24

Nutrient Data Analysis Techniques and Strategies

Alan R. Dyer, Kiang Liu, and Christopher T. Sempos

Overview

Analyses of nutrient data pose special challenges to investigators. In such analyses, investigators need to consider:

1. Possible over- or under-reporting of intakes, leading to "impossible" or extreme values in the data set
2. How to adjust for total energy intake
3. How to model nutrients, e.g., as continuous or categorical variable
4. How to avoid multicollinearity, particularly when nutrients are expressed in absolute amounts, e.g., grams/day
5. How to analyze dietary supplement data
6. How to account for large day-to-day variability in intakes, which can lead to misclassification of individuals with respect to usual intake

The objectives of this section are to examine various approaches to addressing the above issues; to briefly describe the common types of observational and experimental studies that collect nutritional data; and to describe the most common methods of analysis used in the types of studies described.

Quality Control

Whether investigators use a validated food frequency questionnaire or single or multiple 24-hour recalls to collect dietary data, the importance of quality control in such data collection cannot be overemphasized. The phrase GIGO (garbage in, garbage out) serves as a stark reminder of the importance of ensuring that dietary data are of the highest

quality when they are submitted for analysis. No amount of analytic sophistication can make up for poor quality data.

To improve the quality of collected data, investigators should:

- Develop a Manual of Operations for nutrient data collection
- Train and certify dietary interviewers in collection of data and use of the manual
- Tape interviews with the consent of the participant
- Immediately review a printout of the data collected, including nutrient totals if the system being used permits
- Develop range limits for important nutrients that result in careful review of the questionnaire or 24-hour recall with the participant, if limits are exceeded
- Make inquiries to cooks for clarifying information when needed
- Query the participant for a 24-hour recall on whether the amount consumed was typical, and if atypical, the reason the amount consumed was unusually low or high, e.g., lower than usual due to illness
- Use food composition data to estimate nutrient composition for foods not found in a data base when using 24-hour recalls
- Randomly select tape recordings for repeat completion of questionnaires or re-entry of data, with assessment of discrepancies and correction of incorrect data
- Develop criteria for re-certifying interviewers based on the randomly selected recordings

Interviewers may also be requested to indicate whether they believe the participant has provided reliable data. Persons deemed by the interviewer as not providing reliable data should be excluded from the analyses.

Prior to conducting analyses, investigators might wish to set limits on total caloric intake above or below which persons would be excluded. For example, in the Coronary Artery Risk Development in (young) Adults Study (CARDIA),[1] men who reported intake of >8000 kcals or <800 kcals and women who reported intake of >6000 kcals or <600 kcals on food frequency questionnaires were excluded from analyses, because values outside these limits were not considered consistent with a normal lifestyle.[2] The INTERMAP study of macronutrients and blood pressure also established exclusionary cutoffs for caloric intake obtained from 24-hour recalls.[3,4] Food frequency questionnaires generally have larger standard deviations in total energy intake than 24-hour recalls, and thus are more likely to have individuals with "impossible" or extreme values.[5] Hence, investigators using food frequency questionnaires need to be particularly attentive to establishing exclusionary cutoffs for total energy intake, such as those used in CARDIA. Investigators using 24-hour recalls should consider whether or not to exclude persons reporting that their 24-hour intake was unusual.

Identifying Outliers or Extreme Values

Prior to conducting any analyses, investigators should examine the distribution of each variable of interest for outliers or extreme values. The procedure Proc Univariate in SAS is particularly useful in this regard.[6] In addition to providing the standard descriptive

statistics, e.g., mean, median, standard deviation, range, etc., this procedure also identifies the five largest and five smallest values for each variable, and the 1st, 5th, 95th, and 99th percentiles. The user can also request a box plot of the data, which can be very helpful in identifying extreme values. The box plot helps indicate how discrepant the largest and smallest values are from the rest of the data.

