Troiano Giuseppe (Orcid ID: 0000-0001-5647-4414)
NIBALI LUIGI (Orcid ID: 0000-0002-7750-5010)
Petsos Hari (Orcid ID: 0000-0002-8901-8017)
Eickholz Peter (Orcid ID: 0000-0002-1655-8055)
Pasquale Santamaria (Orcid ID: 0000-0003-4102-1759)
Jiao Jian (Orcid ID: 0000-0003-2466-2116)
Shi Shu‐wen (Orcid ID: 0000-0002-1692-0325)
Wang Hom‐Lay (Orcid ID: 0000-0003-4238-1799)

Development and international validation of logistic regression and machine-learning models for the prediction of 10-years molar loss

Giuseppe Troiano[1], Luigi Nibali[2], Hari Petsos[3], Peter Eickholz[3], Muhammad H. A. Saleh[4], Pasquale Santamaria[2], Jao Jian[5], Shuwen Shi[5], Huanxin Meng[5], Khrystyna Zhurakivska[1], Hom-Lay Wang[4], Andrea Ravidà[4,6]

[1] Department of clinical and experimental medicine, University of Foggia, Foggia 71122, Italy

[2] Periodontology Unit, Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral & Craniofacial Sciences, King's College London, London, UK.

[3] Department of Periodontology, Center for Dentistry and Oral Medicine (Carolinum), Johann Wolfgang Goethe-University Frankfurt/Main, Frankfurt/Main, Germany.

[4]Department of Periodontics & Oral Medicine, University of Michigan School of Dentistry, Ann Arbor, MI, USA.

[5]Department of Periodontology, National Engineering Laboratory for Digital and Material Technology of Stomatology, Beijing Key Laboratory of Digital Stomatology, Peking University School and Hospital of Stomatology, Beijing, China.

[6] Department of Periodontics & Oral Medicine, University of Pittsburg, Pittsburgh, PA, USA.

Corresponding author:

Andrea Ravidà: DDS, MS

Department of Periodontics & Oral Medicine, University of Pittsburg, Pittsburg, PA, USA.

E-mail address: Ravidandrea@pitt.edu

Word count: 3502

Tables and figures: tables 5 and figures 1

Supplementary material: 2

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jcpe.13739

Running title: Machine-learning algorithms for molars prediction

Key Words: Furcation involvement, periodontitis, tooth loss, Artificial Intelligence.

Scientific rationale for study: Testing developed models on different populations (external validation) is highly recommended to increase their generalizability.

Principal findings: An ensembled model combining logistic regression and neural network models showed the best performance for prediction of molar loss at 10-year follow-up.

Practical implications: The algorithm was made freely available to clinicians for a widespread use in clinical practice.

Abstract

Aim: To develop and validate logistic regression and artificial-intelligence based models for prognostic prediction of molar survival in periodontally-affected patients.

Material and Methods: Clinical and radiographic data from 4 different centers across 3 continents (2 in Europe, 1 in USA, and 1 in China) including 515 patients and 3157 molars were collected and used to train and test different types of machine-learning algorithms for their prognostic ability of molars over 10 years. The following models were trained: Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, Artificial Neural Network, Gradient Boosting and Naive Bayes. In addition, different models were aggregated by means of Ensembled Stacking method. The primary outcome of the study was related to the prediction of overall molar loss in patients after active periodontal treatment.

Results: The general performance in the external validation settings (aggregating 3 cohorts) revealed that the Ensembled model that combined Neural Network and Logistic Regression showed the best performance among the different models for the prediction of overall molar loss with an AUC = 0.726. The Neural Network showed the best AUC = 0.724 for the prediction of periodontitis-related molar loss. In addition, the Ensembled model showed the best calibration performance.

Conclusion Through a multicenter collaboration, both prognostic models for the prediction of molar loss were developed and externally validated. An Ensembled model showed the best performance in terms of both discrimination and validation, it is made freely available to clinicians for widespread use in clinical practice.

Conflict of interest

All authors declare that they have no conflict of interest related to this manuscript.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

Authors contributions

	GT	LN	HP	PE	MS	PS	JJ	SS	HM	KZ	HLW	AR
Study conception and design	x	x			x							x
Data collection	x	x	x		x	x	x	x	x	x	x	x
Analysis and interpretation of the data	x									x		x
Drafting of the manuscript	x	x	x	x	x	x	x	x	x		x	

Introduction

The prognostic prediction is a forecast of the probable course and outcome of a disease, especially as it pertains to the chances of recovery. Making predictions about the short- and long-term retention of teeth is a necessary step for accurate decision making and treatment planning. Prediction-based decision-making has shown that accurate prediction could lead to not only a reduction in total treatment cost, but also superior and less invasive therapeutics results (Schwendicke, Stolpe, & Graetz, 2017). Due to its exciting ability to predict events, machine learning is a developing field

