QUANTITATIVE RESEARCH SERIES

Regression Modeling Strategies: An Illustrative Case Study From Medical Rehabilitation Outcomes Research

Todd G. Nick, J. Michael Hardin

Key Words: forecasting • research design • statistics

The practice of outcomes research is growing in all segments of the health care industry, yet few practitioners and researchers are prepared to deal with the completion of statistical analyses that characterize the new focus on results. This article discusses basic model formulation and interpretation. It also encourages the use of statistical models that study the simultaneous effects of many variables on an outcome and gives examples of relationships among variables that are not simple and linear. The methods are illustrated with a dataset consisting of stroke rehabilitation inpatients discharged during a 3-year period with an admission date that is within 1 year after stroke.

Nick, T. G., & Hardin, J. M. (1999). Quantitative Research Series—Regression modeling strategies: An illustrative case study from medical rehabilitation outcomes research. *American Journal of Occupational Therapy, 53,* 459–470.

The place of outcomes measurement in health care has been ameliorated recently by new accreditation standards promulgated by the Joint Commission on Accreditation of Healthcare Organizations (JCAHO). Known as the ORYX initiative, JCAHO's "next generation of accreditation" requires hospitals and long-term-care facilities to select a number of performance measurements for JCAHO to begin monitoring in 1998 (JCAHO, 1998). Home health care agencies, ambulatory care facilities, and others must announce their selected indicators in 1999. Although medical rehabilitation organizations have a relatively recent history of monitoring their programs, CARF…The Rehabilitation Accreditation Commission has encouraged accredited organizations to collect data on patient outcomes and to have ongoing program evaluation (Wilkerson & Johnston, 1997). In short, health care organizations and the professionals that staff them now face the daunting task of statistically proving their worth.

However, the measurement of health care outcomes is far from the straightforward description provided by many observers of the health care field. Whyte (1997) asked: If many different variables at one level may influence a variable at a higher level, why not study them all? He asserted that in-depth analysis should be the ultimate goal of theory building; however, he believed that this goal would be a long time in the making. Harrell, Lee, and Mark (1996) stated that if model building is applied uncritically to a dataset, the models will fit the data poorly or they will inaccurately predict outcomes on new subjects. For example, if only simple models (those that only allow for straight-line [linear] relationships between variables) are fitted to a dataset that has complex and nonlinear relationships, then those models will perform poorly and will not fit the data.

An example of a complex model (i.e., one that can be fitted with a multivariable model) is given in Pentland, McColl,

Todd G. Nick, PhD, is Associate Professor, Health Sciences, School of Health Related Professions, University of Mississippi Medical Center, 2500 North State Street, Jackson, Mississippi 39216-4505; tnick@shrp.umsmed.edu.

J. Michael Hardin, PhD, is Professor, Health Informatics and Biostatistics, School of Health Related Professions, University of Alabama at Birmingham.

This article was accepted for publication September 3, 1998.

and Rosenthal (1995). The authors found that as a person lived longer with a spinal cord injury (SCI), he or she felt less financially secure and experienced more symptoms and illnesses. These findings clearly convey to service providers and health care policymakers the added vulnerability of older persons with a disability. By using multivariable regression models, the authors could specify that the joint influence of age and duration of injury and level of lesion are related to long-term health outcomes, such as functional independence.

Many publications in professional journals also do not exhibit a simultaneous use of several variables. For example, only a few articles from *The American Journal of Occupational Therapy* (*AJOT*®) and *Physical Therapy* use statistical models that incorporate many variables simultaneously and consider the joint influence of predictors on a response. Jette and Jette (1996) and Mitchell and de Lissovoy (1997) are exceptions worth noting. Although many uses of analysis of variance (ANOVA), repeated measures ANOVA, and analysis of covariance (ANCOVA) were found, most of the articles include fewer than five total predictor variables and at most two continuous predictor variables. Furthermore, any inclusion of interaction terms is usually only associated with ANOVA, and forced straight-line relationships are common in regression analysis. Although Jette and Jette included many predictor variables, it is apparent that only simple and straight-line relationships were assessed. Mitchell and de Lissovoy assessed some interactions but only allowed for linear relationships among the predictors and outcomes.

We believe that it is important to demonstrate the handling of many predictors and various different types of predictors in one multivariable prognostic model. The methods we outline here are elaborated in greater detail in Harrell et al. (1996). They discussed, in addition to multivariable linear regression models, logistic and survival models and illustrated their methods with a survival analysis in prostate cancer. Our focus is on the multivariable linear regression model, which we will illustrate with the use of patients in a medical rehabilitation facility who have had a stroke. All analyses were done using S-PLUS Version 4.5[1] for Windows in conjunction with the Design library of Microsoft Windows S-PLUS functions (Harrell, 1998).

The application of statistical models relating multiple predictors (independent or explanatory variables, risk factors, treatments, covariates) to a single continuous response (dependent, outcome) variable is referred to as *multivariable modeling*. These models can handle a combination of dichotomous, nominal, ordinal, or continuous predictor variables (see Table 1 for examples of data types). Most textbooks refer to this type of statistical model as a multiple linear regression model. ANCOVA handles both categorical and continuous predictors as does multivariable linear regression models. In fact, the multivariable linear regression model is an ANCOVA-type model. However, ANCOVA models are mainly interested in categorical predictors, such as treatment given, and the continuous predictors, such as age, are introduced primarily to improve the precision of the statistical model (Neter, Kutner, Nachtsheim, & Wasserman, 1996). *Multivariate models* refer to models that simultaneously handle more than one outcome variable. Note that many researchers, including allied health professionals (see Portney & Watkins, 1993), who are not statisticians use the term *multivariate* to describe any statistical technique involving several variables, even if only one dependent variable is considered at a time (Kleinbaum, Kupper, & Muller, 1988).

Other practical multivariate models that are not discussed here include logistic regression models and Cox proportional hazard models (see Harrell et al., 1996). Logistic models are used when the outcome is dichotomous, nominal, or ordinal, such as with many clinical and disease outcomes (Kleinbaum, 1994). Cox proportional hazard models (a form of survival analysis) are used when the outcome is the time until some event occurs, such as with time to return to community in rehabilitation or time to return to work in an industrial rehabilitation setting (Kleinbaum, 1996). See Iwarsson, Isacsson, Person, & Scherstén (1998) for an example in *AJOT.*

Existing Databases

Before describing analytical strategies for outcome assessments, it is important to examine potential issues and problems found in many databases available to a health care organization. Many health care facilities have maintained their own databases for years. These databases often contain a rich variety of clinical information on patients over extended periods. The U.S. Health Care Financing Ad-ministration (HCFA) has also assembled large databases on Medicare patients. Typical information in the HFCA/Medicare database concerns the patient's hospitalization, surgical procedures, and office visits. In similar databases kept by Medicaid in many states, drug data may also be included. Wilkerson and Johnston (1997) provided a history and critique of rehabilitation clinical program–monitoring databases. These types of databases have been the logical starting points for researchers as they have begun to assess the quality of care and outcomes in health care. With the increasing regulatory agency and public demand for such assessments, McDonald and Hui (1991) predicted that funding for both the creation of these databases and their use will increase in the coming years.