The fact that a statistical software package identifies values as large or extreme relative to other values in the distribution should not be taken as prima facie evidence that such values are invalid or that an error was made in data collection or data entry. Values so identified should be examined for such problems. However, if the values are biologically plausible and no error appears to have been made, they should not be arbitrarily excluded from the analysis. Neter, Wasserman, and Kutner[7] suggest that a safe rule is "to discard an outlier only if there is direct evidence that it represents an error in recording, a miscalculation, a malfunctioning piece of equipment, or a similar type of circumstance." When outliers are retained in a data set, the investigator needs to take special steps to assess any influence they may have on the results of the analysis. This can include analyses with and without the outlying value or values, use of nonparametric statistical methods, e.g., the Spearman rank correlation instead of the usual Pearson correlation coefficient, transformations of the data which bring the outlying value closer to the other values, e.g., the log or square root transformation, specific tests for influential observations,[7] or use of robust regression methods.[8]

Adjustment for Total Energy Intake

Adjustment for total energy intake is of particular relevance for epidemiologic studies in which investigators use some form of regression model to examine the associations of specific nutrients with an outcome variable, e.g., blood pressure or cholesterol in multiple linear regression, case-control status in logistic regression, or coronary heart disease incidence in Cox proportional hazards regression. Thorough discussions on adjustment for total energy intake can be found in Willett, Howe, and Kushi,[9] or in Willett.[10] Only the major issues addressed by these authors are described here. The rationale for adjusting for total energy intake is that most nutrients are correlated with total energy intake. This is because they contribute directly to total energy intake, e.g., total fat or carbohydrate, or because persons who consume more kcalories also eat more, on average, of all nutrients, e.g., dietary cholesterol or sodium. For example, in participants of the Multiple Risk Factor Intervention Trial (MRFIT),[11] the baseline correlations of 10 energy contributing nutrients with total energy intake ranged from 0.29 for alcohol intake to 0.87 for total fat intake. Among 24 non-energy contributing nutrients, the correlations ranged from 0.05 for retinol to 0.78 for phosphorus, with a median of 0.52. No nutrient had a negative correlation with total energy intake. Thus, if total energy intake is positively associated with a dependent variable, almost all specific nutrients will also be positively associated with that variable. Hence, in regression analyses involving specific nutrients, there is a need to adjust associations with specific nutrients for the potential confounding effects of total energy intake.

The most common methods of adjustment for total energy intake are typically referred to as the nutrient density method, the standard multivariate method, the residual method, and the multivariate nutrient density method. [9,10,12] The nutrient density method has been the traditional method of adjusting for total energy intake. In this approach, nutrient intake is divided by total energy intake, with energy-contributing nutrients

expressed as percent of kcalories, and non-energy contributing nutrients expressed as intake per 1000 kcal. The strengths of this approach include ease of calculation, familiarity by nutritionists, and use in national guidelines.[9] For example, the Committee on Diet and Health of the National Research Council recommends that total fat intake be less than 30% of total energy intake and that saturated fat intake be less than 10%.[13] The primary problem with the nutrient density method is that it does not completely eliminate potential confounding with total energy intake, since nutrients expressed as nutrient density often remain correlated with total energy intake. For example, in the MRFIT, the correlations of percent kcalories from protein, fat, and carbohydrate intake with total energy intake at baseline were –0.23, 0.18, and –0.11, respectively.[11] However, with these three nutrients expressed as g/day, the corresponding correlations were 0.73, 0.87, and 0.77 in these men.

In the standard multivariate method, total energy intake is included in the multivariate regression model along with the nutrient or nutrients of interest. In this model, the regression coefficient for the nutrient of interest represents the effect of changing the nutrient by one unit while maintaining a constant total energy intake. For energy-containing nutrients this can only be accomplished by making changes in other energy-contributing nutrients equal to the amount of energy contained in one unit of the nutrient of interest. Similarly, the regression coefficient for total energy intake does not represent the effect of changing total energy intake by 1 kcal, but the effect of changing energy intake from all other energy contributing nutrients by 1 kcal. For example, if the nutrient in the model is total protein intake, then total energy intake represents fat and carbohydrate intake. In using this approach, estimates of the effect of changing intake of the nutrient by a specific amount should use variation in the nutrient with total energy intake held constant, i.e., the nutrient residual (see below) as the basis for the estimates of effect. Failure to do so can result in estimates of effect based on unrealistic differences in intake of the nutrient.

In the residual method, the investigator regresses each nutrient of interest on total energy intake, and then computes a nutrient residual for each individual by subtracting from the individual's actual intake of that nutrient, the amount predicted based on his/her total energy intake. Because the mean of these residuals is equal to zero, it may be desirable to add a constant to each residual, e.g., the mean intake for the nutrient. The resulting value does not, however, represent the individual's actual intake, and in fact has no "biological" or public policy meaning. The residual method is simply one means by which investigators can adjust for total energy intake. Nutrient residuals are independent of total energy intake. Models that use nutrient residuals can also include total energy intake. The regression coefficient for a nutrient expressed as a nutrient residual is identical to the regression coefficient for the nutrient in the standard multivariate model. However, the regression coefficient for total energy intake will not be identical to that in the standard multivariate model. In the residual model, the association of total energy intake with the dependent variable is not adjusted for intake of the specific nutrient, which could result in an inaccurate estimate of the association of total energy intake with the dependent variable.

