9-25-2020

PREDICTION OF CLASS III TREATMENT NEED AND SUCCESS

Andrew Levin
University of the Pacific Arthur A. Dugoni School of Dentistry, a_levin2@u.pacific.edu

James Chen
University of the Pacific Arthur A. Dugoni School of Dentistry, jchen8@pacific.edu

PREDICTION OF CLASS III TREATMENT NEED AND SUCCESS

by

Andrew Levin

A Thesis submitted to the

Graduate Orthodontic Program

In partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE IN

DENTISTRY

Arthur A. Dugoni School of Dentistry

University of the Pacific
Arthur A. Dugoni School of Dentistry
San Francisco, California

2020

PREDICTION OF CLASS III TREATMENT NEED AND SUCCESS

by

Andrew Levin

APPROVED BY:

Thesis Advisor: James Chen, DDS, Ph.D.

Thesis Committee: Heesoo Oh, DDS, MSD, Ph.D.

Program Director and Chair: Heesoo Oh, DDS, MSD, Ph.D.

Prediction of Class III Treatment Need and Success

ABSTRACT

Objective: The purpose of the present study is to develop prognostic models for surgical need and treatment success for class III malocclusions.

Material and Methods: This is a retrospective cohort study that evaluated treatment outcomes of consecutively treated patients at UCSF from Jan 1st 2007-Jan 1st 2012 and UoP from May 1st, 2014 – May 1st 2019. Receiver operator curves were used to develop prognostic models for surgical need and treatment success for class III malocclusions. Predictor variables were selected *a priori* (Class III-WITS, U1-PP, IMPA). The prognostic models were validated first using a UCSF validation cohort to show consistency with in one program, and then using consecutively treated patients at UoP from May 1st, 2014 – May 1st 2019 as a second validation group as an outside program.

Results: Derivation model for surgical need of class III malocclusion showed high sensitivity (81.8%); high specificity (94.3%), high positive predictive value (81.8%), high negative predictive value (94.3%), and the model correctly classified 91.3% of the subjects. UCSF validation model for surgical need of class III malocclusion showed moderate sensitivity (63.6%), high specificity (91.4%), high positive predictive value (70.0%), high negative predictive value (88.9%), and the model correctly classified 84.8% of the subjects. UoP validation model for surgical need of class III malocclusion showed moderate sensitivity (46.7%), high specificity (97.4%), high positive predictive value (77.8%), high negative predictive value (90.4%), and the model correctly classified 89.1% of the subjects. Derivation model for treatment success of Class III malocclusions showed moderate

sensitivity (46.7%); high specificity (85.2%), moderate positive predictive value (63.6%), high negative predictive value (74.2%), and the model correctly classified 71.4% of the subjects. UCSF validation model for treatment success of Class III malocclusions showed low sensitivity (35.0%), moderate specificity (69.6%), moderate positive predictive value (50.0%), moderate negative predictive value (55.2%), and the model correctly classified 53.5% of the subjects. UoP validation model for treatment success of Class III malocclusions showed low sensitivity (16.1%), high specificity (87.5%), moderate positive predictive value (41.7%), moderate negative predictive value (65.3%), and the model correctly classified 62.1% of the subjects.

Conclusion: WITS, U1-PP and IMPA were significant predictors of orthognathic surgical need in the derivation group, but only WITS predicted surgical need in the validation groups of Class III Malocclusions. Regarding treatment success, in the derivation group, only U1-PP was significantly associated with treatment success, while no variables were significantly associated with treatment success in the validation groups. Overall, the prognostic models developed in this study are more robust regarding predictions of Class III surgical need, as opposed to treatment success as defined by the ABO Cast and Radiograph examination.

INTRODUCTION:

Orthodontists have recognized early on the importance of the correct diagnosis, and historically numerous publications have provided new and improved methods to better our understanding of the existing malocclusion [1-6]. Identifying the underlying problems and making the correct diagnosis of a dentofacial discrepancy will lead to treatment plans with higher levels of success. Differences in diagnosis may cause the orthodontist to treat with orthodontics alone, versus a combined orthodontic and orthognathic surgical approach. Although there exist numerous cephalometric analyses to aid in decision making for the orthodontist, there is still a lack of specific consensus or recommendation regarding whether the treatment approach should involve orthodontic treatment only, or if it should be combined with orthognathic surgery.