What issues and problems, then, should an outcomes researcher be wary of when analyzing the aforementioned databases? First, it must be remembered that these databases originally were not created for research purposes; that is, patients entered into these databases were not included as part of a designed research project but were included simply because they sought care at the given health care facility. Hence, the data could have biases, including selection bias. Follow-up data may be absent because the patient sought care elsewhere after a bad experience. Patients who

[1]Mathsoft Inc., 101 Main Street, Cambridge, Massachusetts 02142-1521.

Table 1
Examples of Types of Data

Data Type	Examples
Dichotomous	Sex (male/female), bilateral involvement (yes/no), arthritis (yes/no), workers' compensation (yes/no), patient has attorney (yes/no)
Nominal	Impairment group, race (White, Black, Asian, Native American, etc.), prehospital living setting (home, skilled nursing facility, etc.), arthritis (rheumatoid, osteoarthritis, other)
Ordinal	Pain (none, mild, moderate, severe), patient status (unimproved, stable, improved)
Continuous	Age, length of stay, total Functional Independence Measure scores

are sicker tend to have more data than patients who are less sick because they visit the facility more often. Additionally, patient eligibility changes. Moses (1991) and Byar (1991) argued that bias is, in fact, the chief threat to database analyses.

The second problem facing an outcomes researcher was noted by Tierney and McDonald (1991). They claimed that there may be multiple measurements per patient for a particular physiological parameter. Thus, the researcher must determine which of the measures is clinically most appropriate for a given analysis. Similarly, several clinical indicators may be used across patients. Determining equivalence of indicators to allow for comparison across groups of patients can be problematic.

Third, data quality and reliability pose a serious problem for analysis of clinical and claims databases. For example, variation can occur in basic clinical measures, such as blood pressure, because of site differences, clinic differences, clinic focus, and so forth. Similarly, electrocardiogram readings in a coronary care clinic within the facility may vary in accuracy from those in the emergency department.

These various sources of bias and error can be controlled in several ways: stratification techniques, matching schemes that are based on relevant covariates, or adjustment factors and statistical models. Although an in-depth examination of these methodologies is beyond the scope of this article, it is nevertheless important that the reader gain an appreciation of the problems inherent to these kinds of databases and of the limitations that these problems may impose on the resulting analysis.

Case Study From Rehabilitation Outcomes Research

To illustrate various aspects of multivariable regression analyses, we used a sample consisting of 745 stroke rehabilitation inpatients discharged from a facility in Mississippi between January 1994 and December 1996. This dataset included patients within the first year of stroke and those who were admitted to the facility for the first time. Other researchers have addressed the use of this dataset to descriptively and graphically analyze functional status improvement (Nick, Williams, & Barker, 1998). All of the variables presented here are included in the Uniform Dataset, which includes measures of functional status, usually derived from the Functional Independence Measure (FIM℠).[2]

The FIM assesses physical and cognitive disability in terms of burden of care (McDowell & Newell, 1996). This test of functional status is usually administered on admission and discharge from a rehabilitation facility and assessed by various care providers, such as nurses, occupational therapists, physical therapists, and speech–language pathologists. The FIM is composed of an 18-item, 7-level scale (where 1 = total assistance and 7 = independent) of patient performance. By totaling the points on each item, the possible total score ranges from 18 (total dependence) to 126 (highest level of independence). The FIM total score can be separated into two major components—Motor and Cognitive scores. The Motor score represents 13 items for a minimum of 13 and a maximum of 91 and includes the items of self-care, sphincter control, transfers, and locomotion. The Cognitive score represents 5 items for a minimum of 7 and a maximum of 35 and includes the communication and social cognition items. See McDowell and Newell's (1996) review of the FIM for more information on this measurement method along with critiques of more than 80 other measurement methods.

Of the 745 patients in the current study, 711 had either left (right brain) or right (left brain) body involvement, whereas the remaining 34 had either bilateral involvement (n = 10), no paresis (n = 20), or some other stroke (n = 4). Because a predominant proportion (more than 95%) of the patients had only one body side that was impaired, we decided to include only those 711 patients in this analysis.

Formulating a Model

One of the first priorities of the investigator should be to determine the specific research question to be addressed. With large databases, it is often necessary to write special database queries in order to obtain the appropriate dataset for analysis. Without a clearly defined analysis plan, much time and energy can be lost regenerating analysis datasets. It is also important to note that most statistical methods require an assumption that the hypothesis being tested has been produced a priori. Some controversy exists about this assumption, and some new techniques, such as data mining (Mitchell, 1997), offer the promise of assisting investigators in the generation of interesting new hypotheses. However, at present the most secure route for most researchers is to follow the standard paradigm.

For our study, attention will be given to determining the relationships between various patient admission characteristics, including demographics, prehospital vocation, setting from which admitted, payment source, and insur-

[2]FIMSM is a service mark of the Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc.

ance status on length of stay (LOS) in rehabilitation. LOS represents a more traditional outcome in medical rehabilitation outcomes research (Wilkerson & Johnston, 1997). We use it here as a dependent variable for illustrative and instructive purposes. Our particular emphasis is on assessing the effect of payment source on LOS in rehabilitation while controlling for differences in demographic, admission functional status (FIM Motor and Cognitive scores), and other variables. If an effect is demonstrated between insurance status and LOS, this will support the role of payment source as an independent influence on LOS in rehabilitation. Ideally, a facility hopes that LOS is independent of payment source and that patients receive the amount of care, measured by LOS, that they need to achieve rehabilitation benefits or appropriate functional status.

Other important research questions that could be addressed with this dataset are: (a) What are the effects of treatment on LOS? (b) When are demographic and functional status differences among patients controlled for? (c) What are the effects of care on patients' functional status at discharge? Additional treatment variables would have to be collected, such as the amount or type of services they received in the rehabilitation center (e.g., occupational therapy, physical therapy).

In their discussion of conceptual models in rehabilitation, Duncan, Hoenig, Samsa, and Hamilton (1997) stated that an important question in stroke rehabilitation is: Does therapeutic exercise improve motor recovery and subsequently functional independence? To address this question, variables that take into account the time, frequency, and duration of therapy and the progression of exercise intensity would be included in the statistical model. As well as these variables, the type of therapeutic exercise used and the classification of the session (group or individual) would need to be characterized. In the absence of the amount and type of care that patients received, LOS could be used as a predictor variable. In this role, LOS acts as a proxy for the amount and type of care that patients received in a rehabilitation center.

When one is formulating good hypotheses, it is important to collect "good" predictor variables. Predictors are "good" if they are reasonable, appropriate in number, measured reliably, handled well, and characterized adequately. See Lynn, Teno, and Harrell (1995) for a discussion of these aspects of formulating models with regard to prognosticating death. Predictors should be selected on the basis of what is known about the determinants of the outcome (e.g., in the present study, LOS). When a researcher studies a more general population with stroke, SCI, and traumatic brain injury, the type of "impairment group" needs to be included in the statistical model because different impairment groups have different LOSs. One would need to consider what factors, based on clinical and scientific knowledge and the appropriate literature, seem likely to be important in predicting LOS and include those variables in the model. Otherwise, the model will not be accurate and will lead to estimates of effect that are not reflective of the actual population of interest. The other characteristics of "good" predictors will be discussed in the sections that follow.

