Review

Publishing Nutrition Research: A Review of Multivariate Techniques—Part 1

PATRICIA M. SHEEAN, PhD, RD; BARBARA BRUEMMER, PhD, RD; PHILLIP GLEASON, PhD; JEFFREY HARRIS, DrPH, RD, LDN; CAROL BOUSHEY, PhD, MPH, RD; LINDA VAN HORN, PhD, RD

ABSTRACT

This article is the seventh in a series reviewing the importance of research design, analyses, and epidemiology in the conduct, interpretation, and publication of nutrition research. Although there are a variety of factors to consider before conducting nutrition research, the techniques used to conduct the statistical analysis are fundamental for translating raw data into interpretable findings. The statistical approach must be considered during the design phase of any study and often involves the use of multivariate analytical techniques. Multivariate analytical techniques represent a variety of mathematical models used to measure and quantify an exposure–disease or an exposure–outcome association, taking into account important factors that can influence this relationship. The primary purpose of this review is to introduce the more commonly used multivariate techniques, including linear and logistic regression (simple and multiple), and survival analyses (Kaplan Meier plots and Cox regression). These techniques are described in detail, providing basic definitions and practical examples with nutrition relevancy. An appreciation for the general principles within and presented previously in this article series is vital for enhancing the rigor in which nutrition-related research is implemented, reviewed, and published.

J Am Diet Assoc. 2011;111:103-110.

This article is the seventh in a series reviewing the importance of research design, analyses, and epidemiology in the conduct, interpretation, and publication of nutrition research. Other articles in this series focused on topics including study design and hypotheses development (1); sampling techniques, sample size, and critical elements of manuscript preparation (2); nonparametric procedures (3); qualitative research (4); epidemiologic methods (5); and, most recently, measurement and interpretation of nutrition-related outcomes and diagnostic tools (6). Collectively, the aim of this series is to provide the *Journal* readership with tools to enhance general understanding of key concepts inherent in high-quality nutrition research by providing relevant examples and additional resources. These articles are intended to serve as a review for more experienced researchers and also to offer practical, simple explanations for those who are new to the field. An appreciation for the general principles outlined in each of these articles is vital for enhancing the rigor in which nutrition-related research is implemented, reviewed, and published.

Although there are a variety of factors to consider before conducting nutrition research, the techniques used to conduct statistical analysis are fundamental for translating raw data into interpretable findings. The statistical approach must be considered during the design phase of any study and often involves the use of multivariate analytical techniques. By definition, a multivariate analysis (or multivariate modeling) is an efficient analytical tool used to control for confounding effects, to assess effect modification, and to summarize the association of several predictor variables with some outcome variable of interest (7). More simplistically, multivariate analytical techniques represent a variety of mathematical models often used in epidemiologic research to classically measure and quantify an exposure–disease or more practically an exposure–outcome association, taking in account important factors that can influence this relationship. For instance, in most nutrition studies involving human participants, multivariate techniques must be employed to adjust or control for the effects of basic demographic factors (ie, age, race/ethnicity, and sex) on the outcome of interest. Failure to control for these variables, in addition to others, can lead to spurious results and inappropriate inferences regarding the true exposure–disease or exposure–outcome relationship. Essentially

P. M. Sheean is an assistant professor, and L. Van Horn is a professor, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL. B. Bruemmer is a senior lecturer, Graduate Program in Nutrition Sciences, and director, Graduate Coordinated Program in Dietetics, University of Washington, Seattle. P. Gleason is a senior researcher, Mathematica Policy Institute, Inc, Geneva, NY. J. Harris is a professor and Didactic Program director, Department of Health, West Chester University of Pennsylvania, West Chester, PA. C. Boushey is an associate professor and director, Coordinated Program in Dietetics, Department of Foods and Nutrition, Purdue University, West Lafayette, IN.

Address correspondence to: Patricia M. Sheean, PhD, RD, Northwestern University, Feinberg School of Medicine, Department of Preventive Medicine, 680 N Lake Shore Dr, Ste 1400, Chicago, IL 60611. E-mail: p-sheean@northwestern.edu

Manuscript accepted: July 2, 2010.

0002-8223/$36.00

doi: 10.1016/j.jada.2010.10.010

Figure 1. Basic terminology and definitions associated with multivariate statistical techniques. References: [a]Last JM. *A Dictionary of Epidemiology*. 4th ed. New York, NY: Oxford University Press; 2001. [b]Riegelman RD, Hirsch RP. *Studying a Study and Testing a Test*. 3rd ed. Boston, MA: Little, Brown and Co; 1996.

multivariate techniques allow for the analysis of the relationship between more than one independent variable and one or more dependent variables. Thus, the primary purpose of this review is to introduce the more commonly used multivariate techniques, including linear and logistic regression, and survival analyses. These techniques are highly prevalent in the literature and are described in detail, providing basic definitions (Figure 1) and practical examples with nutrition relevancy. Future articles will address other less common, but equally important, multivariate techniques.