Several studies have used cephalometric measurements to develop guidelines or so-called "norms" to help in treatment planning [5 6]. These studies have given the field a wealth of knowledge about normal cephalometric measurements, but often we as practitioners do not treat to specific numbers (Nielsen, 2007). These early studies also lack the ability to distinguish in which case it will be easier or more difficult to achieve an ideal final occlusion. Recent studies have begun to investigate and develop prediction models for the outcome of orthodontic treatment [7-10]. These studies focus particularly on the decision between orthodontic treatment only versus combined orthodontic and orthognathic surgical treatment. These studies have identified, primarily in class III malocclusions, some initial cephalometric measurements that can be important predictors of the need for orthognathic surgical correction. Stellzig-Eisenhauer *et al* used discriminate analysis to develop a

prognostic model for surgical need [9]. They identified WITS, M/M ratio, Lower gonial angle, and mandibular MLD as important variables for determining surgical need. While their study was elegantly designed, the model that was developed is difficult and cumbersome for orthodontists to calculate. Tseng *et al* addressed this drawback in their class III surgical prediction study [10]. The authors used receiver operator characteristic curves (ROC) to determine which cephalometric measurements offered the highest discrimination value for surgical need. Like the Stellzig-Eisenhauer *et al* study, Tseng identified WITS and M/M as important predictors. In addition to these two measurements, they also identified IMPA, gonial angle, overjet and overbite as other important measurements. What came out of their study was an easier to use prognostic model for surgical need.

Other researchers have studied soft tissue responses in Class III malocclusions. Lee, Yun-Sik et al. developed predictive models of soft tissue response after double jaw surgery in class III malocclusions [25]. The researchers tested both the ordinary least squares method and the partial least squares method of model generation, which are two different methods of developing prognostic models. It was found that the partial least squares method was more accurate and could predict soft tissue response to surgery better than some algorithms in commercially available software programs. The ordinary least squares method had a large problem with overfitting, as 226 predictor variables and 64 response variables from only 204 patients were entered into the prediction model. However, the partial least squares method allows variables to be combined, reducing the necessary predictor variables down to approximately 30. Unfortunately, the complexity of this model precludes direct usage by orthodontists. The prognostic models developed here are more

directed towards surgical planning software companies whom could apply the algorithms developed in this study to their products.

Different methods have been used to generate prognostic models by orthodontic researchers. Hodges et al. attempted to predict the lip response from four first premolar extractions [26]. They utilized stepwise multiple regression analyses to identify important variables, and then utilized a validation sample to test the performance of the prognostic model. The researchers found that upper and lower lip retraction could be predicted with moderately high levels of accuracy (62-81% of the variation in horizontal lip movements, and 67-76% of the vertical lip movements) using hard tissue treatment changes and pretreatment soft tissue characteristics. However, the derivation group and the validation group came from the same sample of subjects, wherein 119 subjects were used to derive the model, and 36 subjects were used to validate the model. This raises questions about external validity of the model, as it is not known how the model would perform given a different population of subjects.

Prognostic models are difficult to develop and to validate, as was highlighted in a review article by Fudalej *et al*. Fudalej *et al* identified the most significant drawback to prognostic model development is a lack of a validation group [11]. The tendency will be to use as many variables as possible to fit a prognostic model for surgical need; this often leads to what is called "over fitting" the model. Another weakness to some of these studies is how the subjects were selected to develop the prognostic model. In the Stellzig-Eisenhauer *et al* study, subjects did not go through treatment; only their initial records were used to determine if they should have surgery [9]. This design would make it difficult to translate into practice because the model predicts only what orthodontists possibly would do as

opposed to what actually occurred. The Tseng *et al* study design used patients who completed orthodontic treatment either non-surgically or surgically [10]. This design is an improvement over the Stellzig-Eisenhauer *et al* study, but the use of a case control design can be subject to selection bias in patient selection. Prognostic models are the next logical step in evidence-based decision making for orthodontic treatment.

The purpose of the present study is to develop prognostic models, using receiver-operating characteristic curves, for surgical need and treatment success as determined by the American Board of Orthodontics for class III malocclusion [12 13]. This study is also designed to address both the issues of over fitting and validation for prognostic models.

Study 1 compares two groups within a UCSF patient population, and study 2 compares a UCSF population with a University of the Pacific population.

Null Hypothesis: WITS, IMPA, and U1-PP do not predict surgical need or treatment success.

Alternative Hypothesis: WITS, IMPA, and U1-PP, either alone or in combination, are significantly associated with surgical need and/or treatment success.