Describing the Data

Before conducting formal statistical tests on a dataset, one must explore the variables in one dimension (univariate) both descriptively and graphically. This exploration allows a researcher to uncover the basic structure and information inherent in the data as well as uncover errors in the data. Describing the data includes the computation of simple summaries. For continuous variables, these summaries are provided by statistics, which include means, medians, quartiles, minimums, and maximums. When data are skewed (i.e., do not follow a normal distribution) or when outliers are present, which is generally the case with outcomes data, the center is more meaningfully measured by the median. In other words, if the data are skewed, then the median and quartiles are *better* statistics than the mean and standard deviation, respectively. The standard deviation is often subject to the mistaken belief that 95% of the observations can be expected to fall within two standard deviations from the mean (O'Brien & Shampo, 1981). Because of this common misconception, the standard deviation is used for descriptive purposes far more than it should be. For nominal and ordinal variables, the frequency and percentage of the categories are the appropriate summary statistics. For all types of data, the number of unique (i.e., the number of different scores) and missing values will aid an investigator in the analysis stage.

For the stroke dataset, recall that we computed statistics on only the 711 patients that had just one side of the body impaired. The simple summaries of the categorical variables are depicted in Table 2. For continuous variables, including the outcome of interest, LOS, the mean, and some important percentiles (5th, 10th, 25th, 50th, 75th, 90th) are reported in Table 3. A percentile indicates the percentage of a distribution that is equal to or below that number. The 25th percentile represents the lower, or first quartile, and the 75th percentile represents the upper, or third, quartile. The 50th percentile, or median, represents the middle value of a variable when the data are ordered by size. For example, the 25th percentile for admission FIM scores for patients with stroke from this facility is 43, with the lowest possible value being 18 (each item being rated as 1, with the patient needing total assistance). This means that 25% of the patients with stroke have FIM scores that are at most 43, and 75% have scores that are greater than 43.

Also reported in Table 3 are the number of unique values and the number of missing cases for each variable. Such detail would normally not be published in the final presentation of the results. However, it is crucial to visualize the data in great detail when developing a predictive model. Further discussion of missing values in data is given in the

Table 2
Descriptive Statistics for Nominal and Ordinal Predictor Variables

Variable	Unique (Missing)	Frequency, %
Group	2 (0)	Left brain (341, 48%), right brain (370, 52%)
Sex	2 (0)	Female (352, 50%), male (359, 50%)
Race	3 (1)	Native American (2, 0%), Black (303, 43%), White (405, 57%)
Year	3 (0)	1994 (285, 40%), 1995 (215, 30%), 1996 (211, 30%)
Payment source	5 (2)	Medicaid (55, 8%), Medicare (224, 32%), both (64, 9%), Medicare and private (231, 33%), private (135, 19%)
Prehospital vocation	4 (0)	Employed (129, 18%), not working (19, 3%), retired for age (421, 59%), retired for disability (142, 20%)
Setting admit from	3 (2)	Acute (449, 63%), home (173, 24%), nonacute (87, 12%)

next section.

A measure of variation that is becoming quite popular today, and deservedly so, is the inter-quartile-range (IQR), which is the difference between the 25th and 75th percentiles and contains the middle 50% of the data. It is common to simply report the median along with the lower and upper quartiles, such as the 50th (25th, 75th) percentiles. For example, the FIM scores in this study would be reported as 61 (43, 76).

Note that all of the variables except for chronicity (the number of days between onset of stroke and rehabilitation admission) have similar means and medians, meaning that the distribution of the variable is symmetric (not skewed). When the mean and median of a dataset are not equal, the shape of the distribution is skewed. For example, because the mean of chronicity (51) is significantly higher than the median (24), the shape is said to be positively skewed. The opposite applies to shapes that have a negative skew (see Figure 1 for the actual shapes of the continuous variables of this dataset). Examining the quartiles and other percentiles provide additional, valuable information about the shapes of the distributions.

Dealing With Missing Data

Attention to missing values is an important part of the data description process. Decisions must be made regarding the acceptability of missing values. The frequencies in Table 2 have very few missing data points. Only two patients are missing information for "setting admitted from," and two patients are missing payment source. In Table 3, three patients are missing data for chronicity. Missing values must be tracked down to determine whether, in fact, they were uncollectable data points.

Table 3
Descriptive Statistics for Continuous Predictor Variables and the Outcome Variable

Variable	Unique (Missing)[a]	*M*	Percentile						
			5th	10th	25th	50th	75th	90th	95th
Predictor									
Age	68 (0)	67	42	47	58	69	76	83	85
Chronicity	168 (3)	51	8	10	15	24	51	145	213
FIM at admission	94 (0)	60	24	30	43	61	76	87	94
Motor score (FIM)	72 (0)	39	14	18	26	38	49	60	68
Cognitive score (FIM)	21	21	5	7	15	22	28	32	34
Outcome									
LOS	25	25	10	13	17	24	30	37	42

Note. FIM = Functional Independence Measure; LOS = length of stay.
[a]*Unique*, as used here, simply means a value that produces only one result and is without a like or an equal.

When data are missing for the primary outcome, the patient record is usually deleted from the study. When data are missing for the predictor variables, commonly all observations are inappropriately discarded. For example, in studying the relationships between physical therapy and health outcomes in patients with knee impairment, Jette and Jette (1996) stated that "only data of patients with complete data for the independent variables of interest were included in the analyses" (p. 1179). Of their 426 patients who had complete episode of care and completed both initial and discharge health outcomes questionnaires, only 362 (85%) were included in studying predictors of the bodily pain physical health dimension of the 36-Item Short-Form Health questionnaire. Other dimensions analyzed included up to 405 patients. The disposal of data wastes valuable patient information and usually results in less accurate estimates of effect. If data must be discarded, statistical modeling should be used to characterize the reasons for the missing data (Harrell, 1997).

The most common type of imputation for a missing value or observation is to plug in or fill in the missing value with a descriptive statistic, such as the mean or median. However, if several values or observations are missing, special imputation methods should be used that take explanatory factors into account. See Rubin and Schenker (1991) for an overview of imputation strategies in health care databases. Most statistical packages, including SAS[3], SPSS[4], and S-PLUS, have routines, or at least additional add-on packages, that perform imputation. For the stroke sample, few values were missing; therefore, simple statistics were used to plug in a value.

