# Statistical Analysis of Clinical Variables to Predict the Outcome of Surgical Intervention in Patients with Knee Complaints

**John W. O'Kane**
Department of Orthopedics
Sports Medicine Clinic, Box 354060
University of Washington
Seattle, WA 98195-4060
jokane@u.washington.edu

**Greg Ridgeway and David Madigan**
Department of Statistics
Box 354322, University of Washington
Seattle, WA 98195-4322
{greg, madigan}@stat.washington.edu

## INTRODUCTION

Knee complaints secondary to injury and overuse are common in both general and orthopedic practice. They are particularly common in athletes and other physically active individuals. Knee problems including significant meniscal tears, anterior cruciate ligament (ACL) tears, intraarticular fractures and osteochondritis dessicans are often best managed with surgical treatment. On the other hand, problems including patellofemoral pain, medial collateral ligament sprains, iliotibial band syndrome, and patellar tendonitis are generally best managed non-operatively with appropriate rehabilitation. Based on history and physical exam, it can be difficult to separate those patients with knee pain likely to benefit from early surgical intervention from those in whom initial conservative treatment is more appropriate.

Findings are mixed in prior studies assessing the ability of clinical examination to predict the arthroscopic diagnosis in patients with knee complaints. Gibson, et al. found that clinical exam resulted in an unequivocal diagnosis of internal derangement in only 35% of the cases [Gibson T, 1987 #5]. They report a frequent discordance between clinical diagnosis and arthroscopic findings. Much of the literature focuses on the diagnostic accuracy of clinical exam to predict meniscal tears. While some authors have not found a clinical pattern that would reliably predict meniscal tears [Noble J, 1980 #6], others have found that a combination of historical and physical examination variables can predict meniscal tears with some accuracy [Barry OCD, 1983 #3]. Anderson and Libscomb [Anderson AF, 1986 #2] found that at least one positive mechanical test in 79% of meniscal tears. Alternatively, Curtin, *et al* [Curtin W, 1992 #4] found that for 175 patients taken to arthroscopy, clinical exam and plain radiography demonstrated poor specificity for medial meniscal tears and poor sensitivity for lateral meniscal tears. They demonstrated better specificity for ACL tears and of 30 predicted, 26 were confirmed. There were 7 ACL tears discovered arthroscopically that were not diagnosed on clinical exam. Finally, Abdon, *et al* [Abdon P, 1990 #1] found that while clinical accuracy in detecting meniscal tears was 61%, employing a multivariate analysis of 68 different clinical variables could correctly predict a meniscal tear in 80% of the cases. This last study, like the others, demonstrates that making an accurate diagnosis based on history and physical is difficult. It suggests though, that analysis of the variables statistically may be more accurate than the clinical impression based on those variables.

While the literature sited examines the accuracy of diagnosis following history and clinical exam, an important question not directly addressed is the likelihood that a patient presenting with a knee complaint will benefit from surgical intervention. For the primary care physician contemplating a referral for surgery, or for a surgeon contemplating arthroscopy, the likelihood the patient will benefit from the procedure is of primary concern. This study attempts to determine whether or not historical and clinical

variables at the time of presentation can accurately predict if a patient is likely to have a surgical knee problem. It also examines the role of modern statistical techniques and machine learning to more accurately predict the answer to this clinical question.

METHODS

Data were collected through a retrospective chart review in a university based orthopedic sports medicine clinic. Charts were pulled sequentially in alphabetical order and the record was reviewed for all knee diagnoses. Data were collected for all patients in whom the surgical or non-surgical treatment was satisfactorily completed. The patient's age and gender were noted, and binary data were collected for the historical and clinical variables noted in Table 1.

| History | Physical |
|---|---|
| age | effusion |
| gender | range loss |
| swelling:none | instability MCL |
| swelling: slow | instability LCL |
| swelling: rapid (<12 hours) | instability ACL |
| fracture on XR | patellar crepitus |
| unilateral | McMurray's |
| injury | tender medial joint line |
| locking | tender lateral joint line |
| instability | tender anterior/patella |
| mechanical | |
| anterior pain | |
| localized pain (other than ant.) | |
| sport related | |
| industrial | |
| prior surgery | |
| prior injury | |
| depression | |

**Table 1:** Historical and clinical variables

Data were collected on 99 patients, and analyzed using a boosted naïve Bayes classifier.