Specific Aims:

- Develop prognostic models and utilize receiver operating characteristic curves
- Test if prognostic models can predict surgical need and/or treatment success by comparing ROC curves between derivation and validation groups
- Understand the association between the cephalometric variables WITS, IMPA, and U1-PP, and surgical need and treatment success

MATERIALS AND METHODS:

Study design and Subjects, Study 1:

This study was a retrospective cohort study with patients who completed treatment within the last 5 years (Jan 1st 2007-Jan 1st 2012) the University of California at San Francisco (UCSF). This study was conducted with the approval of the Committee on Human Research (11-08154). Inclusion criteria included the following: completed comprehensive orthodontic treatment at UCSF, complete records (before and after treatment). After initial selection patients were divided into groups based on their molar classification (Class III). Exclusion criteria included: craniofacial anomalies, and orthodontic treatment had not been initiated at UCSF. We initially identified 1100 potential subjects and when we applied our inclusion criteria we ended up with 120 potential subjects. Of these, 92 subjects were included in the class III treatment group, but 6 patients were missing final models, so could not be used in the treatment success portion of the study (TABLE 1). This sampling method of including all patients who fit the inclusion/exclusion criteria, as compared to a case control sampling method, enabled us to determine the direct prognostic value of any potential prediction model on surgical treatment need and treatment outcome. For this class III malocclusion population, we were able to randomly generate two distinct groups, one was used to develop the prognostic model, and the other served as a validation group.

Study design and Subjects, Study 2:

This study was a retrospective cohort study with patients who completed treatment within the last five years (Jan 1st 2007-Jan 1st 2012) at the University of California at San Francisco (UCSF), and the last five years (May 1st 2014-May 1st 2019) at the University of

the Pacific, Arthur A. Dugoni School of Dentistry (UoP). This study was conducted with the approval of the Committee on Human Research (20-45). We used the same derivation group as defined in Study 1. Inclusion criteria for the validation group included the following: completed comprehensive orthodontic treatment at UoP, complete records (before and after treatment). After initial selection patients were divided into groups based on their molar classification (Class III). Exclusion criteria included: craniofacial anomalies, and orthodontic treatment had not been initiated at UoP. We initially identified 528 potential subjects and when we applied our inclusion criteria, we ended up with 323 potential subjects. Of these, 92 were included in the Class III treatment group. 5 Subjects had to be further excluded from this group as they were missing a final set of models.

Common Materials and Methods for Both Studies

Cephalometric pre-treatment variables-

We used an empirical method initially to select the number of variables to be included in the prognostic models, which consisted of 1 predictor variable for roughly every 10 events (outcome). Each event or outcome was defined as the number of surgical patients successfully treated as defined by ABO standards. Based on this method we were restricted to 2 to 3 predictor variables. Based on previous studies, we set out to only focus on a select few cephalometric variables *a priori* (before developing the prognostic model). For our surgical prediction model we chose to use the WITS analysis, the upper incisor inclination to the palatal plane, and lower incisor inclination to the mandibular plane as predictors. The same variables were used for the treatment success prediction model. All cephalometric measurements were made using the Dolphin Imaging program (Chatsworth, CA).

Treatment success based on American board of orthodontics and PAR Index-

The models from before and after treatment were scored using the ABO Cast and Radiograph scoring system [12-15]. For our UCSF sample, two independent observers measured each subject's orthodontic dental models and the average between the two observers were used as the final score. The level of agreement between the two observers was assessed by the Bland-Altman method of repeatability (range of error) [16]. Treatment success was defined as a final occlusion that rated less than 20 points using the ABO Cast and Radiograph scoring system [12]. We measured the change in PAR Index before and after treatment for all subjects and found that a significant majority of subjects scored as improved or greatly improved. This lack of discrimination led us to use the ABO Cast and Radiograph scoring system as the main determinate of treatment success.

Statistical Analysis

All cephalometric and cast scoring data were analyzed using the Stata statistical package (College Station, TX). Both surgery and treatment success were set as dichotomous outcomes. Two different multivariate logistic regression models were used to analyze the association between the outcome variables (surgery or treatment success) and the predictor variables that were determined (*a priori*). For each multivariate logistic regression model, we performed three models checks in order to evaluate the possible "fit" of the model. First, we evaluated for influential points/leverage points, sensitivity analysis, for each model. This first level of model evaluation is used to identify specific outliers that significantly affect the results (positively or negatively), should these "influential points" significantly affect the regression model we would remove them from the analysis. Secondly, we applied the

Hosmer-Lemeshow Goodness of Fit Test to each model to determine whether the observed event rates matched expected rates in the model prediction. All models passed this second step in model checking. Lastly, we applied the Link-Test to determine if the fundamental form of this model is correct and determine if the predictors used in the model were correctly specified.