Categorical Predictors

Whereas continuous variables are easily incorporated into a regression model, dichotomous, nominal, and ordinal predictors require additional attention because multiple regression models cannot handle character strings, such as

[3]SPSS Inc., 233 Wacker Drive, 11th Floor, Chicago, Illinois 60606-6307.
[4]SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

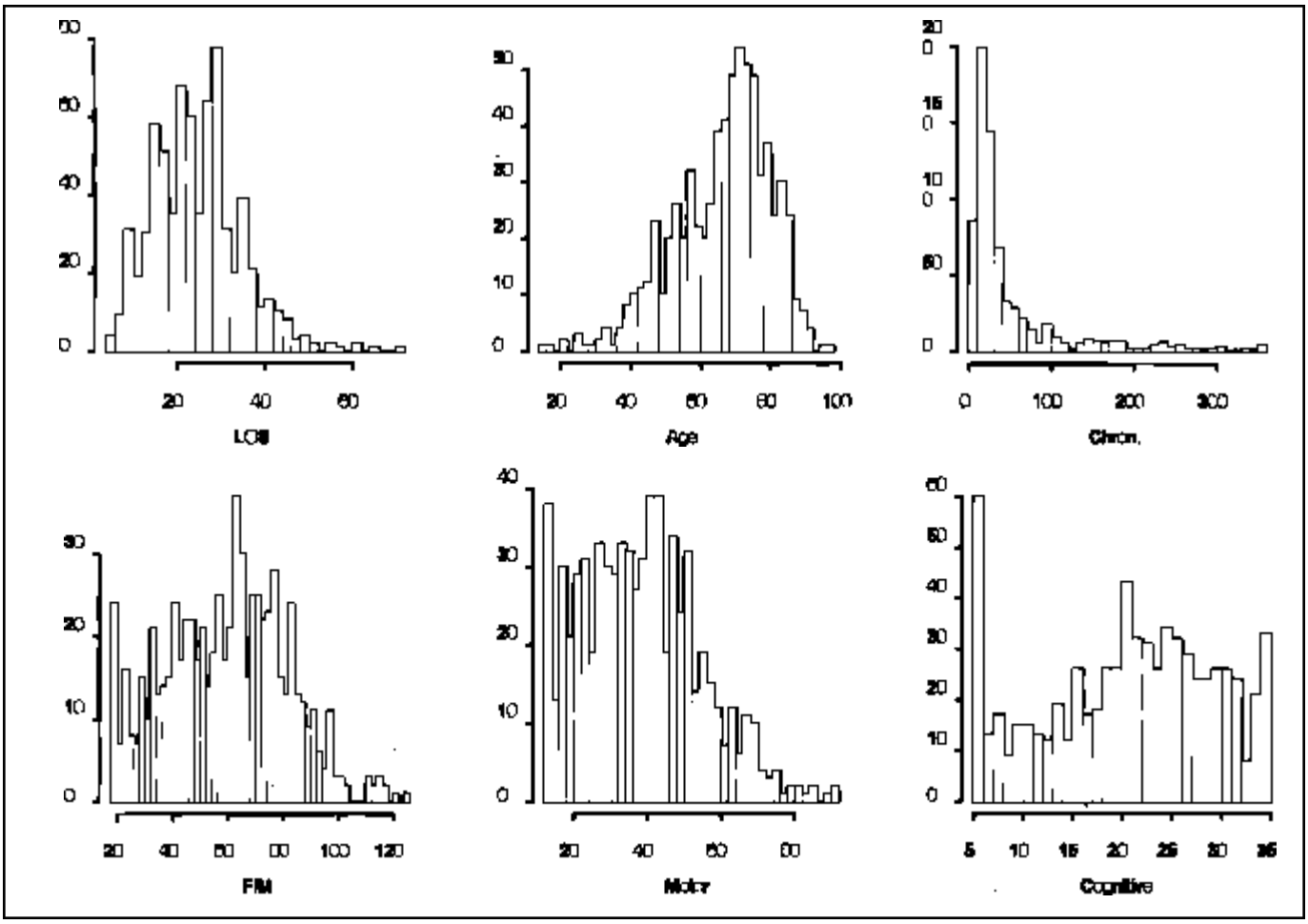

Figure 1. Histograms of all continuous variables under consideration. *Note.* LOS = length of stay; Chron. = chronicity; FIM = Functional Independence Measure.

Male/Female or M/F. This actuality, however, does not mean that only continuous variables should be analyzed in a regression analysis. In particular, when important predictors (be they continuous or categorical) are left out of a model, estimates of effects may be incorrect (Glymour, 1997). For dichotomous, nominal, and ordinal variables, design (dummy or indicator) variables are required. As we will describe, the number of model design variables required to represent a nominal predictor is one less than the number of categories of that predictor. Thus, a reduction in the number of categories of a predictor will reduce the number of variables required in a regression model.

Reducing the Number of Categories

Before coding the categorical predictors for regression modeling, it is important to inspect the frequencies of the categories of the nominal and ordinal predictors (i.e., left brain vs. right brain, male vs. female [see Table 2]). Such inspection will aid the investigator in determining whether a reduction of the categories is warranted. A reduced model would be easier to interpret and validate (see the "Sample Size Requirements" section).

As an example, the FIM has six levels of race (Asian, Native American, Black, Hispanic, White, and other). Assuming a small sample size of 20 patients, a model would not be useful if it were to require five design variables just to describe these six categories. The stroke dataset has only three of these six categories of patients as seen in Table 2 (Native American, Black, White) and requires two variables in a regression model. From the descriptive information, there are only two Native Americans. Forcing a variable for this category would be wasteful. It would be far better to collapse the Native American and the Black categories into a new category (*minority*), thereby reducing the number of total variables by one. Alternatively, the data on the Native Americans could be discarded. In general, it is better to collapse or merge categories into "similar" groups in lieu of deleting them. By "similar" we mean groups that are expected to behave alike with respect to the outcome and that should be supported by subject matter knowledge. It is not appropriate to base the reduction of categories on an inspection of the outcome data. This strategy leads to the reporting of biased models.

The variable group (impairment group) presents another potential category. Recall that we initially had 745 patients in the study. However, 95% of these patients represented single-side impairment; the remaining 5% had bilateral involvement, no paresis, or some other involve-

ment. We opted to delete the patients who did not have single-side involvement. An alternative strategy would have been to create an *other* category consisting of the three groups (bilateral, no paresis, other). Although this puts distinct categories together, it preserves the initial patient population and uses only one additional variable.

A more complex reduction may be required of other variables, such as payment source. The FIM has two variables that represent payment source—primary and secondary—each with 16 categories (e.g., Blue Cross, Medicare non–Managed Care Organization [MCO], Medicaid non–MCO, Commercial Insurance, MCO health maintenance organization, Medicare MCO, Medicaid MCO, workers' compensation]. If both of the two payment source variables are included and the categories are not reduced, 30 variables would be required in the regression model. This becomes impossible to interpret, even if a valid model could be derived. A better "reduced" model would be one that collapsed the two variables (primary and secondary) into one variable and reduced the 16 categories into a manageable number by forming like groups. With regard to the LOS outcome, this may entail forming three simple categories, such as (a) Medicaid only, (b) private only, and (c) other (includes Medicare and Medicaid, Medicare only, and Medicare and private). Note that the Medicaid group is not combined with any other category because it is a very different source from the others. Many other assortments of categories exist, but these three categories, which only require two variables in the regression model, will be used for this article.

Other nominal predictors in this dataset for which categories were collapsed, in addition to race and payment source, were prehospital vocation (employed, not employed [not working, retired for age, retired for disability]), and setting admitted from (acute, other [home, non-acute]). In summary, the 22 categories listed in Table 2 were reduced to 16 categories, which should increase the validity of the regression model.

For ordinal variables, it is valid to collapse a category with only a few patients into a previous category, thus reducing the number of variables required to model the ordinal variable by one. Using an example not found in the dataset, take a variable that measures pain with the ordinal levels none, mild, moderate, and severe. If there were only a few patients in the mild and severe groups, it would be appropriate to have only two categories for this variable: none to mild and moderate to severe.

