JAMA Guide to Statistics and Methods

Logistic Regression
Relating Patient Characteristics to Outcomes

Juliana Tolles, MD, MHS; William J. Meurer, MD, MS

In a recent issue of *JAMA*, Seymour et al[1] presented a new method for estimating the probability of a patient dying of sepsis using information on the patient's respiratory rate, systolic blood pressure, and altered mentation. The method used these clinical characteristics—called "predictor" or explanatory or independent variables—to estimate the likelihood of a patient having an outcome of interest, called the dependent variable. To determine the best way to use these clinical characteristics, the authors used logistic regression, a common statistical method for quantifying the relationship between patient characteristics and clinical outcomes.[2]

Use of the Method

Why Is Logistic Regression Used?

One use of logistic regression is to estimate the probability that an event will occur or that a patient will have a particular outcome using information or characteristics that are thought to be related to or influence such events. Logistic regression can show which of the various factors being assessed has the strongest association with an outcome and provides a measure of the magnitude of the potential influence. It also has the ability to "adjust" for confounding factors, ie, factors that are associated with both other predictor variables and the outcome, so the measure of the influence of the predictor of interest is not distorted by the effect of the confounder.

Although logistic regression can be used to evaluate epidemiological associations that do not represent cause and effect, this article focuses on the use of logistic regression to create models for predicting patient outcomes. In this context, the term *predictors* is used to refer to the independent factors (variables) for which the influences are being quantified, and the term *outcome* is used for the dependent variable that the logistic regression model is trying to predict.

Description of the Method

Patient outcomes that can only have 2 values (eg, lived vs died) are called binary or dichotomous. The outcomes for groups of patients can be summarized by the fraction of patients experiencing the outcome of interest or, similarly, by the probability that any single patient experiences that outcome. However, to understand the results of a logistic regression model, it is important to understand the difference between probability and odds. The probability that an event will occur divided by the probability that it will not occur is called the odds. For example, if there is a 75% chance of survival and a 25% chance of dying, then the odds of survival is 75%:25%, or 3. Logistic regression quantitatively links one or more predictors thought to influence a particular outcome to the odds of that outcome.[2]

The change in the odds of an outcome—for example, the increase in the odds of mortality associated with tachypnea in a patient with sepsis—is measured as a ratio called the odds ratio (OR). If patients with tachypnea have an odds of mortality of 2.0 and patients without tachypnea have an odds of mortality of 0.5, then the OR associated with tachypnea would be 2.0:0.5, or 4. This is the same as an increase in the probability of mortality from 1/3 to 2/3.

In logistic regression, the weight or coefficient calculated for each predictor determines the OR for the outcome associated with a 1-unit change in that predictor, or associated with a patient state (eg, tachypneic) relative to a reference state (eg, not tachypneic). Through these ORs and their associated 95% confidence intervals, logistic regression provides a measure of the magnitude of the influence of each predictor on the outcome of interest and of the uncertainty in the magnitude of the influence.

Logistic regression also enables "adjustment" for confounding factors—patient characteristics that might also influence the outcome and simultaneously be correlated with 1 or more predictors. To accomplish this, both the confounding factors and the predictors of interest are included in the model. For example, when adjusting for the influence of fever in estimating the influence of tachypnea on mortality, both fever and the presence of tachypnea would be included as predictors in the regression model. The result is that the estimate of the association between tachypnea and mortality would not be confounded by a possible correlation between fever and tachypnea.

What Are the Limitations of Logistic Regression?

First, the validity of a regression model depends on the number and suitability of the measured independent predictor variables. Ideally, all biologically relevant factors should be included. When multiple variables convey closely related information (a situation termed *collinearity*), such as would occur when using both serum lactate and anion gap as predictors in patients with septic shock, this can produce serious errors or great uncertainty in the estimates of the effects of these variables on the outcome of interest. When 2 variables provide overlapping information, minor random variation in the data can greatly and unpredictably influence how much of the association is attributed to one factor vs the other in the model.