LINEAR REGRESSION

Linear regression is a commonly applied statistical technique used to assess the relationship between two or more variables where the dependent variable is quantitative. In general, the assumption for linear regression is that the independent variables (x) and the dependent variable (y) are linearly related (Figure 2). Estimating the parameters of a linear regression model is typically done using the least squares method, and linear regression models are often used to assess how well a given set of covariates (or x values) can predict the outcome of interest (or y). For example, suppose an investigator would like to explore the relationship between admission blood glucose and hospital length of stay (LOS). The researcher might hypothesize that patients with higher blood glucose on admission will have longer hospital LOS. To investigate this hypothesis, the researcher collects data from a sample of hospital patients on their admission blood glucose levels and their LOS. A simplistic approach to analyzing these data would be to estimate Pearson linear correlation coefficients between the two continuous variables. However, correlation coefficients can only evaluate the strength of linear relationship between two variables. In this example, they cannot be used for prediction or to tell us what LOS to expect for someone with a given blood glucose level. To estimate how much longer patients remain in the hospital for every unit increase in admission blood glucose, it is essential to find a formula for the best straight line through the observed

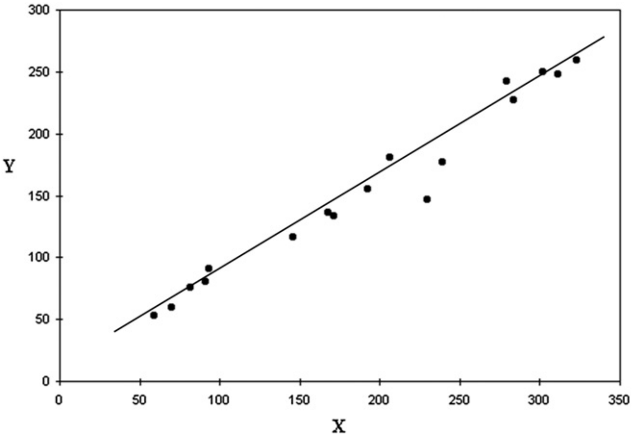

Figure 2. Graphic representation of linear regression.

data points. The results from such a regression can also be used to predict the LOS for a patient with a given set of characteristics including admission blood glucose.

In a simple regression model, a linear relationship between one independent variable (x) and a dependent variable (y) is expressed as: $y=\beta_0+\beta_1 x$, where β_0 is the y-intercept (ie, the estimated value of y when x=0), and β_1 is the regression coefficient (the estimated increase in the dependent variable for every unit increase in the independent variable). Using the previous example, y is the hospital LOS, x is the admission blood glucose, and β_1 represents the additional time spent in the hospital for every 1-unit increase in admission blood glucose. This type of modeling is used to depict simple relationships between two linearly related variables. Clearly, there would be other clinical factors to consider in this relation (eg, age, diagnoses, and comorbidities), which would require more sophisticated modeling.

Multiple linear regression modeling is considered an expansion of the simple linear regression model. In this type of modeling, additional independent variables (or covariates) are added to the regression equation (eg, x_1, x_2, and x_3) to assess the strength of the association between one dependent variable and one independent variable (eg, admission blood glucose), to improve its predictive abilities for the outcome variable and to better fit the data to a straight line. (It should be noted that multiple regression refers to multiple independent variables for a single outcome variable and multivariate regression refers to the analysis of multiple outcome variables in the same model.) Suppose that a simple linear regression reveals a positive relationship between admission blood glucose and the hospital LOS. Such a relationship could conceivably be explained by some other, third variable (eg, age, body mass index [BMI], and diagnosis). For example, patients with high admission blood glucose may also have high BMI values, and it may be their BMI rather than admission blood glucose that helps to explain the longer LOS. In this case, a simple regression model might show a positive relationship between admission blood glucose and LOS, but a multiple linear regression that controlled for BMI and other factors might show no relationship between admission blood glucose and LOS. To explore this relationship further, BMI may be changed from a continuous independent variable to a categorical independent variable and this is often accomplished through the creation of indicator variables (previously referred to as dummy variables). In this example, BMI could be a two-level categorical variable, such as obese vs not obese and this is simple to code (eg, 1=obese, 0=nonobese) and to interpret in your regression model. However, if there is an interest in looking at a variety of levels of BMI, the investigator could create κ-1 indicator variables where κ is the total number of levels of BMI. In most nutrition research using BMI, there are generally four categorizations used; "normal" is most often the reference category to which all other indicator variables are coded and compared. Referring back to our example, indicator variables representing "obese," "overweight," or "underweight" could be created and inserted into the regression model each as an independent variable to compare the influence of these individual BMI categories on the blood glucose and hospital LOS relationship.