In the multivariate nutrient density model, total energy intake is included in the model along with nutrient density. This approach addresses the problem of potential confounding by total energy intake in such analyses. In this model, the regression coefficient for the nutrient estimates the effect of a 1% difference in energy from the nutrient with total caloric intake held constant. As noted by Willett et al.,[9] a major strength of the multivariate nutrient density approach is that it separates diet into two components: composition and total amount.

Modeling Nutrient Intake

Investigators typically model nutrient intake as a continuous variable or a series of dummy variables corresponding to quantiles of the nutrient, e.g., quartiles or quintiles. The advantages of categorizing nutrient intake include reduction of the potential effects of outlying or extreme values, and elimination of the need to assume a linear relation between the nutrient of interest and the dependent variable. Categorization is also more informative to readers since it allows estimation of relative risks in logistic regression and Cox proportional hazards regression for persons in each exposure category relative to a referent category, and in multiple linear regression the mean difference in the dependent variable for persons in each exposure category relative to the referent category. The main weakness in categorizing a continuous variable is that when the relationship is linear, the categorization results in a loss of power. However, regardless of how nutrient intake is modeled in the definitive analysis, categorization is still an extremely useful tool and should be part of any analysis plan. This is because categorization allows the investigator to examine the shape of the relation between the nutrient and the dependent variable, and thus whether the relation is sufficiently linear to support inclusion of the nutrient as a continuous variable in the regression model.

When nutrient intake is categorized, one defines k-1 dummy variables for each individual for the k categories of the variable. For example, if nutrient intake is divided into quartiles, three dummy variables are defined. In defining the dummy variables, it is necessary to define a referent category. This is the category against which the risks or means for the other exposure categories are compared. If nutrient intake is divided into quartiles and the first quartile is to be the referent category, the three dummy variables corresponding to quartiles 2 to 4 are defined as follows:

$$X_1 = \begin{cases} 1 \text{ if intake in second quartile} \\ 0 \text{ otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 \text{ if intake in third quartile} \\ 0 \text{ otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 \text{ if intake in fourth quartile} \\ 0 \text{ otherwise} \end{cases}$$

These definitions produce the following values on each of the variables for individuals in the first through fourth quartiles:

Quartile of intake	X_1	X_2	X_3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

In defining categories for a nutrient, investigators should adjust for total energy intake by defining the categories based on nutrient residuals or nutrient densities, rather than absolute intake.[12] While the standard multivariate method and the nutrient residual

method provide identical regression coefficients for the nutrient of interest when the nutrient is entered as a continuous variable, this is not the case when nutrient intake is categorized.[12] In this case, the standard multivariate method should be avoided. It is also desirable to model total energy intake as a continuous variable in such analyses rather than as a second categorical variable, particularly if nutrient density is the variable being categorized.[10,12]

Multicollinearity

Multicollinearity in a regression model can occur when highly intercorrelated variables are entered simultaneously into the model, or when a linear combination of several variables essentially equals a constant. For example, multicollinearity would occur with nutrient data if the model included percent of kcalories from total fat, protein, and carbohydrate, since the sum of these three variables is often 100 or quite close to 100. Hence, investigators should not attempt to enter more than two of these variables simultaneously into a regression model. Similarly, multicollinearity would also occur if these same three variables were entered into a model as g/day along with total energy intake. In this case, only three of these four variables should be entered simultaneously. In general, investigators need to ensure that they do not include in the same model variables representing total intake for a nutrient and all individual components of that intake, e.g., total fat plus saturated fats, polyunsaturated fats, and monounsaturated fats. However, even if investigators are careful to ensure that the types of multicollinearity described above do not occur, multicollinearity can still be a problem when multiple intercorrelated variables are included in a model, e.g., nutrients that come from the same sources. In this situation it may be impossible to determine the separate and independent associations of the multiple variables with the dependent variable. For example, in a study on the associations of potassium, calcium, protein, and milk intakes with blood pressure, the investigators found that while potassium had a relatively stronger association with blood pressure than the other three dietary factors, the high correlations of potassium intake with intakes of the other three made it impossible to determine the independent association of potassium intake with blood pressure.[14]