applied throughout a large variety of sectors including periodontology (Harrison & Sidey-Gibbons, 2021; Mohammad-Rahimi et al., 2022; Sidey-Gibbons & Sidey-Gibbons, 2019). In periodontics, the introduction of prediction models has opened the doors to the practice of personalized medicine, with one of its goals being to improve the success of treatment outcomes by providing the best approach for each case thus decreasing the rate of future tooth loss. Prognostic tools are usually based on patient-level factors (age, diabetes status and smoking habit, periodontitis staging and grading) (Ravida et al., 2020; Saleh et al., 2022; Schwendicke et al., 2018) and/or tooth-level factors (e.g. clinical attachment level, probing pocket depth and furcation involvement) (Saleh et al., 2021; Shi et al., 2020). A variety of parameters such as the model complexity (e.g. regression models or machine-learning models); sample size; imbalanced class size (e.g. the majority of tooth loss studies, since the vast majority of teeth being retained rather than lost) and prediction period (short, medium or long term) need to be taken into consideration when a prediction model is being created and/or validated (Krois et al., 2019). Furthermore, a key aspect to consider is the training and testing strategy used to validate the model utilized. When validating a model, there are generally two types of validation that can be performed: internal or external. Internal validations (also called in-sample performance) are trained and tested on the same database, while external validations are tested on populations other than those used for training. The developed models are more prone to overperform if they are created and validated on the cohort they are built on (internal validation) because, naturally, the validity of a prediction model may be dependent on a particular population or a specific socioeconomic status. On the other hand, external validations increase the generalizability of prediction models, providing potential for worldwide use. In the dental field, there is often not even an internal validation performed (Du, Bo, Kapellas, & Peres, 2018). To the best of our knowledge there is no study in the periodontal literature simultaneously developing and externally validating a model for the prediction of molar retention/loss.

Machine learning algorithms exploit the computational capacity of modern computer systems to find complex patterns within data (Bates, Saria, Ohno-Machado, Shah, & Escobar, 2014). They may be

trained on population-specific datasets and be used for multiple applications including risk stratification, diagnosis and survival predictions (Kantarjian & Yu, 2015; Ngiam & Khor, 2019). However, it is still unknown whether in the prediction of tooth loss in periodontal patients, such algorithms can overcome the performance of models developed using classical approaches, such as logistic regression (Christodoulou et al., 2019).

Hence the goal of this multicenter (4 cohorts) study was to develop and externally validate both classical logistic regression and artificial-intelligence-based prognostic models capable of predicting 10-year molar loss in periodontitis patients. This tool could help clinicians in everyday practice choose whether to retain or extract a molar based on its chances of survival.

Materials and Methods

This study was performed in accordance with the Transparent Reporting of a multivariable prediction model for Individual prognosis or Diagnosis (TRIPOD): the TRIPOD statement (Collins, Reitsma, Altman, & Moons, 2015). In addition, the ongoing guidance for the development of TRIPOD-AI was also taken into consideration (Collins et al., 2021). This study is both a development and external validation study based on the retrospective data from 4 different universities (2 in Europe, 1 in USA, and 1 in China). The statistical unit of this study were molars in periodontitis affected patients, while the primary outcome was overall molar loss (MLO), intended as teeth extracted for any reasons, after 10 years follow-up from the end of active periodontal treatment (the beginning of the supportive periodontal care (SPC)). The present paper includes data from 4 studies which had previously received ethical approvals by their local Ethics Committees: 1) Institutional Review Board for Human Studies of the Medical Faculty of the Johann Wolfgang Goethe-University approval no. 61/15); 2) University of Michigan, School of Dentistry, Institutional Review Board for Human Studies (HUM00157260); 3) Peking University School and

Hospital of Stomatology (approval number: PKUSSIRB-201310066); 4) and the NRES committee in London (approved as service evaluation, protocol number 14 LO 0629).

Cohorts and Predictors

Models were developed on a cohort of 1475 molars in 222 patients treated at the Ann Arbor University of Michigan in the period between 1966 and 2020. Performance of models in external validation settings was assessed on three other different cohorts, including one cohort of 404 molars in 65 patients treated in a private practice setting at London in the UK, a second cohort of 597 molars in 97 patients treated at University of Frankfurt in Germany and a third cohort of 681 molars in 131 patients treated at the University of Beijing in China; resulting in a total number of 3157 molars included in both development and external validation cohorts. A complete case analysis was performed on three databases (Michigan, Frankfurt, and Beijing), while 7.1% of missing data were present in the London database and were handled by using the most frequent class for the specific variables.

The following predictors were available and selected for the development of a prognostic model in this study, all variables were assessed at baseline:

- Age of the patient,
- Sex of the patient,
- Horizontal furcation involvement (0/1/2/3) (Hamp, Nyman, & Lindhe, 1975),
- Smoking habits (Active Smoker/ Former Smoker / non-Smoker)
- Radiographic Bone Loss (<15%, 15-33%, ≥ 33%),
- Probing Depth (PD),
- Clinical Attachment Level (CAL),
- Mobility (Lindhe & Nyman, 1977) (0/1/2/3),

- Abutment tooth for a crown/bridge (No/Yes).

All the predictors were available in the database. Self-reported smoking status was used in all the four databases [non-smokers (never smoked), former smokers (stopped smoking ≥5 years ago) and active smokers (stopped smoking <5 years ago or currently smoking) (Lang & Tonetti, 2003). Age of the patient at the end of active periodontal treatment, PD and CAL were included in models as continuous variables and were not categorized.

Follow-up and Outcome

A fixed time point at 10-years of follow-up after the end of active periodontal treatment was set for the assessment of the outcome variable. Analysis was performed having MLO as primary outcome, however the performance of the model was also assessed on periodontitis-related molar loss (MLP). Focusing on maintenance, patients who underwent an average of at least 1 maintenance session per year during the 10 years follow-up were considered as compliant.