Coding Nominal Predictors

Design variables are devices used to allow for categorical predictors in statistical modeling. For dichotomous predictors, such as sex (Male/Female), one design variable may be set to 0 if the patient is male and 1 if the patient is female. We therefore would have a new column of zeros and ones and not of males and females in our dataset. Payment source (Medicaid, Private, Other) requires two design variables. One design variable takes on the value of 1 if Medicaid and 0 if otherwise. The other design variable would be defined as 1 if private and 0 if otherwise. These two design variables completely define the three categories in the regression model. In general, one less design variable is required than the number of categories of a predictor. For example, with five different levels of a severity index (None, Mild, Moderate, Severe, Excruciating), four design variables are needed in the regression analysis.

Coding Ordinal Predictors

There are several ways to code ordinal predictor variables. If we code the variable year (1994, 1995, 1996) as 1, 2, and 3, we can only test for a linear relationship between the predictor and the response. Although this is commonly done, it is often incorrect to assume such a linear relationship between the ordinal variable and outcome because the results could be misleading. An assumption with the preceding codings is that the effect on LOS of the years 1994 and 1996 is extreme and that of year 1995 is between them. Here, the trend is not linear but rather a decreasing trend that is more so each year.

The nominal codings work well on ordinal variables with up to five categories. An adequate alternative is the ordinal codings presented in Walter, Feinstein, and Wells (1987), which allows a researcher to see the amount of change occurring from one category to the next. Table 4 applies the nominal and ordinal coding schemes on the year variable. With a software package, one often has to create two new design variables for year, for example, Year1 and Year2. These are usually created with if–then statements such as in SAS [If (Year = 1995) then Year1 = 1; If (Year ~= 1995) then Year1 = 0;] or in SPSS [If (Year = 1995) Year1 = 1. If (Year ~= 1995) Year1 = 0.]. The ~= symbol stands for *not equal to* in both packages. Alternatively, most statistical software packages have user-friendly dialogue boxes to recode variables.

Model Interpretation

To demonstrate the interpretation of regression coefficients and to compare and contrast the coding schemes in Table 4, we again use the Year variable (1994, 1995, 1996) alone in a regression analysis on LOS. The mean LOS for the 3 years are 26.9, 25.2, and 22.1 days, respectively, and the decreasing LOS trend is depicted in Figure 2 with a box plot. Box plots are used to compare commensurate values,

Table 4
Nominal and Ordinal Coding of Design Variables for Year

	Nominal		Ordinal	
Year	Year1	Year2	Year1	Year2
1994	0	0	0	0
1995	1	0	1	0
1996	0	1	1	1

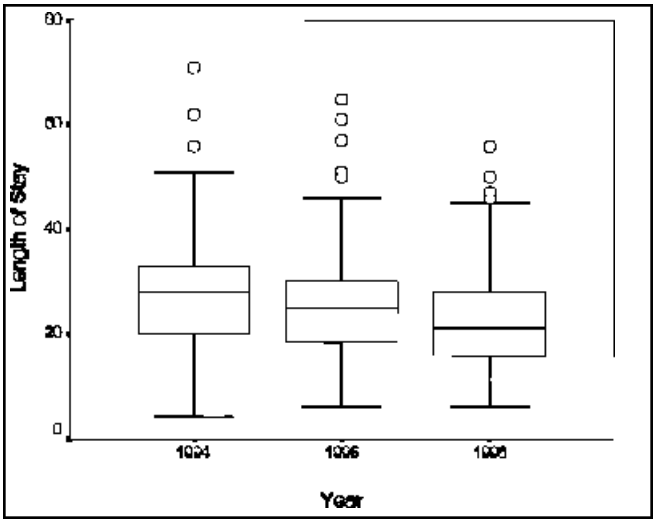

Figure 2. Box plot representing descriptive statistics of length of stay by year.

such as the median and lower and upper quartiles, of patients over various categories. The box extends from the lower to upper quartiles and the horizontal line in the box represents the median. Lines are usually drawn from the rectangle (box) to the 2.5th percentile and 97.5th percentile representing the middle 95% of the data. Outliers are represented with open circles.

Table 5 presents the usual output from a statistical package for the two aforementioned coding schemes. A regression model is usually described with a prediction equation or fitted model, and these can be written on the basis of the output. Using the estimated coefficients from above, we can estimate the LOS for patients with stroke at this facility by inserting the appropriate values for the design variables into the prediction equation. For the nominal coding scheme, the prediction equation is Predicted LOS = 26.9 – 1.7 x Year1 – 4.8 x Year2. The ordinal scheme has a prediction equation of Predicted LOS = 26.9 – 1.7 x Year1 – 3.1 x Year2. By inserting the appropriate values of the design variables into the prediction equation, taking the nominal scheme as an example, we see that 1994 patients with stroke [Year1 = 0 and Year2 = 0] have a predicted LOS of 26.9 days [26.9 – 1.7 x 0 – 4.8 x 0 = 26.9]. For 1995 patients, the predicted value of LOS is easily derived as well [26.9 – 1.7 x 1 – 4.8 x 0 = 25.2]. In fact, as long as the appropriate design codings are used, the codings will produce the same predicted values. The differences in the schemes are with the interpretation of the regression coefficients and their respective *P* values.

To interpret *P* values for the nominal coding scheme, we see that the coefficient and *P* value of the first design variable (Year1) is simply a hypothesis test for differences between mean LOS for patients in 1995 versus 1994. The hypothesis test for the second design variable (Year2) is a test for differences between mean LOS for patients in 1996 and 1994. From the *P* values and regression coefficient (see Table 5), it is apparent that mean LOS for 1995 patients is significantly different from the mean for 1994 patients (*P* value = .05), with 1995 patients having a mean LOS that is 1.7 days less than that of 1994 patients. It is also apparent that 1996 patients have a mean LOS that is significantly different from 1994 patients (*P* value ≤ .001), with a mean LOS that is 4.8 days less than that of the 1994 patients.

Table 5
Regression Coefficients and *P* Values for Coding Schemes

	Nominal		Ordinal	
Term	Coefficient	*P* Value	Coefficient	*P* Value
Intercept	26.9	≤ .001	26.9	≤ .001
Year1	–1.7	.005	–1.7	.05
Year2	–4.8	≤ .001	–3.1	≤ .001

For the ordinal coding scheme, the coefficient and *P* value of the first design variable (Year1) has the same interpretation as that of the nominal scheme. The coding schemes differ with respect to the coefficient and *P* value for the second design variable. The *P* value for the second design variable (Year2) is testing for differences between mean LOS for 1996 versus 1995 patients. From the *P* values and regression coefficient, it is apparent that the mean LOS of 1995 patients is 1.7 days less than that for 1994 patients (*P* value = .05) and that the mean LOS of 1996 patients is 3.1 days less than that for 1995 patients (*P* value ≤ .001). Thus, the regression coefficients are interpreted as amounts of change from the previous category. With this coding scheme, it is easy to see that the facility decreased LOS for 1995 patients by 1.7 days and by 3.1 additional days for 1996 patients.