The use of nutrient residuals and nutrient densities help to reduce the likelihood of multicollinearity, since energy-adjusted nutrients generally have lower intercorrelations than nutrients expressed as absolute amounts.[10] Methods for assessing and detecting multicollinearity, as well as remedial measures, can be found in Neter, Wasserman, and Kutner.[7]

Some investigators may believe that procedures that select variables for inclusion in regression models based on whether or not the variable is significantly related to the dependent variable are appropriate approaches for preventing multicollinearity. Such procedures include forward selection or backward elimination of variables, and stepwise regression. In forward selection of variables, variables are entered into the model one at a time, beginning with the variable that has the strongest association with the dependent variable, followed sequentially by those having the strongest residual associations with the dependent variable, i.e., after taking into account the association of the entering variable with the variables previously entered and their associations with the dependent variable. Variables are entered into the model until no remaining variable would have a statistically significant association with the dependent variable, if it were to enter the model next. In backward elimination of variables, all available variables are entered into

the model, and those with the weakest association are sequentially removed until only variables significantly related to the dependent variable remain. Stepwise regression combines forward selection and backward elimination of variables by removing those that are no longer significant when a new variable is entered into the model, so that the final model only contains variables significantly related to the dependent variable.

These variable selection procedures should be avoided for a number of reasons. First, the final model selected will not necessarily be optimal, e.g., maximize R^2. Second, the hypothesis tests used to determine which variables remain in the model are correlated.[8] Third, if a large number of variables is involved, initial entry of all variables may not be possible if one or more is a linear combination of the other variables. Fourth, stepwise procedures may not select possible confounders that should be included whether or not the confounder has a significant association with the dependent variable, e.g., age and sex, or total energy intake in the multivariate nutrient density approach. Fifth, the results of these procedures are often not unique, i.e., they yield final models that do not include the same variables. For example, in a logistic analysis involving the associations of total energy intake and intakes of protein, fat, and carbohydrate with CHD incidence, McGee, Reed, and Yano[15] found that only carbohydrate intake had a significant association with CHD incidence if forward selection of variables was used. When backward elimination was used instead, the final model included fat intake and total energy intake as the only variables significantly and independently related to CHD incidence.

Dietary Supplements

The use of dietary supplements poses a number of complexities for analyses involving nutrient intake, which are not easily resolved. The first question that must be addressed is whether supplement-based intake should be included or excluded. The approach recommended here is to analyze the data with and without inclusion of the supplement-based intake, since this is likely to provide the most complete information on the association of the nutrient with outcome. It may also be beneficial to treat the intake from supplements and food-based intake as separate nutrients. Such an approach is likely to be particularly appropriate if it is difficult to determine the separate and independent effects of food-based intake of the nutrient, or the separate and independent effects of the supplement due to its high correlation with supplemental intake of other nutrients, or where food-based intake and supplement-based intake have very different correlations with other nutrients. If such an analysis is done, food-based intake should be energy adjusted using nutrient residuals or nutrient densities, with supplement-based intake not energy adjusted, except through inclusion of total energy intake in the model. Hence, the analysis for food-based intake would be based on the residual or nutrient density model, whereas the analysis for supplement-based intake would follow the standard multivariate model. If intake is categorized, the categorization of food-based intake should use the nutrient residuals or nutrient densities, whereas the categories for supplement-based intake would be based on absolute amounts. The reason for not suggesting that supplement-based intake be adjusted for total energy intake through calculation of nutrient residuals or nutrient densities is a likely low correlation between total energy intake and supplement-based intake, particularly if the nutrient does not contribute to energy intake.

If analyses are to be conducted in which food- and supplement-based intakes are combined, the investigator needs to decide whether to simply add the two intakes together

or to add supplement-based intake to the energy-adjusted intake from foods.[9] It is unclear how, or if, results between these two approaches will differ. If the intake from supplements represents a large proportion of the total intake and thus the correlation between total intake of the nutrient and total energy intake is low, the easiest and probably best approach is to simply combine the supplement-based intake with the food-based intake, and then use the standard multivariate model, whether or not nutrient intake is categorized. If the supplement-based intake does not represent a large proportion of the total intake, it may be worthwhile to examine the associations using both approaches, since it is unclear how the results might differ, of if they will differ in any practical way.

