

Number of days, number of subjects, and sources of variation in longitudinal intervention or crossover feeding trials with multiple days of measurement

Gary K. Grunwald^{1,2*}, Debra K. Sullivan⁴, Mary Hise⁴, Joseph E. Donnelly³, Dennis J. Jacobsen³, Susan L. Johnson¹ and James O. Hill¹

¹Center for Human Nutrition, Box C-263, University of Colorado Health Sciences Center, Denver, CO, 80262, USA

²Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, CO, USA

³Energy Balance Laboratory, Department of Health, Sport and Exercise Sciences, The University of Kansas, Lawrence, KS, USA

⁴Department of Dietetics and Nutrition, University of Kansas Medical Center, Kansas City, KS, USA

(Received 2 May 2002 – Revised 7 July 2003 – Accepted 28 July 2003)

Dietary studies are often conducted as longitudinal intervention or crossover trials using multiple days of measurement on each subject during each of several measurement periods, and determining the required numbers of days and subjects is important in designing these studies. Linear mixed statistical models were used to derive equations for precision, statistical power and sample size (number of days and number of subjects) and to obtain estimates of between-subject, period-to-period, and day-to-day variation needed to apply the equations. Two cohorts of an on-going exercise intervention study, and a crossover study of Olestra, each with 14 d of measurement/subject per period, were used to obtain estimates of variability for energy and macronutrient intake. Numerical examples illustrate how the equations for calculating the number of days or number of subjects are applied in typical situations, and sample SAS code is given. It was found that between-subject, period-to-period, and day-to-day variation all contributed significantly to the variation in energy and macronutrient intake. The ratio of period-to-period and day-to-day standard deviations controls the trade-off between the number of days and the number of subjects, and this remained relatively stable across studies and energy and macronutrient intake variables. The greatest gains in precision were seen over the first few measurement days. Greater precision and fewer required days were noted in the study (Olestra) that exerted greater control over the subjects and diets during the feeding protocol.

Dietary assessment: Dietary intake: Dietary variability: Within-subject variation: Between-subject variation

Dietary intervention trials often involve the measurement of each subject for several days during at least two separate periods. This situation arises in many longitudinal intervention studies (for example, see Bowen *et al.* 1996; Lanza *et al.* 1996; Dolecek *et al.* 1997; Lauer *et al.* 2000). Dietary intake for each subject is measured at baseline for several days, subjects are allocated to one of two or more groups (for example, intervention and control), and each subject is again measured for several days during some later period. A similar situation can arise in crossover studies, where subjects are given one of two or more treatments and dietary intake is measured for several days on each subject. Following a washout period, subjects are given a different treatment and dietary intake is again assessed for several days (for example, see Stubbs *et al.* 1995, 1998; Poppitt & Swann, 1998; Johnstone *et al.* 2000).

A substantial literature exists on sample-size estimation in nutrition studies. Several authors have given methods for estimating the number of days of dietary intake data required to estimate a given individual's mean intake to a specified precision (Beaton *et al.* 1979; Basiotis *et al.* 1987; Nelson *et al.* 1989; Hartman *et al.* 1990; Bellu *et al.* 1995). An estimate of the number of days may be needed in order to estimate the mean dietary intake of a group of subjects (Beaton *et al.* 1979; Basiotis *et al.* 1987; Gay, 2000) or to compare the mean intake between two groups (Nelson *et al.* 1989). Also, an estimate of the number of days may be required in order to rank individuals in terms of intake (Nelson *et al.* 1989; Hartman *et al.* 1990), or to classify subjects into percentiles of consumption (Liu *et al.* 1978; Marr & Heady, 1986; Hartman *et al.* 1990). Beaton (1994) has provided a thorough discussion of potential biases in dietary measurement methods

Abbreviations: D , number of days/subject per period; N , number of subjects/group; $R_P = \sigma_P/\sigma_E$, period variation as a fraction of daily variation; σ_E , day-to-day standard deviation; σ_P , period-to-period standard deviation; σ_S , between-subject standard deviation.

*Corresponding author: Dr Gary K. Grunwald, fax +1 303 315 9976, email Gary.Grunwald@UCHSC.edu

and their effects on several of the above research questions. Some authors have also considered trade-offs between increasing the number of subjects and increasing the number of days to estimate the mean dietary intake of a group of subjects (Beaton *et al.* 1979; Gay, 2000), or to estimate correlations between two nutritional variables (Liu *et al.* 1978; Rosner & Willett, 1988). Beaton *et al.* (1979), Beaton *et al.* (1983), Marr & Heady (1986), Nelson *et al.* (1989), Borrelli *et al.* (1989), Hartman *et al.* (1990), and Bhargava *et al.* (1994) have studied and estimated daily and between-subject variability in nutrition studies. Willett (1998) has summarised some of the results from these studies.

Because of the complex design of longitudinal or cross-over studies with multiple measurements, specialised methods for estimating statistical power and sample size are required. The first purpose of the present study is to provide statistical methods for estimating the sources of variation in dietary longitudinal intervention or crossover trials with multiple days of measurement on each subject during each period. The second purpose of the present study is to provide statistical methods for estimating the number of days and number of subjects needed to achieve a given power or detectable effect size in a proposed study. Further purposes of the present study are to study the relative gains of more days *v.* more subjects in these studies, and to provide data on the magnitudes of the sources of variation involved for energy and macronutrient intakes. As a by-product, the proposed statistical models and software describe efficient modern methods for analysing data from such trials, even in the presence of unequal numbers of days or missing days for some subjects.

