Research Pearls: The significance of statistics and perils of pooling. Part 2: Predictive Modeling.

Erik Hohmann

FRCS, FRCS (Tr&Orth), MD, PhD

Medical School, University of Queensland, Australia

Medical School, University of Pretoria, South Africa

Merrick Wetzler

MD

Cherry Hill, New Jersey, USA

Ralph D'Agostino Jr.

PhD

Wake Forest School of Medicine

Winston-Salem North Carolina, 27157. USA

Corresponding Author:
Erik Hohmann
Valiant Healthcare/Houston Methodist Group
PO Box 414296
Dubai
United Arab Emirates
ehohmann@hotmail.com
Tel : +971 4 378 2206

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 the 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 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 datasets and allows a researcher to make predictions about outcomes.

Regression models describe the relations between two 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 post-op) or time to event (i.e., time to surgical revision). The outcome variable is the variable of interest and the predictor variable/s are being 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 studies completion.

Key Terms:

Statistical analysis; predictive modeling; regression analysis; survival analysis

Introduction:

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 rudimentary understanding of statistical concepts especially when more complex analysis are required. [1]

With descriptive statistics data is 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 the spread around the "average" (range, inter-quartile range, variance, standard deviation). The primary difference between the types of measures of central tendency and their corresponding measures of dispersion have to do with whether the data are symmetrically distributed or not. The purpose of

descriptive analysis or modelling 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 is showing, on a variable by variable basis.

An explanatory model describes the effect of an intervention on outcome. [4] In this model one or 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 which 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 therefore (ACLR) 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. Or in other words 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 inference 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 non-randomly received one treatment or another.

In predictive modelling 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 antero-lateral ligament reconstruction (independent variables) on rotational knee stability (dependent variable) is investigated. By changing the two 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, IKDC or similar) intends to assess the influence of gender, 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 influence (or predict) 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 over-simplifying 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 is collected to estimate this parameter (i.e., data is 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 appears 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 inference one can always make an error (i.e., 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-value <0.05), one should reject the Null hypothesis. There are several drawbacks from using this method, in particular two of them are: 1) if the p-value is 0.049 there is still a 4.9% chance that the Null Hypothesis is true and a Type 1 error could be made and 2) statistical significance does not always directly link to clinical significance – meaning a p-value of less than <0.05 does not imply that the actual difference between groups is at all meaningful in real clinical practice.

In Bayesian statistics the researcher begins with a prior distribution that describes their current hypotheses concerning the question to be studied. If for instance previous studies had already taken place looking at this question, then the previous results of those studies could be used to generate a prior distribution or estimate for plausible outcome of the new study. This generation of a prior distribution to be used in the research occurs prior to collection of the data in. [9] These prior probabilities allow researchers to make estimates about the efficiency of a particular treatment and allows the researcher to incorporate all information of both the treatment arm and control group prior to data collection. If there is only anecdotal evidence about a particular treatment effect, these uncertainties can also be incorporated into the analysis. [9] In principle analysis entails four steps. In step one prior evidence is collected from the existing literature. In step two data is collected. In contrast to classical hypothesis

testing an a-priori sample size calculation is not necessarily needed, although there are methods for determining an appropriate sample size to be collected. In step 3 the collected data is used to revise the pre-estimates ("priors") using Bayes' theorem and in a final step the posterior or post study estimates are used to interpret the collected data. [9] In contrast to classical hypothesis testing there is no arbitrary cut-off of probabilities to call something statistically significant (i.e., no focus on whether a p-value is less than 0.05). Bayesian analysis rather describes probabilities that a certain treatment has an effect on outcome. For example: "there is a 95% probability that arthroscopic assisted ACL reconstruction with hamstring grafts results in a stable knee".

Here is another example to make it easier to understand the Bayesian approach. Let's say that there we have a simple blood test to determine whether a patient will develop rheumatoid arthritis (RA) in his lifetime. Let's also say that the know prevalence is 1/1000; this is the prior distribution or probability. The known false positive rate of this test is 10 percent. When we apply this test in a study including 1000 patents we will therefore find that 101 patients test positive. In classical statistics our results would therefore indicate that the chance of RA in our population group is 10.1 percent with a clinician raising fear in these 101 patients. With Bayesian statistics the prior distribution would be included and now we would conclude that only 1/101 will be positive allowing making better sense of the collected data.

Data mining – machine learning

Simply speaking machine learning uses algorithms to identify specific patterns in datasets to make predictions about outcomes. The variables (predictors) of interest and outcomes are identified; the software then applies these variables to make predictions about outcome. This

approach often makes no assumptions about the underlying distributions of the data being examined, whereas both classical and Bayesian models do make assumptions about the data (i.e., they usually assume that continuous data follow a normal distribution). The major advantage of this technique is that no specific hypothesis in contrast to conventional regression analysis is needed to predict that predictor 'A' is associated with outcome variable 'B'. [10] One of the major disadvantages of this technique is that generally large datasets are needed to allow useful conclusions. This is because machine learning approaches often involve fitting relatively complex models from the data that would involve multiple interaction terms in a traditional modeling framework. In general, these techniques have arisen out of the field of computer science and not statistical science, and therefore implement optimization algorithms found often in that field, without specific connections to modeling assumptions that are prevalent in statistical models. One criticism in this method is that often the optimal algorithm may appear to be "over-fitting" the data meaning that include more parameters are included in these models than would be considered appropriate for the sample size used and this limitation can only be addressed with large sample sizes. Further and similar to regression analysis the principle of Occam's razor is followed with many algorithms assuming that predictor variables are independent of one another. [10] However with machine learning non-linear relationships and interdependent variables can be examined in a more unstructured approach which may lead to innovative predictive models that may have been difficult to identify using more conventional approaches. [11] A simple example of machine learning are algorithms that allow a system to find patterns and correlations within a set of large data, i.e. identifying groups of friends in social network data.