Continuous Predictors

Model Interpretation

For continuous predictors, such as FIM Motor score, regression coefficients are simply interpreted as the change in the predicted LOS per unit change in admission FIM Motor score. Using Motor score as the only predictor for illustrative purposes, the predicted LOS = 32.2 – .20 x Motor. The regression coefficient of – .20 can be interpreted as LOS is decreased by .2 days for a 1-unit increase in Motor score at admission. The intercept is 32.2 days, and there is usually no interest in this parameter. A regression coefficient of 0 would imply no relationship whatsoever between the predictor and the outcome. A useful description is the 95% confidence interval for the coefficient. Because the sample is large, this formula is approximately [–.20 ± 1.96 (.02)] = [–.16, –.24]. The .02 is the standard error of the coefficient and is shown in typical regression output. Thus, we are 95% confident that a 1-unit increase in Motor score relates to an LOS decrease of between .16 to .24 days.

Although coefficients are most often reported in terms of a 1-unit increase in the predictor variable, a more useful description would be to determine a meaningful Motor score change, such as a change of 10-units. By simply multiplying the coefficient or confidence interval by this

change, the new statistics can be easily interpreted. For a 10-unit change in Motor score, the coefficient is –2.0 days with a 95% confidence interval of (–1.6, –2.4). Thus, comparing a patient with a Motor score at admission that is 10 units higher than another results in a shorter LOS by 2 days. It is often difficult to determine a meaningful change on the basis of data that have a tendency to be rather arbitrary in their fluctuations. A simple strategy is to use the change from the 25th to the 75th percentiles as a difference. This difference is commonly referred to as the IQR coefficient (Harrell, 1997). Here, the lower quartile is 26 and the upper quartile is 49 for a total change of 23 units, producing a coefficient of 4.6 with a 95% confidence interval of (3.7, 5.5). A patient whose Motor score is at the upper quartile at admission should have an LOS that is 4.6 days shorter than a patient whose Motor score is in the lower quartile at admission.

Because we only regressed Motor score on LOS, the model is assumed to be linear in Motor score. However, this linearity is often not the case, and curved relationships should be allowed for if these trends are expected. Graphically, the relationship between Motor score and LOS is depicted in Figure 3. These graphs also show the differences in three different cognitive values (low [less than 18], medium [18–25], high [≥ 26]). Plots A and C illustrate the differences in assuming a linear versus curved relationship. From Plot C, patients with Motor scores between 20 and 40 apparently have about the same stay. However, once beyond a Motor score of 40 at admission, there is a dramatic change in slope (change in LOS per 1-unit change in Motor score), which demonstrates shorter LOS for patients with higher Motor scores at admission.

Functional recovery can be nonlinear, as Johnston, Stineman, and Velozo (1997) have shown. Additionally, because existing outcome instruments commonly measure function over a limited range, there can be ceiling and floor artifacts. Allowing for nonlinear trends in the relationships between predictors and outcomes would take some of these artifacts into account.

The easiest way to incorporate curved relationships into a model is to include a squared term. For example, in the regression model, we would have to create a new variable, $Motor^2$, and include both Motor and $Motor^2$ in the regression model. Doing this produces a more complex prediction equation and affects our ability to interpret the coefficients. Now, the Predicted LOS = 24.3 + .25 Motor – .005 $Motor^2$. Thus, for a Motor score of 26, the lower quartile, we could plug in the values (Motor = 26 and $Motor^2 = 26^2 = 676$) into the equation to derive a predicted LOS of 27.4. For a Motor score of 49 (the upper quartile) the predicted LOS is 24.6. The difference in predicted LOS values [24.6 – 27.4 = –2.8] represents the new IQR coefficient. That is, a patient at the 75th percentile of Motor scores at admission would stay about 3 (exactly 2.8) days shorter than a patient at the 25th percentile.

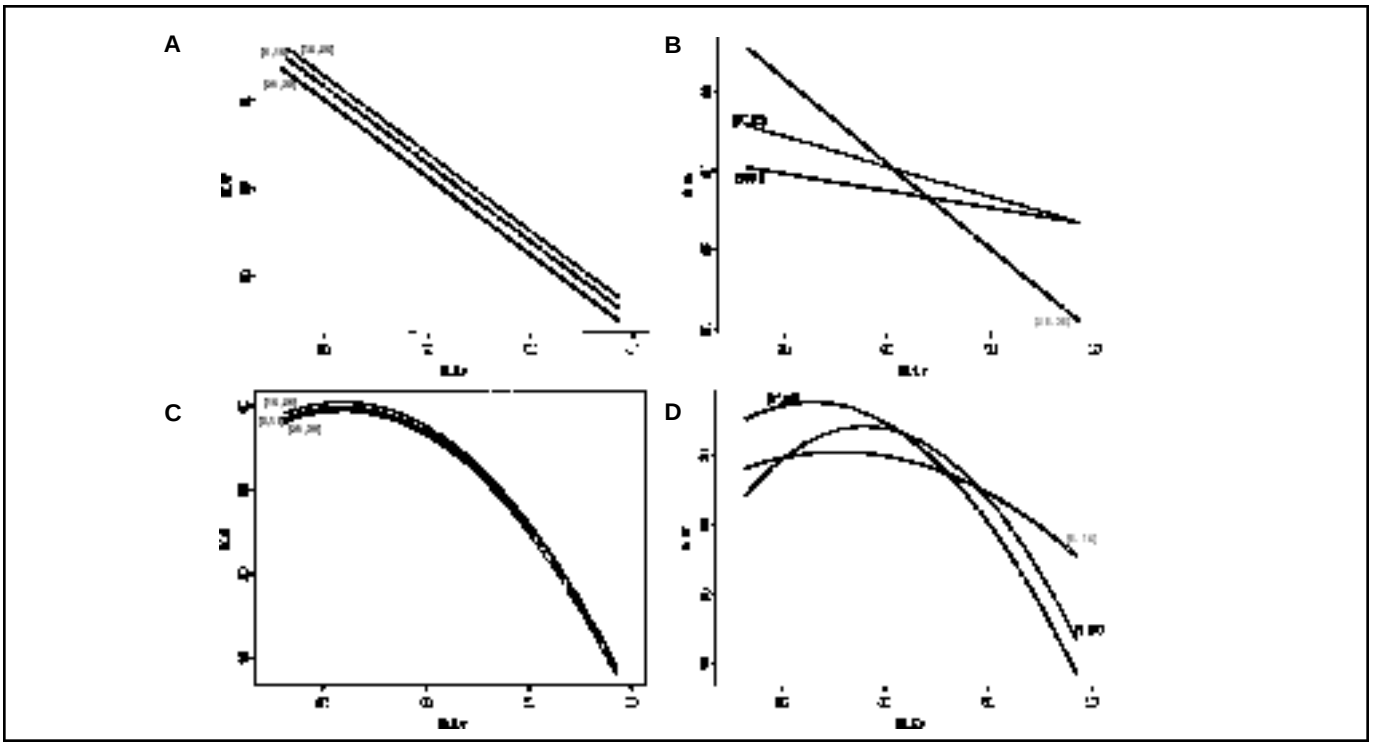

Figure 3. Length of stay predictions of Functional Independence Measure's Motor and Cognitive scores. Cognitive scores collapsed for graphical reasons into three equal categories: (a) < 18; (b) < 26 and ≥ 18; and (c) ≥ 26.These categories are represented in the graphs as [5,18), [18, 26), [26,35), respectively. The four graphs represent different scenarios of the assumptions of linearity and no interaction. From left to right: Plot A = assumption of linearity and no interaction; Plot B = assumption of linearity; Plot C = assumption of no interaction; Plot D = assumptions relaxed.