**Naïve Bayes classification**

Naïve Bayes classification, a statistical technique with a moderate history in medical applications [Spiegelhalter DJ, 1984 #7], seemed a well-suited approach for this scenario. The literature at times refers to naïve Bayes classification as simple, idiot's, or independence Bayes classification. Using this model we attempted to construct an accurate predictor of the necessity of knee surgery from historical and clinical variables. The appendix contains an overview of the naïve Bayes model.

Furthermore, empirical studies have shown that building a sequence of classification models and merging them together to form one model often increases the predictive performance of the classifier. An interesting class of such methods consists of the "Adaptively Resample and Combine" algorithms [Breiman L, 1996 #11]. These algorithms sequentially generate classifiers where the observations in the training set that the current classifier predicted poorly receive a higher weight on the next iteration. Adaptively reweighting the training set in this manner forces successive classifiers to work harder on the regions of the sample space that are difficult to classify. After a fixed number of iterations the set of classifiers vote on the final prediction. We reweighted the observations as defined by the AdaBoost algorithm [Freund, 1995 #9] and applied the voting scheme developed in Ridgeway, *et al* [Ridgeway G, 1998 #12]. This drives the misclassification rate on the training set to zero exponentially quickly and provides an interpretable model with improved generalization accuracy. Elkan [Elkan C, 1997 #8] applied boosting to the naïve Bayes classifier and showed that it is mathematically equivalent to a non-

parametric, non-linear generalization of logistic regression. We applied the boosted naïve Bayes classification model to the prediction of the necessity of knee surgery.

We evaluated our model according to procedures common in the statistics and machine learning communities. We first randomly divided the patient sample of 99 observations into two groups, a training set and a test set. We estimated the parameters of the boosted naïve Bayes model as well as the optimal number of boosting iterations using only the training set. The estimated model was then used to predict the necessity of knee surgery on the patients in the test set and compute the misclassification rate of the test set. Repeating these steps for several partitions of the sample and averaging the misclassification rate for each partition yields an estimate of the accuracy of the classifier on future observations. The standard jackknife procedure is a special case of this evaluation method where the test set contains only one patient and the remaining patients compose the training set. We also varied the proportion of the observations used in the training set to estimate a "learning curve" for the knee injury classification problem that helps determine the number of patients needed for constructing an accurate model.

**Weights of evidence**

In making clinical decisions, the physician needs to know how the states of the individual variables contribute to the classifier's final diagnosis. That is, knowledge of the extent to which the presence of a symptom is evidence for or against a diagnosis is critical to the utility of any medical decision support system. Spiegelhalter and Knill-Jones [Spiegelhalter DJ, 1984 #7] advocate extensive use of weights of evidence in medical diagnosis and propose evidence balance sheets as a means of viewing the reasoning process of the naïve Bayes classifier (also Seymour, et al ) [Seymour DG, 1990 #14]. We found weights of evidence to be a simple and transparent way of visualizing the boosted naïve Bayes classifier's reasoning process for knee surgery recommendation.

A weight of evidence is the logarithm of the odds in favor of knee surgery. Let *Y* represent the necessity of knee surgery (*Y=0* indicates no surgery, *Y=1* indicates surgery). Let *X* represent the collection of *d* indicants (symptoms, history, medical exam variables, etc.). For the non-boosted naïve Bayes classifier, writing the log-odds in favor of *Y=1* we obtain the following

$$\log \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)} = \log \frac{P(Y=1)}{P(Y=0)} + \sum_{j=1}^{d} \log \frac{P(X_j \mid Y=1)}{P(X_j \mid Y=0)}$$

$$= w_0 + \sum_{j=1}^{d} w_j(X_j)$$

The $w_j$ are the weights of evidence described by Good [Good IJ, 1965 #16]. A positive $w_j(X_j)$ indicates that the state of $X_j$ is evidence in favor of the hypothesis that *Y=1*. A negative weight is evidence for *Y=0*. More recently, Madigan, *et al* [Madigan D, 1996 #17] and Becker, *et al* [Becker B, 1997 #15] further discuss and develop the explanatory strengths of weights of evidence. Ridgeway, *et al* [Ridgeway G, 1998 #12] propose an extension of the weight of evidence for the boosted naïve Bayes classifier, discuss its properties, and demonstrate its performance on several data sets.