Statistical Analysis and Machine Learning

Due to the low event rate of MLO and MLP at 10 years leading to an imbalanced dataset, Synthetic Minority Oversampling Technique (SMOTE) methods was applied to the Michigan (development cohort) database using the R software (Blagus & Lusa, 2013). Subsequently, data were loaded into the Orange software (https://orangedatamining.com) and machine learning analyses were carried out. A Ranking of the included variables was then carried out by applying different ranking methods, including Info gain, Gain Ratio, Gini Index, ANOVA, X2, ReliefF, Principal Component Analysis (PCA) and Fast Correlation Based Filter (FCBF). Subsequently various machine learning models were applied using the clinical predictors including Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest, Artificial Neural Network, Gradient Boosting and Naive Bayes. In addition, different models were aggregated by means of Ensembled Stacking method to assess possible combined models. Cross-validation 10-folds was applied on models developed on

the Michigan database after oversampling by means of the SMOTE technique (training set) as an internal validation method. External validation was then performed by making predictions for the occurrence of both MLO and MLP on the other 3 cohorts. The predictive performance of these models was assessed by calculating the area under the ROC curve, sensitivity, and specificity between predicted and actual events. Calibration was assessed by analyzing slope and intercept of calibration plots, the command *pmcalplot* on the STATA 16.1 software (StataCorp LLC, USA) was used for this analysis. All the statistical analyses were performed by one author (GT) who was not blinded on outcomes and predictors. Third molars were excluded from the analysis.

Results

Predictors, Outcome Occurrence and Event Rate

More details of frequencies and average of predictor values in the included cohorts are provided in *Table 1.* A total of 3157 teeth were included. 77 out of 1412 molars (5.22%) were lost for any reason at 10-years follow-up in the development cohort (Michigan); while 135 out of 681 (19.8%) in the Beijing cohort, 59 out of 597 (9.88%) in the Frankfurt cohort and 31 out of 404 (7.67%) in the London cohort were lost for any reason. An aggregated total number of 302 molars were lost for any reason in the cohorts. Of these, 218 (218/302, 72.18%) were lost due to periodontal reasons. The level of compliance differed significantly (chi-squared p-value < 0.001) among the different cohorts. A full compliance (100%) was achieved in the Michigan cohort, while 89.10% in the London Cohort, 67.50% in the Frankfurt cohort and 1.62% in the Beijing cohort.

The rapport between the number of events and the number of predictors for the developed cohort was 8.55% (77 events / 9 predictors). The best performance of machine learning models was obtained including all 9 predictors. In fact a weaker performance was detected in external validation settings after applying different feature selection methods (Table 2).

Model Development and internal validation

Different models were developed by combining both feature ranking methods and machine learning algorithms and were initially internally validated by means of cross-validation 10-folds. Results of the model development phase when all the predictors were included was promising with all the models showing AUC values over 0.70. In particular: Naïve Bayes showed AUC = 0.969, Random Forest AUC = 0.929, Gradient Boosting AUC = 0.924, K-Nearest Neighbors AUC = 0.924, Logistic Regression AUC = 0.787, Neural Network AUC = 0.757 and Support Vector Machine AUC = 0.755. In addition, an Ensembled model by combining Neural Network and Logistic regression showed an AUC = 0.759. Due to the possible presence of overfitting, the performance was then assessed in external validation settings by applying prediction on the other 3 cohorts of patients.

Models' discrimination in external validation setting

The discriminative performance on the aggregate data for the 3 cohorts included in external validation revealed that the Ensembled model showed the best discriminative performance with an AUC = 0.726, followed by Neural Network AUC = 0.724, Naïve Bayes AUC = 0.695, Logistic Regression AUC = 0.647, Support Vector Machine AUC = 0.626, Random Forest AUC = 0.590, K-Nearest Neighbors = 0.569 and Gradient Boosting AUC = 0.659 (Table 3). The same workflow was also applied for the MLP as outcome (taking into consideration molar lost only for periodontal reasons). As shown on Table 3, the average performance resulted to be equal for Neural Network and the Ensembled model with an AUC = 0.702, followed by Random Forest AUC = 0.683, Naïve Bayes AUC = 0.649, Logistic Regression AUC = 0.611, K-Nearest Neighbor AUC = 0.565, Gradient Boosting AUC = 0.527 and Support Vector Machine AUC = 0.512. The performance of both Ensembled and Neural Network models was the most stable when applied on the different 3 cohorts used for external validation with their performance never lower than 0.70 in AUC. More details about the performance of the validated models and information about sensitivity and specificity with difference among single cohorts is provided in Table 4 and 5. Another important

metrics to take into consideration according to TRIPOD is model calibration, which evaluates the degree to which numerical predictions are too high or too low compared to the observed outcome. As already reported for models developed on a database applying SMOTE, the calibration is usually not perfect (Dhiman et al., 2022). This was also the case for many of the models developed and validated in this study, however the Ensembled model showed a good improvement in calibration metrics (slope = 0.589 and calibration-in-the-large = 1.165; Supplemental Figure 1) compared to the single models it was derived from (slope of respectively 27 for Neural Network and 0.47 for Logistic Regression) (Van Calster et al., 2019) .

