Understanding Logistic Regression Analysis in Clinical Reports: An Introduction

Richard P. Anderson, MD, Ruyun Jin, MD, and Gary L. Grunkemeier, PhD

The Virginia Mason Medical Center, Seattle, Washington, and Providence Health System, Portland, Oregon

Much of our understanding of biological effects and their determinants is gained through statistical regression analysis. Linear and nonlinear regression methods are often applied in the basic sciences. Clinical studies that evaluate the relative contribution of various factors to a single binary outcome, such as the presence or absence of death or disease, most often employ the method of logistic regression. The purpose of this article is to provide an introduction sufficient to permit clinicians who are unfamiliar with regression methodology to understand and interpret its results. We will begin by describing linear regression techniques in order to present basic concepts. We will then consider logistic regression at greater length because of its importance and increasing use by cardiothoracic surgeons.

The calculations involved in logistic regression are complex, but currently available personal computers and ubiquitous statistical software have brought the capability for performing the analysis to the desktop of virtually all clinicians. Consequently, one can hardly find a recent medical journal that does not include at least one report that employs this technique. Figure 1 illustrates the increasing use of logistic regression in studies appearing in three thoracic surgical journals during the last decade. After a description of logistic regression, we will present a clinical example illustrating the technique.

Simple Linear Regression

The term "regression" had its origin with the 19th century statistician Francis Galton. He used it to describe his observation that the sons of short fathers tended to be taller and sons of tall fathers shorter, so that the height of sons "regressed" to the mean height of all men. Galton's friend Karl Pearson developed the mathematical basis for what has come to be termed "regression analysis," a statistical technique used to describe and quantify the relationship between two or more variables. In linear regression, the term "simple" refers to the fact that only two variables are to be related. The technique is therefore said to be bivariate. The term "linear" indicates that the relationship can be described by a straight line. The relationship between variables is one of change, that is, as one variable increases or decreases in magnitude, the other also changes in magnitude.

The data employed in the analysis are assumed to meet certain conditions. First, the dependent variable must be continuous; that is, it must be measured on a numerical scale (eg, blood pressure, age). Second, the relationship between the variables is one of functional dependence. This means that the dependent variable, which is usually plotted on the *y,* or vertical, axis of an *xy* plot, is determined by, or is a function of, the independent variable plotted on the *x,* or horizontal, axis. Third, the values of the *x* variables are either prescribed or measured with negligible error, whereas the values of the *y* variables are assumed to be a random sample from a normally distributed population. Moreover, the standard deviation of the *y* values for each value of *x* is the same.

We will not consider the calculations necessary to fit a straight line to a data set, leaving our computer software to do the math. Instead, we will concentrate on the computer output and what it tells us about the relationship between the dependent and independent variables. Virtually all statistical software packages will perform simple linear regression and yield the information we seek. The labels of these values may differ but are usually defined in the software documentation. We will employ commonly used terms to describe these values.

The computer output in simple linear regression contains four classes of information. The first is the information that defines the straight line that best describes or "fits" the values of the dependent variable for each value of the independent variable. The second is a measure of how strongly the two variables are related in the data that we have available. The third tells us how likely it is that our data, although showing a relationship, have been drawn from a population where, in fact, no such relationship exists. The fourth class of information contains estimates of the variability of the regression line and its coefficients, termed standard errors, and some values used in the calculation of statistical significance. This information is not required for the intuitive understanding of regression sought in this paper, but interested readers will find a thorough discussion of these matters in the book by Glantz and Slinker [1]. We will consider the first three classes of information in detail.

The line that determines the regression of *y* on *x* is determined by two coefficients. The first coefficient, the y intercept, is the value of *y* where *x* is equal to 0. The second coefficient is the slope of the line or the change in *y* for each unit change in *x.* The computer determines the regression line through a process that minimizes the squared difference between the line and the observed values of *y.* If the slope is positive, then *y* increases as *x*

Address reprint requests to Dr Grunkemeier, St. Vincent Hospital and Medical Center, 9155 SW Barnes, Suite 33, Portland, OR 97225; e-mail: ggrunkemeier@providence.org.

