Explanatory Versus Predictive Modeling

Kristin L. Sainani, PhD

INTRODUCTION

When building multivariate statistical models, researchers need to be clear as to whether their goals are explanatory or predictive. Explanatory research aims to identify risk (or protective) factors that are causally related to an outcome. Predictive research aims to find the combination of factors that best predicts a current diagnosis or future event. This distinction affects every aspect of model building and evaluation. Unfortunately, researchers often conflate the two, which leads to errors [1]. This article reviews the differences between explanatory and predictive modeling.

EXPLANATORY MODELING

The aim of explanatory research is to establish causal relationships. For example, in the article by Yu et al [2], the researchers attempted to identify risk factors that would predispose amputees to falling during the postoperative period. A better understanding of these risk factors could help researchers design interventions to prevent falls. Explanatory model building is primarily concerned with identifying individual risk factors that are associated with the outcome as well as ruling out confounding (extraneous variables that are related to both the risk factor and outcome may create spurious associations between them). Explanatory modelers need to worry about chance findings, unmeasured confounding, and residual confounding.

CANDIDATE VARIABLES

Testing too many candidate variables may lead to type I errors (a statistically significant finding that is due to chance). Researchers can limit the number of candidate variables by focusing on a few key hypotheses (eg, Do opioids increase the risk of falls? Do benzodiazepines increase the risk of falls?). Often, however, the goal is broader: to identify multiple risk factors for an outcome. In this case, researchers should select candidate variables a priori based on biologic plausibility, previous research, or clinical experience. For example, Yu et al [2] chose potential factors "on the basis of clinical relevance and experience," including comorbidities, cognitive deficits, and medications that might increase the risk of falls.

VARIABLE SELECTION

The process of selecting variables for the final multivariate model should be driven by the specific hypotheses being tested. Risk factors are included if they have a significant or near-significant relationship with the outcome; confounder variables are included if they change the relationship between a risk factor of interest and the outcome, regardless of statistical significance. Some variables, such as age and gender, may be included for face validity even if they are not significantly related to the outcome. Explanatory modelers should avoid automated variable selection procedures, for example, stepwise regression, because these optimize overall model fit with no regard to the roles of individual variables.

Yu et al [2] included 7 clinically relevant variables in their multivariate logistic regression (a multivariate regression technique used when the outcome is binary [eg, fall/no fall]): etiology of amputation, level of amputation, side of amputation, presence of cognitive

K.L.S. Division of Epidemiology, Department of Health Research and Policy, Stanford University, Stanford, CA. Address correspondence to: K.L.S.; e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose

impairment, presence of chronic renal failure, use of opioid analgesics, and use of benzodiazepines. They manually removed nonsignificant variables to arrive at a final model that included the type of etiology (dysvascular versus nondysvascular), level of amputation (transtibial versus nontranstibial), and side of amputation (right versus left). The remaining 4 risk factors were not independently associated with falling once the 3 amputation characteristics were taken into account.

MODEL ASSESSMENT AND VALIDATION

For explanatory models, researchers should focus on the individual β coefficients and P values for the risk factors of interest. For example, in the study by Yu et al [2], a transtibial level of amputation had an odds ratio of 2.127 ($P < .05$) for falls compared with a nontranstibial amputation. Measures of overall model performance, such as R^2 values, are less important. Similarly, researchers who attempt to validate the findings need to confirm individual causal relationships rather than overall model performance. Authors and readers should consider the potential role of chance in the findings, particularly if a large number of risk factors were tested and if the resulting P values achieve only a moderate level of significance ($.01 < P < .05$). They should also consider whether the apparent relationships could be explained by unmeasured or residual confounding.

PREDICTIVE MODELING

Prediction models aim to accurately estimate the probability that a disease is present (diagnosis) or that a future event will occur (prognosis). For example, Bates et al [3] built a model to predict 1-year mortality of veterans with stroke. Knowing a patient's mortality risk can help patients, physicians, and caregivers to better plan postdischarge care and priorities. In predictive modeling, overall predictive accuracy is paramount and the role of individual variables is less critical. Variables may be included in the final model even if they are not causally related to the outcome. For example, stroke patients discharged to an acute care facility are more likely to die, but this variable is a marker of poor health rather than a cause of death.

Predictive modelers need to consider several aspects of model performance. They also need to worry about overfitting and generalizability. Overfit models are tuned to the idiosyncrasies of a particular sample and thus have high predictive accuracy for the sample but not for new observations. Because of the problem of overfitting, prediction models should always undergo validation.

CANDIDATE VARIABLES

Predictive modelers typically start with a larger pool of candidate variables than explanatory modelers. Bates et al [3] considered 6 types of candidate variables: demographics, type of stroke, comorbidities, procedures received during hospitalization or intensive care unit stays, length of stay in the hospital, and discharge location. When the pool of candidate predictors is large relative to the sample size, overfitting is likely. Thus, predictive modelers may screen out candidate variables or apply data reduction techniques, such as principal components analysis, before final model building. For example, Bates et al [3] first screened out candidate variables that appeared unrelated to death in bivariate analyses ($P > .20$), which left 38 variables.

VARIABLE SELECTION

Predictive modelers often use automated selection procedures. For example, Bates et al [3] fit a logistic regression model by using automated backward selection (retaining variables with $P < .05$) to arrive at a final model with 17 variables. Automated selection procedures help researchers to find the combination of predictors that optimizes overall model fit. However, these methods cause considerable overfitting and thus should be used cautiously and only in conjunction with validation [4]. Newer automated selection procedures that incorporate shrinkage, for example, LASSO ("least absolute shrinkage and selection operator"), have considerable advantages over traditional methods [4,5].

The final prediction model may be translated into a clinical scoring rule. Bates et al [3] assigned scores for different risk factors based on the size of their β coefficients in the final logistic regression model. For example, patients younger than 60 years old receive 0 points, patients 60-69 years old receive 2 points, patients 70-79 years old receive 5 points, and patients 80 years old and older receive 8 points. Higher risk scores correspond to a higher probability of death.

MODEL PERFORMANCE

Predictive models should be assessed for discrimination, calibration, and goodness of fit. Unfortunately, many predictive modelers fail to report any metrics beyond discrimination [6]. Additional statistics (called reclassification statistics) are needed if the goal of an analysis is to compare 2 prediction models. Prediction models also should be evaluated for their clinical utility, although this may require further studies. A model is said to discriminate well if it systematically assigns higher predicted probabilities to those who have the outcome compared with those who do not. Discrimination is typically measured by using receiver operating characteristic (ROC) curves and the related C-statistic (equal to the area under the ROC curve). Bates et al [3] reported an area under the ROC curve of 0.785, which represents moderate discrimination.

Calibration addresses the accuracy of the estimated probabilities. A model may discriminate well but still be poorly calibrated. For example, a model that assigns predicted