Discussion

Classification by means of machine-learning models is not a novel approach, but it can be considered as an up-and-coming field due to the contemporary improvements in the computational capabilities of processors (Deo, 2015). For the field of periodontics, it is well known that few teeth are lost in patients under SPC (Leow et al., 2022). Consequently, it is common to work on "imbalanced datasets" where the target class has an uneven distribution of observations, as tooth loss occurs much less often than tooth retention. In unbalanced datasets where the outcome event is rare, models with higher specificity tend to show a higher value of AUC. However, this reflects the fact that if a model always predicts the outcome 'no loss' it will be right in about 90% of cases, but the model is clinically useless. In this study we had a total MLO rate of $\approx$ 9.5%. Consequently, the class "survived" is called the majority class, and the much smaller sized outcome class is called the minority class. The main issue to consider with prediction on imbalanced datasets is how accurately the model predicts both the majority and minority class. To illustrate this, let us assume that our model predicts that 3157/3157 molars will be retained at 10 years follow-up. As 302 teeth were lost during follow-up, the model will be right in its prediction in 2855 cases, resulting in an accurate prediction for 90.43% of molars with an excellent specificity but very low sensitivity. Hence, in periodontics, it is very common to develop

models with very poor sensitivity while also resulting in an apparently good general prognostic performance. In support of this concept, a recent study evaluated prospectively four different periodontal prognostic systems and found very high values of specificity but a very low sensitivity ranging between 3 to 12% (Saydzai et al., 2022b). The challenge in developing accurate prediction models for the prognostic prediction in periodontology is to have a good balance between specificity and sensitivity. To fix the above-mentioned issue with imbalanced datasets, we applied an oversampling technique (SMOTE) for the development phase. This led to an improved performance in discriminative models, but not a perfect calibration due to the artificial event rate created using this approach (Dhiman et al., 2022). To improve such performance metrics and combine different algorithms, we applied the stacking ensemble method. This resulted in the development of a new model with better calibration and a similar performance in discrimination (Kim, You, Reps, Cheong, & Park, 2021; Zhai & Chen, 2018). In the present study, such combined models displayed an overall sensitivity of 40.9% for MLP and 66,7% for MLO during external validation (Table 3). It is still doubtful, in the current literature, if and in which cases machine learning models outperform classical logistic regression models (Christodoulou et al., 2019). A recent study (Bashir, Rahman, & Chen, 2022), applied different pre-processing methods and machine learning algorithms to develop diagnostic models of periodontitis. Results were disappointing, showing a collapse in predicting performance when an external validation was performed. In their conclusions, authors encouraged larger sample sizes, accurate predictors, and external validation first to consider the use of these models in the clinical practice (Bashir et al., 2022). The present study utilized these recommendations in an international collaboration with the aim of aggregating different cohorts, increasing the sample size for model development and validation (Rischke et al., 2022). These cohorts were previously used for other prognostic studies published in periodontology (Petsos et al., 2021; Saleh et al., 2021; Saydzai et al., 2022b; Shi et al., 2020). A higher tooth-loss rate appeared in the Beijing dataset, as well as some differences in predictor rates. This could have been due to the overall worst compliance in this specific population compared to the other groups. Furthermore, the unique dental anatomical

characteristics of the Chinese population may be an additional factor to be considered. Previous studies showed that they had narrow furcation entrance diameter (FED)(Bower, 1979), a higher prevalence of cervical enamel projection CEP (Zee, Chiu, Holmgren, Walker, & Corbet, 1991), and shorter root trunk length (Hou & Tsai, 1997). Consequently, this brings immense challenges to the management of periodontitis in this population, leading to a high rate of tooth loss (Guo et al., 2018). The presence of more than one center is a strength due to higher external validity, however it also leads to some limitations. First, there are differences in the maintenance regimen among the cohorts. The compliance was 100% in the Michigan cohort, 89.10% in the London Cohort, 67.50% in the Frankfurt cohort and 1.62% in the Beijing cohort, showing a statistically significant difference among the included cohorts. However, sensitivity, specificity and AUC (table 4) of the Beijing cohort (which is the least compliant) was very similar to the Frankfurt and London cohort. This result shows that the external validly is maintained despite the discrepancy in the compliance. Second, some centers collected data from patients treated in private practice (English cohort) while others at a dental school (Germany, USA, and China) by a variety of operators. This could lead to different subjective criteria for the need for extractions, which hinges on a variety of factors that are not all related solely to the periodontal health status of the tooth, such as economic considerations and overall treatment plan.

on the prognosis of the individual tooth. Some patient level variables can influence the tooth survival rate. For example, in some stage IV periodontitis cases, the whole dentition must be rehabilitated due to already missing teeth, and healthy/not-hopeless molars may be removed due to prosthetic reason to make way for a full arch implant-supported prosthesis. These circumstances cannot be predicted with machine learning as it is not possible to do a multilevel analysis considering both patient related and tooth related factors It should also be noted that predicting long term tooth loss/retention based on baseline data can never attain a perfect prediction for all teeth evaluated. Being limited to baseline data implies that it cannot weigh for factors occurring during the follow-up period (ie. changes in smoking habits, compliance with maintenance, etc.) which may influence the long-term prognosis. There are also situations where more strict criteria need to be met to meet the requirements for molar

retention. Retaining a molar in a complete or even shortened arch (stage I to III periodontitis) may be easy even if the molar has poor or questionable prognosis. If it does not cause pain/discomfort and is a functioning unit in the dentition, it may easily be retained and managed during SPC for as long as it can function (Eickholz et al., 2021). However, if after periodontal treatment there is need for prosthetic reconstruction, the respective molar with poor or questionable prognosis may have to become an abutment tooth of fixed or removable dental protheses which may deteriorate its prognosis (Pretzl, Kaltschmitt, Kim, Reitmeir, & Eickholz, 2008) necessitating more strict criteria for retention.