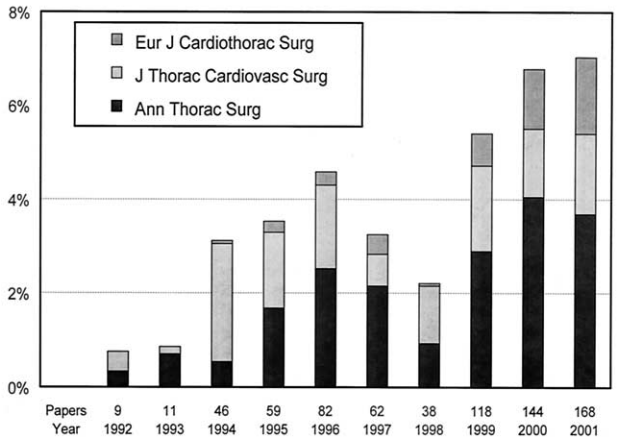

Fig 1. The percentage of papers appearing in three thoracic surgical journals, according to the year of publication. Numbers were obtained by searching the CTSNet Web site using the text string "logistic regression."

increases. If the slope is negative, then y decreases as x increases.

The strength of the relationship between the dependent and independent variables is given by the correlation coefficient, usually designated as R. R varies from -1 (a perfect inverse relationship), through 0 (no relationship), to $+1$ (a perfect direct relationship). R^2, the coefficient of determination, is the proportion or percentage of the total variation in the values of the dependent variable accounted for by the regression.

It is theoretically possible for a data set to show a strong relationship when in fact no such relationship exists. If this were the case, then the best estimate of y at any value of x would be the mean value in the population of y values, and the slope of a line describing this relationship would be 0. This possibility forms the basis for statistical significance tests. A Student's t test of the hypothesis of a zero slope is calculated by dividing the regression coefficient by its standard error. The t statistic may be used to determine confidence intervals for the regression coefficient. If the 95% confidence interval for the slope does not include 0, then it is generally considered that a true relationship between dependent and independent variables exists.

In those biological systems where linear regression provides a good fit to the data, the prediction of the dependent variable from the independent variable can provide valuable insights. However, it is not appropriate to extend predictions beyond the limits of the measured variables because of potential errors. Of particular note is the fact that the mathematical dependence of one variable upon another does not prove causation. Causation may be inferred from knowledge of the system under evaluation but can never be proved by statistics alone. Even when a dependent relationship appears to exist, it may be due to the influence of additional variables. A description of how such additional variables enter regression analysis follows below.

Multivariable Linear Regression

Multiple linear regression is a generalization of simple linear regression. It describes the relationship between two or more independent variables and a single dependent variable. The same constraints on the nature of the data required in simple linear regression apply. One can visualize a plot of two independent variables and the single dependent variable by imagining a third (z) axis projected perpendicular to the other two axes into the plane of the xy plot. A regression defines a plane in three-dimensional space, with one dimension for the dependent variable and one dimension each for the two independent variables. Any point on this plane identifies a value of the dependent variable for corresponding values of both independent variables.

The numbers of independent variables that can be applied are not limited in multiple regression analysis. Three or more independent variables may determine the values of the dependent variable, but our three-dimensional space precludes plotting of such relationships. These relationships are, nevertheless, valid. Many of the same computer-generated results of multiple regression are the same as those produced for simple regression. Their interpretation is also similar. A regression coefficient is produced for the y intercept and for each of the independent variables. The regression coefficients for each independent variable represent the change in the dependent variable for a one-unit change in the independent variable, with the other independent variables held constant.

The strength of the relationship between the dependent variable and all of the independent variables is given by the coefficient of determination, R^2. Just as with simple regression, R^2 measures the percentage of variation in the dependent variable accounted for by the regression. The computer output may also produce a value termed "adjusted R^2." The addition of multiple independent variables increases R^2 regardless of the influence of the additional variables on the dependent variable. The adjustment is the "price" paid for additional variables, and corrects for this inflation by producing a better overall estimate. Just as in simple regression, t statistics, p values, and confidence intervals are generated for the regression coefficients to test whether they are significantly different than 0.

Logistic Regression

Multivariable logistic regression is the statistical technique used when we wish to estimate the probability of a dichotomous outcome such as the presence or absence of a disease or of death. The probability of the outcome is the dependent variable and the various factors that influence it are the independent variables, sometimes termed risk factors. One may think of the probability of the outcome as a proportion or a percentage. For example, suppose in a series of 500 aortic valve replacements there are 25 deaths. The proportion of deaths is 25/500, or 0.05 or 5%. However, the results of logistic regression are

presented in terms of the odds, rather than the probability, of the outcome. There is a direct relationship between probabilities and odds: the odds of the occurrence are simply the probability of the outcome occurring divided by the probability of the outcome not occurring. In the above example we determine the odds of death by dividing 0.05, the proportion of deaths, by 0.95, the proportion of survivors, and obtain 1 to 19. To obtain the probability of dying from the odds, simply divide the odds by 1 plus the odds or $(1/19)/(1 + 1/19) = 0.05$.