To enhance understanding of multiple regression modeling, consider a well known example—the Harris-Benedict equation. Based on experiments with indirect calorimetry, multiple regression models were developed to estimate resting energy expenditure (REE) in adult men and women as predicted by a set of characteristics of 239 men and women (8). For illustrative purposes, we will consider the following equation for women, where resting energy expenditure is the dependent variable (y):

$$\text{REE}=655+9.6\times(\text{weight in kg})+1.8\times(\text{height in cm})-4.7\times(\text{age}).$$

The regression coefficients in the Harris-Benedict equation can be interpreted as follows:

- β_0 (the intercept) is equal to 655; corresponding to the estimate of REE when an individual has 0 weight, 0 height and is 0 years of age. The intercept, as in this case, if often biologically implausible and simply reflects a mathematical extrapolation with no meaningful interpretation.
- The regression coefficient for weight ($\beta_1=9.6$) reflects the estimated average increase in a person's REE for each 1 kg increase in their weight, while holding height and age constant.
- The regression coefficient for height ($\beta_2=1.8$) represents the estimated average increase in REE for each 1 cm increase in a person's height, while holding weight and age constant.
- The regression coefficient for age ($\beta_3=-4.7$) has a slightly different interpretation due to the negative sign, reflecting the estimated average decrease in REE for each 1-year increase in age, while holding weight and height constant.

An important assumption of the Harris-Benedict equation is that there is no effect modification (or interaction) between the covariates in the model (ie, weight×height, weight×age, or weight×sex). That is to say, changes in REE per unit change in weight are assumed to be constant for individuals of all heights, ages, and sex. If this assumption is incorrect, then regression models based on the effect modifier (ie, weight, height, and sex) must be conducted and presented. For example, a researcher could estimate separate regression models for individuals with different values of the variable that is believed to modify the effect (ie, separate models for men and women). Alternatively, the researcher could estimate a single regression model, but interact the independent variable of interest (eg, weight) with the covariate believed to be the effect modifier (eg, sex). Considering there are sex-specific Harris-Benedict equations, it is likely the REE varied significantly by sex; thus, two separate formulae were developed to better reflect more precise estimates of the indirect calorimetry data.

Although simple and multiple linear regression modeling techniques are used extensively in nutrition research, it is important to remember that these estimates are subject to random error and uncertainty. This uncertainty is recognized in regression equations in a couple of different ways. First, researchers should estimate the standard error of each of the regression coefficients and use the standard errors to calculate the confidence intervals (CIs) around these regression coefficient point estimates (7). The standard error and associated CI reflect the uncertainty of the estimation process by providing information on the extent to which the true value of the regression coefficient could differ from the point estimate and still remain consistent with the data used to estimate the model. Given this uncertainty in the point estimate of a regression coefficient, researchers often conduct tests of the statistical significance of a regression coefficient to determine whether a claim can be confidently made that the regression model provides evidence of a true relationship—positive or negative—between the covariate and outcome. If a regression coefficient is positive and statistically significant, the researcher can claim evidence of a positive relationship between the covariate and outcome, even after controlling for the other covariates in the model.

Second, researchers should consider the coefficient of determination, or the R^2, in linear regression modeling. The R^2 will relay how well the regression line approximates the real data points; the higher the R^2 the better the agreement between the observed and modeled values—an R^2 value of 1.0 indicates perfect agreement, an R^2 value of 0.0 indicates no agreement. Another way of interpreting R^2 values is that they indicate the proportion of the variation in the dependent variable explained collectively by the independent variables in the model. In the previous example, an R^2 value of 0.50 would indicate that half of the variation in adults' REE can be explained by differences or variation in their heights, weights, and ages. When additional covariates are added to a linear regression model (eg, x_1 and x_2), the R^2 value will either remain constant or more likely increase, since additional information should help predict values in the dependent variable, even if only by chance. Thus, in addition to the simple R^2 value, adjusted R^2 values are generated in multiple linear regression models. These values will increase only if the new covariate improves the model more than would be expected by chance.

There are also additional considerations that are important in conducting and reporting results from multiple linear regression models. First, when interpreting the results of any regression equation, it is vital to report the unit to which the regression function corresponds. For example, if an investigator were interested in examining the relationship between sodium intake and diastolic blood pressure, it would be meaningless to report the independent variable in one mg sodium increments. Increments of perhaps 1,000 mg sodium would likely be more interpretable. Second, it is also important to consider the context of the regression equation. For example, the Harris-Benedict equation was developed nearly a century ago and may not reflect the body weight and racial diversity prevalent in society today; thereby limiting its applicability for certain clinical populations. As with any statistical tool, the reader should be well informed of its inherent and contextual limitations.