Methods

Subjects and measurements

The present study used data for energy intake (MJ/d), macronutrient intake (g/d), and macronutrient composition (percentage of energy) from two sources. Cohorts 1 and 2 are part of an ongoing 16-month supervised exercise intervention trial conducted at the Department of Health, Sport and Exercise Science at the University of Kansas. The subjects were weight stable (± 2.27 kg for 6 months), moderately overweight (defined as BMI of 27.3–32.3 kg/m² for females and 27.8–31.1 kg/m² for males) healthy, sedentary college-age students. The study was a two-group longitudinal intervention study with the subjects randomly assigned to a control group or an exercise intervention group. The exercise intervention subjects engaged in 45 min of supervised, verified exercise at 70% VO_{2max} for 5 d/week. Energy and nutrient intake were measured at six separate 14 d periods throughout the 16-month trial. Data from the baseline and 4-month measurements were used. The subjects were advised to maintain normal *ad libitum* eating habits throughout the study. At least two meals/d were consumed in the university cafeteria. The cafeteria is a large food court offering ten entrees/meal plus numerous stations with additional items. Dietary intake was measured by observer-recorded plate waste. Research staff interviewed the subjects at each meal to ascertain the foods

and beverages consumed outside the cafeteria. The subjects were advised to bring in labels from any foods consumed outside the cafeteria. The subjects were also instructed on routine household portion sizes, and three-dimensional food models were available for subjects to aid in portion-size estimations for snack recalls. Data were used from two cohorts; these cohorts began the study around the start of 1997 and 1998 respectively.

The third dataset arose from a randomised double-blind crossover trial of the fat replacer Olestra conducted at the Center for Human Nutrition, University of Colorado Health Sciences Center, Denver (Hill *et al.* 1998). The subjects were recruited from the Fort Collins, Colorado area and were required to have a BMI between 19 and 35 kg/m², to be weight stable (± 2.27 kg for 6 months), to be between 18 and 65 years old, and to be in good health and not pregnant or lactating. Energy and nutrient intake were measured during two 14 d periods separated by a 7 d washout period. During each 14 d period each subject consumed foods containing either a covert Olestra supplement or a placebo (triacylglycerol), and they received the opposite treatment during the second period. The order of treatment was determined randomly for each subject. Breakfast and dinner were consumed on-site at the feeding laboratory. The subjects were required to eat core menu food items, to which Olestra was substituted during the intervention period but which were normal foods during the control period. Core foods comprised about 25% of the subjects' daily energy intake. The subjects could self-select additional items from a buffet. The core food item at breakfast was a muffin or biscuit, and optional buffet items included cereal, bread, eggs, hash browns, sausage, fruit, milk, and condiments. Core food items at dinner were one entrée (for example, chicken pot pie, corn muffin, pizza) and one snack or dessert item (for example, cookies, tortilla chips), and optional buffet items included vegetables, chicken, lasagne, ice-cream, yoghurt, milk, soda, fruit, and cake. Lunch and snacks were provided to subjects in backpacks for off-site consumption, and contents were pre- and post-weighed to determine consumption. These additional food items were provided in large quantities to ensure that availability did not limit intake. Lunch items included sandwiches, microwave entrées, soups, fruit, crackers, crisps (potato chips), candy, cookies, milk, juice, and soda. All foods were served in discrete, weighed portions, and the amounts of foods eaten were recorded by laboratory personnel. More detailed information on the study is given in Hill *et al.* (1998).

Subject characteristics are shown in Table 1. The subjects in the two exercise cohorts were younger college-aged students and were more overweight than those in the Olestra study.

Statistical analysis

Data format. Studies can be considered with *M* and *N* subjects in groups 1 and 2 respectively, with *C* and *D* days of measurement on each subject during periods 1 and 2 respectively. Each subject is in only one group but is measured for several days during each period. Various means are denoted as follows: *p*₁ is the period 1 mean,

Table 1. Subject characteristics at baseline for three studies*
(Mean values and standard deviations)

	Cohort 1				Cohort 2				Olestra			
	Females		Males		Females		Males		Females		Males	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Number of subjects (<i>N</i>)	11		8		20		17		28		23	
Missing days	5		3		13		6		4		2	
Age (years)	21	4	22	3	23	5	23	4	44	11	41	10
Weight (kg)	77	10	101	11	82	9	95	13	72	14	83	16
Body fat (%)	38	4	29	4	36	5	27	4	21	4	11	2
BMI (kg/m ²)	29	3	31	3	30	3	30	3	26	5	25	4
Energy intake (MJ/d)	9.97	2.72	14.45	3.18	10.90	2.92	14.04	4.71	10.81	2.07	14.55	3.10
Fat intake (g/d)	79	32	137	43	91	37	121	57	87	26	114	28
Dietary fat (%)	30	9	35	7	31	8	32	8	30	6	30	5

* Cohorts 1 and 2 are from the longitudinal exercise intervention trial (for a description see p. 1088). The Olestra data are from the crossover trial (Hill *et al.* 1998). In all three cases dietary intake was measured for 14 d during each of two periods, though occasional missing days occurred as shown.

g_1 is the group 1 mean, gp_{12} is the period 2 mean for group 1, etc. Data of this form can arise from a two-period, two-group longitudinal intervention trial where periods 1 and 2 denote pre-intervention and post-intervention periods. Subjects are allocated into one of two groups (for example, intervention and control) and are measured for several days during each period.

Data from a two-treatment (A or B), two-period (2×2) crossover design (Jones & Kenward, 2003) can also be arranged in this form, by letting period 1 represent the period where treatment A was given and period 2 represent the period where treatment B was given, regardless of treatment order. Each subject is measured for several days under each treatment condition. In this case, ‘group’ represents the order of treatment administration (A then B or B then A, sometimes called ‘sequence’ in the crossover study literature), which is typically random. It is assumed that there are no systematic differences between the first and second measurement period, and that carry-over effects are eliminated by the use of a sufficient washout period.

