JAMA Guide to Statistics and Methods

Logistic Regression Diagnostics
Understanding How Well a Model Predicts Outcomes

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

In the March 8, 2016, issue of *JAMA*, Zemek et al[1] used logistic regression to develop a clinical risk score for identifying which pediatric patients with concussion will experience prolonged postconcussion symptoms (PPCS). The authors prospectively recorded the initial values of 46 potential predictor variables, or risk factors—selected based on expert opinion and previous research—in a cohort of patients and then followed those patients to determine who developed the primary outcome of PPCS. In the first part of the study, the authors created a logistic regression model to estimate the probability of PPCS using a subset of the variables; in the second part of the study, a separate set of data was used to assess the validity of the model, with the degree of success quantified using regression model diagnostics. The rationale for using logistic regression to develop predictive models was summarized in an earlier JAMA Guide to Statistics and Methods article.[2] In this article, we discuss how well a model performs once it is defined.

Use of the Method

Why Are Logistic Regression Model Diagnostics Used?

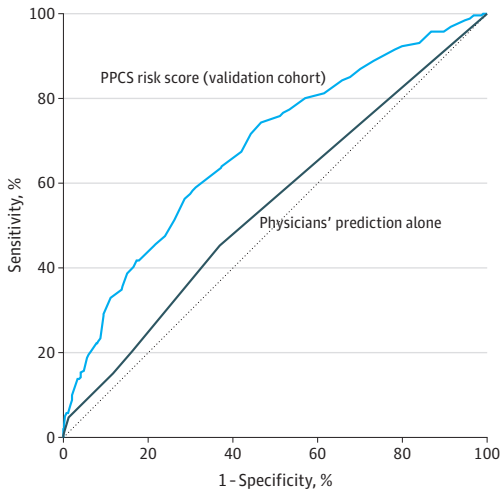
Logistic regression models are often created with the goal of predicting the outcomes of future patients based on each patient's predictor variables.[2] Regression model diagnostics measure how well models describe the underlying relationships between predictors and patient outcomes existing within the data, either the data on which the model was built or data from a different population.

The accuracy of a logistic regression model is mainly judged by considering *discrimination* and *calibration*. Discrimination is the ability of the model to correctly assign a higher risk of an outcome to the patients who are truly at higher risk (ie, "ordering them" correctly), whereas calibration is the ability of the model to assign the correct average absolute level of risk (ie, accurately estimate the probability of the outcome for a patient or group of patients). Regression model diagnostics are used to quantify model discrimination and calibration.

Description of the Method

The model developed by Zemek et al discriminates well if it consistently estimates a higher probability of PPCS in patients who develop PPCS vs those who do not; this can be assessed using a receiver operating characteristic (ROC) curve. An ROC curve is a plot of the sensitivity of a model (the vertical axis) vs 1 minus the specificity (the horizontal axis) for all possible cutoffs that might be used to separate patients predicted to have PPCS compared with patients who will not have PPCS (**Figure**).[1] Given any 2 random patients, one with PPCS and one without PPCS, the probability that the model will correctly rank the patient with PPCS as higher risk is equal to the area under the ROC curve (AUROC).[3] This area is also called the *C statistic*, short for "concordance" between model estimates of risk and the observed risk. The C statistic is discussed in detail in a previous JAMA Guide to Statistics and Methods article.[4] A model with perfect sensitivity and specificity would have an AUROC of 1. A model that predicts who has PPCS no better than chance would have an AUROC of 0.5. While dependent on context, C statistic values higher than 0.7 are generally considered fair and values higher than 0.9 excellent; those less than 0.7 generally are not clinically useful.[5]

Figure. Receiver Operating Characteristic Curves



PPCS indicates persistent postconcussion symptoms. The area under the curve was 0.71 (95% CI, 0.69-0.74) for the derivation cohort and 0.68 (95% CI, 0.65-0.72) for the validation cohort. Based on Figure 2 from Zemek et al.[1]

A particular model might discriminate well, correctly identifying patients who are at higher risk than others, but fail to accurately estimate the absolute probability of an outcome. For example, the model might estimate that patients with a high risk of PPCS have a 99% chance of developing the condition, whereas their actual risk is only 80%. Although this hypothetical model would correctly discriminate, it would be poorly calibrated. One method to assess calibration is to compare the average predicted and average observed probabilities of an outcome both for the population as a whole and at each level of risk across a population. The patients are commonly divided into 10 groups based on their predicted risk, so-called deciles of risk. In a well-calibrated model, the observed and predicted proportions of patients with the outcome of interest will be the same within each risk category, at least within the expected random variability (see Table 6 in the article by Zemek et al). The Hosmer-Lemeshow test measures the statistical significance of any differences between the observed and predicted outcomes over the risk groups; when there is good agreement, the Hosmer-Lemeshow statistic will not show a statistically significant difference, suggesting that the model is well calibrated.[6] Another way to assess calibration is through a calibration plot (eFigure 3 in the article by Zemek et al) in which the observed

proportion of the outcome of interest is plotted against the predicted probability.