As reported in a recently-published study: having high quality predictors and cohorts is a fundamental requirement in order to have a performance advantage with trained machine learning models compared to classical logistic regression (Bashir et al., 2022). Results of this study support the findings of a recent publication where an improved performance (superior to classic logistic regression) was obtained with both the Neural Network and the Ensembled models by using well standardized cohorts from different parts of the world (Bashir et al., 2022). Efforts to share data among different groups of research will be more and more fundamental for the development of very accurate models moving forward. Future studies should compare the performance of artificial intelligence-based models with the traditional prognostic models. Indeed, although efforts have been put toward creating and comparing intelligence-based models (Bashir et al., 2022) as well as traditional prognostic methods (Saleh et al., 2021; Saleh et al., 2022; Saydzai et al., 2022a), to the best of our knowledge no comparison between traditional vs. intelligence-based models has been published.

Aiming to favor a widespread use of the introduced model, we are offering the developed model for usage free of charge on the open-source software Orang (Supplementary Material).

Conclusion

In this study different machine-learning models for prognostic prediction of molar teeth were developed and validated. The performance of an Ensembled model which combines neural network

and logistic regression models resulted in the highest and the most stable algorithm on the 3 different cohorts utilized to validate the model.

REFERENCES

Bashir, N. Z., Rahman, Z., & Chen, S. L. (2022). Systematic comparison of machine learning algorithms to develop and validate predictive models for periodontitis. *J Clin Periodontol*. doi:10.1111/jcpe.13692

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood), 33*(7), 1123-1131. doi:10.1377/hlthaff.2014.0041

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics, 14*, 106. doi:10.1186/1471-2105-14-106

Bower, R. C. (1979). Furcation morphology relative to periodontal treatment. Furcation entrance architecture. *J Periodontol, 50*(1), 23-27. doi:10.1902/jop.1979.50.1.23

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol, 110*, 12-22. doi:10.1016/j.jclinepi.2019.02.004

Collins, G. S., Dhiman, P., Andaur Navarro, C. L., Ma, J., Hooft, L., Reitsma, J. B., . . . Moons, K. G. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open, 11*(7), e048008. doi:10.1136/bmjopen-2020-048008

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ, 350*, g7594. doi:10.1136/bmj.g7594

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation, 132*(20), 1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593

Dhiman, P., Ma, J., Andaur Navarro, C. L., Speich, B., Bullock, G., Damen, J. A. A., . . . Collins, G. S. (2022). Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol, 22*(1), 101. doi:10.1186/s12874-022-01577-x

Du, M., Bo, T., Kapellas, K., & Peres, M. A. (2018). Prediction models for the incidence and progression of periodontitis: A systematic review. *J Clin Periodontol, 45*(12), 1408-1420. doi:10.1111/jcpe.13037

Eickholz, P., Runschke, M., Dannewitz, B., Nickles, K., Petsos, H., Kronsteiner, D., & Pretzl, B. (2021). Long-term prognosis of teeth with class III furcation involvement. *J Clin Periodontol, 48*(12), 1528-1536. doi:10.1111/jcpe.13551

Guo, J., Ban, J. H., Li, G., Wang, X., Feng, X. P., Tai, B. J., . . . Wang, S. C. (2018). Status of Tooth Loss and Denture Restoration in Chinese Adult Population: Findings from the 4th National Oral Health Survey. *Chin J Dent Res, 21*(4), 249-257. doi:10.3290/j.cjdr.a41083

Hamp, S. E., Nyman, S., & Lindhe, J. (1975). Periodontal treatment of multirooted teeth. Results after 5 years. *J Clin Periodontol, 2*(3), 126-135. doi:10.1111/j.1600-051x.1975.tb01734.x

Harrison, C. J., & Sidey-Gibbons, C. J. (2021). Machine learning in medicine: a practical introduction to natural language processing. *BMC Med Res Methodol, 21*(1), 158. doi:10.1186/s12874-021-01347-1

Hou, G. L., & Tsai, C. C. (1997). Types and dimensions of root trunk correlating with diagnosis of molar furcation involvements. *J Clin Periodontol, 24*(2), 129-135. doi:10.1111/j.1600-051x.1997.tb00479.x

Kantarjian, H., & Yu, P. P. (2015). Artificial Intelligence, Big Data, and Cancer. *JAMA Oncol, 1*(5), 573-574. doi:10.1001/jamaoncol.2015.1203

Kim, C., You, S. C., Reps, J. M., Cheong, J. Y., & Park, R. W. (2021). Machine-learning model to predict the cause of death using a stacking ensemble method for observational data. *J Am Med Inform Assoc, 28*(6), 1098-1107. doi:10.1093/jamia/ocaa277

Krois, J., Graetz, C., Holtfreter, B., Brinkmann, P., Kocher, T., & Schwendicke, F. (2019). Evaluating Modeling and Validation Strategies for Tooth Loss. *J Dent Res, 98*(10), 1088-1095. doi:10.1177/0022034519864889

Lang, N. P., & Tonetti, M. S. (2003). Periodontal risk assessment (PRA) for patients in supportive periodontal therapy (SPT). *Oral Health Prev Dent, 1*(1), 7-16.