**RESULTS**

**Estimated weights of evidence**

Table 2 shows the estimated weights of evidence, $\hat{w}_j(X_j)$. The point estimates shown are the expected value of the boosted weight of evidence. The number shown in parentheses is a bootstrap estimate of the standard deviation of $\hat{w}_j(X_j)$. If the weight of evidence estimates were normally distributed then the ratio of the estimate to the estimated standard deviation would provide a standard normal test statistic for testing whether the weight of evidence differed significantly from 0. However, the bootstrap distribution of the estimates were often skewed, indicating substantial departures from normality. According to the

bootstrap distribution, if $\hat{P}(\hat{w}_j(X_j) > 0)$ is less than 0.025 or exceeds 0.975 (evidence that the weight of evidence is strongly negative or strongly positive respectively) then we boldfaced the variable in Table 2. This amounts to a $\alpha$=0.05 test based on the bootstrap percentile interval [Efron B, 1993 #13]. This is analogous to the $p < .05$ used to determine statisical significance in other statistical models. The bold faced weights of evidence indicate that the associated variable is an independent significant predictor for or against surgery.

| | |
|---|---|
| Prior | -1 (11) |

| Variable | negative | positive |
|---|---|---|
| Unilateral | -29 (64) | 5 (7) |
| Injury | **-50 (23)** | **39 (14)** |
| Locking | **-6 (3)** | **172 (50)** |
| Instability | **-14 (5)** | **88 (50)** |
| Mechanical | -1 (31) | -4 (25) |
| Anterior pain | 0 (17) | -4 (38) |
| Local pain | 23 (32) | -10 (12) |
| Sports related | 14 (15) | -19 (15) |
| Industrial | -5 (8) | 34 (96) |
| Prior surgery | 12 (8) | -26 (41) |
| Prior injury | -7 (9) | 20 (33) |
| Depression | -5 (8) | 14 (64) |
| Effusion | **-72 (24)** | **85 (29)** |
| Range loss | -8 (15) | 12 (22) |
| Instability MCL | 6 (4) | **-140 (53)** |
| Instability LCL | **-4 (2)** | **133 (36)** |
| Instability ACL | **-34 (7)** | **298 (19)** |
| Patella crepitus | -11 (29) | 13 (34) |
| McMurray's | **-38 (25)** | 56 (52) |
| Tender med JL | **-30 (14)** | **49 (18)** |
| Tender lat JL | **-20 (9)** | **50 (29)** |
| Tender anterior | **13 (6)** | **-72 (89)** |

| | Male | Female |
|---|---|---|
| Sex | -6 (13) | 8 (14) |

| | ≤ 24 | 25 - 32 | 33 - 46 | ≥ 47 |
|---|---|---|---|---|
| Age | -3 (37) | 28 (22) | -1 (26) | -12 (23) |

| | None | Slow | Rapid |
|---|---|---|---|
| Swelling | **-41 (40)** | 58 (37) | -29 (23) |

| | No | Yes | Arthritis |
|---|---|---|---|
| XR-fracture | 64 (75) | -22 (14) | 45 (60) |

**Table 2:** Estimated weights of evidence

The weights in Table 2 are additive. That is, to determine the total weight of evidence in favor of a patient needing knee surgery the physician merely needs to elicit a collection of indicants, sum the associated weights, and note the magnitude and sign of the total weight. The total weight may be converted into a probability by the following formula

$$p = \frac{1}{1 + \exp(-\hat{w}/100)}$$

or by the conversion table shown in Table 3 as well as in Figure 1.

