saturated fat intake (g/d)</td>
<td>28.8 ± 21.3</td>
<td>32.9 ± 19.2</td>
<td>26.1 ± 15.0</td>
<td>23.4 ± 14.4</td>
</tr>
<tr>
<td>Energy intake (kJ/d)</td>
<td>7606.1 ± 4380.2</td>
<td>8510.6 ± 3792.5</td>
<td>7518.5 ± 3254.5</td>
<td>7084.1 ± 3638.3</td>
</tr>
<tr>
<td>LDL cholesterol (mmol/L)</td>
<td>3.47 ± 1.03</td>
<td>3.57 ± 0.99</td>
<td>3.40 ± 0.92</td>
<td>3.33 ± 0.92</td>
</tr>
</tbody>
</table>

* x ± SD.
TABLE 2
Data file structure for three subjects in a random-effects analysis with unequal numbers of observations per subject

<table>
<thead>
<tr>
<th>Identification number</th>
<th>Age¹</th>
<th>Sex (male = 1, female = 2)²</th>
<th>BMI¹</th>
<th>Energy¹</th>
<th>Saturated fat¹</th>
<th>LDL¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>35.8</td>
<td>2</td>
<td>25.0</td>
<td>6559</td>
<td>25.6</td>
<td>4.61</td>
</tr>
<tr>
<td>001</td>
<td>40.1</td>
<td>2</td>
<td>25.8</td>
<td>5573</td>
<td>16.7</td>
<td>4.22</td>
</tr>
<tr>
<td>002</td>
<td>58.2</td>
<td>2</td>
<td>32.9</td>
<td>9224</td>
<td>38.6</td>
<td>4.18</td>
</tr>
<tr>
<td>003</td>
<td>41.3</td>
<td>1</td>
<td>22.5</td>
<td>11189</td>
<td>47.8</td>
<td>3.08</td>
</tr>
<tr>
<td>003</td>
<td>45.9</td>
<td>1</td>
<td>23.5</td>
<td>12092</td>
<td>36.1</td>
<td>3.27</td>
</tr>
</tbody>
</table>

¹ Time-varying covariate.
² Non-time-varying covariate.
³ Time-varying response.

Models A and B were run with the PROC MIXED code above but without the REPEATED statement. The model for subject i is:

\[
LDL_{ij} = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_3 \text{BMI}_i + \\
\beta_4 (\text{energy intake})_{ij} + \\
\beta_5 (\text{saturated fat})_{ij} + \epsilon_{ij}
\]

where the coefficient \( \beta_5 \) estimates the strength of the association between saturated fat intake and LDL concentrations, adjusting for age, sex, BMI, and energy intake, and \( \epsilon_{ij} \) is the error. Models C and D incorporated the REPEATED statement with a random subject effect:

\[
LDL_{ij} = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_3 \text{BMI}_i + \\
\beta_4 (\text{energy intake})_{ij} + \\
\beta_5 (\text{saturated fat})_{ij} + \gamma_i + \epsilon_{ij}
\]

where the model represents subject i at time j and \( \gamma_i \) is the level of the random subject effect for subject i and \( \epsilon_{ij} \) is the error. Sex is a non-time-varying attribute whereas the remaining variables were assessed at each observation time. In random-effects models C and D, saturated fat intake was also a time-varying covariate and the parameter estimate reflected the combined contribution of between-subject and within-subject associations with the response variable LDL. The mean saturated fat intake for a given subject carries between-subject information whereas the variation about the mean intake for the subject carries within-subject information.

Investigators may believe that even in a random-effects model, which can account for lack of independence between observations on the same subject, the data set should be balanced; correspondingly, model C includes 552 subjects with two observations, ie, 1104 records, in a random-effects model. Currently available methods and software, however, allow investigators to take advantage of information from all visits; thus, model D includes all 1480 available visits.

Power curves were estimated by using PASS 6.0 software (16). Power calculations as a function of effect size for a one-sample, two-tailed \( t \) test with an alpha level of 0.05 require a sample size and an SD. The sample size is used only to convert the SD to an SE. By using the estimated SEs of the \( \beta \) coefficients for saturated fat in models A–D (Table 3), power curves were calculated by entering a sample size of 100 and entering an SD obtained by multiplying the SEs by 10 (the square root of 100).