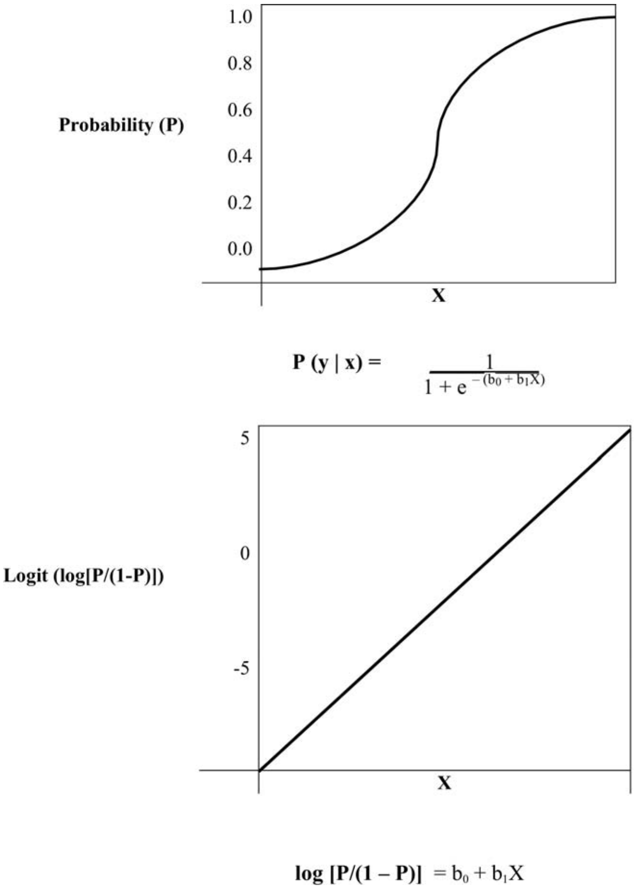

Figure 3. Graphic and mathematic equivalent formulations of the logistic regression function.

LOGISTIC REGRESSION

Although linear regression is appropriate when the dependent variable is quantitative, logistic regression modeling is used in epidemiologic studies or other nutrition-related research where the dependent variable is binary (eg, 0 and 1) or dichotomous (eg, yes/no or alive/deceased.) For example, the probability that an individual will develop coronary heart disease by a certain age might be predicted by family history, serum cholesterol, body mass index, and diet quality. The basic logistic regression model can be expressed mathematically as follows:

$$\log (\mathrm{P}/1-\mathrm{P})=\log \text{odds}=\beta_0+\beta_1\mathrm{X},$$

where P denotes the probability of the outcome, the intercept (β_0) is an estimate of the log odds of the outcome when X=0, and the coefficient (β_1) reflects the estimated increase in log odds of the outcome per 1-unit increase in X (7). A graphic depiction of this mathematical formulation is depicted in Figure 3. Fundamentally, it is critical to understand that this type of modeling is designed to describe the probability of an outcome given a set of numerical or categorical risk factors. The logistic function, or (P/1−P), describes a probability and, therefore, has a limited range between 0 and 1. The end value reflects the natural logarithm of the odds of the outcome, also called the logit or the log odds. This type of modeling may be used in case-control, cross-sectional and prospective studies and is considered the most popular modeling procedure (9). This is due, in part, to the ease and interpretability of its regression estimates that can easily translate into an odds ratio (OR). The OR expresses the probability of having the exposure among those with the outcome (cases) divided by the odds of having the exposure among those without the outcome, given a set of risk factors (5). An OR should always be accompanied by a 95% CI. The 95% CI is similar but distinctly different from the coefficient of determination used in linear regression. Specifically, the 95% CI relays the reliability of the point estimate and is calculated using information from the variance-covariance matrix. Although this derivation is beyond the scope of this review, it is important to understand that a narrow CI (eg, 2.2 to 2.8) implies high precision and a wide confidence interval (eg, 2.2 to 34.8) implies poor precision.

To better illustrate logistic regression modeling, suppose an investigator wanted to examine the relationship between the presence or absence of infection and parenteral nutrition administration in a cohort of patients with cancer. Here, the dependent variable is infection (eg, yes/no) and parenteral nutrition is the independent variable, or the exposure variable. This is simply presented as:

$$\text{OR} = \beta_0 + \beta_1(\text{parenteral nutrition}),$$

where parenteral nutrition=1 or 0. Inherent in this example is the need for a non-exposed group, or a control group where parenteral nutrition=0. This group should have similar baseline clinical characteristics to the other group, but did not receive parenteral nutrition. Limiting the analyses to this simple relationship reveals a univariate logistic regression model, which is often referred to as the crude model. However, most analytical approaches require several other characteristics of the participants be considered to help increase the predictive capabilities of the model and to control for specific confounders (ie, factors known to influence the exposure–disease relationship). Similar to multivariate linear regression, this is known as multivariate logistic regression modeling. There are many approaches to model building, all of which include practical and mathematical decisions and some consideration of additional characteristics of the participants. For example, depending on the research question or population under study, most models control for age and sex. Some statistical programs can provide the best model when a set of given variables are provided; however, these best models should be reviewed carefully because they may not reflect a model that includes all expected variables. Often, some variables will need to be forced into the model based on convention and the best approach may not take this into consideration (eg, age, sex, and race/ethnicity).