| Probability | Total weight of evidence |
|---|---|
| 10% | -220 |
| 20% | -139 |
| 30% | -85 |
| 40% | -41 |
| 50% | 0 |
| 60% | 41 |
| 70% | 85 |
| 80% | 139 |
| 90% | 220 |

**Table 3:** Conversion from weight of evidence to probability

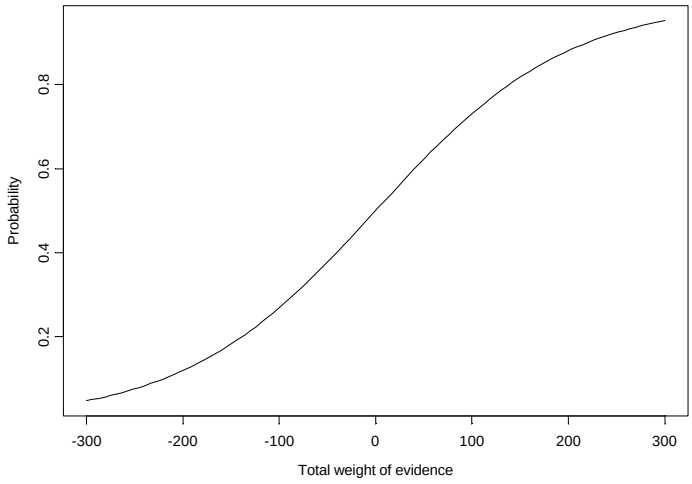


**Figure 1:** Conversion from weight of evidence to probability

For a new patient a physician could easily create an evidence balance sheet (Spiegelhalter and Knill-Jones) [Spiegelhalter DJ, 1984 #7] from Table 2. Table 4 shows an example of such an evidence balance sheet.

| **Evidence in favor of knee surgery** | | **Evidence against knee surgery** | |
|---|---|---|---|
| Female | +8 | Age 50 | -12 |
| Knee is unstable | +88 | No effusion | -72 |
| Knee locks | +172 | Negative McMurray's | -38 |
| Tender med JL | +49 | | |
| Total positive evidence | +317 | Total negative evidence | -122 |
| | | | |
| **Total evidence** | | +195 | |
| **Probability of knee surgery** | | 88% | |

**Table 4:** Example evidence balance sheet

**Performance**

As described in the methods, during the construction of the boosted naïve Bayes classifier the classifier has access to the records of a subset of the patients. The model, using parameter estimates obtained from these observations, then attempts to predict which of the remaining patients will require knee surgery. Restricting the number of patients in the training set and estimating the misclassification rate on the test set yields the model's "learning curve". Figure 2 shows the estimated learning curve for the necessity of knee surgery. With our training set of 66 patients we predicted knee surgery in the test group of 33 patients with 87% accuracy.

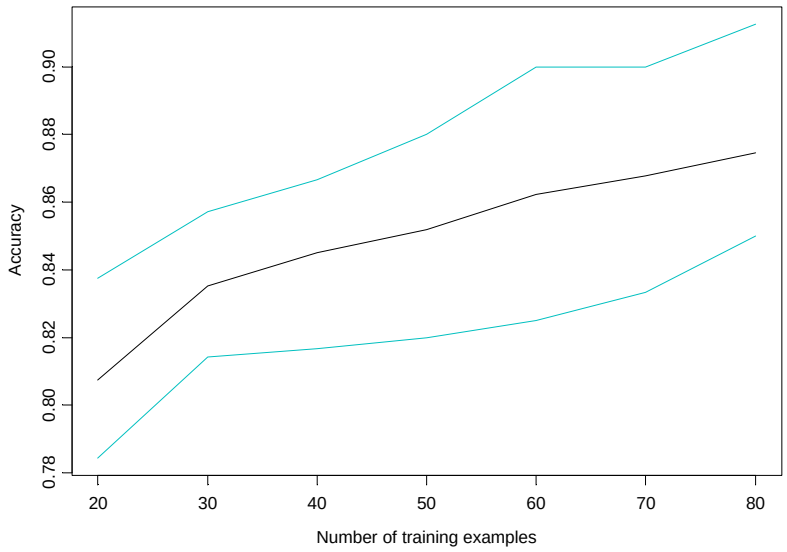


**Figure 2:** Learning curve for predicting the necessity of knee surgery (mean misclassification with 1st and 3rd quartiles)