TABLE 3
Parameter estimates from longitudinal data analyses predicting LDL cholesterol (mmol/L) through use of four analysis strategies

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model A: 870 subjects, 870 visits, 870 records²</th>
<th>Model B: 552 subjects, 1104 visits, 552 records¹</th>
<th>Model C: 552 subjects, 1104 visits, 552 records¹</th>
<th>Model D: 928 subjects, 1480 visits, 1480 records¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>( \beta ) = 2.38</td>
<td>( \beta ) = 2.37</td>
<td>( \beta ) = 2.48</td>
<td>( \beta ) = 2.57</td>
</tr>
<tr>
<td>Age (10 y)</td>
<td>( 0.17 \pm 0.03^6 )</td>
<td>( 0.16 \pm 0.03 )</td>
<td>( 0.08 \pm 0.03 )</td>
<td>( 0.08 \pm 0.02 )</td>
</tr>
<tr>
<td>Sex (F = M)</td>
<td>( -0.14 \pm 0.07 )</td>
<td>( -0.08 \pm 0.09 )</td>
<td>( -0.04 \pm 0.08 )</td>
<td>( -0.04 \pm 0.06 )</td>
</tr>
<tr>
<td>BMI (5 units)</td>
<td>( 0.08 \pm 0.04 )</td>
<td>( 0.06 \pm 0.05 )</td>
<td>( 0.10 \pm 0.04 )</td>
<td>( 0.09 \pm 0.03 )</td>
</tr>
<tr>
<td>Energy (100 kJ/d)</td>
<td>( -0.02 \pm 0.03 )</td>
<td>( 0.04 \pm 0.05 )</td>
<td>( -0.02 \pm 0.02 )</td>
<td>( -0.04 \pm 0.06 )</td>
</tr>
<tr>
<td>Saturated fat (20 g/d)</td>
<td>( 0.05 \pm 0.07 )</td>
<td>( 0.12 \pm 0.11 )</td>
<td>( 0.11 \pm 0.05 )</td>
<td>( 0.14 \pm 0.04 )</td>
</tr>
</tbody>
</table>

¹ Model A included only baseline visits, model B included only subjects with two visits and time-varying data averaged across those visits, model C included only subjects with two visits in a random-effects model, and model D included all visits in a random-effects model. Number of records is number of records included in the analysis.
² Total model variance estimate = 0.962
³ Total model variance estimate (equal to the between-variance plus one-half the within-variance when averaging over two replicates.) = 0.750.
⁴ Model variance estimates as follows: total = 0.896, between = 0.638, and within = 0.258.
⁵ Model variance estimates as follows: total = 0.932, between = 0.671, and within = 0.261.
⁶ ± SE.
RESULTS

Models A and B estimated the association between saturated fat intake and serum LDL concentrations across subjects. With use of a single observation per subject (model A, Table 3), 20 g saturated fat (1 SD in the data) was associated with an increase in LDL of 0.05 mmol/L. This increase in LDL was not statistically different from zero. In model B, in which time-varying variables were averaged, the parameter estimate was in the opposite direction and statistically nonsignificant. Use of both models C and D resulted in a highly significant positive association between intake of saturated fat and serum LDL concentration. If saturated fat intake was reduced from 30 to 10 g while energy intake was held constant, LDL would be expected to decrease by 0.14 mmol/L (model D).

How the estimated SEs for models A–D altered study power as a function of the effect size is shown in Figure 1. The smallest SE in model D resulted in the greatest study power. The larger effect size and lower SE in model D are graphically presented in Figure 2, where the estimated mean and 95% CI (shaded area) for serum LDL concentration are shown at various intakes of saturated fat.

DISCUSSION

In this observational data set, associations between saturated fat intake and serum LDL concentrations were detectable only in the random-effects models that included repeated measurements in individuals (models C and D). These findings suggest that strong within-subject effects between saturated fat and serum LDL are not apparent in cross-sectional data (such as in model A) and are detectable only when longitudinal within-subject associations are measured (models C and D). This conclusion is consistent with earlier hypotheses that between-subject differences in genetic (or unmeasured environmental) determinants of serum LDL concentrations were limiting our ability to see associations between diet and serum lipids in cross-sectional studies. The data in the present study show that associations between diet and serum lipids can be observed by using the 24-h diet recall method and in an observational study without a dietary change experiment.