To see the utility of working arithmetically in odds rather than probabilities, recall that the probability of an event can only vary between 0 and 1. Suppose we have a probability of 30% that an event will occur. We can speak of doubling the probability to 60%, but doubling a probability of 60% has little meaning because probability cannot exceed 100%. By working with odds, we remove the restriction on an upper limit. If the probability is 60%, then the odds are 0.6/0.4, or 1.5. Doubling the odds to 3 yields a probability of 3.0/4.0, or 0.75. Any multiple of the odds will yield a probability of the outcome that is less than 1.

Logistic regression uses the past experience of a group of patients to estimate the odds of an outcome by mathematically modeling or simulating that experience and describing it by means of a regression equation. A key feature in modeling a clinical experience is the selection of the independent variables that influence the outcome. This process has many variations but will not be further considered here. Clinicians who are interested in learning more about the development of a regression model will find the monograph by Katz [2] an excellent source. The method of calculation for the regression coefficients takes into consideration all possible combinations of the independent variables. It then maximizes the probability that, for any given individual with a particular combination of independent variables, the odds of the outcome will be close to the actual or observed outcome of all other individuals possessing the same combination of independent variables.

The general form of the logistic regression equation is similar to that of multivariable linear regression; however, the logarithm of the odds of the outcome, termed the logit or log odds, is used as the dependent variable. The regression coefficients are also expressed as natural logarithms. Just as with linear regression, the value of each coefficient is added to the constant coefficient whenever there is a one-unit change in the value of the independent variable. Coefficients may have positive or negative values depending on whether they increase or decrease the logit of the outcome. Dichotomous independent variables such as presence or absence of a risk factor are coded as 1 or 0; if 1, the coefficient is added, and if 0, it is not added. Continuous variables such as age would have the coefficient multiplied by the value of the variable and then added. The logit of the outcome can be converted to the odds of the outcome by exponentiation, raising *e,* the base of the natural logarithms, to the power of the logit. One may do this easily on a scientific calculator by entering the logit and pressing the key that raises *e* to the entered value.

Clinical Example

To illustrate the application of logistic regression analysis, we present results using the coronary artery bypass grafting (CABG) database of the Providence Health System. The data originate from nine hospitals in four Western states covering a period from January 1997 to December 2001, and include 13,341 patients who underwent isolated CABG. The dependent variable selected for modeling is death. Although many risk factors are involved in this outcome, we have selected two for purposes of demonstration: patient age at operation (AGE) and a history of acute or chronic renal insufficiency (RENAL). AGE is a continuous variable measured in years and RENAL is a dichotomous variable coded as either 0 (absent) or 1 (present). RENAL is defined as a history of acute or chronic renal insufficiency or a history of a serum creatinine > 2.0 recorded in the clinical record.

There were 605 records excluded because data concerning one or more of the independent variables were missing, leaving 12,736 patients for the analysis. Overall mortality in this group was 2.37%, RENAL was present in 4.0%, and AGE ranged from 24 to 97 years, with a mean of 65.3 years. Logistic regression analysis will help to provide answers to the following questions. What are the odds of death for individuals of given age and renal status? What are the odds ratios or relative importance of each independent variable in determining the outcome? Are AGE and RENAL independent predictors of death? How well does the model perform in assigning appropriate risk? What is the ability of the model to discriminate between those who live and those who die?

If we had no information about the independent variables in our series of patients, the best estimate of mortality for any patient would simply be the average mortality for the group. We know from experience that older patients who have RENAL are at greater risk than younger patients who lack RENAL. To assess and quantify this difference, we enter our data into a logistic regression program and obtain the values listed in Table 1. Figure 2 shows a picture of this regression. The horizontal axis is the patient's age and the vertical axis is the probability of dying. This picture has some similarity to the familiar linear regression, but the regression lines are curved (logistic) rather than linear. The solid line is for NO RENAL and the dashed line is for RENAL. The numbered symbols are the observed data points, the average mortality for patients grouped by decade of age.