Interactions

One of the assumptions of multiple linear regression is that predictor variables are additive, meaning that the effect of a predictor on the outcome does not depend on another predictor. This assumption needs to be verified; if the assumption fails, we say that the two predictors are interacting with each other or that there is a statistical interaction present.

What happens if two predictor variables are not independent of each other? That is, suppose the effect of the patient's Motor score on LOS depends on the values of the patient's cognitive score. Whenever it is reasonable to believe that predictors have this joint influence on an outcome, specific interaction terms should be included in the statistical model (Gunst & Mason, 1980). Harrell et al. (1996) listed interactions that have consistently been found to be important in predicting clinical outcomes and, thus, should be included or prespecified in a regression model. One plausible interaction term is that of calendar time by study center. For example, if there are two study centers, it is possible that LOS is decreasing over time for one of the centers but is remaining constant for the other. Yet another important interaction is the quality and quantity of therapy. If patients receive individual therapy sessions, they might derive greater benefit, even if they are only seen once a week, whereas patients who receive only group therapy might profit from the treatment only if they are seen every day. Whyte (1997) gave examples of plausible interactions as well; for example, weakness may be relevant to gait dysfunction only if it is greater than some level, or pain may interact with weakness in ways that differ from those experienced by persons who have pain or weakness alone.

Interactions can be incorporated into a statistical model simply by multiplying the two terms together to form a cross-product term. For example, we would include a variable named MotCog = Motor x Cognitive in the regression model. Although the interpretation of interaction coefficients will not be addressed here, Figure 3 shows the effect of not allowing for interaction as well as the effect of forcing a straight-line relationship. A comparison of Plots A and B demonstrates the impact that an interaction term may have on a relationship. Assuming no interaction between Motor and Cognitive scores forces the three cognitive levels to have the same Motor and LOS relationship (see Figure 3, Plot A). As one would expect, patients with high Cognitive scores and low Motor scores stayed longer for inpatient rehabilitation than patients with high Motor and low Cognitive scores (see Figure 3, Plot B) at admission. It might be assumed that patients with good motor and cognitive function would need less rehabilitation and that patients with good cognition but poor motor function would have greater needs for and a larger capacity to benefit from inpatient rehabilitation. Plot D of Figure 3 illustrates how Cognitive and Motor scores together relate to LOS by including both interaction and nonlinear terms.

Sample Size Requirements

To produce valid statistical models, one must have appropriate sample sizes for the given number of variables in the model. Although there are many software packages for planning sample sizes that include routines for multiple linear regression models, these packages do require specific a priori hypotheses. For example, by using nQuery software (Elashoff, 1997), it is possible to determine that when there are 60 patients, the multiple linear regression test of no relationship (alpha = .05) for 5 predictor variables will have 80% power to detect an R^2 of .20. R^2 is the amount of variance of the outcome that is accounted for by the predictors. R^2 is a very useful measure of the model's predictive accuracy or specifically the model's ability to discriminate, which is its ability to separate patients' outcomes (Harrell, 1997). Additionally, for a multiple linear regression model that already includes 5 predictors with an R^2 of .20, we could determine that a sample size of 60 will have 80% power (alpha = .05) to detect an increase in R^2 of .10 due to including 1 additional predictor.

Although we encourage the preceding power computations whenever appropriate, in some instances, there is usually little a priori information available. Besides, the main concern of a study usually is whether the regression model is reliable or accurate. A general rule of thumb for multiple regression models is that there should be *at least* 10 participants per degree of freedom in the model (Harrell, 1996). *Degrees of freedom* does not simply mean the number of predictors, but the number of continuous predictors and their nonlinear terms (usually at least 3 degrees of freedom per continuous variable), design variables for categorical predictors (one less than the number of categories), and interaction terms. The interaction terms alone could potentially use an exorbitant number of degrees of freedom. Note that this rule of thumb is often quoted as 10 subjects per "variable," which often erroneously leads an investigator to have too few patients.

For example, suppose that a model is being developed from a sample containing 200 patients. The 10:1 rule suggests that we can examine, at most, 20 degrees of freedom. We wish to analyze payment source (5 categories); sex; race (Black, White, other); nonlinear effect of age, chronicity; FIM Motor score; FIM Cognitive score; and the possible interactions between race and age, race and chronicity, sex and age (only linear term), and sex and chronicity (only linear term). We would need 4 degrees of freedom for payment source, 1 for sex, 2 for race, 3 each for the 4 continuous predictors to allow for nonlinear trends, and 6 total for the interaction terms because race has 2 design variables. There is a total of 25 degrees of freedom, which is 5 more than the 10:1 rule. Now what? A simple strategy is to collect more data. If this is not possible, one might forego some of the nonlinear or interaction terms upon further investigation of the body of research on this outcome. Alternatively, the number of degrees of freedom could be

reduced by combining "like" categories or variables in a clinically meaningful way (see the Categorical Predictors section) by using summary scores. One could also use some sophisticated statistical methods, such as principal components analysis or variable clustering (Harrell, 1996).

Many researchers have used stepwise variable selection (stepwise regression), even when sample size is lacking. This type of variable selection produces R^2 values that are too high (higher than they should be) and P values that are too small (smaller than they should be). Hence, the results are better than they should be. Therefore, these stepwise procedures are not recommended for determining significant predictors. See Derksen and Keselman (1992) for problems associated with the use of stepwise regression. Additionally, simple bivariable analyses (which define the relationship between one independent variable and one dependent variable), such as t tests and correlation analyses, should not be used for selecting variables to be used in a multivariable analysis. Although simple analyses are commonly used for this purpose in the medical sciences, this use is inappropriate: It wrongly rejects potentially important variables when the relationship between an outcome and an independent variable is confounded by any confounder variable and when this confounder is not properly controlled (Sun, Shook, & Kay, 1996). In summary, associations should not be performed using the outcome variable to determine data reductions, and stepwise regression should be used with caution. See Sun et al. (1996) for some recommendations about variable selection and the use of stepwise regression.

If the data are already collected, a simple but informative strategy to determine whether there are too many variables, given the number of patients, is to run all of the variables (dummy variables, interactions, and the like) and determine the R^2 and adjusted R^2. The adjusted R^2 is computed in most linear regression outputs (Neter et al., 1996) and calibrates the R^2 by the number of variables. The adjusted R^2, and not the regular R^2, will be an accurate estimate of the model's predictive ability. If there are a sufficient number of patients given the number of variables, the adjusted R^2 should be within 90% of the R^2. For example, if the full model (all predictors) produces an R^2 of .50 and an adjusted R^2 of .40, there should be some concern.