Statistical models. A linear mixed model appropriate for the studies being considered is:

$$Y_{ijkl} = m + g_i + p_j + gp_{ij} + s_{ik} + ps_{ijk} + e_{ijkl}, \quad (1)$$

where Y_{ijkl} is the response measurement on day d ($d = 1, \dots, C$ or D) of period j ($j = 1, 2$) for subject k ($k = 1, \dots, M$ or N) in group i ($i = 1, 2$). The effects m , g_i , p_j and gp_{ij} are fixed effects for the overall mean, group, period and group \times period interaction (the difference in period effect between the two groups). The fixed effects describe systematic differences that affect all subjects in the same way (for example, group differences, changes over time, or differences between groups in the change over time). The effects s_{ik} , ps_{ijk} and e_{ijkl} are random effects that describe subject-specific patterns, and are assumed to be independent and normal with mean 0 and standard deviations σ_s , σ_p and σ_e respectively. The effect e_{ijkl} and σ_e describe the day-to-day variation for a given subject during a given period, due to natural variation in behaviour or food supply. Thus, about 95 % of days for a given

subject during a given period would lie within about $\pm 2\sigma_e$ of that subject’s mean level during that period. The effect ps_{ijk} and σ_p describe the variation between a subject’s true mean response during period 1 and period 2, adjusting for overall period and group differences. This variation is due to changes in behaviour, season, etc. The effect s_{ik} and σ_s describe the subject-to-subject variation in mean response, adjusting for period and group differences. This variation is due to differences in behaviour, environment, body size and composition, etc. Model (1) was also used by Hartman *et al.* (1990) for estimating individuals’ dietary intake and for ranking and classifying individuals.

A major benefit of repeated-measures designs is that estimates of treatment effects can be based on within-subject changes from one period to another. The period variance is the variation in these within-subject changes above and beyond the daily variation. In contrast, the fixed effects for period and group describe systematic changes for all subjects in a group. Observing each subject for multiple days during each period ‘averages out’ the daily variation but does not eliminate the period variation (see the factor $1/D$ in equation (2)). Another way of looking at the period variance is that it allows measurements on a given subject during a given period to be more highly correlated than measurements on a given subject during different periods. This is typically desirable because the periods are separated by more time than the daily measurements. When analysing studies with different numbers of days/period per subject, or designing studies that are to have a different number of days than were used in obtaining the prior variance estimates for sample size, it is necessary to separate the period and daily variances.

The fixed effect mean parameters and random effect variance parameters in model (1) were estimated using linear mixed models (Laird & Ware, 1982; Verbeke & Molenberghs, 2000). SAS statistical software (SAS Institute Inc., Cary, NC, USA) was used, and details are given in Appendix 1. Assumptions of normality and equal variance were examined graphically. CI for standard deviation parameters are based on Chi-square distributions (SAS Institute Inc., Cary, NC, USA). Significance tests of

standard deviation parameters equal to zero were carried out using a mixture of Chi-square distribution tests (Verbeke & Molenberghs, 2000).

Power and sample size. Studies can be first considered which use equal numbers of subjects in the groups; $M = N$, and equal numbers of days/period; $C = D$. It is most useful and interpretable to base sample-size estimates on the comparison of most interest in a given study. For longitudinal studies the most useful estimate of the effect of the intervention is generally the interaction effect describing the difference between the intervention and control groups in the increase from pre-intervention to post-intervention; $(gp_{22} - gp_{21}) - (gp_{12} - gp_{11})$. The variance of this interaction effect is:

$$V = \frac{4}{N} \sigma_E^2 \left[R_P^2 + \frac{1}{D} \right]. \quad (2)$$

The period variation can be expressed as a fraction of the daily variation ($R_P = \sigma_P / \sigma_E$) since the latter represents the natural 'background' variation. The widths of CI and the magnitudes of detectable differences in power analysis are proportional to the standard error (square root of V in equation (2)). The variance V in equation (2) can be rewritten as:

$$V = \frac{4}{N} \left[\sigma_P^2 + \frac{\sigma_E^2}{D} \right]$$

showing it to be a combination of the period and daily variances, and the latter term decreasing with more days.

Consider designing a longitudinal intervention study to achieve $100 \times (1 - \beta) \%$ power to detect a difference between groups in the change from pre- to post-intervention (interaction) of magnitude Δ using two-sided level α tests. Using N subjects/group, the required number of days/subject per period is:

$$D = \frac{1}{\frac{N \times \Delta^2}{4 \times (Z_{1-\beta} + Z_{1-\alpha/2})^2 \times \sigma_E^2} - R_P^2}. \quad (3)$$

(For example, for 80 % power, $Z_{1-\beta} = Z_{0.80} = 0.84$ and for power 90 %, $Z_{1-\beta} = Z_{0.90} = 1.28$; for two-sided level 0.05 tests, $Z_{1-\alpha/2} = Z_{0.975} = 1.96$.) Using D days/subject per period, the required number of subjects/group is:

$$N = \frac{4 \times (Z_{1-\beta} + Z_{1-\alpha/2})^2 \times \sigma_E^2 \times (R_P^2 + \frac{1}{D})}{\Delta^2}. \quad (4)$$

Equations (3) and (4) can be solved as usual for power $1 - \beta$ or for detectable difference Δ if those are of interest. There are many ways to rewrite these equations, but equations (3) and (4) involve only quantities that can usually be specified (for example, N or D and Δ) or estimated (for example, σ_E) easily by investigators designing studies, as well as the ratio R_P (see examples, pp. 1091–1092).

For crossover designs, the most useful estimate of the effect of the treatment will typically be the overall treatment effect, $p_2 - p_1$ in the present notation. In 2×2 crossover studies, the group effect represents the order of treatments (A then B v. B then A). The interaction effect

represents a combination of the effect of the period of treatment administration (first or second), any differences in the effects of treatments during the first and second periods, plus any carry-over effect of treatments from the first to the second period of treatment administration. These effects are typically of little or no practical interest and tend to be small in comparison with the treatment effect, and so here the interaction term is omitted from equation (1) for crossover studies. Equations (2), (3) and (4) apply to estimating the treatment effect for crossover designs, with the slight change that the factor of '4' that appears in each equation is omitted.

Note that equation (3) for the required number of days can yield a negative value. In this case, the number of subjects N and the detectable effect size Δ are too small and it is not possible to measure enough days/subject to achieve the necessary power. Equation (4) for the required number of subjects always gives a positive result, so it is always possible to increase the number of subjects to achieve the required power regardless of the size of the effect Δ to be detected or the number of days D of measurement/subject.

Equations for studies with unequal numbers of days and/or subjects are given in Appendix 2. These can be useful in designing unbalanced studies or for exploring the sensitivity of power and sample size to various designs under consideration.

