Level V Evidence

Research Pearls: The Significance of Statistics and Perils of Pooling. Part 2: Predictive Modeling

Erik Hohmann, M.D., Ph.D., F.R.C.S., Merrick J. Wetzler, M.D., and Ralph B. D'Agostino Jr., Ph.D.

Abstract: The focus of predictive modeling or predictive analytics is to use statistical techniques to predict outcomes and/or the results of an intervention or observation for patients that are conditional on a specific set of measurements taken on the patients prior to the outcomes occurring. Statistical methods to estimate these models include using such techniques as Bayesian methods; data mining methods, such as machine learning; and classical statistical models of regression such as logistic (for binary outcomes), linear (for continuous outcomes), and survival (Cox proportional hazards) for time-to-event outcomes. A Bayesian approach incorporates a prior estimate that the outcome of interest is true, which is made prior to data collection, and then this prior probability is updated to reflect the information provided by the data. In principle, data mining uses specific algorithms to identify patterns in data sets and allows a researcher to make predictions about outcomes. Regression models describe the relations between 2 or more variables where the primary difference among methods concerns the form of the outcome variable, whether it is measured as a binary variable (i.e., success/failure), continuous measure (i.e., pain score at 6 months postop), or time to event (i.e., time to surgical revision). The outcome variable is the variable of interest, and the predictor variable(s) are used to predict outcomes. The predictor variable is also referred to as the independent variable and is assumed to be something the researcher can modify in order to see its impact on the outcome (i.e., using one of several possible surgical approaches). Survival analysis investigates the time until an event occurs. This can be an event such as failure of a medical device or death. It allows the inclusion of censored data, meaning that not all patients need to have the event (i.e., die) prior to the study's completion.

Statistical methods are important tools to determine whether results from a research study are significant and can be applied to the general population. Statistical models can be used to describe data, explain the significance of data or predict outcomes, and establish, or at least suggest, causality. The statistical methods used are an important part of any research study and are essential for the correct design of a research project.[1] However, many authors have only a rudimentary understanding of statistical concepts, especially when more complex analysis is required.[1]

With descriptive statistics, data are summarized in a more compact manner. The focus is to describe measured outcome variables and/or demographic characteristics of the study population quantitatively.[2,3] In general, measures of central tendency describe the data "average" (mean, median, mode) and measures of dispersion that spread around the "average" (range, interquartile range, variance, standard deviation). The primary difference between the types of measures of central tendency and their corresponding measures of dispersion has to do with whether the data are symmetrically distributed or not. The purpose of descriptive analysis or modeling is not to establish causal relationships between variables or predict outcome but rather to allow a researcher to have a general sense of what the data are showing, on a variable by variable basis.

An explanatory model describes the effect of an intervention on outcome.[4] In this model one or

From the Medical School, University of Queensland, Australia, and Medical School, University of Pretoria (E.H.), South Africa; South Jersey Orthopedics (M.J.W.), Vorhees, New Jersey; and Wake Forest School of Medicine (R.B.D.), Winston-Salem, North Carolina, U.S.A.

The authors report the following potential conflicts of interest or sources of funding: M.J.W. received support from Arthroscopy: The Journal of Arthroscopic and Related Surgery *and Storz.*

Received November 14, 2016; accepted January 31, 2017.

Address correspondence to Erik Hohmann, M.D., Ph.D., F.R.C.S., Valiant Healthcare/Houston Methodist Group, P.O. Box 414296, Dubai, United Arab Emirates. E-mail: ehohmann@hotmail.com