To demonstrate how these curves were computed from Table 1, we now use these values to compute the odds of dying for a 65-year-old (patient A) with RENAL. We multiply the coefficient for AGE, 0.073, by 65 and obtain 4.745, add 1.162 for the presence of RENAL, and add the constant value, -8.868, for a total of -2.961, the logit or log odds of death. The logit is then converted to the odds by exponentiation, yielding a value of approximately

Table 1. Logistic Regression Analysis of Risk of Death in Coronary Bypass Patients

	Coeff.	Standard Error	*p* Value	Odds Ratio	CI Lower Limit	CI Upper Limit
AGE	0.073	0.006	<0.001	1.076	1.062	1.090
RENAL	1.162	0.177	<0.001	3.198	2.259	4.526
Constant	−8.868	0.471	<0.001			

AGE = age in years; RENAL = history of renal insufficiency; Coeff. = coefficient expressed in logits; CI = 95% confidence interval for the odds ratio.

0.052. The probability of death would be 0.052/0.948, or 5.5%. The effect of RENAL on outcome can be seen by comparing the above result to that of another 65-year-old (patient B) who lacks this risk factor. The coefficient for RENAL is not added because it is absent. The value for AGE 65, 4.745, combined with the constant value, −8.868, yields a log odds of −4.123, an odds of 0.016, and a probability of death of 0.016/0.984, or 1.6%. Thus, the odds of dying for patient A is just over three times that of patient B, as the odds ratio in Table 1 shows. The predicted mortality for these 2 patients is shown by the boxes in Figure 2.

The exponentiated values of the coefficients are termed "odds ratios." These values are generally provided together with their 95% confidence intervals as the primary results of logistic regression analysis in many published studies. As the name implies, an odds ratio is simply the ratio of one odds to another. For example, the odds ratio of a dichotomous variable is the ratio of the odds of the outcome in the presence of the variable to the odds in its absence. To understand the numeric derivation of an odds ratio, recall that logistic regression yields values expressed in natural logarithms. To obtain the odds, we exponentiate the log values. Therefore, the odds ratio for RENAL is 3.198:1, which, as noted above, triples the odds of dying when present.

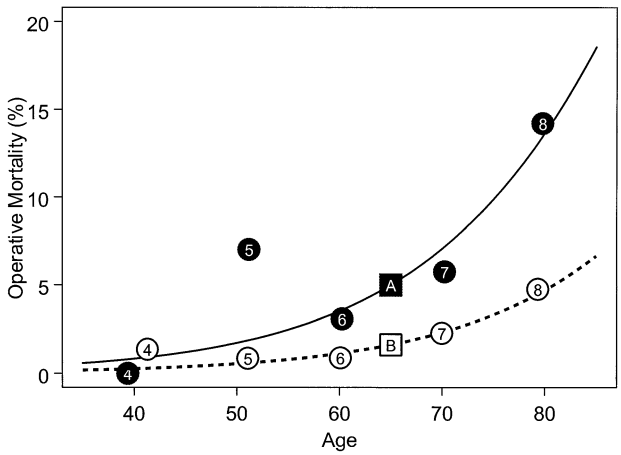

Fig 2. Logistic regression model from the example in the text. The smooth curves *are plots of the probability of operative mortality for patients with* (solid line) *and without* (dashed line) *renal failure, according to age. The* numbered circles *are the observed mortality by age group (4 = < 45 years, 5 = 45 to 55, 6 = 55 to 65, 7 = 65 to 75, 8 = > 75) for the patients with* (black symbols) *and without* (white symbols) *renal failure. The predicted mortalities for the 2 patients* (A *and* B) *discussed in the text are shown by the* boxes.

Our computer program provides several statistics that determine the significance of individual coefficients. In Table 1, we show the standard error, which may be reported in some publications. In general, a coefficient needs to be at least twice the size of its standard error to be statistically significant. The coefficients in our example are highly significant. Statistical significance may also be inferred from inspection of the 95% confidence intervals of the odds ratios. If a value of 1 is not included within the upper and lower bounds of the confidence interval, then the odds ratios are significant at least at the 0.05 level. A value of 1 included within these bounds indicates that the odds ratio is not significantly different from 0.