Leow, N. M., Moreno, F., Marletta, D., Hussain, S. B., Buti, J., Almond, N., & Needleman, I. (2022). Recurrence and progression of periodontitis and methods of management in long-term care: A systematic review and meta-analysis. *J Clin Periodontol, 49 Suppl 24*, 291-313. doi:10.1111/jcpe.13553

Lindhe, J., & Nyman, S. (1977). The role of occlusion in periodontal disease and the biological rationale for splinting in treatment of periodontitis. *Oral Sci Rev, 10*, 11-43.

Mohammad-Rahimi, H., Motamedian, S. R., Pirayesh, Z., Haiat, A., Zahedrozegar, S., Mahmoudinia, E., . . . Schwendicke, F. (2022). Deep learning in periodontology and oral implantology: A scoping review. *J Periodontal Res*. doi:10.1111/jre.13037

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *Lancet Oncol, 20*(5), e262-e273. doi:10.1016/S1470-2045(19)30149-4

Petsos, H., Ramich, T., Nickles, K., Dannewitz, B., Pfeifer, L., Zuhr, O., & Eickholz, P. (2021). Tooth loss in periodontally compromised patients: Retrospective long-term results 10 years after active periodontal therapy - tooth-related outcomes. *J Periodontol, 92*(12), 1761-1775. doi:10.1002/JPER.21-0056

Pretzl, B., Kaltschmitt, J., Kim, T. S., Reitmeir, P., & Eickholz, P. (2008). Tooth loss after active periodontal therapy. 2: tooth-related factors. *J Clin Periodontol, 35*(2), 175-182. doi:10.1111/j.1600-051X.2007.01182.x

Ravida, A., Troiano, G., Qazi, M., Saleh, M. H. A., Lo Russo, L., Greenwell, H., . . . Wang, H. L. (2020). Development of a nomogram for the prediction of periodontal tooth loss using the staging and grading system: A long-term cohort study. *J Clin Periodontol, 47*(11), 1362-1370. doi:10.1111/jcpe.13362

Rischke, R., Schneider, L., Muller, K., Samek, W., Schwendicke, F., & Krois, J. (2022). Federated Learning in Dentistry: Chances and Challenges. *J Dent Res*, 220345221108953. doi:10.1177/00220345221108953

Saleh, M. H. A., Dukka, H., Troiano, G., Ravida, A., Galli, M., Qazi, M., . . . Wang, H. L. (2021). External validation and comparison of the predictive performance of 10 different tooth-level prognostic systems. *J Clin Periodontol, 48*(11), 1421-1429. doi:10.1111/jcpe.13542

Saleh, M. H. A., Dukka, H., Troiano, G., Ravida, A., Qazi, M., Wang, H. L., & Greenwell, H. (2022). Long term comparison of the prognostic performance of PerioRisk, periodontal risk assessment, periodontal risk calculator, and staging and grading systems. *J Periodontol, 93*(1), 57-68. doi:10.1002/JPER.20-0662

Saydzai, S., Buontempo, Z., Patel, P., Hasan, F., Sun, C., Akcali, A., . . . Nibali, L. (2022a). Comparison of the efficacy of periodontal prognostic systems in predicting tooth loss. *J Clin Periodontol, 49*(8), 740-748. doi:10.1111/jcpe.13672

Saydzai, S., Buontempo, Z., Patel, P., Hasan, F., Sun, C., Akcali, A., . . . Nibali, L. (2022b). Comparison of the efficacy of periodontal prognostic systems in predicting tooth loss. *J Clin Periodontol*. doi:10.1111/jcpe.13672

Schwendicke, F., Schmietendorf, E., Plaumann, A., Salzer, S., Dorfer, C. E., & Graetz, C. (2018). Validation of multivariable models for predicting tooth loss in periodontitis patients. *J Clin Periodontol, 45*(6), 701-710. doi:10.1111/jcpe.12900

Schwendicke, F., Stolpe, M., & Graetz, C. (2017). Cost comparison of prediction-based decision-making for periodontally affected molars. *J Clin Periodontol, 44*(11), 1145-1152. doi:10.1111/jcpe.12796

Shi, S., Meng, Y., Li, W., Jiao, J., Meng, H., & Feng, X. (2020). A nomogram prediction for mandibular molar survival in Chinese patients with periodontitis: A 10-year retrospective cohort study. *J Clin Periodontol, 47*(9), 1121-1131. doi:10.1111/jcpe.13343

Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Med Res Methodol, 19*(1), 64. doi:10.1186/s12874-019-0681-4

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Topic Group 'Evaluating diagnostic, t., & prediction models' of the, S. i. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Med, 17*(1), 230. doi:10.1186/s12916-019-1466-7

Zee, K. Y., Chiu, M. L., Holmgren, C. J., Walker, R. T., & Corbet, E. F. (1991). Cervical enamel projections in Chinese first permanent molars. *Aust Dent J, 36*(5), 356-360. doi:10.1111/j.1834-7819.1991.tb01356.x

Zhai, B., & Chen, J. (2018). Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. *Sci Total Environ, 635*, 644-658. doi:10.1016/j.scitotenv.2018.04.040

Figure Legends

Figure 1: Area Under the Receiver Operating Characteristics Curve for the different machine-learning models on the aggregate external validation cohort of molars predicting the outcome overall-molar loss.

Table 1: Demographic and clinical data at tooth-level of the analyzed cohorts.