Classical hypothesis testing – Regression Models

A more conventional approach to predictive modeling includes classic regression models. [8] It is important to understand that there are two fundamental differences in interpretation between the more conventional explanatory theory in regression and predictive modeling using regression. Fundamentally a difference between these two approaches is what is the underlying goal of the research being performed. In explanatory models, the goal of the research is often to understand specifically the relationship a particular independent variable to a particular dependent (outcome) variable. Thus, the goal is to understand, for example what the effect of BMI is on functional outcomes following ACLR. The reason to do this type of research is to determine (or recommend) what kind of modification of BMI may lead to what level of improvement (or worsening) of functional outcomes following ACLR. The specific relationship of BMI and functional outcomes is of interest. In predictive modeling one is not specifically interested in the relationship of any individual predictor (independent) variable and the dependent (outcome) variable, rather one is interested in finding a group of predictor variables that best allow one to predict what the outcome will be in the future. Thus, explanatory models focus on a particular relationship between the predictor and outcome, where it is assumed that there is a cause-effect relationship where 'Y' is caused by 'X'. It is retrospective testing of an already existing hypothesis. [4] Predictive models focus on understanding the predicted value of the outcome conditional on a set of predictor variables.

In both types of models, when the outcome of interest is measured on a continuous scale, the statistic, R^2 (R-square), can be used to measure the "goodness-of-fit" of a particular model. This statistics represents the proportion of the variability in the outcome measure that is explained by the predictor variable(s). In explanatory models, the focus often is on whether

there exists a high R^2 when one looks at the relationship between the independent variable of interest (i.e. BMI) and the outcome of interest (functional outcomes following ACLR). If for instance we observed a direct and statistically significant relationship between these two variables with an R^2 of 0.32 (meaning that 32% of the variability in functional outcomes following ACLR is explained by a patient's BMI) we may not believe that there is a fully causal relationship between BMI and the outcome, since there would be 68% of the variability not explained by BMI. However, if the calculated R^2 would have been 0.94 then the results would have a different meaning.

In predictive modeling the relationship between 'X' and 'Y' is examined in a prospective fashion establishing the relationship between two variables. [4] Often the goal of predictive modeling is to determine the best set of variables to make an accurate prediction of an outcome of interest, where there is no goal of understanding any particular variables role in the model, but rather the overall impact of the variables included. Therefore, the inclusion of many variables, even some not thought to be "statistically significant" is often thought to be appropriate in predictive modeling since the goal is to get the best predictive value of the outcome. Thus a high R^2 is more critical in predictive modeling than examining the specific impact of any particular variable in the model. The challenge faced in these models is that it can be shown mathematically that R^2 must increase as more variables are included in the model, however the inclusion of variables with little relationship with the outcome can also lead to overfitting problems, similar to those mentioned above in machine learning algorithms. . In theory, the lack of association cannot be compensated with a larger sample size as these predictions should be independent of the sample size. In contrast the lack of a

strong relationship in explanatory models may be due to a type II error and an increase in the sample may change the associations significantly. [1]

For simplification we will only outline the more classical regression model with explanatory modeling. As practicing clinicians we are far more familiar with these techniques. In principle, regression analysis examines the relationship or correlation between variables.

Simple Linear Regression

Simple linear or univariable regression is a mathematical technique that describes the relationship between two variables. [1,12] The relationship between the outcome variable 'y' and the predictor variable 'x' can be plotted on a scatter diagram (Figure 1). [12] When looking at the scatterplot it is often possible to visualize a line which passes through the midst of all points. [1] The regression line can be calculated by using a simple mathematical formula:

$$y = kx + c$$

'*k*' is the coefficient that describes the slope or gradient of the linear relationship and 'c' is a constant which describes where 'x' crosses the 'y' axis. [12] For inference or significance testing four assumptions about the relationship must be met: [12]

(1) A linear relationship between the two variables exists. If the points scatter randomly and do not center around a straight line, a relationship does not exist

(2) The variation around the regression line must be constant. In other words the distance from the regression line for all points should be similar.

(3) The data follows a normal distribution

(4) The deviation from the regression line for each data point is independent of other data points.