Once all the model checks were completed, separate multivariate logistic regression model was run to evaluate the association between outcome variables and predictor variables. Receiver operating characteristic curves (ROC) were developed for each model, where the graph is a plot of sensitivity versus 1-specificity of each model. The area under the curve (AUC) is a measure of the discriminatory ability of each model, with perfect discrimination as 1 and a complete lack of discrimination as 0.5. Discrimination is defined as the ability of a prognostic model to tell the difference between two possible outcomes (yes or no). The higher the discriminatory value the better the prognostic ability of the model. Using the same variables, multivariate logistic regression models were fit for both groups along with ROC curves. We evaluated the "fit" of the model by comparing the AUC between the derivation and validation groups. A well-fitted model would have roughly similar AUC's in both derivation and validation groups, with the validation group always being slightly lower than the derivation group.

Analytical Approach:

ROC Curves

Receiver operating characteristic curves (ROC curves) achieved original popularity in the diagnostic testing field. Utilizing sensitivity and specificity, the area under the ROC curve can be calculated, and can give an overall impression of the quality of a particular diagnostic

test. Additionally, ROC curves can be used to predict specific outcomes such as disease risk, or the likelihood that an orthodontic patient received surgery or not. In the present study, the area under the curve (AOC) of the ROC curve will be used to compare performance of a prognostic model between two groups. If the AOC's of the two groups are very close in value, then it can be concluded that the prognostic model performs equally well in both groups. However, if the AOC values differ greatly between the two groups, it is likely that the prognostic model generated is not generalizable between the two populations.

A drawback to utilizing ROC curves is that sensitivity and specificity values can be skewed based on prevalence of a specific type of outcome. For example, if 75% of all patients in a sample do not receive surgery, the ROC curve will be skewed towards a higher specificity value. As real life populations very rarely have equal prevalence of the two outcomes being studied (i.e. surgery vs non surgery), the quality of a given test may be over or underestimated.

Ultimately, provided one utilizes calibration and discrimination checks (such as comparing the Hosmer-Lemeshow statistic with the AOC of the prognostic models), and the models pass these checks, the utilization of ROC curves is a reasonable estimate for true performance of a prognostic model.

A priori Variable Selection

How one selects variables, as well as how many variables are selected, can have a profound influence on the outcomes of a study. Both inclusion and exclusion of specific independent variables can alter variance and induce confounding bias into other coefficients. Since sample sizes in orthodontic literature are often small, there is a tendency, especially in cephalometric research, to study far too many variables relative to the sample size. To avoid

problems related to a shotgun approach to variable selection, we defined a set of three variables a priori, based on background knoweldge from other studies in the same field. By selecting variables before seeing our results, we reduce the bias that may have resulted from less controlled variable selection.

RESULTS

Study 1 RESULTS:

Subject demographics-

For the class III treatment group there were no significant differences between the derivation group and the validation group (Table 2).

Baseline Cephalometric measurements-

There were no significant cephalometric differences pre-treatment between the derivation and the validation groups.

Prediction model for surgical treatment:

We had a sufficient number of class III patients to allow us to develop a prediction model and create both a derivation group and a validation group. The subjects were randomly assigned to each group and the same multivariate logistic model (WITS, U1-PP, IMPA) was applied to each group. The AUC for the derivation sets of patients was 0.9532 and for the validation set was 0.9091 (Figure 1). There was no statistical difference between these two groups, which indicates that the model fits equally well for both groups. After establishing the discrimination ability and the fit of the multivariate logistic model, we determined the sensitivity (81.8%), specificity (94.3%), positive predictive value (81.8%), negative predictive value (94.3%), and the model correctly classified 91.3% of the subjects

in the derivation group (Figure 2). In the validation group, we determined the sensitivity (63.6%), the specificity (91.4%), the positive predictive value (70.0%), the negative predictive value (88.9%), and the model correctly classified 84.8% of the subjects in this group. In the derivation group, WITS (p=.014), IMPA (p=.045), and U1-PP (p=.036) were all significantly associated with surgical need, while in the validation group, only the WITS value was significantly associated with surgical need (Figure 3).