To illustrate an approach to fitting a multivariate logistic regression model, we will continue with the previous example using the output shown in the Table. Each row of the table represents a different logistic regression model reflecting the odds of infection and a given set of independent variables; parenteral nutrition is considered the main exposure variable. The α, or the intercept only model, is often conducted to assess the background odds of the outcome. In this example, any patient admitted for this type of cancer treatment has a 40% likelihood of developing an infection, regardless of the exposure to parenteral nutrition. Model 2, or the crude model, reflects the odds of infection when only the exposure variable is added (when $X_1=1$). The crude model is often thought of as the comparative model because it serves as the baseline for which all other variables are considered for inclusion. In this model, patients admitted for this type of cancer treatment who receive parenteral nutrition are approximately 2.2 times more likely to develop an infection than patients who do not receive parenteral nutrition, not taking into account other clinical characteristics (ie, β_2, β_3, β_4, β_5, and β_6). In Models 3 through 6, we observe very little movement of the OR and 95% CI when age, sex, and race/ethnicity are included independently, respectively, or when added together (Model 8). In Models 6 and 7, there is a slight decrease in the OR and 95% CI when diagnosis or treatment, respectively, are added to the crude model. These changes reflect a confounding association of these independent variables on the exposure-disease relationship. In Models 9 and 10, we continue to observe a slight alteration in the OR and 95% CI when diagnosis and then treatment are included along with age, sex, and race/ethnicity. To interpret either of these models, one can say that patients admitted to the hospital for this cancer treatment who receive parenteral nutrition are about two times more likely to develop an infection after controlling for age, sex, race/ethnicity, diagnosis (Model 9), and Treatment (Model 10).

However, in model building it is always critical to assess if there is evidence of effect modification (or interaction). This is necessary when the investigator suspects that the outcome may be significantly altered by the presence of another variable in the model. Typically, there is a rationale or biological plausibility underlying why the two independent variables may be interacting to greatly alter the point estimate. To test for interaction, entails adding product terms to your multivariate logistic regression model. For example, in the previous example, perhaps the investigator suspected that the risk of infection was significantly altered by diagnosis or a specific treatment modality. Using multivariate logistic regression, one can test these effects by inserting the product terms of parenteral nutrition×diagnosis and parenteral nutrition×treatment into the regression equation. As seen in Models 11 and 12, significant decreases in the ORs and accompanying 95% CIs occur. In this cohort a particular diagnosis confers a specific treatment (coded as 0 or 1); therefore, a more logical approach to presenting these data would be to present the results separately for each treatment regimen. Based on the models presented in the Table, the researcher can conclude that the odds of infection increase with parenteral nutrition administration in this cohort of patients with cancer and that this relationship is modified by which particular cancer treatment the patient receives and is largely unaffected by age, sex, or race/ethnicity. Additional modeling and analysis would then be required for each treatment strata to best optimize the model reflecting the exposure-disease relationship of interest.

SURVIVAL ANALYSIS

Cohort studies allow us to follow an exposure, such as a specific diet or behavior, forward to an outcome such as death, hospitalization, or having a body mass index >30. The most powerful analyses of this type of study examine

Table. Sample logistic regression models, parameter estimates, standard errors (SEs), odds ratios (ORs) and 95% confidence intervals (CIs) examining the associations of infection (the outcome) and parenteral nutrition (PN) administration (the exposure)

Model[a]	Intercept±SE	Exposure±SE	β_2±SE	β_3±SE	β_4±SE	β_5±SE	β_6±SE	OR (95% CI)
Model 1: Intercept	.3336±.1073							1.40 (1.25-1.55)
Model 2: PN	−.0904±.1608	.7687±.2192						2.16 (1.40-3.30)
Model 3: Age	−.1366±.4548)	.7712±.2204	.000921±.00847					2.16 (1.40-3.33)
Model 4: Sex	−.0255±.1835	.7749±.2196	−.1627±.2216					2.17 (1.41-3.33)
Model 5: Race/ethnicity	−.0916±.1951	.7689±.2197	.00234±.2197					2.16 (1.40-3.32)
Model 6: Diagnosis	−.3641±.2076	.7090±.2219	.4849±.2285					2.03 (1.32-3.14)
Model 7: Treatment	−.2353±.1731	.7186±.2215	.5761±.2448					2.05 (1.33-3.17)
Model 8: Age+sex+race	−.0673±.4959	.7777±.2216	.000705±.00856	−.1625±.2220	.0126±.2220			2.18 (1.41-3.37)
Model 9: Age+sex+race+diagnosis	−1.0521±.6508	.7356±.2238	.0111±.00967	−.0796±.2264	.1609±.2327	.6294±.2656		2.09 (1.35-3.24)
Model 10: Age+sex+ race+diagnosis+treatment	−1.2433±.6583	0.7115±.2252	0.0131±.00972	−0.0590±.2277	0.1898±.2343	0.5237±.2713	.5066±.2580	2.04 (1.31-3.17)
Model 11: PN+diagnosis+ PN×Diagnosis	−.1769±.2435	.3105±.3553	.1539±.3245	.6486±.4550				1.36 (.96-1.95)
Model 12: PN+treatment+ PN×treatment	−.1035±.1859	0.4640±.2581	.0522±.3704	.9397±.5031				1.59 (1.23-2.06)

[a]Variable definitions: age is continuous; sex is coded as 1=woman, 0=man; race is coded as 1=white, 0=other; diagnosis is coded as 1=multiple myeloma, 0=other diagnosis; and treatment is coded as 1=aggressive chemotherapy, 0=non-aggressive chemotherapy.

factors that influence the length of time until the outcome occurs. So in this context, survival is remaining in the disease-free state or being free from whatever physical condition (eg, obesity) is being analyzed.