Results

Estimated standard deviations for sources of variation in energy and macronutrient intake

Table 2 shows the estimated random effect standard deviations $\hat{\sigma}_S$, $\hat{\sigma}_P$, and $\hat{\sigma}_E$ obtained as in Appendix 1 for energy intake (MJ/d), macronutrient intake (g/d), and macronutrient composition (percentage of energy) for each of the three studies. These are of interest in showing the relative magnitudes of the sources of variation, and also in providing estimates of σ_E and R_P , which are needed in designing future studies. Between-subject standard deviations $\hat{\sigma}_S$ and day-to-day standard deviations $\hat{\sigma}_E$ were larger than period standard deviations $\hat{\sigma}_P$ for all outcomes in all studies. Day-to-day variation as quantified by $\hat{\sigma}_E$ was smaller in the Olestra study (Hill *et al.* 1998) than in the two exercise study cohorts for all energy and macronutrient variables. Period standard deviations $\hat{\sigma}_P$ were also smaller for the Olestra study (Hill *et al.* 1998) than for the two exercise cohorts for all energy and macronutrient variables. There are no apparent patterns in between-subject standard deviations between the studies. Period variation was significantly different from 0 ($P < 0.05$) for all outcomes in all studies except for energy intake, protein intake, and protein percentage in Cohort 1 and fat percentage and protein percentage in Cohort 2. This means that in most cases there was a greater difference between the averages for the two measurement periods on a given subject than would result from the natural variation of daily measurements within each period.

Table 2 also shows estimates of the ratio $R_P = \sigma_P / \sigma_E$. Estimates of R_P are between 0.16 and 0.46 for all outcomes in all studies (i.e. period standard deviations are 16 to 46 %

Table 2. Variability of between-subjects (σ_S), period (σ_P), day-to-day (σ_E) and ratio of σ_P/σ_E (R_P) for seven energy and macronutrient intake variables in three datasets*

(Estimated standard deviations and 95% confidence intervals)

	Energy (MJ/d)	FAT (g/d)	CHO (g/d)	PRO (g/d)	FAT (% energy)	CHO (% energy)	PRO (% energy)
Cohort 1							
σ_S	2.75	31.1	81.6	24.4	3.38	4.98	1.77
95 % CI	2.02, 4.30	22.7, 49.4	59.5, 129.8	18.0, 38.1	2.24, 6.79	3.55, 8.34	1.28, 2.87
σ_P	0.82	11.7	29.8	7.1	2.82	2.53	0.55
95 % CI	0.52, 1.93	7.8, 23.1	19.2, 65.9	4.4, 17.0	1.95, 5.05	1.66, 5.20	0.28, 3.85
σ_E	2.70	30.9	91.7	23.8	6.14	7.08	2.89
95 % CI	2.54, 2.88	29.1, 33.0	86.3, 97.8	22.3, 25.3	5.78, 6.55	6.66, 7.56	2.72, 3.08
R_P	0.30	0.38	0.32	0.30	0.46	0.36	0.19
95 % CI	0.19, 0.72	0.25, 0.75	0.21, 0.72	0.19, 0.72	0.32, 0.82	0.23, 0.73	0.10, 1.33
Cohort 2							
σ_S	2.61	30.1	82.4	23.7	3.77	4.48	1.86
95 % CI	2.05, 3.60	23.8, 40.8	64.8, 113.3	18.8, 32.1	2.95, 5.22	3.45, 6.36	1.46, 2.57
σ_P	1.26	12.0	35.1	8.0	1.14	2.27	0.58
95 % CI	0.94, 1.92	8.7, 19.0	25.0, 58.9	5.4, 15.4	0.61, 6.02	1.57, 4.10	0.32, 2.51
σ_E	3.22	34.2	112.6	30.9	7.17	8.18	3.45
95 % CI	3.08, 3.37	32.7, 35.8	107.7, 117.9	29.6, 32.3	6.86, 7.51	7.83, 8.57	3.30, 3.61
R_P	0.39	0.35	0.31	0.26	0.16	0.28	0.17
95 % CI	0.29, 0.60	0.26, 0.56	0.22, 0.52	0.17, 0.50	0.09, 0.84	0.19, 0.50	0.09, 0.73
Olestra							
σ_S	2.82	21.5	111.7	27.1	3.65	4.65	1.72
95 % CI	2.34, 3.54	17.8, 27.2	92.9, 140.2	22.5, 33.9	3.02, 4.62	3.84, 5.89	1.42, 2.18
σ_P	0.45	4.7	19.9	3.5	0.78	1.12	0.45
95 % CI	0.32, 0.73	3.2, 8.6	14.9, 30.0	2.3, 6.6	0.51, 1.72	0.79, 1.93	0.32, 0.73
σ_E	1.64	20.4	62.6	15.4	3.95	4.49	1.66
95 % CI	1.59, 1.71	19.6, 21.2	60.3, 65.1	14.9, 16.0	3.81, 4.11	4.32, 4.66	1.60, 1.72
R_P	0.27	0.23	0.32	0.22	0.20	0.25	0.27
95 % CI	0.20, 0.44	0.16, 0.42	0.24, 0.48	0.15, 0.42	0.13, 0.43	0.18, 0.43	0.19, 0.44

CHO, carbohydrate; PRO, protein.

*For details of datasets and procedures, see Table 1 and p. 1088.

of daily standard deviations), with most between 0.2 and 0.4 and median value 0.275. There are no consistent differences between variables or studies. A small value of R_P indicates that subject averages between two periods vary mainly due to the variability in daily measurements, and that there is relatively little tendency for a subject's true mean value to vary from period to period (aside from treatment effects).

The estimation of the standard deviations (Table 2) was repeated separately for men and women (results not shown). There was a tendency for the day-to-day standard deviations for energy (MJ/d) and macronutrient (g/d) intakes to be slightly greater for men than for women, probably because of the tendency toward greater intakes for men. No other consistent sex differences were noted. All of the patterns previously noted were still evident, although reduced precision due to fewer subjects made the estimates more variable and patterns less consistent. In particular no consistent differences were noted between men and women in estimates of R_P .