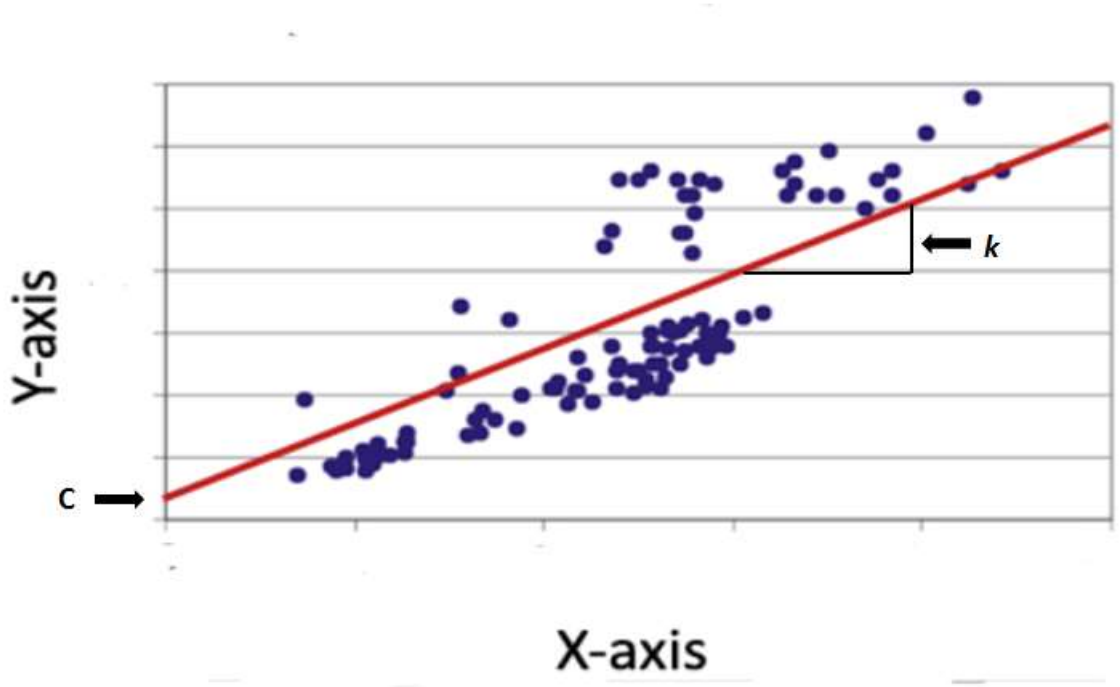

Figure 1 : Scatter Plot

The scatter plot is also called the X-Y graph. Each observation has two coordinates. The X-coordinate is the predictor variable and defines the distance from the Y-axis. Vice versa the Y-coordinate is the outcome variable and defines the distance from the X-axis. The regression line can often be visualized and should pass through the midst; alternatively a statistical software program can be used to draw the regression line. The regression line quantifies an inexact relationship meaning that the two variable are related to each other. The correlation coefficient measures the strength of the relationship between the two variables and falls between (-)1 and (+)1. A correlation coefficient of zero means that there is no relationship at all and the observations scatter all over the graph. If the correlation coefficient is 1 all observations are perfectly linear and located directly on the regression line. With correlation coefficients between 0 and 1 the regression line represents the best fit. '*k*' is the gradient and simply describes the steepness of the regression line. 'c' describes where the regression line crosses the 'y' axis which is not always at zero.