Table 2: Performance of different machine learning models for the forecast of overall-molar loss with different set of predictors ranked by means of some feature selection methods; Fast Correlation Based Filter (FCBF) – Area Under the ROC Curve (AUC).

Table 3: Predictive performance of different machine-learning algorithms in external validation settings for both Periodontitis-Related Molar loss and Overall-Molar loss.

Table 4: Predictive performance of the trained machine-learning models for overall molar loss on the different cohorts included in this study.

Table 5: Predictive performance of the trained machine-learning for periodontitis-related molar loss on the different cohorts included in this study.

Supplemental Figure 1: Calibration plots of different machine-learning models on the unified external validation cohorts for overall-tooth loss prediction. a) Neural Network; b) Logistic Regression; c) Ensembled Model (logistic regression + neural network); d) Naive Bayes; e) Random Forest; f) K-Neigtherest Neighbor (KNN); g) Support Vector Machine; h) Gradient Boosting.

Supplemental Materials: Files and instructions for using the model in the clinical practice.

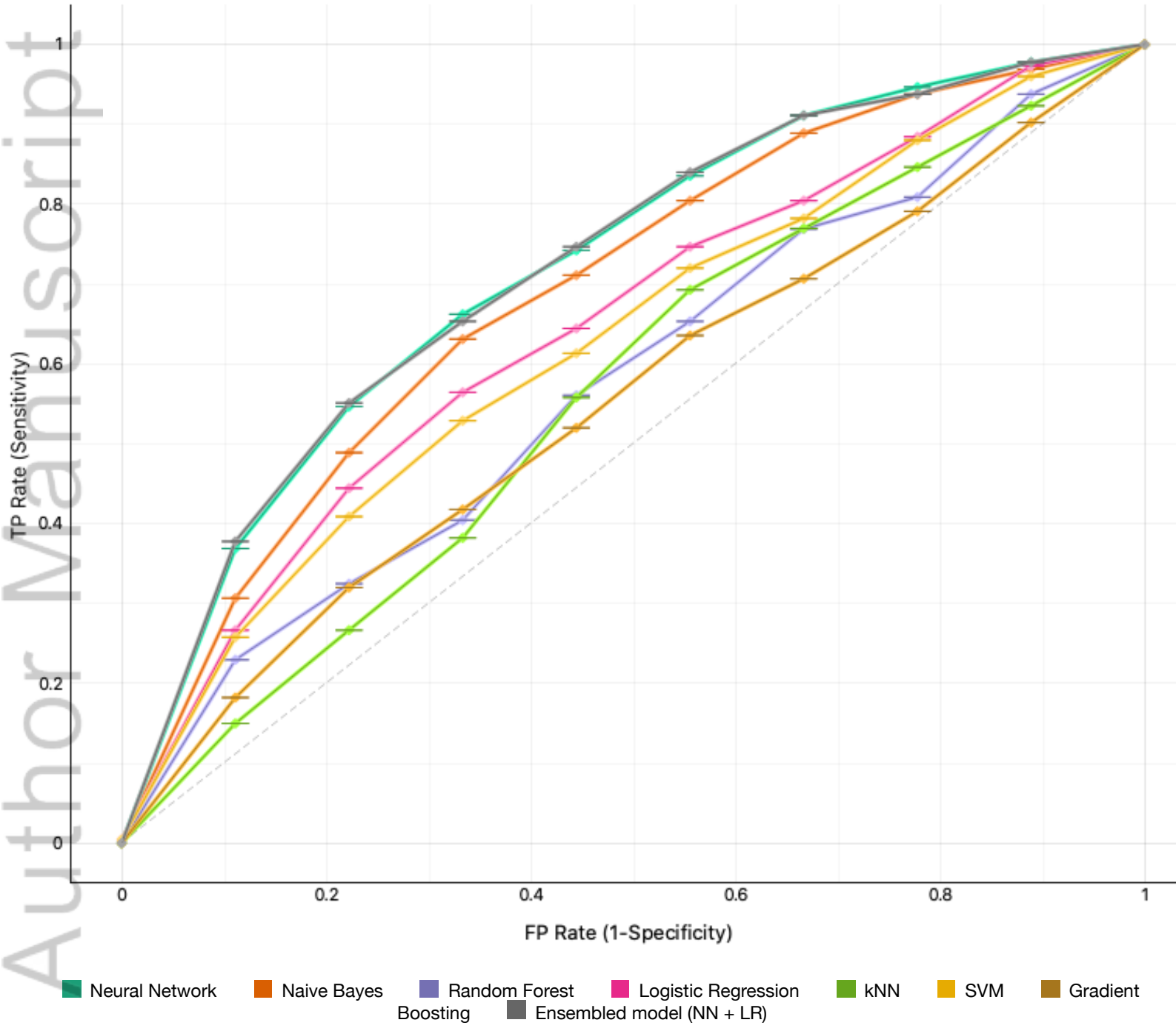

Table 1: Demographic and clinical data at tooth-level of the analyzed cohorts.