0749-8063/161092/$36.00
http://dx.doi.org/10.1016/j.arthro.2017.01.054

more variables can be controlled by the researcher to a certain extent.[4] For example, a study design investigating the effect of anterior cruciate ligament reconstruction (ACLR) on the incidence of meniscus injuries compared to a control group that received conservative treatment investigates the effect of surgery on a specific condition. This would be an example of a comparative study. Let us assume that meniscal injuries are significantly lower in the ACLR group. The intervention (ACLR) therefore explains the lower incidence of meniscal injuries in the intervention group. A causal relationship between surgery and meniscus injury could be suggested if this study were designed properly, meaning if the patients were randomized to receive either treatment being examined and if the patients included in the study represented a random sample of all possible patients who could receive a meniscus injury—in other words, if the intervention has had an effect on the measured outcome variable. Explanatory statistics can be used for both experimental studies or observational data.[4] In general, it is more challenging to make causal inferences in observational studies since patients are not randomized to receive a treatment, and thus it is difficult to determine whether a difference between treatments is due to the treatment itself or the difference in patients who nonrandomly received one treatment or another.

In predictive modeling, observations are used to predict outcome and/or the results of an intervention or observation.[5] This model investigates associations between one or more (dependent) variables of interest and the independent predictor variables.

In a basic scientific experiment, the independent variables can be controlled to investigate their effect on the dependent variable. For example, in a cadaver model, the effect of varying the femoral and tibial tunnel position with or without anterolateral ligament reconstruction (independent variables) on rotational knee stability (dependent variable) is investigated. By changing the 2 independent variables (predictors), the outcome will change. In clinical studies, these predictors may not be controlled. A study investigating the effect of ACLR on functional outcome (dependent variable) with a validated scoring system (Lysholm, International Knee Documentation Committee, or similar) that intends to assess the influence of gender, body mass index (BMI), age, mechanism of injury, time to surgery, chondral and meniscal injuries, previous ACLR, and other associated injuries (independent variables) on outcome would be an example of a clinical study. Here it is not possible to easily vary or change the independent variables. When applying a predictive model to this study, predictions about the “future” are possible. The results of the analysis can help the researcher understand which of the independent variables influences (or predicts) the outcome.

Predictive Modeling

Prediction research aims to predict outcomes based on a set of independent variables and can provide information about the risk of developing a certain disease or predict the course of a disease based on the analysis of these predictor variables.[6,7]

Predictive modeling uses statistical techniques to predict outcomes, and several statistical models can be used.[5,7] Prediction research is any model that produces predictions[5] and includes such approaches as Bayesian techniques, data mining techniques such as machine learning, and classical statistical models of regression, logistic, linear, and Cox proportional hazards models, depending on the number and character of outcome variable(s).[8]

Bayesian Statistics

To describe all the differences between a classical frequentist approach to statistical inference and a Bayesian approach to statistical inference goes beyond the scope of this paper. Therefore we now give a brief overview of the differences in the approaches, recognizing that we are oversimplifying many of the details.

The main difference between classical hypothesis testing and Bayesian statistics is that in classical (frequentist) methods, a null hypothesis is constructed about a specific parameter (i.e., the mean value of a distribution) and then data are collected to estimate this parameter (i.e., data are collected and an estimate of population mean is made by calculating a sample mean from the data). The frequentist approach will then examine the data collected and the hypothesis made and determine whether (1) the data appear to contradict the null hypothesis, leading to rejecting the null hypothesis, or (2) the data seem consistent with the null hypothesis, leading to not rejecting the null hypothesis. In this framework of statistical modeling, the assumption is that what is observed during a particular experiment is only one plausible set of outcomes from a possibly much larger set of all possible outcomes. The frequentist tries to determine the likelihood that this one set of outcomes observed is consistent with a hypothesis that was previously stated (the null hypothesis), recognizing that when making inferences one can always make an error, that is, rejecting a null hypothesis when it was true (type 1 error) or failing to reject a null hypothesis when it is false (type 2 error). Prior to the data being collected, a researcher using this approach should specify the criteria for rejecting or not rejecting the null hypothesis. In general, researchers often use a 0.05 (5%) threshold to determine whether to reject the null hypothesis or not—meaning that if the data suggest that there is less than a 5% chance that the null hypothesis is true given the data observed (i.e., $P < .05$), one should reject the null hypothesis. There are several drawbacks to using this method, in particular