Some statistical programs also report a *pseudo-R^2* regression diagnostic for logistic regression models. The pseudo-R^2 is meant to mimic the R^2 calculated for linear regression models, a measure of the fraction of the variability in the outcome that is explained by the model. However, because is there no direct equivalent to R^2 in logistic regression, many variations of pseudo-R^2 have been developed by different statisticians, each with a slightly different interpretation.[7]

What Are the Limitations of Logistic Regression Diagnostics?

It is easy to interpret extreme values of the AUROC statistic—those close to 1 or 0.5—but it is a matter of judgment to decide whether a value of 0.75, for example, represents acceptable discrimination. The AUROC is therefore subject to interpretation and comparison with the AUROC values of competing diagnostic tests. Additionally, using the AUROC alone as a metric assumes that a false-positive result is just as bad as a false-negative result. This assumption is often not appropriate in clinical scenarios, and more sophisticated metrics such as a *decision curve analysis* may be needed to appropriately account for the different costs of different types of misclassification.[8]

With large sample sizes the Hosmer-Lemeshow statistic can yield false-positive results and thus falsely suggest that a model is poorly calibrated. In addition, the Hosmer-Lemeshow statistic depends on the number of risk groups into which the study population is divided. There is no theoretical basis for the "correct" number of risk groups into which a population should be divided. Also, with sample sizes smaller than 500, the test has low power and can fail to identify poorly calibrated models.[9]

Why Did the Authors Use Logistic Regression Diagnostics in This Particular Study?

Logistic regression model diagnostics, and model diagnostics generally, are essential for judging the usefulness of any new prediction instrument. A model is unlikely to improve practice if it performs no better than chance or currently available tests. However, in particular clinical applications, physicians may be interested in using models that perform well on only one of these metrics or perform well only at a particular cut point. For example, consider a clinical screening test for which the intended use is to discriminate between patients with very low risk of a particular outcome and all others. Such a model might discriminate well at a particular screening cut point but have poor calibration, or it may have inaccurate estimation of risk for patients who are not classified as very low risk but still be completely appropriate for its intended use.

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

The ROC curve plotted by Zemek et al (Figure) demonstrates modest discrimination; in the initial derivation cohort, the AUROC was 0.71. In the validation cohort, the combination of physician judgment with the final prediction model produced an AUROC of 0.68. While this AUROC value might seem low, it was substantially better than physician estimation alone for predicting PPCS (AUROC of 0.55). As the authors pointed out, this difference indicated that the model generally outperformed clinical judgment alone, although it provided only fair discrimination at best.

The model used by Zemek et al appears well calibrated. The Hosmer-Lemeshow statistic associated with the comparison between predicted and observed rates of PPCS (Table 6 in the article by Zemek et al) across all deciles of risk was not significant. Furthermore, the sample size in this study is large enough that the Hosmer-Lemeshow statistic should have reasonable power to detect poor calibration. The intercept and slope of the calibration plot on the validation cohort were 0.07 and 0.90, respectively, closely approaching their respective ideals of 0 and 1.

Caveats to Consider When Assessing the Results of Logistic Regression Diagnostics

Whenever possible, all metrics of model quality should be measured on a data set separate from the data set used to build the model. Independence of test data is crucial because reusing the data on which a model was built (the "training data") to measure accuracy will overestimate the accuracy of the model in future clinical applications. Zemek at al used an independent validation cohort, recruited from the same centers as the training cohort. Therefore, although the model was tested against data other than from which it was derived, it still may lack external validity in patient populations seen in other settings.[10]

ARTICLE INFORMATION

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

Corresponding Author: William J. Meurer, MD, MS, Department of Emergency Medicine, 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: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Zemek R, Barrowman N, Freedman SB, et al; Pediatric Emergency Research Canada (PERC) Concussion Team. Clinical risk score for persistent postconcussion symptoms among children with acute concussion in the ED. *JAMA*. 2016;315(10):1014-1025.

2. Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. *JAMA*. 2016;316(5):533-534.

3. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.

4. Pencina MJ, D'Agostino RB Sr. Evaluating discrimination of risk prediction models: the C statistic. *JAMA*. 2015;314(10):1063-1064.

5. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988;240(4857):1285-1293.

6. Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. New York, NY: Wiley; 2013.

7. Cameron AC, Windmeijer FAG. An *R*-squared measure of goodness of fit for some common nonlinear regression models. *J Econom*. 1997;77(2):329-342. doi:10.1016/S0304-4076(96)01818-0

8. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. *JAMA*. 2015;313(4):409-410.

9. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*. 1997;16(9):965-980.

10. Efron B. How biased is the apparent error rate of a prediction rule? *J Am Stat Assoc*. 1986;81(394):461-470. doi:10.2307/2289236