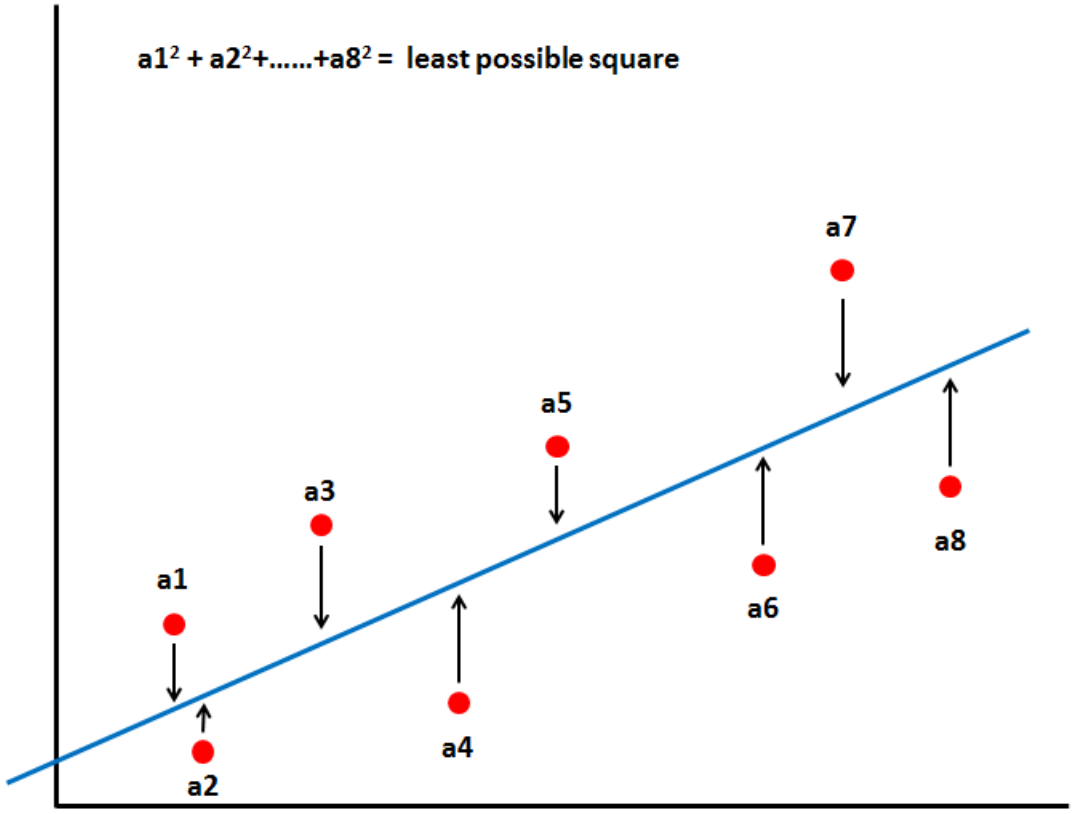

Figure 2 : Best Fit: method of least squares

The method of least square measures the distance of all data points from the regression line and the smallest vertical distance from the regression line is established by calculating the sum of the squares of the vertical distances.

If these four assumptions are met, the model is valid. To establish the best fit of these regression lines, a visual approach may help to "get an idea" where the line should be drawn but is more accurate to use a more scientific mathematical approach for best fit. Several estimation methods have been described but the most commonly used technique to find the best fit in linear regression is called the *method of least squares*. [13] The technique is based on the following two formulas and calculates the least square estimates for the constant 'c' and the coefficient '*k*':

$$c' = M_y - kM_x$$

$$k' = \frac{\Sigma (Y_i - M_y)(X_i - M_x)}{\Sigma (X_i - M_x)^2}$$

M_y and M_x are the mean values of both variables and Y_i and X_i are pairs of observations. What does this all mean? The least squares model establishes the smallest vertical distance between datapoints (Figure 2) and the regression line reducing error and creating the best fit for the regression line for all datapoints in the scatterplot.

A classical measure for linear relationships between two variables is Pearsons moment correlation. The correlation coefficient '*r*' ranges between -1 to +1. A positive relationship indicates an upward slope, whereas a negative relationship indicates a downward slope on the scatterplot. A value of 1 means that the relationship between the two variables is perfect (and linear) and the regression line moves through every datapoint. In contrast, if the relationship is zero, there is no linear relationship between the two variables. An example of a simple regression would be a study design that wants to establish whether posterior tibial slope is related to the amount of knee flexion. If the assumed relationship between the two variables is $r = +0.95$, the results would suggest that an increased posterior slope is associated with more knee flexion. In contrast if the relationship is $r = -0.95$ an increased posterior slope is related to less knee flexion. With Pearsons moment correlations the variables must be normally distributed and the relationship must be linear. The square of the correlation coefficient is R^2, the measure of goodness of fit described above, which represents the proportion of the variability in the outcome variable in the simple linear regression model that is explained by the predictor variable. In intuitive understanding of R^2 is the following, if one has a single continuous outcome variable measured that has a normal distribution the best

"guess" for any future measure of that outcome is the average (mean) value of the data already observed. However, if a predictor variable can be used in a regression model to "predict" the outcome, R^2 represents how much better the prediction is than just guessing the mean value.

If the variables in a regression model are not normally distributed or the relationship is not linear, then linear regression or Pearson correlations may be inappropriate to use. A non-parametric approach to examining correlation is to use as the Spearman's rank correlation, which essentially estimates the correlation between the ranks of the data (rather than actual observed values in the data set) However, since the actual values of the data are transformed into their ranks, the Spearman correlation coefficient provides an assessment of association rather than a linear association. [1]

For simple linear regression it is advisable to produce graphs to inspect data visually to ensure that the assumptions are met and outliers are checked. [12] For significance testing a parametric test such as a t-test can be used to determine whether the slope of the regression line is equal to 0 or not.

Multiple Linear Regression

In orthopedic surgery as in most other fields of medicine, it is unlikely that one variable determines outcome of a particular disease or intervention. When there is more than one predictor, different tests must be employed. Multiple linear regression or multivariable linear regression is a mathematical technique that allows to investigate the relationship between multiple independent predictor variables and a single dependent outcome variable. [12] It is an

extension of the simple linear regression and the same four assumptions must be met. The predictor variables can range from two to a large number depending on who many patients are included in the research study. . Similar to simple linear regression the regression line can be calculated by using a simple mathematical formula:

$$y = k_1x_1 + k_2x_2 + k_nx_n + c$$

To establish the best fit for multiple linear regression the method of least squares can also be used. If there are many predictor variables or covariates it is absolutely critical to have a large sample size. As a general rule there should be at least 10 times as many observations or patients per predictor variable. [1] For example if we would like to determine whether age, gender, BMI, sporting code and weekly exercise hours (5 predictor variables) influence the functional outcome of ACL reconstruction, a minimum of 50 patients are needed to make useful predictions. However it must be remembered that sample size has a distinct effect on what R^2 can be detected with statistical significance. Subsequently an increase in observations (patients) may change the associations significantly; a fact that needs to be considered when designing these type of studies and also when interpreting the results. Another important consideration is collinearity between predictor variables. It is not uncommon that predictor variables are related (correlated) to each other. In the above example it maybe that the higher BMI is highly correlated to the weekly exercise hours. This phenomenon is called collinearity and means that one predictor also predicts another predictor. Collinearity can have a significant effect on the outcome of the analysis and complicates the interpretation of the results. An obvious warning sign would be a substantial increase or decrease of R^2 when either removing or adding a predictor variable. In addition,