A second limitation of logistic regression is that the variables must have a constant magnitude of association across the range of values for that variable. For example, in examining the relationship between age and mortality, if the odds ratio for mortality is 2 for each 10-year increase in age, this association needs to be the same when comparing 30- and 40-year-olds as it is when comparing 70- and 80-year-olds if the model is to be used across this entire age range. If the association is not consistent over the age range, then age may be stratified into ranges (eg, 21-50, 51-65, and ≥66) based on the assumption that within each category, the influence of age will be similar. The age category would then

be used as the independent variable, usually with the lowest-risk age group the reference category.

A third limitation is that many logistic regression analyses assume that the effect of one predictor is not influenced by the value of another predictor. When this is not true and the value of one predictor alters the effect of another, there is said to be an "interaction" between the 2 predictors. Such interactions need to be explicitly included in the analysis to ensure the estimated associations are valid.

Why Did the Authors Use Logistic Regression in This Study?

Seymour et al likely selected logistic regression for its familiarity and interpretability. More complex prognostic models may produce algorithms that are difficult to use clinically.

How Should the Results of Logistic Regression Be Interpreted in This Particular Study?

Seymour et al used logistic regression to derive a new clinical tool for assessing the risk of mortality in patients with sepsis, called the quick Sequential [Sepsis-related] Organ Failure Assessment (qSOFA).[1] The qSOFA model is used to estimate the likelihood of in-hospital mortality in patients with suspected infection using respiratory rate, systolic blood pressure, and Glasgow Coma Scale score. Rather than using the precise OR coefficient for each predictor in their final model, the authors simplified the model by assigning the same 1-point value to each predictor. By assigning all coefficients equal value, the authors created a simplified model that could be applied to individual patients by counting the number of positive clinical predictors. The authors then determined how well the qSOFA score estimated mortality relative to other models for estimating mortality in sepsis, demonstrating that a qSOFA score of 2 or more produced a 3- to 14-fold increase in the probability of in-hospital mortality over baseline risk in patients with sepsis. They also found that for patients not in intensive care, the qSOFA predicted mortality in patients with sepsis better than systemic inflammatory response syndrome criteria or the usual Sequential [Sepsis-related] Organ Failure Assessment score.

Caveats to Consider When Assessing the Results of a Logistic Regression Analysis

The associations found through logistic regression models are intended to provide insights into what might happen in a similar population of future patients. Certain combinations of patient characteristics and factors may have been sparsely represented in the data set (eg, young patients with sepsis and a low Glasgow Coma Scale score but a normal blood pressure and respiratory rate), and the estimates of the model for mortality among such patients should be considered with caution.

Because probabilities are more intuitive than ORs, it is important to avoid confusing them. For example, an increase in probability from 25% to 75% would correspond to a risk ratio (RR) of 3 but an OR of 9. However, when probabilities are very close to zero, the OR and the RR will be nearly equal. Thus, ORs and RRs are practically interchangeable when the outcome of interest is rare. However, when the outcome of interest is a common event (eg, occurring in >20% of patients in any group), it is important to recognize that ORs do not approximate RRs.

Reported ORs for the effects of predictors should be accompanied by 95% confidence intervals; intervals that include an OR of 1 would indicate a non–statistically significant relationship between that predictor and the outcome of interest.

The predictors included in logistic regression models should be selected to avoid redundancy in the information they provide (collinearity). It is also important to consider the possibility that the value of one predictor might alter the effect of another (interactions). Both of these situations can adversely affect the validity of the resulting logistic regression model.

ARTICLE INFORMATION

Author Affiliations: Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California (Tolles); Los Angeles Biomedical Research Institute, Torrance, California (Tolles); David Geffen School of Medicine, University of California, Los Angeles (Tolles); Department of Emergency Medicine, University of Michigan, Ann Arbor (Meurer); Department of Neurology, University of Michigan, Ann Arbor (Meurer).

Corresponding Author: William J. Meurer, MD, MS, Department of Emergency Medicine, University of Michigan, 1500 E Medical Center Dr, Ann Arbor, MI 48109-5303 (wmeurer@med.umich.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016;315(8):762-774.

2. Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley; 2013.