Within-Person Variability in Intake

The goal of examining associations of nutrients with an outcome is to estimate the association of usual intake with that outcome. However, information on nutrient intake collected from a single 24-hour recall is subject to substantial within-person variability due to day-to-day variability in intake in most individuals. Hence, nutrient intake in a single 24-hour recall often does not reflect the individual's average or usual intake. This day-to-day variability in nutrient intake is often referred to as "measurement error." Measurement error typically results in underestimation of associations of nutrients with outcomes. For example, for MRFIT men[11] it was estimated that with one 24-hour recall, the association between a dependent variable and percent of kcalories from total fat would be underestimated by 77.7% in simple linear regression.

Error can generally be divided into two types: random and systematic. If error is random, the average of a large number of repeated measurements approaches the true value, or for nutrient intake, the individual's usual intake. To reduce measurement error, studies will often collect multiple 24-hour recalls. Liu et al[16] describe methods for estimating the number of 24-hour recalls required to achieve a suitable degree of accuracy. In MRFIT men[11] it was estimated that the association between a dependent variable and percent of kcalories from total fat would be underestimated by 46.7% with four 24-hour recalls, compared to the 77.7% underestimation with one 24-hour recall.

If error is systematic, for example due to systematic over- or underreporting of intake, the average of a large number of repeated measurements will not reflect the individual's usual intake. Food frequency questionnaires may also have systematic error if specific foods eaten by an individual are not included in the questionnaire.

Methods are available for correcting or adjusting regression coefficients for measurement error. However, the assumptions that underlie such corrections can be quite strong and may not be strictly applicable to nutrient data. In particular, it is typically assumed that error is random and independent of the true value or usual intake, and that error and usual intake are normally distributed. However, for nutrient data it is likely that error is correlated with usual intake. For example, an individual with a usual intake of 3000 kcal is likely to vary more about his/her usual intake than an individual with a usual intake of 1000 kcal. Correcting for measurement error should be done with care and caution, and with attention to the assumptions underlying such corrections. A thorough discussion on correcting for measurement error in linear regression models is given by Fuller.[17] Willett[10] and Clayton and Gill[18] also discuss measurement error in the context of nutrient data. Spiegelman, McDermott, and Rosner[19] describe the regression calibration method for adjusting point and interval estimates for measurement error in linear regression, logistic

regression, and Cox proportional hazards regression. The regression calibration method is appropriate when a gold standard is available in a validation study and a linear measurement error with a constant variance applies, or when replicate measurements are available in a reliability study and linear random within-person error can be assumed. These authors also describe SAS macros that can be used to adjust regression coefficients in these models when the assumptions underlying use of the regression calibration method appear appropriate.

Types of Epidemiologic Studies

A discussion of types of epidemiologic studies with particular reference to nutrition can be found in Sempos, Liu, and Ernst,[20] while a more general review of the topic is given in Hennekens and Buring.[21] There are generally two types of epidemiologic studies: observational and experimental. The main difference between an experimental and an observational study is the control that the investigator exercises over participants, procedures, and exposures. In an experiment, the investigator controls who enters the study, what drugs or procedures are given to participants, and how the study is carried out. In a nutritional intervention study, the investigator would manipulate or attempt to manipulate some or all participants' dietary intake. An observational study does not involve an intervention or manipulation. In such a study, the investigator does not control who enters the study or the factors or drugs to which participants are exposed. Observational studies of individuals include cross-sectional, case-control, and prospective studies, while studies of groups are referred to as ecologic studies. In nutritional epidemiologic studies, nutrient intake is measured but not manipulated, the frequency and pattern of outcomes observed, and associations between nutrients and outcomes estimated using statistical methods.

In a cross-sectional study the question asked is, "What is the correlation or association between nutrient intake and the outcome?" Individuals are included in the study without regard to their status on the outcome or nutrient intake. In these studies, nutrient intake and the outcome are both measured at the same point in time. For example, INTERMAP is a cross-sectional study of the associations of macronutrients with blood pressure.[3,4] In this study, each participant had blood pressure measured twice on each of four occasions and completed a 24-hour recall on each day that blood pressure was measured.

Case-control studies, also referred to as retrospective and case-referent studies, are designed to answer the question, "Do persons with disease (cases) have different nutrient intake than persons who have not been diagnosed with the disease (controls)?" For example, do persons with heart disease consume more dietary cholesterol and saturated fatty acids than persons without heart disease? In case-control studies, recently diagnosed persons with the disease and a set of persons without the disease are interviewed concerning their dietary habits. The goal is to determine usual nutrient intake before the onset of disease.