Prediction of treatment success-

Subjects were randomly assigned to each group and the same multivariate logistic model (WITS, U1-PP, IMPA) was applied to each group. 7 subjects did not have final dental models and were therefore excluded from the study leaving the total population for treatment success to be 85. The AUC for the derivation set of patients was 0.7778 and for the validation set was 0.6891 (Figure 4). There was no statistical difference between these two groups, and this suggests that the model fits equally well for both groups, but the low value for each AUC suggests poor discrimination ability. We also determined the sensitivity (46.7%); specificity (87.5%), positive predictive value (63.6%), negative predictive value (74.2%), and the model correctly classified 71.4% of the subjects in the derivation group (Figure 5). In the validation group, we determined the sensitivity (35.0%), the specificity (69.6%), the positive predictive value (50.0%), the negative predictive value (55.2%), and the model correctly classified 53.5% of the patients in this group. In the derivation group, only U1-PP (p=.03) was significantly associated with treatment success, while in the validation group no variables were significantly associated with treatment success. (Fig 6).

Summary:

In study 1, it was found that WITS was a strong predictor of surgical need in both derivation and validation groups, and the prognostic model performed reasonably well. In terms of treatment success, only U1-PP was associated with treatment success in the derivation group, while no variables were associated with treatment success in the validation group. The prognostic model did not perform well in terms of predicting treatment success. In the following study, we will examine the fit of the prognostic models in the patient population of a different institution, utilizing UoP patients as the validation group.

Study 2 RESULTS:

Baseline Subject demographic and cephalometric measurements-

In our class III treatment group, the only cephalometric characteristic that differed at baseline was the IMPA. The derivation group showed lower incisors that were approximately three degrees more upright pre-treatment. Additionally, the derivation group had significantly fewer female patients than the validation group. (Table 2)

Prediction model for surgical need:

The same multivariate logistic model (WITS, U1-PP, IMPA) was applied to both the derivation (UCSF) and validation (UoP) groups. The AUC for the derivation set of patients was 0.9532 and for the validation set was 0.8848 (Figure 7). There was no statistical difference between these two groups, which indicates that the model fits equally well for both groups. After establishing the discrimination ability and the fit of the multivariate

logistic model, we determined the sensitivity (81.8%), specificity (94.3%), positive predictive value (81.8%), negative predictive value (94.3%), and the model correctly classified 91.3% of the subjects in the derivation group (Figure 8). In the validation group, we determined the sensitivity (46.7%), the specificity (97.4%), the positive predictive value (77.8%), the negative predictive value (90.4%), and the model correctly classified 89.1% of the subjects. In the derivation group, WITS (p=.014), IMPA (p=.045), and U1-PP (p=.036) were all significantly associated with surgical need, while in the validation group, only the WITS (p=.001) value was significantly associated with surgical need (Figure 9).

Prediction of treatment success-

The same multivariate logistic model (WITS, U1-PP, IMPA) was applied to the derivation (UCSF) and validation (UoP) groups. 7 UCSF subjects and 5 UoP subjects did not have final dental models and were therefore excluded from this portion of the study, leaving a total derivation population for treatment success of 39 and a total validation population of 87. The AUC for the derivation set of patients was 0.7778 and for the validation set was 0.6457 (Figure 10). There was no statistical difference between the two groups, which suggests that the model fits equally well for both groups, but the low values for each AUC suggest poor discrimination ability. For the derivation group, we determined the sensitivity (40.0%); specificity (85.2%), positive predictive value (60.0%), negative predictive value (71.9%), and the model correctly classified 69.1% of the subjects. For the validation group, we determined the sensitivity (16.13%), specificity (87.50%), positive predictive value (41.67%), negative predictive value (65.33%), and the model correctly classified 62.07% of the subjects (Figure 11). In the derivation group, only U1-PP (p=.03) was significantly

associated with treatment success, while in the validation group no variables were significantly associated with treatment success. (Fig 12).