Gains in precision from additional days

The precision of a design depends on both the daily variation σ_E and the period variation expressed as $R_P = \sigma_P/\sigma_E$, as in equation (2). Smaller values of R_P lead to greater relative gains from using more days, because when R_P is small, most of the variation in subject averages

between periods is due to daily variation, and this is reduced by averaging more days. For example, when $R_P = 0.25$ the standard error (square root of V in equation (2)) using 7 d compared with using 1 d is:

$$\sqrt{\left[0.25^2 + \frac{1}{7}\right]} / \sqrt{\left[0.25^2 + \frac{1}{1}\right]} = 0.44$$

so a CI for the interaction effect is 44 % as wide using 7 d as the same interval would be using 1 d. The corresponding value for 14 d compared with 7 d is 0.81 (again assuming $R_P = 0.25$), so the CI using 14 d will be 81 % as wide as those using 7 d. Experimentation with these equations shows that relative gains/d are greatest for the first few days of measurement, but adequate precision may require many days of measurement. General statements about the effects of the number of days and the number of subjects are more difficult because they depend on the daily variation σ_E and on the difference to be detected, but equations (3) and (4) can be used to explore various scenarios, as seen later (p. 1092).

Example: designing a future study

As an example of estimating the sample size for a future study, consider designing a two-group, two-period longitudinal intervention study with a protocol similar to that used for Cohort 1. Equal numbers of subjects/group

($M = N$) and equal numbers of days/subject per period ($C = D$) are assumed. Suppose the number of days D needs to be found that are required to obtain 80% power using two-sided level 0.05 tests to detect a difference of 1.25 MJ/d ($\Delta = 1.25$ MJ (300 kcal)/d) between groups in the change in energy intake from baseline to follow-up. Assuming the data from Cohort 1 are available, the estimated values of $\hat{R}_P = 0.30$ and $\hat{\sigma}_E = 2.70$ MJ/d in Table 2 are substituted into equation (3) for the estimated number of days D . Then, with $M = N$, i.e. thirty subjects/group, about $C = D = 9$ d of measurement/subject per period would be required. With twenty subjects/group, about 22 d/subject per period would be needed, and with ten subjects/group it would not be possible to obtain 80% power, regardless of the number of days measured.

Now suppose the same study is to be designed, but the only data available are from a single period of several days of measurement/subject rather than the two separate periods available in Cohort 1. Without further assumptions, it is not possible to estimate the required number of days without an estimate of $R_P = \sigma_P/\sigma_E$, which requires a second period of measurement. However, the results in Table 2 indicate that for studies similar to those in the present paper, R_P tends to be roughly in the range 0.2 to 0.4 with a median value of 0.275 for energy and macronutrient intakes. The other required quantity, σ_E , can be estimated as usual from a single period with multiple days of measurement on several subjects, using a linear mixed model similar to that in Appendix 1. The quantity σ_E can also be estimated as the root mean squared error from a one-way ANOVA model with a single factor for subject, or more simply by pooling the within-subject variances for each subject. For example, suppose the available data are the first 5 d on the six female control subjects in Cohort 1, shown in Table 3. The pooled day-to-day standard deviation $\hat{\sigma}_E$ is calculated as:

$$\sqrt{\frac{(4-1) \times 1.22^2 + (5-1) \times 1.76^2 + (5-1) \times 1.38^2 + (5-1) \times 2.44^2 + (5-1) \times 2.65^2 + (4-1) \times 1.14^2}{(4-1) + (5-1) + (5-1) + (5-1) + (5-1) + (4-1)}} = 1.91 \text{ MJ/d.}$$

Using this in equation (3) shows that with thirty subjects/group, between about 3 (if $R_P = 0.2$) and 4 (if $R_P = 0.4$) d of measurement/subject per period would be needed. With

Table 3. The first 5 d of energy intake (MJ/d) during the pre-intervention period for six female control subjects in Cohort 1, for use in the example study design

Day	Subject					
	1	2	3	4	5	6
1	4.48	7.19	10.49	11.12	10.41	11.86
2		9.23	8.92	13.51	13.71	9.82
3	6.63	6.45	10.56	11.88	15.26	
4	6.47	6.25	9.04	15.23	9.48	9.30
5	7.31	10.19	12.32	8.78	14.91	9.76
Subject SD	1.22	1.76	1.38	2.44	2.65	1.14

twenty subjects/group, between about 5 d (if $R_P = 0.2$) and 9 d (if $R_P = 0.4$) would be needed, and with ten subjects/group about 11 d (if $R_P = 0.2$) would be needed, but 80% power if $R_P = 0.4$ would not be able to be achieved. A conservative approach would use the larger value of R_P , and it should be emphasised that using any value for R_P not estimated from a protocol similar to that being considered represents a substantial assumption. Note that the example (p. 1092) refers to energy intake. For other variables the calculations would be modified with the appropriate standard deviations from Table 2.

Discussion

Summary

Statistical methods have been given for estimating power, detectable effect magnitude, and sample size for dietary intervention or crossover studies with multiple measurements on subjects in two or more groups during each of two or more periods. These methods were applied to data from two cohorts of a dietary intervention study and data from a crossover trial (Hill *et al.* 1998) to obtain estimates of the magnitudes of the sources of variability inherent in energy and macronutrient intakes in such studies and to provide ranges of some quantities necessary in designing future studies. These results may be useful to researchers designing and interpreting multi-day feeding trials.

Patterns in variance components

There was substantial between-subject variation as shown by the generally large and significant values of $\hat{\sigma}_S$, but there were no apparent patterns between studies, probably because this variation depends on subject variability within the sample, and on subject selection in a particular study. This variation does not appear in the precision (standard errors) or sample-size equations for the effects of most interest. This is a consequence and a major advantage of measuring each subject during each period. Day-to-day variation ($\hat{\sigma}_E$) was highly significant and of similar magnitude as between-subject variation in all three studies and for all energy and macronutrient intake variables. Other studies have also noted that day-to-day variation in dietary intake variables is substantial (Beaton *et al.* 1979, 1983; Marr & Heady, 1986; Borrelli *et al.* 1989; Nelson *et al.* 1989; Hartman *et al.* 1990; Bhargava *et al.* 1994). This is somewhat striking because at least some of the between-subject variation is due to differences in body size and composition and to subject-specific behaviours. However, the daily variation simply represents the natural daily variation in the behaviour of individuals, and this is seen to be very substantial. Period standard deviation was typically significant, and relative to day-to-day standard deviation (the ratio $\hat{R}_P = \hat{\sigma}_P/\hat{\sigma}_E$) remained fairly stable across studies and dietary intake variables, ranging from 16 to 46% of daily variation (R_P between 0.16 and 0.46), and in most cases between 20 and 40%.