Survival analysis has many of the same parameters as previously described for logistic regression but adds additional information to that approach on the time to the event. Using our previous example of parenteral nutrition exposure and infection, we could use survival analysis to examine time until first infection in a group of patients who received parenteral nutrition compared to a group who did not receive parenteral nutrition. The investigator could test the hypothesis that infections would occur more rapidly in the parenteral nutrition exposed group based on the higher incidence of hyperglycemia observed in parenteral nutrition recipients. This type of analysis enhances the previous logistic regression findings further by reinforcing the strength of the association in a different context and provides a foundation to support a cause and effect premise. However, the difficulty in this example is the varying time points of actual exposure since the non–parenteral nutrition group did not actually receive parenteral nutrition. To determine these temporal changes in infections, standardized time frames would need to be created to allow for adequate comparisons for the groups across time.

Another more straightforward example of survival analyses could be an investigator who wants to examine the association between dietary cholesterol and heart disease in individuals over a 10-year period. Logistic regression could be used to test the association between the dietary data, including dietary cholesterol intake, and the incidence of myocardial infarction and produce an OR to summarize the estimated risk. However, survival analysis would take advantage of additional information that is not captured in the logistic regression model. The time to the event may contribute important information on the difference between two diets (eg, defined Western diet and low [total and saturated] fat, low cholesterol diet). In theory, such research could conclude that individuals on one diet had a longer disease-free period or time until a cardiovascular event compared to the other diet group. Survival analysis is also conducted using a regression model, but the model always includes a factor representing the timing of the outcome (10). The actual statistical test is the Cox regression analysis test and is represented as:

$$\mathrm{Log}(t)=\beta_0(t)+\beta_1 x_i$$

In this mathematical formula t=time, β_0 is the intercept or constant, and β_1 is the beta or slope. The inclusion of a time factor allows for the examination of the risk of the outcome based on estimates at any one point in time following initial enrollment. The risk is actually evaluated in a cumulative fashion over the time of the study. The test yields a hazard ratio that is a relative risk estimate for the association between the exposure or risk factor and the health outcome of interest (11).

Censored Values

To conduct a survival analysis, the investigator creates two variables, one variable for the endpoint that indicates whether the event or condition ever occurs and a second variable for the time factor. The endpoint in this example could be one of two options: either participants are diagnosed with a myocardial infarction (endpoint=1) during the follow-up period, or they are not diagnosed with a myocardial infarction (endpoint=0). Of those who do not have the endpoint, some participants may be lost to follow-up or some may die from another cause. These are termed censored outcomes because from that time forward the participant is no longer at risk of the outcome myocardial infarction and should be essentially removed from further analysis. The time variable is generally counted from the day of enrollment in the study and is usually recorded in days. In this example the potential day of the endpoint would then be between 0 and 3,652 representing the 10 years of follow-up. An extension of our example, including the endpoint and time factors is an individual who has a myocardial infarction at day 456. The endpoint would be one and the time would be 456.