DISCUSSION

The results from this study and other similar research endeavors offer small steps towards improving evidence based practice and decision making. Elucidating this important information can provide the ability for practitioners to theoretically make better clinical choices, and to guide and advise their patients regarding the paths to the best possible outcomes. Our results confirmed the value of WITS as a major predictor for surgical need for Class III malocclusions across both derivation and validation groups. This result was similar to other models described in the literature [7 9 10]. One new predictor that was not evaluated previously and that our study showed to be valuable was the upper incisor inclination relative to the palatal plane. Our results showed that when the upper incisor was increasingly proclined, indicating an increased level of dental compensation, the need for surgical correction was greater. However, this is was only true in the derivation group, as the only significantly associated variable for surgical need in the validation groups was the WITS value. Separately, our study attempted to identify prognostic cephalometric variables associated with treatment success as determined by the American Board of Orthodontics. Our study found that in the derivation group, the U1-PP value was significantly associated with treatment success, but no variables were significant in the validation groups. Despite the lack of statistical significance in the validation group, there was a mild association between an increased WITS value and an ABO passing score in the UoP population, indicating

that cases with a milder skeletal discrepancy may result in better treatment success as defined by the ABO Cast and Radiograph Examination.

Orthodontic correction for class III malocclusions includes several options: orthodontics only, orthodontics with orthopedics at the appropriate age, and orthodontics combined with orthognathic surgery [17-19]. The difference between each treatment modality depends on several factors, such as age of the patient, degree of skeletal disharmony, and any dento-alveolar compensation. Until recently, there has been very little evidence to aid orthodontists in making this decision. Some of the earliest work on predicting surgical need has come from the field of craniofacial anomalies, where unilateral and bilateral cleft lip and palate patients were used to predict the need for orthognathic surgical correction of their craniofacial related malocclusions [20 21]. The findings from these studies, although well done, cannot be applied directly to the more typical orthodontic patient. Stellzig-Eisenhauer *et al.* used general orthodontic patients for their study and developed a prognostic model via discriminant analysis [9]. They identified WITS, maxilla/mandible ratio, lower gonial angle, and anterior cranial base length as useful predictors and developed a mathematical formula using these variables to determine the cut-off for non-surgical versus surgical correction. The authors recently published a follow-up article adding mandibular midline deviation and saddle angle (SN-Ar) to the previous four cephalometric measurements, while removing anterior cranial base length from the original model [7]. In their first study, four cephalometric variables, the identification accuracy was 86.4%, compared to the improved 92.7% accuracy of the second study with six cephalometric measurements [7]. The authors cautioned that the second, more comprehensive model was not checked with another separate set of subjects, which is

important in order to validate the effectiveness of the model. Moreover, in both studies the authors did not evaluate the discrimination ability of their model. Tseng *et al* in 2011 sought to identify a new prognostic model for class III surgical prediction using ROC curves and AUCs [10]. The authors identified 6 cephalometric variables that fit the best with their case-control study: overjet, maxilla/mandible ratio, IMPA, overbite, gonial angle, and WITS [10]. Although, the prognostic model Tseng *et al* developed was highly discriminatory, they also did not validate their model with another set of subjects. A recent review article by Fudalaj *et al* highlighted this distinct caveat to the aforementioned studies [11]. The major issue with developing a prognostic model without testing the model on a validation set of different subjects is the possibility of over-fitting. Over-fitting often occurs because the authors desire to achieve the best fitting model to best explain their subject population. We illustrated this important difference when we applied the prognostic model Tseng *et al* had developed to our derivation set of subjects and found that their model under-performed compared to their original article (Data not shown).

The results of our study showed only a moderate level of prognostic ability for predicting surgical need, and a low level of prognostic ability for predicting treatment success. Some of the factors that could have contributed to a reduced prognostic ability are the inappropriate choice of cephalometric predictors, limited sample size, and individuality of treatment. The selection of which cephalometric measurements to input into our prognostic model are dependent variables that we felt would best predict the outcome based on our clinical judgment. Because of the numerous cephalometric measurements available, this task is very difficult. For class III subjects, we wanted to incorporate a combination of skeletal and dental measurements. Both the combination of small sample size and limited

number of cephalometric variables may have affected the discriminatory ability of both prognostic models. The cephalometric variables we chose are measurements that are commonly used in orthodontic diagnosis, but perhaps the moderate level of prognostic ability was not due to misspecification of the predictor variable but rather an inadequate number of predictors. Increasing the sample size will not only improve the power of our study, but also give us the opportunity to include more cephalometric variables *a priori* into the models. The lack of standardized treatment mechanics, as well as the individuality of treatment plans are potential avenues of great variability in our results. Additionally, our study was retrospective in nature and included patients of record from two different orthodontic clinics, and our inclusion/exclusion criteria did not stratify by resident or attending faculty member. This variation in the experience, skill, and philosophy of the treating orthodontist represents undoubtedly resulted in increased variability. Without standardizing the treatment for each patient, a significant amount of variation can creep into the study, which could strongly confound the results. The gold standard for clinical studies are randomized clinical trials (RCT) where all aspects of the study can be controlled, thereby limiting variations. However, an RCT to develop a prognostic model would be expensive and difficult to conduct. Our current study design may not be able to control for all the variation, but it does offer a reasonably effective method to sample the population while still addressing prognostic model development.