## DISCUSSION

For a patient with a knee complaint at initial presentation, the results support the hypothesis that applying statistical methods to analyze historical and physical exam variables can accurately predict the probability that the patient will need knee surgery. In our patient population with a training set of 66 patients, the model predicts the likelihood of surgery with 87% accuracy employing all of the variables collected. As evidenced by the learning curve, for up to about 50 patients the accuracy increases steeply and beyond that the curve begins to level off. However, extrapolating the learning curve beyond 80 patients seems to indicate that the accuracy can potentially improve further with the addition of more patients. This information is useful in planning an appropriate sized training set which any practice could employ to develop a model based on its own patient population.

The greatest advantage of this tool is that a practitioner can develop a model using his/her own training set so the predictions will reflect their practice experience. Our patient training set allows accurate estimates for our clinic, but can not estimate surgical probability with the same accuracy over a wide range of patient populations. For the data collection period, our clinic was staffed by two orthopedic surgeons and one primary care sports physician. A proportion of our patients had ACL tears and were referred for surgery. As a result, in this patient set, the finding of anterior laxity had a significant weight of evidence toward surgery. Were the data set collected in a non-surgical, non-sports clinic, a greater number of ACL tears would likely be managed successfully without surgery. Our weights of evidence

would not be accurate for that patient population, but the classification system using those patients as a training set would accurately predict surgical probability for that practice.

A second useful feature of the model is the ability to provide weights of evidence for individual variables. When faced with conflicting clinical variables, it is useful to have an understanding of the significance of those variables relative to a given outcome. For example, in a patient with localized medial tenderness to palpation, an effusion, and a negative McMurray's; the surgical probability is 70% to 80%. The negative McMurray's goes against a surgical diagnosis, but it is outweighed by the other two findings. Adding a locked knee to the clinical picture increases the surgical probability to greater than 90%. On the other hand, adding medial instability to the original findings drops the probability of surgery to just under 40%. Clinically this makes sense because with medial instability, an MCL tear, generally managed non-operatively becomes a more likely diagnosis. The effusion, more likely with internal derangement than an isolated MCL tear, adds to the probability of surgery. If the effusion is absent, leaving only medial joint line tenderness and medial instability, the likelihood of an isolated MCL tear increases, and the probability of surgery drops to less than 30%. The model achieves its high accuracy incorporating all of the clinical variables. Practically speaking though, without the aid of a computer, the process of clinical decision making involves assessing the relative importance of a handfull of clinical variables. The model is also useful in this arena, by assigning specific weights to individual variables, thereby assisting the practitioner in unraveling seemingly conflicting findings.

The surgical probabilities also provide useful information regarding further diagnostic tests. Many surgeons feel pressured by patients or their insurance companies to order an MRI as additional "proof" that a patient needs a surgical intervention for their suspected clinical diagnosis. In a patient with a high probability of surgery, it may be more cost effective to proceed directly to arthroscopy than to invest time and money obtaining an MRI. In patients with an intermediate probability of surgical intervention, an MRI may provide useful additional information regarding diagnosis and subsequent treatment.

For primary care providers the probabilities are useful regarding medical decision making and referrals. For patients with high surgical probabilities, an orthopedic referral may be more cost effective than further diagnostic testing or a course of non-surgical treatment. With lower probabilities of surgery, further diagnostic testing may establish an accurate diagnosis and then based upon the diagnosis, a rehabilitation program may be prescribed. Some of these patients may eventually come to surgery, but the probability is that they will not, and an unnecessary orthopedic referral may be avoided. This assumes of course that the primary care provider is able to make a working diagnosis on which to base treatment. Regardless of the surgical probability, referral to a musculoskeletal specialist is always warranted if the diagnosis remains elusive and the patient is not seeing improvement over a reasonable time course.