when counterintuitive regression coefficients appear in the same model (i.e., a predictor variable that alone should have a positive correlation with the outcome, but in a multiple linear regression model it has a negative slope in the model) this is often a signal that collinearity may exist in the model. Possible solutions are to remove highly correlated predictors or possibly perform a stepwise regression procedure, which allows variables to enter one at a time into the model, and therefore highly correlated variables will likely not enter into the same model. Another approach is use the *partial least squares regression method*. In principle partial least squares regression reduces predictors to the uncorrelated variables and then performs *least squares regression* on the remaining predictors. [14]

Logistic Regression

If the outcomes (dependent variables) are ordinal or categorical, simple linear and multiple regression should not be applied. For example if the dependent variable is a 'yes' or 'no', a logistic regression model is more suitable. Logistic regression describes the relationships between one or multiple numerical independent variables and one dependent categorical (yes/no) variable. There are several assumptions in such models. These include:

1) The outcome is measured on a binary (2-level) or ordinal scale.
2) The units (patients) included in the model are independent of each other
3) The independent variables and the outcomes are linearly related on the log odds scale.

This last assumption, of the linearity on the log odds scale is more technical to explain than is needed in this paper, however, in most cases with continuous predictors that have a somewhat symmetric distribution (i.e., approximately normal) the linearity assumption will be met. This

method uses logistic transformations to establish the probability of outcomes in a binary fashion. The outcome is then expressed as the odds ratio as a "yes" or "no" response. For example if the risk of ACL injury in males soccer player is '1' (control) and the odds ratio for females performing the same sports is '5', the results would indicate that females have a five times higher risk of ACL injury when playing soccer. If one would assess specific risk factors in the female cohort like coronal and sagittal knee flexion angles during a landing task, phases of menstrual cycle, quadriceps strength, radiological alignment of the lower extremity, logistic regression can estimate the odds, confidence intervals and significance (*p-value*) of each variable.

As with any analysis the findings and conclusions drawn from the analysis depend on whether an appropriate model has been used and whether the assumptions of the model have been satisfied. A critical step is to assess how well the model describes the observed data. [15] One of the traditional approaches to assess goodness of fit in logistic regression is to use Pearson's chi-squared test to examine the sum of the squared differences between the expected and observed number of cases divided by the standard error. One of the major problems with this test is its dependence on sample size. A smaller sample size may lead to the wrong conclusion of non-significance and increasing the sample size often leads to significance. [16] In addition, often a C-statistic is calculated from a logistic regression model and this measures the predictive accuracy of the logistic regression model. A C-statistics of 1.0 would suggest that the model used perfectly predicts the outcome of interest, whereas a C-statistic of 0 would suggest that the model could not predict the outcome of interest.

Survival Analysis

Survival analysis investigates the time until an event occurs. [17] This outcome can be described as failure or survival time (i.e., time to re-operation) or death. Failure or survial time is also called event time and data examined is always positively valued. A typical example in orthopedic surgery is the survival of total joint arthroplasty in joint registries around the world. [18,19] For patients who survive their arthroplasty and require revision surgery for septic or aseptic loosening, the event time is known exactly and the observation is complete. If patients cannot be followed up until failure occurs (i.e. death, lost to follow-up or withdraws), survival or event time is not fully observed. These incomplete observations are defined as censored data. [17] Censored data also occurs if a study ends and some of the included patients did not have an event during the study period. This type of censoring is called right censoring and occur when a participant does not have an event during the study period or drops out before the study ends. [20,21] Left censoring occurs when the event has already occurred before the study period. [22] This is very rarely encountered in orthopedic studies. For example a cross-linked polyethyelene insert is tested in the laboratory with cyclic loading and checked every 2 hours for failure. The first checkpoint occurs at two hours but the insert fails at 20 minutes already, long before the first check occurs. If the insert fails between two checkpoints, i.e. at 4.5 hours this is defined as interval-censored and means that there is uncertainty as to when the insert fails as the status is only checked every two hours. For this particular example the insert then fails between two and four hours.

In contrast to standard regression models, survival analysis allows inclusion of both censored and uncensored data. In some ways survival analysis is the combination of the linear and logistic regression in one technique. This is because it accounts for the outcome using a

continuous and binary form – specifically the "time" until the event occurs is a continuous measure and whether the event occurs (yes/no) is a binary measure. The challenging part of this model is that when an event is censored, the time variable is a lower estimate of the time to event and this must be accounted for in the model. The most commonly used approach for analyzing survival data in orthopedic surgery is the Kaplan Meier approach which is a non-parametric statistical approach. [23] The Kaplan Meier test uses lifetime data to estimate the probability of survival. Basic assumptions are used in this analysis: censored patients have the same prognosis as those who continue to be followed-up or are uncensored and survival probabilities are the same for all patients irrespective whether they were included early or late.