In the random-effects models with time-varying covariates, the regression coefficients combined information across subjects (cross-sectional) with information within subjects over time and under different conditions (longitudinal). In addition, separate variance components were estimated for the within- and between-subject effects (Table 3). Study power can be improved by increasing sample size, effect size, or precision of the parameter estimates of either the across-subject or within-subject effects. With use of models C and D, we expect enhanced precision because of the multiple measures of saturated fat and LDL in the same subject. In addition, this increased precision removes some of the bias toward zero in the parameter estimate otherwise resulting from measurement error. This results in a larger observed effect size. The random subject effect also allows more precise and accurate within-subject comparisons by controlling for other sources of variability and potential confounding. This occurs because the subject is acting as his or her own control with respect to unmeasured non-time-varying factors that influence fasting LDL concentrations and that vary from subject to subject (eg, genetic determinants of LDL). Confounding will be prevented if these unmeasured factors are also associated with saturated fat intake.

To the extent that subjects change their diet (day-to-day or systematically over time), we have a natural experiment. With two observations and two different saturated fat intakes, the analysis would be similar to a paired t test. However, instead of two treatments (high saturated fat compared with low saturated fat), in observational data we have a continuum of intakes with respect to nutrient composition. Consequently, these longitudinal analyses take advantage of varying conditions at different points in time and allow us to evaluate whether changes in diet are associated with changes in LDL concentrations within subjects.

Model D also takes advantage of the additional within-subject information from individuals without losing the between-subject information for subjects with only a single visit. The negligible difference in the magnitude and statistical significance of the association between saturated fat and LDL in models C and D suggested that the increased power of the random-effects analysis was driven primarily by the within-subject comparisons, in which the subjects served as their own controls. The approach used in model D will be increasingly more powerful than that used in model C as the number of subjects with only one observation increases (conversely, if the data set has relatively few subjects with missing observations, the small increase in sample size by inclusion of those subjects will likely have a negligible effect) and when the effect size across subjects is consistent with the within-subject effect.

It is interesting to consider why the parameter estimates for age were so different in models A and B than in models C and D. One possible explanation is that the earlier birth cohorts had higher total and LDL cholesterol concentrations than the later birth cohorts yet within-subject cholesterol concentrations were decreasing with age. Data from the National Health and Nutrition Examination Surveys do show increases in total and LDL cholesterol concentrations with age within any single cross-sectional survey and decreases when an aging cohort is followed over time (17). We suggest that the strong age effect across subjects in models A and B represents primarily a birth cohort effect and that declining cholesterol concentrations with aging attenuate this birth cohort effect in models C and D, in which the information across subjects was combined with information within subjects over time. We are currently working to develop methods to distinguish effects across subjects from within-subject effects.
Potential bias introduced by unequal numbers of observations across subjects in model D (due to nonresponse, missing data, or exclusion of visits because of use of medications known to affect serum lipids) was evaluated by including an indicator variable for number of visits. This indicator variable was not related to the response variable (LDL) and did not change coefficients or $P$ values for between-subject effects. These findings argue against a systematic bias that would occur if missing observations responsible for the unbalanced design were related to saturated fat intake and LDL concentrations across subjects.

We recommend using currently available methods for analyzing longitudinal data that incorporate random subject effects even when replicate measurements are available on only a subset of subjects and when time between replicates may be long. Use of standard regression methods that force the investigator to use only one measure per subject or to average measures that vary across visits does not take advantage of information about the covariation of variables within subjects. In the analyses presented here in which replications were used, measurement error was reduced in the estimation of cross-sectional (across subject) effects and information on the strong within-subject effects was captured by using time-varying covariates in a random-effects model.

We acknowledge the staff and participants of the San Luis Valley Diabetes Study who made this study possible; the Nutrition Coordinating Center at the University of Minnesota for interviewer training, data coding, and nutrient analysis of the 24-h recalls; and the laboratory of the University of Colorado Health Sciences Center General Clinical Research Center for conducting blood chemistry analyses.

REFERENCES