Generalisability of the present results

The equations given in the statistical analysis section (pp. 1089–1090) and appendices apply to any numerical

variable that is roughly normally distributed. These methods apply to longitudinal or crossover designs with more than two groups and/or periods when power and sample sizes are to be based on estimating particular comparisons between groups and/or periods. This approach can be more useful and interpretable than basing estimations on ANOVA and *F* tests (for example, Helms, 1992; Verbeke & Molenberghs, 2000), where the detectable treatment effect is difficult to specify and interpret.

It could be possible, but the data are not here to show, that the estimates in Table 2 would not be greatly affected by the methodology used to estimate intake (for example, weighed food records, estimated food records or 24 h recalls), since measurement effects would be expected to be relatively small in comparison with the very large day-to-day variation. Since the subject-to-subject variability does not appear in power and sample-size equations for these study designs, the estimates of σ_P , σ_E and R_P in Table 2 would also not be expected to be greatly affected by different normal adult subject populations.

Day-to-day variation plays an important role in sample-size estimation, and this would be expected to depend on the feeding protocol. Day-to-day variation $\hat{\sigma}_E$ was smaller in the Olestra study (Hill *et al.* 1998) than in the exercise cohorts for all energy and macronutrient intake variables. The consistency and standardisation of the laboratory feeding along with core food requirements of about 25 % of energy in the Olestra study would probably reduce a subject's day-to-day variation compared with the unrestricted cafeteria feeding for the exercise cohorts, and the present study's data are consistent with that notion. Thus, in the Olestra study, fewer days of measurement are required to achieve the same level of precision. However, the reduced variation may also indicate a less realistic reflection of free-living behaviour. The variation observed in the exercise cohorts represents a closer estimation of free-living dietary intake as the subjects are presented with a wide variety of different foods at each meal. Slightly larger estimates of daily variation for men were noted though (data not shown), so it may be advisable to use gender-specific estimates of this quantity if possible when designing studies that will involve only one gender.

Period variation $\hat{\sigma}_P$ was also smaller in the Olestra study (Hill *et al.* 1998) than in the exercise cohorts, perhaps due to the much shorter length of time between periods in the Olestra study (7 d) compared with the exercise cohorts (4 months). Despite the larger daily and period standard deviations in the exercise cohorts compared with the Olestra study, the ratio $R_P = \sigma_P/\sigma_E$ remained fairly stable across studies and intake variables, ranging from 0.16 to 0.46 and in most cases between 0.20 and 0.40. It is not known how generally this would hold, since it will probably be affected by at least the feeding protocol and the time between periods. In designing a future study without data relevant to the particular protocol, the range that was observed in the present study could be used though that represents a substantial assumption. The Olestra study and the two exercise studies were carried out in settings that differed in several ways and were not designed to study differences in variation due to different study designs

and feeding protocols, so no definitive causes may be determined for differences between the studies.

Relation to other literature

Hartman *et al.* (1990) considered models similar to that given in equation (1), but for different purposes. However, some similarities with the present results can be derived. Hartman *et al.* (1990) noted period standard deviations to be less than daily standard deviations (these values can be derived from their Tables 1 and 2). They had estimated ratios $R_P = \sigma_P/\sigma_E$ of 0.50, 0.45, 0.53, 0.39, and 0.43 for energy, fat, carbohydrate, protein, and percentage fat by energy (the variables in common with those in the present study). These values are slightly higher than those in the present study, but are in general agreement considering the differences in location (Finland), time period (1984), procedure (self-reported food-use questionnaires for twelve periods of 2 d, separated by 16 d each, with n 162), and statistical methodology and software.

For the designs considered in the present study, Frison & Pocock (1992) and Schouten (1999) gave equations similar to equations (2) and (4) but in a different form. In their formulation, investigators must specify three quantities: the correlation between two observations on a subject during the same period, the correlation between two observations on a subject during different periods, and the total variance $\sigma_S^2 + \sigma_P^2 + \sigma_E^2$. In most nutritional settings, it can be expected that specifying these will be more difficult than specifying σ_E^2 and R_P as in the formulation in the present paper. In addition, the correlations and total variance depend on the subject variance, which depends on the population variance and sample selection methods and this may limit the usefulness of data from previous studies. However, if those quantities are easier for investigators to specify, or if complete data from a previous study using a similar population and sample selection methods are available (the correlations and total variance can be calculated from Table 2), using them along with the formulae of Frison & Pocock (1992) or Schouten (1999) will give the same power and sample-size estimates as the formulae in the present paper.

Frison & Pocock (1992) also consider an alternative analysis for longitudinal designs, using the baseline mean values as covariates in an analysis of the follow-up means. In some situations this analysis can be substantially more powerful than the analysis using model (1), but their methods along with the data from Table 2 indicate the gain in precision of this analysis compared with model (1) is at most 10–15 % and typically much less for the datasets and variables in the present study. The more complex analysis of covariance is correspondingly more difficult to interpret, particularly in studies with more than two periods. Those methods also do not apply to crossover designs.

Hsieh (1988) has given equations for the number of units/cluster and the number of clusters/group for intervention studies involving clusters of subjects which are similar to equations (3) and (4). When N is small, the *Z* values in equations (3) and (4) can be replaced by *t* values (Hsieh, 1988), though the equations then require iterative solution.