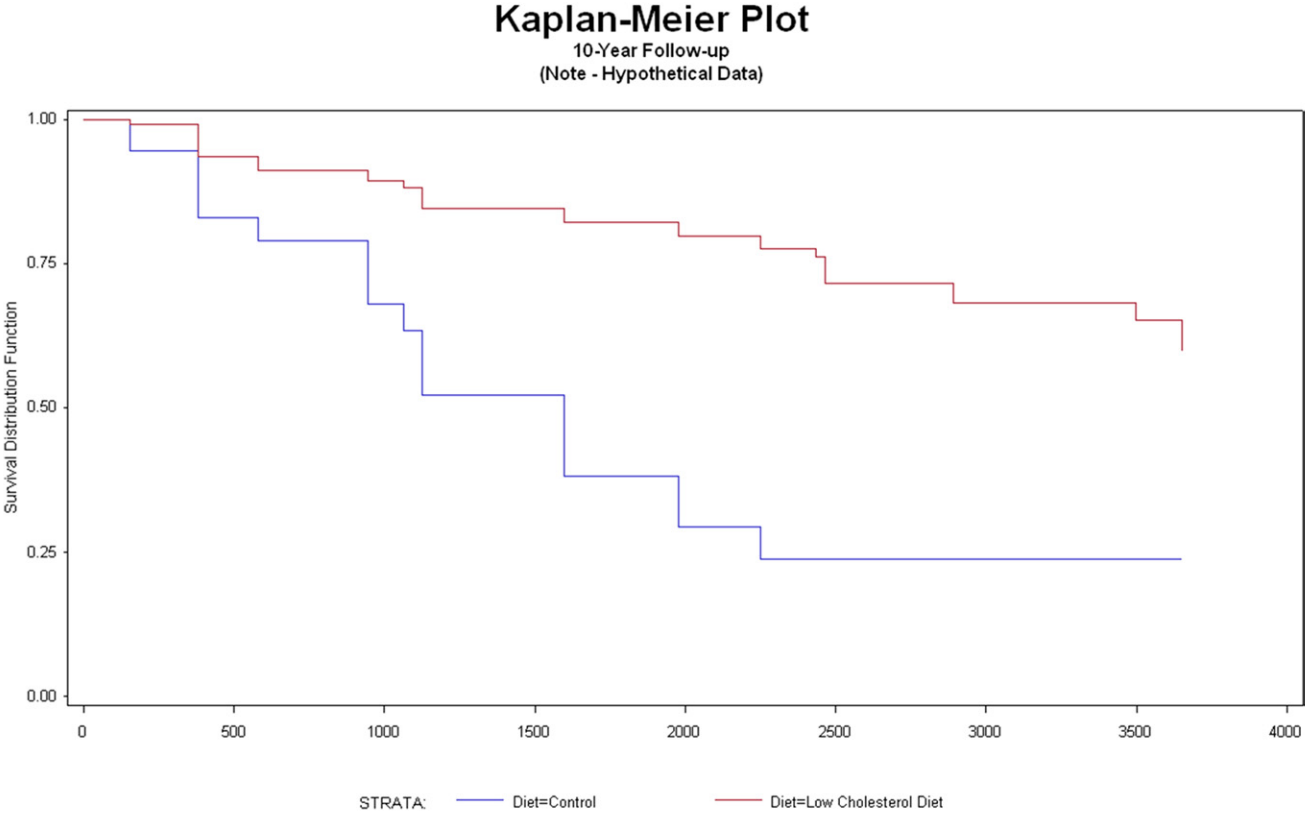

Figure 4. Kaplan-Meier Plot using a hypothetical example reflecting a 10-year follow-up period.

For a participant lost to follow up during the sixth year the endpoint would be zero and the time would be the day of last contact, such as day 2,242. At day 2,243 this person is censored from the analysis because this person can no longer provide an estimate on the association between dietary cholesterol and the incidence of a myocardial infarction. Those who complete the study without a myocardial infarction would have an endpoint of zero with a time of 3,650. This regression analysis provides the overall estimate of the association between the dietary groups and the outcome with consideration for the time to the event and the use of censoring to account for nonevents during the study. This is a more refined analysis and makes use of all of the information on the individuals. The Cox regression analysis has provided an estimate of the survival or hazard analysis over time (12).

Kaplan-Meier Plots

In addition to the regression analysis, a survival analysis may include a Kaplan-Meier plot, which is an illustration of the change in the survival curve from the initial time point of enrollment at Day 0 to the closure date of the endpoint at 10 years. This curve is a very useful visual of the relationship. An example of a Kaplan-Meier curve is presented in Figure 4. To orient to the plot, consider the x axis a time continuum from the start of the study (Day 0) to the last day of the study (Day 3,650). The y axis represents the participants still at risk of the endpoint throughout the study with an estimate of the association. At enrollment, all data points are at the upper left of the graph showing all participants free of the endpoint. Over the course of the study, as events occur (a myocardial infarction) (lost to follow-up or death from another cause) the curves will slope down. However, if a participant has been censored the slope does not drop at the date when the individual was censored but at the next date when there is an event. Therefore, a drop in the curve always indicates that there has been at least one event but it may also reflect any censored participants since the last event. At the end of the study, the final position of the line represents both loss to events and to censored participants and reflects only those individuals still at risk for the outcome at the end of the study. If there were no events (a myocardial infarction) or censored observations both curves would be straight across. Every drop in the slope represents one or more events (occurrences of the outcome among sample members) at that point in time. In the example where the question compares two diet groups, there would be one curve representing the low-cholesterol diet and one for the control diet. In addition to the graphic presentation, the Kaplan-Meier method estimates a log-rank test, which will compare the difference between these two slopes at the end of the study. A more common use of a log-rank test is the Mantel-Haenszel test, which also compares two samples with a nonparametric distribution. In the example of the Kaplan-Meier method the data are nonparametric because the distribution is right censored as opposed to a normal curve. (See Figure 4 to confirm the shape of the distributions.) The actual log-rank test that was conducted on this hypothetical data produced a P value of <0.001. The interpretation of the information presented on the plot, combined with the finding of the log-rank test, could be summarized as: There was a significant difference between the two diets in this study in the time to a myocardial infarction with the individuals on the low cholesterol diet remaining free of a myocardial infarction longer than the individuals on the defined Western diet.