The data also may be useful for patient triage in a system partitioning patients between surgical and non-surgical providers. Using our weights of evidence; patients with a history of injury, locking, or instability would best be referred to a surgeon for evaluation because the probability that they will need surgery is high. Alternatively, patients who deny swelling, injury, locking, or instability could start with a non-surgical provider because the probability that they will need to be referred on for surgery is low.

Limitations of this study include retrospective data collection. We are currently in the process of collecting prospective data in the same clinic to validate the accuracy of the model. Because two of the three providers in the clinic were orthopedic surgeons, one can argue that the patient population was biased toward surgical intervention. While this may be true, this model can accurately reflect any patient population by simply using that population as the training set. We are currently in the process of designing an interface through which different training sets can be easily entered. We are also working on a modification of the classifier that would allow the prediction of a diagnosis in addition to a surgical probability. A given diagnosis, unlike the probability of surgery, is not a binary variable complicating the statistical analysis. A multidimensional weight of evidence table would be cumbersome, but the model may be able to pick a most likely diagnosis out of a group of possibilities. This work is ongoing.

One potential inconsistency in the results is that swelling is not significant as an isolated variable but the presence or absence of effusion is highly significant. Clinically we do not feel that this data is inconsistent because effusion is an objective finding and swelling a subjective report. Many times, patients erroneously think their knees are swollen or do not recognize an effusion that is clearly present on exam. Not surprisingly, swelling has large weights of evidence, but also large standard deviations resulting in its lack of significance as an isolated variable.

In conclusion, our findings support the notion that statistical analysis of clinical variables can predict the probability of surgery with close to 90% accuracy. Analysis of the variables using a boosted naïve Bayes classifier can estimate the probability of a surgical outcome, as well as provide weights of evidence for the individual variables. An attractive characteristic of this model is that it has the potential to "train" on any patient population for which the clinical variables and surgical outcome are compiled, and as a result, the predictions can be made directly applicable to any given practice. In a medical workplace with increasing utilization of computer technology and a practice environment with increasing pressures to control costs, referrals, and diagnostic procedures; computer assisted diagnostic tools will likely gain in popularity. The boosted naïve Bayes classifier is one example of a computer diagnostic tool with a role in clinical practice.

***Appendix A: Naïve Bayes classification***

If the random variable $Y$ represents the necessity of knee surgery (0 – not necessary, 1 – necessary) then Bayesian classification predicts $Y$ to be the label $k$ which maximizes $P(Y=k \mid X_1, \ldots, X_d)$. That is, given a list of observed features $X_1, \ldots, X_d$ consisting of historical and clinical exam variables, predict that knee surgery is necessary if $P(Y=1 \mid X_1, \ldots, X_d) > P(Y=0 \mid X_1, \ldots, X_d)$ and otherwise predict that knee surgery is not necessary. The naïve Bayes assumption is that the observed features, $X_i$, are independent of one another given the necessity or non-necessity of knee surgery. Although this is almost never satisfied in practice, the model has repeatedly proved itself to be robust to dependencies (Domingos and Pazzani) [Domingos, 1996 #10]. Furthermore, this assumption greatly simplifies the estimation of the parameters and the classification rule.

***Naïve Bayes classification model***

Classify an observation to the class $k$ that maximizes

$$
\begin{aligned}
& P(Y = k \mid X_1, \cdots, X_d) && \text{where } k = 0 \text{ or } 1 \\
& \propto P(X_1, \cdots, X_d \mid Y = k) P(Y = k) && \text{Bayes' Theorem} \\
& = P(X_1 \mid Y = k) \bullet \cdots \bullet P(X_d \mid Y = k) \bullet P(Y = k) && \text{Naïve Bayes assumption}
\end{aligned}
$$

The parameters of the naïve Bayes classification model, the $P(X_i \mid Y=k)$'s and $P(Y=k)$'s, are trivial to estimate from a training data set even in the presence of missing $X$'s.