The Kaplan-Meier survival approach allows one to construct a curve that is a graphical tool demonstrating the results of the analysis (Figure 3). [23] The horizontal axis measures time and the vertical axis measures the proportion of patients free of the event. Thus, using the graph one can estimate the time it takes for a certain proportion of events to occur. [24] When interpreting the survival curves it is important to identify the units of measurement along the X-axis. Small steps with shorter intervals in general means larger patient cohorts whereas large steps have limited patient numbers. [24] This can typically be seen at the right aspect of the graph if either a large group had their event or data was censored during earlier study intervals. Large steps should be interpreted with caution and are not very accurate. Poor study design or ineffective treatment may result in large numbers of censored events and may also result in large steps and again requires caution with analysis. It should be noted that in these analyses, the only data that contributes to the actual statistical modeling is when events occur,

thus censored data does not directly contribute to the estimates of the statistics needed to estimate the survival curves.

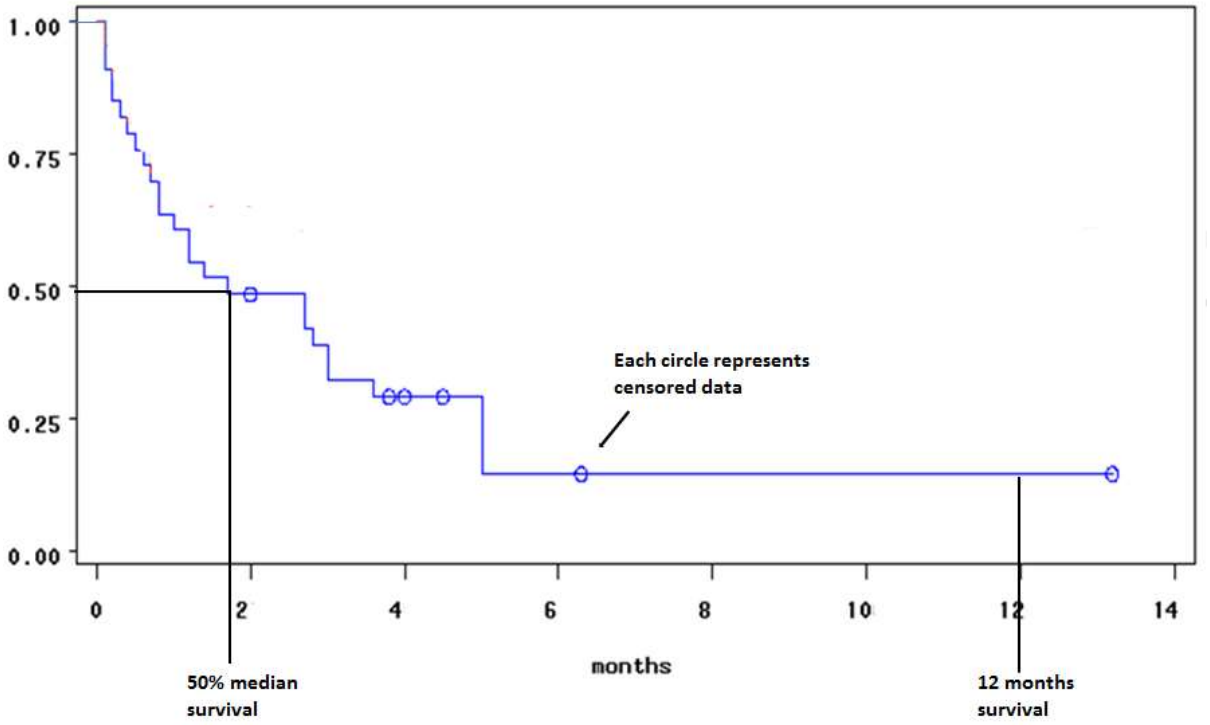

Figure 3 : Kaplan-Meier Survival Curve represents survival times

The x-axis denotes time and the y-axis denotes the percentage of a particular subject or object of interest has survived. Drops only occur at event times and the curve does not go to zero if there is no event at the last checkpoint or when the study has finished. The circle or dotpoints along the curve represent censored data. The curve allows to plot the 50% median survival time and check survival at specific time points. In medical research especially in cancer research one and five year survival rates are used to establish the effect of treatment on survival.