Note in Table 1 that the coefficient and odds ratio for AGE are much smaller than the corresponding values for RENAL. This occurs because a single unit change in AGE is only 1 year. If we were to change the scale of measurement from 1 year to 1 decade, as is often done in clinical studies, then the resulting coefficient and standard error would be 10 times greater (0.732 and 0.063), and the odds ratio, corresponding to a 1-decade increase in age, would become exp 0.732, or 2.08. This has several implications for the interpretation of odds ratios and their relative importance to the outcome. The values for odds ratios of continuous variables are not directly comparable with one another or to the odds ratios of dichotomous variables in terms of their relative importance to the outcome because of measurement on different scales. However, because they are measured on the same scale (0 equals absent and 1 equals present), the values for the odds ratios of dichotomous variables are directly comparable and indicate their relative importance to the outcome.

The two curves in Figure 2 appear to be proportional to one another. If there were a potentiation of the AGE effect due to the presence of RENAL, then another variable, called an interaction term (AGE × RENAL), would be needed. A second regression was done, including this term, but it was not significant ($p = 0.664$), so the simpler model was sufficient.

Determining Quality of the Logistic Regression Model

Several properties are useful for an assessment of accuracy, but the results of their evaluation in published articles are inconsistent. The first property is termed

"calibration" and is defined as the ability of the model to assign appropriate risk among the patients whose experience the model simulates. Calibration of a model may be demonstrated qualitatively by a table or graph of estimated versus observed outcomes in various patient groups. It may also be determined quantitatively by the Hosmer-Lemeshow goodness-of-fit test [3]. This test yields a modified χ^2 statistic, where a small value is desired, yielding a nonsignificant p value for the difference between observed and estimated outcomes. Models can be adjusted during their development to overcome poor calibration, and many articles will not refer to calibration measures.

A second property relating to the accuracy of a model is termed "discrimination" and is defined as the ability of the model to distinguish between those patients having and those not having the outcome. Discrimination is evaluated using receiver-operating characteristic curve analysis [4]. This analysis is often used in evaluating diagnostic tests and is based on the properties of specificity and sensitivity, and on the probability that a randomly chosen death will have a higher risk than a randomly chosen survivor. A curve (receiver operating characteristic [ROC]) is generated, and the area under the curve (termed the c-index or c-statistic) is calculated. A c-index of 0.5 would mean that the probability of correctly identifying a death from a randomly chosen pair would be 50%, or no better than chance. Perfect discrimination would yield a c-index of 100% accuracy. Many clinical risk models will report this statistic, which will usually be found to range from 65% to 85%.

A third, and perhaps the most important, property relating to the accuracy of a model, is "validation." Validation may be defined as the demonstration that the predictive accuracy of the model is similar when it is applied to a different group of patients than those used in the construction of the model. Most clinicians will not trust predictive models that are not validated. A common method of validation with large data sets randomly assigns half of the patients to learning and test sets, develops the risk model using the learning set, and applies the resulting model to the test set. The standard measures of calibration and discrimination are then determined for the test set, and if little degradation is seen compared with the values obtained for the learning set, then the model is said to be validated. This is the procedure followed for the development of the 1996 coronary artery bypass risk model of The Society of Thoracic Surgeons Adult Cardiac National Database [5].

Comment

No one knows better than surgeons how multiple factors can combine to produce patient outcomes. Logistic regression analysis is a powerful tool for assessing the relative importance of factors that determine outcome. As such, it is increasingly used in clinical medicine to develop diagnostic algorithms and evaluate prognosis. Yet, this tool is both imperfect and subject to misuse. A recent article by Shahian and colleagues [6] in *The Annals* describes the deficiencies of the method as currently employed in the production of "report cards." A basic understanding of logistic regression analysis is the first step to appreciating both the usefulness and the limitations of the technique.

We are grateful to the Providence Health System hospitals in Alaska, Washington, Oregon, and California for use of their cardiac surgery data in the example. We also thank Ying Xing Wu, MD, for her help with graphics and analysis.

References

1. Glantz SA, Slinker BK. Primer of applied regression and analysis of variance. New York: McGraw-Hill, 1990.
2. Katz MH. Multivariable analysis: a practical guide for clinicians. New York: Cambridge University Press, 1999.
3. Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, 1989.
4. Grunkemeier GL, Jin R. Receiver operating characteristic curve analysis of clinical risk models. Ann Thorac Surg 2001;72:323–6.
5. Shroyer AL, Plomondon ME, Grover FL, Edwards FH. The 1996 coronary artery bypass risk model: The Society of Thoracic Surgeons adult cardiac national database. Ann Thorac Surg 1999;67:1205–8.
6. Shahian M, Norman S-L, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. Ann Thorac Surg 2001;72:2155–68.