There is still an important question of how true treatment success is defined, and who decides. In this study, we defined treatment success as a passing score (<20) based on the American Board of Orthodontics Cast and Radiograph scoring system. In the ABO scoring system, eight different factors are measured to evaluate clinical expertise: 1. Alignment, 2.

Marginal ridge height, 3. Buccolingual inclination, 4. Occlusal relationships, 5. Occlusal contacts, 6. Overjet, 7. Interproximal contacts, and 8. Root angulation. Clearly, the ABO scoring system does not consider any soft tissue measurements, nor does it assess smile esthetics. Brian J. Schabel *et al* reviewed the relationship between post-treatment smile esthetics and the ABO Objective Grading System [22]. Extraoral smiling photographs of 48 patients were taken and were then rated by both orthodontists and by the parents of orthodontic patients. Extremely weak positive and negative relationships were found between all factors of the ABO scoring system and perceived smile attractiveness. Additionally, neither total scores nor individual components of the ABO scoring system predicted the attractiveness of smiles. In another study, Espeland and Stenvik found that most laypeople do not use occlusal outcomes to define treatment success, but rather utilize the attractiveness of the smile [23]. In our study, we did not employ an index to evaluate smile esthetics, but rather focused on using the ABO scoring system. It may be possible that even though a patient may have not passed via the ABO scoring system, the result was highly esthetic regarding smile attractiveness, soft tissue features, and facial profile. A future direction could be to examine how cephalometric characteristics predict successful treatment as defined by smile and facial esthetics, rated by orthodontists, patients, or both.

CONCLUSIONS

- For Class III malocclusions, WITS, IMPA and upper incisor inclination relative to the palatal plane were strong predictors of surgical need in the derivation group, but only WITS predicted surgical need in the validation groups

- For Class III malocclusions, U1-PP was significantly associated with treatment success in the derivation group, but no variables studied were significantly associated with treatment success in the validation groups.
- The prognostic models developed are of moderate utility to predict surgical need
- The prognostic models developed are of low utility to predict treatment success as defined by the ABO scoring system

REFERENCES

1. Dimitriades AG, Sassouni V, Sotereanos GC. [Skeletal Class II deep bite or open bite correction with surgical orthodontic management]. Odontostomatol Proodos 1975;**29**(6):339-55
2. Dimitriadis AG, Sassouni V, Sotereanos GC. [Skeletal Class II and Class III deep-bite or open-bite correction with surgical-orthodontic management (genioplasty)]. Odontostomatol Proodos 1974;**28**(5):261-73
3. Holdaway RA. A soft-tissue cephalometric analysis and its use in orthodontic treatment planning. Part I. Am J Orthod 1983;**84**(1):1-28
4. Holdaway RA. A soft-tissue cephalometric analysis and its use in orthodontic treatment planning. Part II. Am J Orthod 1984;**85**(4):279-93
5. Steiner CC. [Importance of cephalometry in orthodontic treatment]. Inf Orthod Kieferorthop 1969;**1**(2):3-12 passim
6. Tweed CH. The Frankfort-mandibular plane angle in orthodontic diagnosis, classification, treatment planning, and prognosis. Am J Orthod Oral Surg 1946;**32**:175-230
7. Kochel J, Emmerich S, Meyer-Marcotty P, Stellzig-Eisenhauer A. New model for surgical and nonsurgical therapy in adults with Class III malocclusion. American Journal of Orthodontics and Dentofacial Orthopedics 2011;**139**(2):e165-e74 doi: http://dx.doi.org/10.1016/j.ajodo.2010.09.024%5Bpublished Online First: Epub Date]|.
8. Schuster G, Lux CJ, Stellzig-Eisenhauer A. Children with class III malocclusion: development of multivariate statistical models to predict future need for orthognathic surgery. Angle Orthod 2003;**73**(2):136-45 doi: 10.1043/0003-3219(2003)73<136:CWCIMD>2.0.CO;2[published Online First: Epub Date]|.
9. Stellzig-Eisenhauer A, Lux CJ, Schuster G. Treatment decision in adult patients with Class III malocclusion: orthodontic therapy or orthognathic surgery? Am J Orthod Dentofacial Orthop 2002;**122**(1):27-37; discussion 37-8
10. Tseng Y-C, Pan C-Y, Chou S-T, et al. Treatment of adult Class III malocclusions with orthodontic therapy or orthognathic surgery: Receiver operating characteristic analysis. American Journal of Orthodontics and Dentofacial Orthopedics

2011;**139**(5):e485-e93 doi: http://dx.doi.org/10.1016/j.ajodo.2010.12.014%5Bpublished Online First: Epub Date]|.