The Kaplan-Meier approach is a useful approach to examine survival curves and compare these curves among groups. If the main interest is to investigate the influence of risk factors on survival a Cox Proportional Hazards Regression (often referred to as Cox Regression) allows analysis for the relationship between time to event outcomes and one or more predictors to be made. [23,24] For example Cox regression could investigate the influence of

age, gender and radiological malalignment of a total joint arthroplasty on survival of the implant. Cox regression uses a non-parametric approach to fit the model. [23] The basic assumption that must be met is that the hazard or risk must be proportional. For example if women have twice the risk of ACL injury compared to men at age 20, they also must have twice the risk at age 30. In addition, the risk of an event occurring over time must be comparable similar between groups, so if women had twice the probability of an event occurring after 12 months of follow-up post-surgery when compared to men, then women should also have twice the probability of an event occurring after 24 months of follow-up post-surgery when compared to men. Generally speaking Cox regression allows one to estimate the risk for a particular individual to have an event considering all potential variables that can result in the event. The hazard function is a way to express the probability of an event occurring for a pre-determined time interval. [25] The hazard function can be expressed as:

$$h(t) = \frac{\text{Number of individuals with an event occurring during the time interval}}{\text{Number of individuals without an event during the time interval}}$$

The hazard ratio is an expression of the chance of an event occurring during a specific time interval. [23,25] For example if 1000 patients are surveyed during the month of September and October for ACL injury and 50 patients sustain an injury in both September and October the hazard ratio for September is 0.05 (50/1000) and 0.053 (50/950) for October. The hazard ratio can also be used to assess risk in more than one group. For example survival rates for ACL reconstruction over a specific time interval or two different surgical techniques could be evaluated.

Table 1: Terms and Definitions

Dependent Variable	The dependent variable is also called the outcome variable. The variable responds to the independent variable/s and changes if the independent variable/s are manipulated.
Independent Variable	The independent variable is also called the experimental or predictor variable. It can be manipulated and represents input and directly effects the dependent variable.
Parametric Data	Follows a normal Gauss distribution and displays homogeneous variance
Non Parametric Data	The distribution follows any pattern and variance
Variance	Is a measure of how the data is spread. t is defined as the average of the squared differences from the mean
Standard Deviation	Is also a measure of how the data is spread. It is defined as the square root of the variance
Z Score	Is a measure of how many standard deviations below or above a raw score is The raw score is the original score from one test/individual
Sample Size Calculation	If a sample size calculation is done before data collection it is called a-priori and is used to establish that the sample is sufficiently powered. Post hoc power or the observed power of the sample is performed once the data has been collected and is based on the effect size estimate
Type I error	Type I errors occur when the null hypothesis is rejected when the null hypothesis is true. It is also referred to as "false positive". Alpha levels (p-values) are the probabilities of a type I error occurring. For example a p=0.05 means that there is a 5% chance that a true null hypothesis will be rejected
Type II error	Type II errors occur when the null hypothesis is false but accepted. This is also referred to as "false negative" This error most often occurs when there is no difference in outcome because the sample is too small.
Type III error	Type III error occur when the right answer to the wrong question is provided. For example it is correctly concluded that the two groups are different but sampling results in a variable to be lower in one group. With more samples the variable then increases and results in no differences between the two groups
Occam's razor	Is derived from the philosophical principle by William of Occam stating that entities should not be multiplied without necessity. In data mining it means that when there are two models with the same error, the simpler should be preferred because it has possibly a lower generalization error.
Correlation coefficient **r**	R measures the strength and direction of a linear relationship between two variables. A value of +1 is perfect positive relationship and a value of -1 is a perfect negative relationship. Values can range between +1 and -1
Coefficient of determination **R-squared**	R squared is a measure of how close the data is to the regression line and explains the variability of data around the mean. 100% indicates that the model explains all the variability and 0% means that the model does not explain the variability. In other words it is the percentage of variability that is explained by the model.

Table 2: Which Statistical Tests To Use For Classical Predictive Modelling

	Parametric Data (Normal Gauss Distribution)	Non-Parametric Data (Non-Gauss Distribution	Two Possible Outcomes (Binominal)	Survival Analysis
Descriptive	Mean, Standard Deviation	Median, Interquartile Range, Range	Proportions	Kaplan-Meier Estimate and Survival Curve
Relationship between two variables	Pearson Moment Correlations	Spearman Rank Correlation	Contingency Tables and Coefficients	
Predict outcome from one measured variable	Simple Linear Regression	Non-Parametric Regression	Logistic Regression	Cox Proportional Hazards Regression
Predict outcome from multiple measured variables	Multiple Linear Regression		Multiple Logistic Regression	Cox Proportional Hazards Regression

Discussion

Thus we have presented a brief overview of several possible statistical techniques that can be used in predictive modeling research. In table 1 we have summarized the most commonly used terms and definitions with regards to these statistical tests and table 2 summarizes the statistical tests typically used for classical predicitive modeling. Each method has potential strengths and limitations and researchers should be aware of these prior to initiating such a project. Among the methods described above, the ones that are most often used in current research are those that focus on either binary outcomes (i.e., 2-year revision rates) or time-to-event outcomes (i.e., median time to joint failure after surgery). When examining these types of outcomes, Bayesian methods can be used in conjunction with classical regression techniques (logistic regression or Cox regression). In addition, machine learning approaches can also be used to examine these type of outcome models. Finally, the classical (frequentist) approaches of logistic and Cox regression models can be used without a Bayesian framework.

Regardless of which method is used, one should focus on the primary goal in such analyses, which is to best predict the likelihood of the event of interest (i.e., the occurrence of a

revision surgery within 2 years or joint failure after surgery). Therefore, it is imperative that all pertinent predictor variables are measured during the study. These include patient level characteristics such as age, gender, race, bmi, smoking status, as well as other risk factors and co-morbid conditions which could influence the outcome such as diabetes, hypertension, number of previous surgeries, etc.