Practical and statistical considerations

Adjustments for between-subject effects such as sex, age, or body composition and/or within-subject effects such as weekday v. weekend can be made by including fixed effects in model (1), as shown in Appendix 1. The dependence of daily observations on the previous days' observations has also been reported in daily dietary data, typically in the form of auto-correlation (Hartman *et al.* 1990; Bhargava *et al.* 1994; Morgan *et al.* 1987; Tarasuk & Beaton, 1991). Model (1) can again be modified to account for this as shown in Appendix 1. However, it would be difficult to incorporate this effect into the design of future studies, mainly because the magnitude of the effect has tended to be relatively small and not in a consistent direction. It has also been noted that there tends to be greater correlation between consecutive days than between non-consecutive days (Hartman *et al.* 1990; Tarasuk & Beaton, 1992), and that this has consequences in designing dietary assessments. The methods provided in the present study apply to either consecutive or non-consecutive days of assessment, but the timing of assessment to be used in the actual study should also be used in obtaining preliminary variances for sample-size estimation.

Conclusions

The statistical methods that have been provided (for example, equations (2) to (4)) for estimating the number of days and the number of subjects required in longitudinal or crossover dietary trials with multiple measurements can help investigators assess the potential for a successful study. Two quantities describing variability in the proposed study (daily variation σ_E and period variation relative to daily variation $R_P = \sigma_P/\sigma_E$) must be estimated from a similar previous study if available. Estimates have been provided of these quantities for energy and macronutrient intake variables in two studies (Table 2), which can be used as in the examples on pp. 1091–1092 when no previous data are available, though this represents a substantial assumption.

Acknowledgements

This work was supported by NIH grant DK42549, by the Clinical Nutrition Research Unit at the University of Colorado Health Sciences Center, Denver, Colorado, USA (DK48520), and by NIH grant DK49181 awarded to the University of Kansas.

References

- Basiotis PP, Welsh SO, Cronin FJ, Kelsay JL & Mertz W (1987) Number of days of food intake records required to estimate individual and group nutrient intakes with defined confidence. *J Nutr* **117**, 1638–1641.
- Beaton GH (1994) Approaches to analysis of dietary data: relationship between planned analyses and choice of methodology. *Am J Clin Nutr* **59**, Suppl., 253S–261S.
- Beaton GH, Milner J, Corey P, *et al.* (1979) Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. *Am J Clin Nutr* **32**, 2546–2559.
- Beaton GH, Milner J, McGuire V, Feather TE & Little JA (1983) Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. Carbohydrate sources, vitamins, and minerals. *Am J Clin Nutr* **37**, 986–995.
- Bellu R, Ortisi MT, Riva E & Giovannini M (1995) Determination of intra- and inter-individual variability and its effect on the number of days required to assess the usual intake of a 1-year-old infant population. *Paediatr Perinat Epidemiol* **9**, 98–104.
- Bhargava A, Forthofer R, McPherson S & Nichaman M (1994) Estimating the variations and autocorrelations in dietary intakes on weekdays and weekends. *Stat Med* **13**, 113–126.
- Borrelli R, Cole TJ, Di Biase G & Contaldo F (1989) Some statistical considerations on dietary assessment methods. *Eur J Clin Nutr* **43**, 453–463.
- Bowen D, Clifford CK, Coates R, *et al.* (1996) The Women's Health Trial Feasibility Study in Minority Populations: design and baseline descriptions. *Ann Epidemiol* **6**, 507–519.
- Dolecek TA, Stamler J, Caggiula AW, Tillotson JL & Buzzard IM (1997) Methods of dietary and nutritional assessment and intervention and other methods in the Multiple Risk Factor Intervention Trial. *Am J Clin Nutr* **65**, Suppl., S196–S210.
- Frison L & Pocock SJ (1992) Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Stat Med* **11**, 1685–1704.
- Gay C (2000) Estimation of population distributions of habitual nutrient intake based on a short-run weighed food diary. *Br J Nutr* **83**, 287–293.
- Hartman AM, Brown CC, Palmgren J, *et al.* (1990) Variability in nutrient and food intakes among older middle-aged men: implications for design of epidemiologic and validation studies using food recording. *Am J Epidemiol* **132**, 999–1012.
- Helms RW (1992) Intentionally incomplete longitudinal designs: I. Methodology and comparison of some full span designs. *Stat Med* **11**, 1889–1913.
- Hill JO, Seagle HM, Johnson SL, *et al.* (1998) Effects of 14 d of covert substitution of Olestra for conventional fat on spontaneous food intake. *Am J Clin Nutr* **67**, 1178–1185.
- Hsieh FY (1988) Sample size formulae for intervention studies with the cluster as unit of randomization. *Stat Med* **8**, 1195–1201.
- Johnstone AM, Shannon E, Whybrow S, Reid CA & Stubbs RJ (2000) Altering the temporal distribution of energy intake with isoenergetically dense foods given as snacks does not affect total daily energy intake in normal-weight men. *Br J Nutr* **83**, 7–14.
- Jones B & Kenward MG (2003) *Design and Analysis of Cross-over Trials*, 2nd ed., Boca Raton, FL: Chapman and Hall/CRC.
- Laird N & Ware JH (1982) Random effects models for longitudinal data. *Biometrics* **38**, 963–974.
- Lanza E, Schatzkin A, Ballard-Barbash R, *et al.* (1996) The polyp prevention trial II: dietary intervention program and participant baseline dietary characteristics. *Cancer Epidemiol Biomarkers Prev* **5**, 385–392.
- Lauer RM, Obarzanek E, Hunsberger SA, *et al.* (2000) Efficacy and safety of lowering dietary intake of total fat, saturated fat, and cholesterol in children with elevated LDL cholesterol: the Dietary Intervention Study in Children. *Am J Clin Nutr* **72**, Suppl., S1332–S1342.
- Littell RC, Pendergast J & Natarajan R (2000) Modelling covariance structure in the analysis of repeated measures data. *Stat Med* **19**, 1793–1819.
- Liu K, Stamler J, Dyer A, McKeever J & McKeever P (1978) Statistical methods to assess and minimize the role of intra-individual variability in observing the relationship between dietary lipids and serum cholesterol. *J Chronic Dis* **31**, 399–418.