Prospective studies are also referred to as cohort, incidence, follow-up, and longitudinal studies. The question asked in prospective studies when a nutrient is thought to be related to increased risk of disease is, "Do persons with higher intake develop or die from the disease more frequently or sooner than persons with lower intake?" Alternatively, if a nutrient is thought to be related to decreased risk of disease, the question asked is, "Do persons with lower intake develop or die from the disease more frequently or sooner than persons with higher intake?" For example, are persons who consume more than 50 g/day

of alcohol more likely to have a stroke than persons who consume less alcohol? Persons found to be disease free at the time of the cross-sectional survey are followed over time to determine who develops the disease and when the disease occurs.

Ecologic studies compare aggregate data representing entire populations. A common example of this type of study is one in which disease-specific mortality rates for different countries are correlated with nutrient measurements based on food disappearance data.[22] The INTERSALT study included ecologic analyses on associations of urinary electrolytes and other factors with blood pressure, as well as cross-sectional analyses on electrolyte-blood pressure associations within individuals.[23,24]

Experimental studies involving nutritional interventions include feeding or metabolic ward studies and randomized clinical trials. Feeding studies involve feeding groups of individuals precisely measured diets with one or more components varied, with an effect on a biologic variable then measured. The Keys equation for predicting change in total cholesterol from changes in intakes of saturated and polyunsaturated fatty acids and dietary cholesterol was determined from a metabolic ward study.[25] A common design for feeding studies is the crossover design, in which each participant serves as his/her own control. The randomized clinical trial is a prospective study in which individuals are randomly assigned to intervention and control groups. After randomization, both groups are followed over time to assess the efficacy and safety of the intervention. For example, the trial on the Primary Prevention of Hypertension was a randomized, controlled clinical trial on the effects of weight loss, reduction in sodium intake, decreased alcohol intake, and increased exercise on the five-year incidence of hypertension in men and women with high normal blood pressure.[26]

Methods for Comparing Groups in Cross-Sectional Studies

Table 24.1 lists methods of analysis that can be used to compare nutrient intake between two groups, e.g., men and women, or among three or more groups, e.g., African-Americans, Hispanics, and whites. For nutrient intake considered as a continuous variable, the goal of the analysis is to determine whether mean or median intake differs significantly between or among groups. For such analyses, the table indicates the usual method of

TABLE 24.1

Methods for Comparing Nutrient Intake among Groups in Cross-Sectional Studies

	Number of Groups (k)	
Description	**k = 2**	**k > 2**
Nutrient Intake Continuous		
Usual method	Two-sample t-test	Analysis of variance
Nonparametric alternative	Wilcoxon rank-sum test	Kruskal-Wallis test
Adjustment for other variables	Analysis of covariance or multiple linear regression	
Nutrient Intake Categorical (c categories)		
Usual method	Chi-square test for 2 x c contingency table	Chi-square test for k x c contingency table

analysis, the nonparametric alternative, and methods that can be used to adjust for potential confounders of differences between groups, e.g., age or total energy intake. Nonparametric tests make fewer assumptions about the shape of the distributions of variables than parametric tests such as the two-sample t-test or analysis of variance. In the Wilcoxon rank-sum test and the Kruskal-Wallis test, the actual observations are replaced by their ranks in the combined sample of all observations. If nutrient intake is divided into categories, the goal of the analysis is usually to determine whether the distributions of intake are homogeneous across groups. The methods listed in Table 24.1 can also be used to compare nutrient intake at baseline in an experimental study; for example, to determine whether in a randomized clinical trial randomization has provided comparable groups with respect to intake of specific nutrients. A useful text on these methods and those described below is that of Rosner.[27]

Methods for Comparing Cases and Controls in Case-Control Studies

Table 24.2 lists methods of analysis that can be used to compare nutrient intake between cases and controls in unmatched and matched case-control studies. Matching is often done in case-control studies to make cases and controls comparable on variables that could confound associations of the variable of interest with disease. For unmatched case-control studies, the methods listed are identical to those for comparing nutrient intake between two groups in cross-sectional studies. For matched case-control studies, the methods of analysis need to take into account the matching. Hence, for a simple comparison of means between cases and controls, the investigator should use a paired t-test rather than a two-sample t-test, or the Wilcoxon signed-rank test rather than the Wilcoxon rank sum test. When multiple regression is used to adjust the mean difference between cases and controls for other variables in matched case-control studies, the investigator needs to ensure that the dependent and independent variables in the model are defined correctly. In such studies, the dependent variable is the difference in nutrient intake for each case-control