The Final Model

Once the full model produces an adjusted R^2 that is close to the R^2 (90% of R^2), then interpretation of the full model can begin. With simple terms (i.e., terms not including nonlinear and interaction terms), IQR coefficients can be easily computed with almost any software. However, this is not the case with complex terms because there are multiple coefficients associated with each complexity. One way to compute IQR coefficients easily is by using S-PLUS Version 4.5 for Windows in conjunction with the Design library of Microsoft Windows S-PLUS functions (Harrell, 1998). As of June 1999, S-PLUS released S-PLUS 2000 for Windows. This recent release now includes the Design libraries of Harrell (1998). Of course, graphics should always be used for interpreting complex terms, such as in Figure 3.

Summary

In the published literature, little attention is given to the interpretation and presentation of the simultaneous effects of many variables on an outcome, and usually only relationships that are simple and linear are described. In this article, we addressed these issues and presented some strategies for handling complex terms.

We acknowledge that some of the methods and concepts in this article may be difficult to implement because they depend on the statistical experience of the researcher and the availability of statistical software. We encourage all researchers, preferably at the outset, to form a collaborative relationship with a statistician. See Moses and Louis (1992) on how to effectively collaborate with a statistician. Medical statisticians can be found at health science centers or through the American Statistical Association's (ASA's) Web site at www.amstat.org. By selecting ASA Directories, then Sections, and then Section on Statistical Consulting, consulting centers can be found at Centers & Facilities. Although one-on-one meetings facilitate a collaborative effort, long-distance efforts are almost as effective with the use of e-mail, fax, and telephone. ▲

Acknowledgments

We thank Robin Davis, MS, OTR/L, and John Barker, PhD, for input, critical comments, and thoughtful discussion of concepts in this article. We also thank Kathleen Savage for support in the preparation of this article.

References

Byar, D. P. (1991). Problems with using observational databases to compare treatments. *Statistics in Medicine, 10,* 663–666.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45,* 265–282.

Duncan, P. W., Hoenig, H., Samsa, G., & Hamilton, B. (1997). Characterizing rehabilitation interventions. In M. J. Fuhrer (Ed.), Assessing medical rehabilitation practices: *The promise of outcomes research* (pp. 307–317). Baltimore: Brookes.

Elashoff, J. D. (1997). *NQuery Advisor Version 2.0 user's guide.* Los Angeles: Dixon Associates.

Glymour, C. (1997). Social statistics and genuine inquiry: Reflections on The Bell Curve. In B. Devlin, S. E. Fienberg, D. P. Resnick, & K. Roeder (Eds.), *Intelligence, genes, and success: Scientists respond to The Bell Curve* (pp. 257–280). New York: Springer-Verlag.

Gunst, R. F., & Mason, R. L. (1980). *Regression analysis and its application.* New York: Marcel Dekker.

Harrell, F. E. (1997). *Predicting outcomes: Applied survival analysis and logistic regression.* Charlottesville, VA: University of Virginia.

Harrell, F. E. (1998). Design: S functions for biostatistical/epidemiological modeling, testing, estimation, validation, graphics, and prediction. (Program available from www.med.virginia.edu/medicine/clinical/hes/biostat.htm)

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions

and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15,* 361–387.

Iwarsson, S., Isacsson, Å., Persson, D., & Scherstén, B. (1998). Occupation and survival: A 25-year follow-up study of an aging population. *American Journal of Occupational Therapy, 52,* 65–70.

Jette, D. U., & Jette, A. M. (1996). Physical therapy and health outcomes in patients with knee impairments. *Physical Therapy, 76,* 1178–1187.

Johnston, M. V., Stineman, M., & Velozo, C. A. (1997). Outcomes research in medical rehabilitation: Foundations from the past and directions for the future. In M. J. Fuhrer (Ed.), *Assessing medical rehabilitation practices: The promise of outcomes research* (pp. 1–41). Baltimore: Brookes.

Joint Commission on Accreditation of Healthcare Organizations. (1998). *1998 comprehensive accreditation manual for hospitals: The official handbook.* Oakbrook Terrace, IL: Author.

Kleinbaum, D. G. (1994). *Logistic regression: A self-learning text.* New York: Springer-Verlag.

Kleinbaum, D. G. (1996). *Survival analysis: A self-learning text.* New York: Springer-Verlag.

Kleinbaum, D. G., Kupper, L. L, & Muller, K. E. (1988). *Applied regression analysis and other multivariable methods* (2nd ed.). Boston: PWS-KENT.

Lynn, J., Teno, J. M., & Harrell, F. E., Jr. (1995). Accurate prognostications of death—Opportunities and challenges for clinicians. *Western Journal of Medicine, 163,* 250–257.

McDonald, C. J., & Hui, S. L. (1991). The analysis of humongous databases: Problems and promises. *Statistics in Medicine, 10,* 511–518.

McDowell, I., & Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires* (2nd ed.). New York: Oxford University Press.

Mitchell, J. M., & de Lissovoy, G. (1997). A comparison of resource use and cost in direct access versus physician referral episodes of physical therapy. *Physical Therapy, 77,* 10–18.

Mitchell, T. M. (1997). Does machine learning really work? *AI Magazine, 18,* 11–20.

Moses, L. E. (1991). Innovative methodologies for research using databases. *Statistics in Medicine, 10,* 629–633.

Moses, L. E., & Louis, T. A. (1992). Statistical consultation in clinical research: A two-way street. In J. C. Bailar & F. Mosteller (Eds.), *Medical uses of statistics* (pp. 349–356). Boston: NEJM Books.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin.

Nick, T. G., Williams, J. M., & Barker, J. R. (1998). Descriptive and graphical strategies for assessing change: A case study on functional status in stroke patients. *Topics in Health Information Management, 18*(3), 8–17.

O'Brien, P. C., & Shampo, M. A. (1981). Statistics for clinicians. *Mayo Clinic Proceedings, 56,* 45–46.

Pentland, W., McColl, M. A., & Rosenthal, C. (1995). The effect of aging and duration of disability on long term health outcomes following spinal cord injury. *Paraplegia, 33,* 367–373.

Piantadosi, S. (1997). *Clinical trials: A methodologic perspective.* New York: Wiley.

Portney, L. G., & Watkins, M. P. (1993). *Foundations of clinical research: Applications to practice.* Norwalk, CT: Appleton & Lange.

Rubin, D. B., & Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine, 10,* 585–598.

Sun, G., Shook, T. L., & Kay, G. L. (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology, 49,* 907–916.

Tierney, W. M., & McDonald, C. J. (1991). Practice databases and their uses in clinical research. *Statistics in Medicine, 10,* 541–557.

Walter, S. D., Feinstein, A. R., & Wells, C. K. (1987). Coding ordinal independent variables in multiple regression analyses. *American Journal of Epidemiology, 125,* 319–323.

Wilkerson, D. L., & Johnston, M. V. (1997). Clinical program monitoring systems: Current capability and future directions. In M. J. Fuhrer (Ed.), *Assessing medical rehabilitation practices: The promise of outcomes research* (pp. 275–305). Baltimore: Brookes.

Whyte, J. (1997). Distinctive methodologic challenges. In M. J. Fuhrer (Ed.), *Assessing medical rehabilitation practices: The promise of outcomes research* (pp. 43–59). Baltimore: Brookes.