11. Fudalej P, Dragan M, Wedrychowska-Szulc B. Prediction of the outcome of orthodontic treatment of Class III malocclusions,Äîa systematic review. The European Journal of Orthodontics 2011;**33**(2):190-97 doi: 10.1093/ejo/cjq052[published Online First: Epub Date]|.
12. Casko JS, Vaden JL, Kokich VG, et al. Objective grading system for dental casts and panoramic radiographs. American Board of Orthodontics. Am J Orthod Dentofacial Orthop 1998;**114**(5):589-99
13. Greco PM, English JD, Briss BS, et al. Posttreatment tooth movement: for better or for worse. Am J Orthod Dentofacial Orthop 2010;**138**(5):552-8 doi: 10.1016/j.ajodo.2010.06.002[published Online First: Epub Date]|.
14. Richmond S, Shaw WC, O'Brien KD, et al. The development of the PAR Index (Peer Assessment Rating): reliability and validity. Eur J Orthod 1992;**14**(2):125-39
15. Richmond S, Shaw WC, Roberts CT, Andrews M. The PAR Index (Peer Assessment Rating): methods to determine outcome of orthodontic treatment in terms of improvement and standards. Eur J Orthod 1992;**14**(3):180-7
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;**1**(8476):307-10
17. Campbell PM. The dilemma of Class III treatment. Early or late? Angle Orthod 1983;**53**(3):175-91 doi: 10.1043/0003-3219(1983)053<0175:TDOCIT>2.0.CO;2[published Online First: Epub Date]|.
18. Chang HP, Tseng YC, Chang HF. Treatment of mandibular prognathism. J Formos Med Assoc 2006;**105**(10):781-90 doi: 10.1016/S0929-6646(09)60264-3[published Online First: Epub Date]|.
19. Rabie AB, Wong RW, Min GU. Treatment in Borderline Class III Malocclusion: Orthodontic Camouflage (Extraction) Versus Orthognathic Surgery. Open Dent J 2008;**2**:38-48 doi: 10.2174/1874210600802010038[published Online First: Epub Date]|.
20. Atack NE, Hathorn IS, Semb G, Dowell T, Sandy JR. A new index for assessing surgical outcome in unilateral cleft lip and palate subjects aged five: reproducibility and

validity. Cleft Palate Craniofac J 1997;**34**(3):242-6 doi: 10.1597/1545-1569(1997)034<0242:ANIFAS>2.3.CO;2[published Online First: Epub Date]|.

21. Mars M, Plint DA, Houston WJ, Bergland O, Semb G. The Goslon Yardstick: a new system of assessing dental arch relationships in children with unilateral clefts of the lip and palate. Cleft Palate J 1987;**24**(4):314-22
22. Schabel, Brian J., et al. "The relationship between posttreatment smile esthetics and the ABO Objective Grading System." The Angle Orthodontist 78.4 (2008): 579-584.
23. Espeland, Lisen V., and Arild Stenvik. "Perception of personal dental appearance in young adults: relationship between occlusion, awareness, and satisfaction." *American Journal of Orthodontics and Dentofacial Orthopedics* 100.3 (1991): 234-241.
24. Dolatabadi, Nora. "Validity and Reliability of Computerized Dental Landmarks." Thesis Manuscript, 2017.
25. Lee, Y. S., Suh, H. Y., Lee, S. J., & Donatelli, R. E. (2014). A more accurate soft-tissue prediction model for Class III 2-jaw surgeries. *American Journal of Orthodontics and Dentofacial Orthopedics*, *146*(6), 724-733.
26. Hodges, A., Rossouw, P. E., Campbell, P. M., Boley, J. C., Alexander, R. A., & Buschang, P. H. (2009). Prediction of lip response to four first premolar extractions in white female adolescents and adults. *The Angle Orthodontist*, *79*(3), 413-421.