- Marr JW & Heady JA (1986) Within- and between-person variation in dietary surveys: number of days needed to classify individuals. *Hum Nutr Appl Nutr* **40**, 347–364.
- Morgan KJ, Johnson SR & Gougetas B (1987) Variability of food intakes: an analysis of a 12-day data series using persistence measures. *Am J Epidemiol* **126**, 326–335.
- Nelson M, Black AE, Morris JA & Cole TJ (1989) Between- and within-subject variation in nutrient intake from infancy to old age: estimating the number of days required to rank dietary intakes with desired precision. *Am J Clin Nutr* **50**, 155–167.
- Poppitt SD & Swann DL (1998) Dietary manipulation and energy compensation: does the intermittent use of low-fat items in the diet reduce total energy intake in free-feeding lean men? *Int J Obes Relat Metab Disord* **22**, 1024–1031.
- Rosner B & Willett WC (1988) Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. *Am J Epidemiol* **127**, 377–386.
- Schouten HJA (1999) Planning group sizes in clinical trials with a continuous outcome and repeated measures. *Stat Med* **18**, 255–264.
- Stubbs RJ, Johnstone AM, O'Reilly LM, Barton K & Reid C (1998) The effect of covertly manipulating the energy density of mixed diets on ad libitum food intake in 'pseudo free-living' humans. *Int J Obes Relat Metab Disord* **22**, 980–987.
- Stubbs RJ, Ritz P, Coward WA & Prentice AM (1995) Covert manipulation of the ratio of dietary fat to carbohydrate and energy density: effect on food intake and energy balance in free-living men eating ad libitum. *Am J Clin Nutr* **62**, 330–337.
- Tarasuk V & Beaton GH (1991) The nature and individuality of within-subject variation in energy intake. *Am J Clin Nutr* **54**, 464–470.
- Tarasuk V & Beaton GH (1992) Statistical estimation of dietary parameters: implications of patterns in within-subject variation – a case study of sampling strategies. *Am J Clin Nutr* **55**, 22–27.
- Verbeke G & Molenberghs G (2000) *Linear Mixed Models for Longitudinal Data*. New York, NY: Springer.
- Willett W (1998) *Nutritional Epidemiology*, 2nd ed., Oxford: Oxford University Press.

Appendix 1

Statistical analysis for model (1) using SAS

Data must be entered as one row/measurement per subject, with column variables identifying the subject, period, and group of each measurement.

```
PROC MIXED METHOD = REML CL;
CLASS GROUP PERIOD SUBJ;
MODEL ENERGY = GROUP PERIOD GROUP
* PERIOD/DDFM = SATTERTH;
RANDOM SUBJ SUBJ*PERIOD/TYPE = SIMPLE;
RUN;
```

Covariance Parameter Estimates

Cov Parm	Estimate	Lower	Upper
SUBJECT	7.5766	4.0978	18.4897
SUBJECT * PERIOD	0.6787	0.2712	3.7381
RESIDUAL	7.3092	6.4708	8.3226

REML (REstricted Maximum Likelihood) is considered preferable to ML for estimating variances (Verbeke & Molenberghs, 2000). The present study used REML in all analyses. The option CL provides 95 % CI for variance components. The option / DDFM = SATTERTH gives correct degrees of freedom for tests and estimates of fixed effects (Verbeke & Molenberghs, 2000).

The interaction (for longitudinal studies) and the period effect (for crossover studies) can be estimated, along with their standard errors, by adding the following statements below the RANDOM statement:

```
ESTIMATE (gp22 - gp21) - (gp12 - gp11)' PERIOD*GROUP 1 -1 -1 1;
```

```
ESTIMATE 'p2-p1' PERIOD -1 1;
```

To include within-subject effects such as weekday v. weekend, or between-subject effects such as sex, modify the MODEL statement as follows:

```
MODEL ENERGY=GROUP PERIOD GROUP
* PERIOD WEEKDAY SEX/DDFM = SATTERTH;
```

To allow first-order autoregressive (AR(1)) day-to-day correlation for days on a given subject within a given period, add the following statement after the RANDOM statement (Littell *et al.* 2000):

```
REPEATED/TYPE = AR(1) SUBJECT = SUBJ * PERIOD;
```

Appendix 2

Power and sample-size estimation with unequal group sizes and/or numbers of days

Relevant quantities are shown for estimating the interaction effect in longitudinal studies using a two-sided level α test. N is the number of subjects in group 1 and $N \times k_N$ is the number of subjects in group 2. D is the number of days/subject in period 1 and $D \times k_D$ is the number of days/subject in period 2. The day-to-day variance is σ_E^2 and $R_P = \sigma_P/\sigma_E$ where σ_P^2 is the period variance. The cumulative normal probability function is denoted by $\Phi(x)$, and $Z_{1-\alpha/2}$ denotes the $100 \times (1 - \alpha)$ percentage quantile of the standard normal distribution. For example, for level = $\alpha = 0.05$, $Z_{1-\alpha/2} = 1.96$. For power = $1 - \beta = 80\%$, $Z_{1-\beta} = 0.84$.

For crossover designs, replace the term $(1 + \frac{1}{k_N})$ by $(\frac{1}{1+k_N})$ in all equations. Do not change the term $(\frac{1}{1+k_D})$.

Quantity	General equation
Variance	$V = \frac{1}{N} \left(1 + \frac{1}{k_N} \right) \left[2R_P^2 + \frac{1}{D} \left(1 + \frac{1}{k_D} \right) \right] \sigma_E^2$
Detectable difference	$\Delta = (Z_{1-\beta} + Z_{1-\alpha/2})\sqrt{V}$
Power	$1 - \beta = \Phi(-Z_{1-\alpha/2} + \frac{\Delta}{\sqrt{V}})$
Number of days/subject per period	$D = \frac{\left(1 + \frac{1}{k_D} \right)}{\left(\left(1 + \frac{1}{k_N} \right) (Z_{1-\beta} + Z_{1-\alpha/2})^2 \sigma_E^2 \right) - 2R_P^2}$
Number of subjects/group	$N = \frac{\left(1 + \frac{1}{k_N} \right) (2R_P^2 + \frac{1}{D} \left(1 + \frac{1}{k_D} \right)) (Z_{1-\beta} + Z_{1-\alpha/2})^2 \sigma_E^2}{\Delta^2}$