TABLE 24.2

Methods for Comparing Nutrient Intake between Cases and Controls in Case-Control Studies

Description	Unmatched	Matched
Nutrient Intake Continuous		
Usual method	Two-sample t-test	Paired t-test
Nonparametric alternative	Wilcoxon rank-sum test	Wilcoxon signed-rank test
Adjustment for other variables	Analysis of covariance or multiple linear regression	Multiple linear regression*
Nutrient Intake Categorical (c categories)		
Usual method	Chi-square test for 2 x c contingency table	McNemar's test for c = 2

* In this model, differences in each variable for the case-control pair are used, with the difference in the nutrient of interest serving as the dependent variable. The test of significance for the adjusted mean difference is the test of the hypothesis that the intercept of the model is equal to zero.

TABLE 24.3

Methods for Assessing Associations in Epidemiologic Studies

Dependent Variable	Nutrient Intake	Unadjusted	Adjusted for Other Variables
Cross-Sectional or Ecologic Study			
Continous	Continous	Pearson correlation Spearman correlation Linear regression	Partial correlation Linear regression
Continous	Categorical	Linear regression	Linear regression
Dichotomous	Continuous, categorical	Logistic regression	Logistic regression
Unmatched Case-Control Study			
Case-control status	Continuous, categorical	Logistic regression	Logistic regression
Matched Case-Control Study			
None	Continuous, categorical	Conditional logistic regression	Conditional logistic regression
Prospective Study			
Time to event	Categorical	Log rank test Cox regression	Cox regression
Time to event	Continuous	Cox regression	Cox regression

pair, while the independent variables are the within-pair differences for the potential confounding variables. The test of significance for the adjusted mean difference is the test of the hypothesis that the intercept in the model is equal to zero.

Methods for Assessing Associations in Epidemiologic Studies

Table 24.3 lists methods for assessing associations of nutrient intake with outcome variables in cross-sectional or ecologic studies, matched and unmatched case-control studies, and prospective studies. For each type of study, the table indicates methods that can be used when nutrient intake is modeled as a continuous variable or as a categorical variable. The table also lists the dependent variable for each type of analysis. For example, in Cox proportional hazards regression, the dependent variable is the time to some event, e.g., death from coronary heart disease. In unmatched case-control studies, the dependent variable is typically case-control status. Since cross-sectional and ecologic studies can have both continuous and dichotomous dependent variables, methods are listed for both types of dependent variable. No dependent variable is listed for conditional logistic regression, since there is no outcome variable that varies from individual to individual in this model. In conditional logistic regression, the independent variables are the case-control difference in each variable, and the model does not include a constant term. A useful text on logistic and Cox regression methods is that of Kahn and Sempos.[28]

The Spearman correlation is listed for use in cross-sectional and ecologic studies, since it is the nonparametric alternative to the Pearson product-moment correlation coefficient. The Pearson-product moment correlation coefficient should not be used if either nutrient

intake or the second variable has a very skewed distribution, since the assumption underlying its use is that each variable has a normal distribution for each value of the other variable.

In analyses involving linear regression, interest focuses on the difference in the mean of the dependent variable for a one-unit or greater difference in the independent variable. Hence, the focus is on the regression coefficient. In logistic and Cox regression, interest focuses on estimates of relative risk. In logistic regression, the relative risk is given by the odds ratio, and in Cox regression the hazard ratio. In both models, relative risk estimates are obtained by exponentiation of the regression coefficient or the regression coefficient times some convenient multiplier. For example, if total energy intake is the dietary variable of interest, exponentiation of the regression coefficient gives the relative risk of the outcome for two persons who differ in total energy intake by 1 kcal. Since this is not a particularly meaningful difference for calculating relative risk, an investigator might multiply the regression coefficient by 500 to obtain the relative risk of the outcome for two persons who differ in total energy intake by 500 kcal. When nutrient intake is categorized and dummy variables are included in the regression model, exponentiation of the regression coefficient for a dummy variable gives the risk of the outcome for those in the category corresponding to the dummy variable relative to the referent category, e.g., quartile 4 relative to quartile 1.

In analyses based on Cox regression, true associations between diet and disease may not be found if there are substantial changes in nutrient intake between the baseline assessment of diet and the development of disease, or if there are substantial changes in the rank ordering of study participants with respect to intake over the course of followup.