Ultimately, one goal of developing predictive models is to be able provide to clinicians decision support systems that can eventually provide real-time pertinent information concerning their patients and recommendations on treatment decisions that should be made in order to optimize long-term results on procedures to be performed.

Conclusion/Summary

Predictive modeling is a technique that can use several different statistical techniques to predict future outcomes. There are two principle approaches. When the relationship is examined in a prospective fashion the relationship between two or more variables are established to predict future outcomes. With classical hypothesis testing regression models are applied to retrospectively test an already existing hypothesis. Simple and multiple or multivariable regression models are used for continuous data and logistic regression for categorical data.

For all regression analysis it is worthwhile creating scatterplots and visually inspect for goodness to fit and outliers. Goodness to fit tests should be used to create the best fit for the regression line representing all datapoints in the plot and reducing error.

Survival analysis uses censored and non-censored data and is a useful statistic to analyze survival. If the main interest is how risk factors influence survival the Cox Proportional Hazards Regression can be used to investigate the effect of predictor variables on survival.

As quality metrics are becoming part of the evaluation of performance for surgeons and will likely be linked to reimbursement rates, it will be more important to have accurate predictive models available to assist with evaluating performance and also guide decisions to be made in the clinical setting. To do this accurately, one needs to use valid statistical methods and understanding these methods will provide a higher probability of success in these endeavors.

References

1. Petrie A. Statistics in orthopedic papers. *J Bone Joint Surg Br* 2006; 88-B:1121-1136

2.Larson MG. Descriptive statistics and graphical displays. *Circulation* 2006; 114 (1):76-81

3. Nick TG. Descriptive statistics *Methods Mol Biol* 2007; 404:33-52

4. Shmueli G. To explain or to predict? *Statistical Science* 2010; 25 (3): 289-310

5. Waljee AK, Higgins PDR, Singal AG. A primer on predictive models. *Clin Transl Gastroenterol* 2014; 5, e44

6. Toll DB, Janssen KJ, Vergouwe Y, Mooms KG. J *Clin Epidemiol* 2008; 61 (11):1085-1094

7. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009; 338:b605

8. Cerrito PB. The difference between predictive modeling and regression. Proceedings of the 2008 Midwest SAS Users Group. S03

9. Lewis RJ, Wears RL. An introduction to the Bayesian analysis of clinical trials. *Ann Emerg Med* 1993; 22 (8):1328-1336

10. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. *Am J Gastroenterol* 2010; 105 (6):1224-1226

11. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics* 2006; 2:59-77

12. Marill KA. Advanced statistics: linear regression, part 1: simple linear regression. *Acad Emerg Med* 2004; 11 (1):87-93

13. Abdi H. The methods of least squares. In: Neil Salkind (Ed.) 2007. Encyclopedia of measurement and statistics. Thousand Oaks (CA); Sage

14. Abdi H, Williams LJ. Partial least squares methods: partial least squares correlation and partial least square regression. *Methods Mol Biol* 2013; 930:549-579

15. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *Am J Public Health* 1991; 81:1630-1635

16. Hosmer DW, Hosmer T, Le Chessie S, Lemeshow S. A comparison of goodness-to-fit tests for the logistic regression model. *Statistics Med* 1997; 16:965-980

17. Altman DG, Bland JM. Time to event (survival data). *BMJ* 1998; 317 (7156):468-469

18. Patel A, Pavlou G, Mujica-Mota RE, Toms AD. The epidemiology of revision total knee and hip arthroplasty in England and Wales: a comparative analysis with projections for the United States. A study using the National Joint Registry dataset. *Bone Joint J* 2015; 97B (8):1076-1081

19. Fennema P, Lubsen J. Survival analysis in total joint replacement: an alternative method of accounting for the presence of competing risk. *J Bone Joint Surg Br* 2010; 92 (5):70-706

20. Zhang J, Heitjan DF. Nonignorable censoring in randomized clinical trials. *Clin Trials* 2005; 2 (6):488-496

21. Lagakos SW. General right censoring and its impact on the analysis of survival data. *Biometrics* 1979; 35 (1):139-156

22. Prinja S, Gupta N, Verma R. Censoring in clinical trials: Review of survival analysis techniques. *Indian J Community Med* 2010; 35 (2):217-221

23. Abd ElHafeez S, Torino C, D'Arrigo G, Bolignano D, Provenzano F, Mattace-Raso F, Zocacali C, Tripepi G. An overview on standard statistical methods for assessing exposure-outcome link in survival analysis (Part II): the Kaplan-Meier analysis and the Cox regression method. *Aging Clin Exp Res* 2012; 24 (3):203-206

24. Rich JT, Neely JG, Paniello RC, Voelker CCJ, Nussenbaum B, Wang EW. A practical guide to understanding Kaplan-Meier curves. *Otolaryngol Head Neck Surg* 2010. 143 (3):331-336

25. Fleming TR, Lin DY Survival analysis in clinical trials: past developments and future directions. *Biometrics* 2000; 56 (4):971-983