http://dx.doi.org/10.2106/JBJS.OA.20.00128

The following content was supplied by the authors as supporting material and has not been copy-edited or verified by JBJS.

Appendix I

A. Machine Learning

This study was based on the use of machine learning to recognize various clusters of values ("patterns") in the preoperative data to predict the surgical outcomes for patients who exhibit those values. The predictions usually consist of probabilities of the occurrence of possible outcomes for a patient whose data matches the pattern. A collection of patterns is termed a "predictive model" if it yields a unique prediction for any patient for whom the preoperative data is known. The specific machine learning models used in the present study were created by OBERD – Universal Research Solutions, LLC. (Columbia, MO; www.oberd.com) from the patient data described in the text.

A desirable characteristic of the approach adopted is that it is self-correcting: if a useless variable is hypothesized, it will not end up in the model. Thus, the analyst is aided in avoiding redundant variables, no matter their initial plausibility or usefulness in other settings. Furthermore, there is no *a priori* assumption about normality (or other parametric forms) of the distribution of variables, errors, or outcomes. It can be applied to non-numerical data such as ethnicity, gender, comorbidities, and the like. Both sampling error (the target of many statistical tests) and flaws in the model are reflected in the results so that the overall accuracy can be assessed.

A drawback to the approach is that it requires relatively large data sets of known outcomes to build an accurate model and a relatively large data set to ascertain the realm of applicability of the model. While this is true of virtually any study, the machine learning approach may be more sensitive to outliers or "bad" data. Nonetheless, if the baseline data set is large and known to be

representative of a population of interest, then the model can be applied with confidence to new cases from the population.

Once a model is constructed, it may be joined to a computer program that takes cases as input, feeds the values to the model, and outputs the prediction of the model. In this scenario, there is no requirement for human knowledge of the patterns or the specific details of the model. Human interaction is solely to provide input and utilize output. In brief, the program is a "black box" that answers the question, "What is this patient's likely outcome?"

Further insight is gained if one can look inside the box to determine the nature of the clusters of pre-operative patient characteristics that predict the outcome. To this end, tree-based, support vector, nearest neighbor, or rule extraction methods of machine learning are preferred for building the models. The machine learning methods used in this study, although based on such methods, employed an iterative refinement technique to maximize accuracy which made more difficult the elucidation of human-interpretable patterns. Work in this direction is continuing.

The models constructed exhibited high accuracy in classifying the cases available for training the model. However, there was not enough data to form a statistical assessment of the generality of the models within the universe of shoulder patients. These models would apply to any new patient who resembles a patient in the study group as long as there was no filtering or any known circumstance that would bias the composition of the study group.

Study inclusion criteria were any patient who underwent a total shoulder arthroplasty (anatomic or reverse) by the senior author from January 1, 2007 through December 31, 2015 for a primary diagnosis of glenohumeral osteoarthritis. Our initial query yielded 995 cases. Exclusion criteria were patients who had no preoperative CT scan (n=165) and those with insufficient clinical follow up (<2 years, n=358). The input data set consisted of 472 distinct

patients with glenohumeral osteoarthritis and underwent either a total shoulder arthroplasty or reverse shoulder arthroplasty. All patients had complete data for all independent variables as well as a post-operative ASES score which determined the predictive target. The patients were divided into three improvement tiers of approximately equal size designated as A (improvement ≤28 points, n= 137), B (28 < improvement ≤ 55 points, n=172), and C (improvement > 55, points n=163). The slight difference in size was allowed to provide somewhat sharper boundaries between the groups. The goal was to successfully predict the range in which a patient would fall at a follow-up interval of roughly two years or more. The actual follow-up range in the model was 21-99 months

B. Using Data Collected at different time points

Two approaches were taken to assess the possible influence of the follow-up interval on the post-op ASES score. One approach was based on defining a "two-year range" of 21 to 30 months, where it was assumed that all patients had reached an equivalent level of healing and would be comparable to previous studies. This range contained 300 patients. The second approach included the follow-up interval in months as an independent variable so that any effect would be explicitly part of the model. This approach utilized all 472 shoulders and made possible predictions for any time interval between 21 and 99 months.

C. Predictive Effectiveness of Variables

We conducted experiments to address the relative predictive efficacy of the variables for three follow-up regimes. Model 1 utilized all the explanatory variables; Model 2 omitted Walch

and Goutallier variables; Model 3 omitted the ASES individual questions and ASES overall score. The results are recorded in Tables **4** and **5**.