MULTIVARIATE COX REGRESSION ANALYSIS

The examples above were univariate approaches in which the estimates were made with only the exposure (low [total and saturated] fat, low cholesterol diet and the control diet (Western diet) and the outcome (myocardial infarction or no myocardial infarction) in the model. As described in the earlier section there may be extraneous factors that influence the basic relationship of interest (ie, confounding). Therefore, it may be necessary to include possible covariates in these analyses. However, although Cox regression analysis is well suited to the inclusion of covariates, the Kaplan-Meier plot and the log-rank test will only test the univariate model. For this reason, the combination of the two approaches provides an easily interpretable graphic of the overall shape of the differences over time and a test result on the relationship, which may include confounding factors (10). In this hypothetical example, the actual multivariate Cox regression analysis, including adjustment for age and sex, yielded a hazard ratio of 0.22 (95% CI of 0.16 to 0.29; $P<0.0001$). The conclusion is that there is a significant protective effect in the time to a myocardial infarction with the low (total and saturated) fat, low cholesterol diet compared to the time until myocardial infarction with the defined Western diet after adjustment for age and sex.

Survival analysis methods have been used in many large prospective studies involving diet and health such as the Nurse's Health Study, the Women's Health Initiative, and the Framingham Heart Study. Therefore, it is very useful for nutrition researchers and dietetics practitioners to have the skills to interpret these statistical tests.

CONCLUSIONS

Before embarking on the aforementioned statistical procedures, it is assumed that careful attention has been paid to all data collection methods. Although the topic of data management is not addressed in this review, it is important to highlight that the statistical analyses, no matter how rigorous, will be fraught with inherent errors and lead to meaningless or flawed results when recorded data are not reflective of the study conditions. Quality assurance techniques and standardized approaches to data gathering and data management should be sufficiently detailed within the methods section of every article before the description of the statistical approach. This review has provided a basic explanation of commonly used multivariate techniques in nutrition-related research, simple and multiple linear and logistic regression, as well as Kaplan-Meier and Cox Proportional Hazards regression. Understanding the constructs and interpretation of these techniques can help considerably when designing nutrition research, and also when analyzing, interpreting, and publishing study findings. Consultations and collaboration with a biostatistician is highly recommended because these types of advanced analyses involve sophisticated statistical programming and statistical expertise. Future articles in this series will introduce other multivariate statistical techniques encountered when conducting and publishing nutrition-related research.

STATEMENT OF POTENTIAL CONFLICT OF INTEREST: No potential conflict of interest was reported by the authors.

References

1. Boushey C, Harris J, Bruemmer B, Archer S, Van Horn L. Publishing nutrition research: A review of study design, statistical analysis, and other key elements of manuscript preparation, Part 1. *J Am Diet Assoc.* 2006;106:89-96.
2. Boushey C, Harris J, Bruemmer B, Archer S. Publishing nutrition research: A review of sampling, sample size, statistical analysis, and other key elements of manuscript preparation, Part 2. *J Am Diet Assoc.* 2008;108:679-688.
3. Harris J, Boushey C, Bruemmer B, Archer S. Publishing nutrition research: A review of nonparametric methods, Part 3. *J Am Diet Assoc.* 2008;108:1488-1496.
4. Harris J, Gleason P, Sheean P, Boushey C, Beto J, Bruemmer B. An introduction to qualitative research for food and nutrition professionals. *J Am Diet Assoc.* 2009;109:80-90.
5. Bruemmer B, Harris J, Gleason P, Boushey C, Sheean P, Van Horn L. Publishing nutrition research: A review of epidemiological methods. *J Am Diet Assoc.* 2009;109:1728-1737.
6. Gleason PM, Harris JE, Sheean PM, Boushey CJ, Bruemmer B. Publishing nutrition research: Validity, reliability, and diagnostic test assessment in nutrition-related research. *J Am Diet Assoc.* 2010;110: 409-419.
7. Szklo M, Nieto FJ. Stratification and adjustment: Multivariate analysis in Epidemiology. In: *Epidemiology Beyond the Basics*. 1st ed. Gaithersburg, MD; Aspen Publishers; 2000.
8. Harris JA, Benedict FG. *A Biometric Study of Basal Metabolism in Man*. Washington, DC: Carnegie Institute; 1919. Publication No. 279.
9. Kleinbaum DG. Introduction to logistic regression. In: Dietz K, Gail M, Krickeberg K, Singer B, eds. *Logistic Regression: A Self Learning Text*.1st ed. New York, NY: Springer-Verlag; 1994.
10. Royston P, Parmar MKB, Altman DG. Visualizing length of survival in time-to-event studies: A complement to Kaplan-Meier plots. *J Natl Cancer Inst*. 2008;100:92-97.
11. Dekker FW, de Mutsert R, van Dijk PC, Zoccali C, Jager KJ. Survival analysis: Time-dependent effects and time-varying risk factors. *Kidney Int.* 2008;74:994-997.
12. van Dijk PC, Jager KJ, Zwinderman AH, Zoccali C, Dekker FW. The analysis of survival data in nephrology: Basic concepts and methods of Cox regression. *Kidney Int.* 2008;74:705-709.