Analyses of Intervention Studies with Change in Nutrient Intake as Outcome

In nutritional intervention studies, investigators often wish to examine the effects of the intervention on intakes of specific nutrients following completion of the intervention. Investigators can use three approaches to determine whether intake of specific nutrients changed in an intervention group relative to a control group or among three or more groups:

1. Compare intake among groups at followup, ignoring pre-intervention intake using the methods for cross-sectional studies in Table 24.1.
2. Compare the change in intake from pre-intervention to followup among groups using the methods for cross-sectional studies.
3. Compare intake among groups at followup, adjusting for pre-intervention intake with multiple linear regression or analysis of covariance.

Investigators typically use the second approach for intervention studies, even though it tends to be less powerful than analysis of covariance. Assumptions in regard to the analysis of covariance may or may not be met in an intervention study. The first approach may, however, be preferable to the second, if there are no differences in intake among the groups compared at the pre-intervention assessment, and the correlation between the pre-intervention and followup assessments for the nutrient of interest is less than 0.5. Correlations smaller than 0.5 are not uncommon for many nutrients assessed on two occasions.[11]

Hence, for nutrient intake the best approach may be to ignore pre-intervention intake. Prior to conducting analyses in nutritional intervention studies, investigators should examine the correlations of the nutrients from pre-intervention to followup and be prepared to ignore pre-intervention intake in the analyses.

References

1. Slattery ML, et al. *J Am Coll Nutr* 14: 635; 1995.
2. Goldberg GR, et al. *Eur J Clin Nutr* 45: 569; 1991.
3. INTERMAP Cooperative Research Group *Canad J Cardiol* 13: 235B; 1997.
4. INTERMAP Research Group. *Canad J Cardiol* 13: 80B; 1997.
5. Liu K. *Am J Clin Nutr* 59: 262S; 1994.
6. SAS Procedures Guide, Release 6.03 Edition. Cary, NC: SAS Institute, 1988.
7. Neter J, Wasserman W, Kutner MH. *Applied Linear Regression Models.* Homewood, IL: Richard D. Irwin, Inc., 1983.
8. Ryan TP. *Modern Regression Methods.* New York, NY: John Wiley & Sons, 1997.
9. Willett WC, Howe GR, Kushi LH. *Am J Clin Nutr* 65: 1220S; 1997.
10. Willett W. *Nutritional Epidemiology.* New York, NY: Oxford University Press, 1990.
11. Grandits GA, Bartsch GE, Stamler J. In: *Dietary and Nutritional Methods and Findings: The Multiple Risk Factor Intervention Trial (MRFIT)* (Stamler J, et al. Eds.) *Am J Clin Nutr* 65: 211S; 1997.
12. Brown CC, et al. *Am J Epidemiol* 129: 323; 1994.
13. National Research Council. *Diet and Health: Implications for Reducing Chronic Disease Risk.* Washington, DC: National Academy Press, 1989.
14. Reed D, McGee D, Yano K, Hankin J. *Hypertension* 7: 405; 1985.
15. McGee D, Reed D, Yano K. *J Chron Dis* 37: 713; 1984.
16. Liu K, et al. *J Chron Dis* 31: 399; 1978.
17. Fuller WA. *Measurement Error Models.* New York, NY: John Wiley & Sons, 1987.
18. Clayton D, Gill C. In: *Design Concepts in Nutritional Epidemiology.* Margetts BM, Nelson M, Eds, Oxford, UK: Oxford University Press, 1991, pp. 79-96.
19. Spiegelman D, McDermott A, Rosner B. *Am J Clin Nutr* 65: 1179S; 1997.
20. Sempos CT, Liu K, Ernst N. *Am J Clin Nutr* 69: 1S; 1999.
21. Hennekens CH, Buring JE. *Epidemiology in Medicine* (Mayrent SL, Ed) Boston, MA: Little, Brown, 1987.
22. Stamler J, Shekelle R. *Arch Pathol Lab Med* 112: 1032; 1988.
23. The INTERSALT Cooperative Research Group. *J Hypertens* 4: 781; 1986.
24. The INTERSALT Cooperative Research Group. *Br Med J* 297: 319; 1988.
25. Keys A, Anderson JT, Grande F. *Metabolism* 65; 776; 1965.
26. Stamler R, et al. *JAMA* 262: 1801; 1989.
27. Rosner B. *Fundamentals of Biostatistics,* 5th ed. Belmont, CA: Duxbury Press, 1999.
28. Kahn HA, Sempos C. *Statistical Methods in Epidemiology.* New York, NY: Oxford University Press, 1989.