The models constructed exhibited high accuracy in classifying the cases available for training the model. However, as noted in the paper, there was not enough data to both build a strong model and to test it on enough independent cases to create confidence in its general applicability. We can say these models would apply to any new patient who resembles any patient in the study group, that the treatment regimens have been described, the practice was described, and that there was no filtering or any known circumstance that would bias the composition of the study group.

If the study cohort is small fraction of the universe of similar cases, as is true of this study, the binomial distribution can be used to provide additional insight about generality (otherwise use the hypergeometric distribution).We can think of the study group as equivalent to a sample drawn with replacement from the general population, and we can feel comfortable about the model if each pattern is based on about 8 or more instances found in the training set. The binomial distribution then permits the conclusion that, for our present sample size, 8 instances will be found with probability 0.98 for any pattern that occurs in more than 2% of the general population. This quantifies the notion that a sample might miss rare situations, by defining “might miss” and “rare” in probabilistic terms.

D. Associations between muscular and glenoid characteristics

Other information can be extracted from this data that was considered too far afield for inclusion in this paper, but which can be the subject of future reports. For instance, the independence of Walch classification from the Goutallier variables could also be studied from

this dataset. For this purpose, as indicated in Table 1, "Walch" is considered as a single variable which can assume any one of five values. Each muscle defines another variable which can assume any one of 5 values, reflecting the Goutallier scale of fatty infusion. Association can potentially. occur between items, between variables, or between groups of variables.

Item association occurs when a particular Walch class and a particular Goutallier value for a particular muscle are both found in the same patient with greater than chance frequency. This can be determined from a frequency table. Two variables are said to be associated when each value of one variable has an item association with a value of the second variable.

An association between variables is said to occur when there are item associations throughout the ranges of the variables. For numerical variables this can be quantified by the correlation coefficient; for categorical variables, the "Goodman-Kruskal Tau" (GKT) is an appropriate measure. GKT is based upon a statistical analysis of how the patient counts are distributed over the joint values of two variables compared to their separate distributions. Like the correlation coefficient, GKT varies between -1 and +1, with 0 indicating no association.

The association of Walch to the entire group of Goutallier variables can be addressed using the present modeling method. The accuracy with which the Walch classification of a patient can be predicted from all the patient's Goutallier variables, taken together, provides a measure of the independence of glenoid morphology and fatty infiltration of the muscles. If the Goutallier variables do not predict Walch value more accurately than chance, then there is no comprehensive relationship between glenoid morphology and fatty infiltration.

Appendix II

Operative Technique and Rehabilitation

All surgeries were performed by the senior author (M.A.F.). The deltopectoral approach with lesser tuberosity osteotomy was utilized in all anatomic total shoulder arthroplasties and subscapularis peel was utilized in all reverse shoulder arthroplasties. The implants used include the Foundation® anatomic shoulder system (DJO Surgical, Austin, Texas), the Turon™ modular anatomic shoulder system (DJO Surgical, Austin, Texas), the Reverse® Shoulder Prosthesis (RSP®; DJO Surgical, Austin, Texas), RSP® Monoblock (DJO Surgical, Austin, Texas), and AltiVate Reverse® (DJO Surgical, Austin, Texas). A standardized approach to the treatment of biconcave glenoids and glenoids with significant retroversion was performed in each case. The glenoid was gently reamed until a completely congruent surface was obtained, while removing as little bone as possible to preserve the subchondral plate; no attempt at version correction was performed with asymmetric reaming.

If the surgeon was in the process of performing anatomic total shoulder arthroplasty (TSA) and it was noted that the glenoid component would not sit flat against the prepared bone and rocking was present, the intra-operative decision was made to convert to reverse shoulder arthroplasty (RSA). Additionally, if during trialing it was noted that the shoulder did not balance well, as evidenced by poor centering of the humeral head relative to the glenoid while passively ranging the shoulder, inability to reduce the lesser tuberosity fragment to its native surface on the proximal humerus, or the ability to dislocate the humeral head posteriorly more than 50% when testing translation, the intra-operative decision was made to convert to RSA.

http://dx.doi.org/10.2106/JBJS.OA.20.00128

All patients were immobilized in a sling for 6 weeks, with pendulum exercises started on post-operative day 1. After 6 weeks, gentle active-assisted range of motion was initiated with progression to active range of motion as tolerated. Strengthening was started at 3 months post-operatively.