Variables	Michigan Cohort		Beijing Cohort		Frankfurt Cohort		London Cohort	
Patients	222		131		97		65	
Molars	1475		681		597		404	
10-years' MLO								
No	1398	94.78%	546	80.18%	538	90.12%	373	92.33%
Yes	77	5.22%	135	19.82%	59	9.88%	31	7.67%
Sex (Tooth-level)								
Males	753	51.05%	293	43.02%	271	45.39%	267	66.09%
Females	722	48.95%	388	56.98%	326	54.61%	137	33.91%
Smoking								
Never Smoker	806	54.64%	561	82.38%	486	81.41%	202	50.00%
Former Smoker	405	27.46%	12	1.76%	27	4.52%	146	36.14%
Actual Smoker	264	17.90%	108	15.86%	84	14.07%	56	13.86%
Furcation Index								
0	958	64.95%	275	40.38%	155	25.96%	207	51.24%
1	344	23.32%	96	14.10%	231	38.69%	118	29.21%
2	160	10.85%	253	37.15%	143	23.95%	51	12.62%
3	13	0.88%	57	8.37%	68	11.39%	28	6.93%
Bone Loss								
< 15%	691	46.85%	15	2.20%	36	6.03%	43	13.36%
15 – 33%	529	35.86%	153	22.47%	265	44.39%	278	68.81%
> 33%	255	17.29%	513	75.33%	296	59.58%	83	20.54%
Mobility								
0	1310	88.81%	489	71.81%	476	79.73%	366	90.59%
1	130	8.81%	123	18.06%	69	11.56%	32	7.92%
2	30	2.03%	54	7.93%	41	6.87%	5	1.24%
3	5	0.34%	15	2.20%	11	1.84%	1	0.25%
Retainer								
No	1463	99.19%	.678	N.R.99.56%	397	66.50%	391	96.78%
Yes	12	0.81%	3	0.44%	200	33.50%	13	3.22%
Age	46.24±11.65		41.51±10.46		53.06±11.03		54.22±8.26	
Probing Depth	4.79±1.63		6.48±1.74		4.07±1.18		5.41±1.96	
Clinical Attachment Level	4.72±1.92		3.08±3.02		4.55±1.50		6.23±2.28	

MLO: overall molar loss

Aggregated External Validation Cohorts (n = 1682)	Gini Inequality index			FCBF			Chi-Squared			ReliefF		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Neural Network	0.609	23.6%	91.1%	0.665	84.9%	34.6%	0.683	78.2%	43.0%	0.621	64.4%	55.5%
Naive Bayes	0.619	27.6%	83.1%	0.656	39.1%	81.4%	0.666	52.4%	65.5%	0.603	22.2%	89.0%
Random Forest	0.550	36.9%	69.7%	0.545	36.4%	70.3%	0.585	56.7%	54.2%	0.535	36.9%	69.7%
Logistic Regression	0.660	76.0%	34.9%	0.616	76.0%	35.9%	0.636	68.4%	46.1%	0.616	61.3%	57.2%
K-Nearest Neighbors	0.530	31.1%	78.1%	0.510	29.3%	77.8%	0.538	21.3%	82.5%	0.540	35.6%	70.6%
Gradient Boosting	0.520	7.1%	91.8%	0.488	9.3%	93.8%	0.521	20.4%	84.2%	0.531	10.2%	87.6%
Support Vector Machine	0.611	34.8%	76.0%	0.589	76.0%	34.9%	0.635	72.9%	44.5.%	0.613	63.6%	54.7%
Ensembled model (neural network + logistic regression)	0.609	0.4%	99.7%	0.669	100.0%	0.4%	0.682	95.1%	9.5%	0.621	99.1%	1.6%

Table 2: Performance of different machine learning models for the forecast of overall-tooth loss with different set of predictors ranked by means of some feature selection methods; Fast Correlation Based Filter (FCBF) – Area Under the ROC Curve (AUC). The five best ranked predictors were included in each model.

External validation aggregated cohorts (n = 1682)	Periodontitis-related molar loss			Overall molar loss		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Neural Network	0.702	40.9%	85.7%	0.724	66.7%	63.4%
Naive Bayes	0.649	31.6%	84.4%	0.695	48.9%	77.1%
Random Forest	0.683	16.4%	89.7%	0.590	36.4%	70.7%
Logistic Regression	0.611	52.6%	61.0%	0.647	74.7%	44.0%
K-Nearest Neighbors	0.565	24.6%	83.7%	0.569	66.6%	38.2%
Gradient Boosting	0.512	5.8%	95.6%	0.559	16.9%	89.2%
Support Vector Machine	0.517	50.3%	54.3%	0.626	75.6%	38.2%
Esembled model (neural network + logistic regression)	0.702	40.9%	85.7%	0.726	40.4%	87.4%

Table 3: Predictive performance of different machine-learning algorithms in external validation settings for both Periodontitis-Related Molar loss and Overall-Molar loss.

Overall Tooth loss	Beijing Cohort (n = 681)			Frankfurt Cohort (n = 597)			London Cohort (n = 404)		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Neural Network	0.728	65.9%	62.8%	0.732	67.8%	64.3%	0.707	67.7%	63.0%
Naive Bayes	0.703	51.9%	72.7%	0.676	47.5%	84.4%	0.635	38.7%	72.9%
Random Forest	0.486	34.8%	64.3%	0.625	40.7%	68.0%	0.733	35.5%	83.9%
Logistic Regression	0.614	65.2%	44.1%	0.717	88.1%	39.6%	0.732	90.3%	50.1%
K-Nearest Neighbors	0.591	30.4%	74.1%	0.579	47.5%	63.2%	0.620	54.8%	59.2%
Gradient Boosting	0.610	10.4%	95.6%	0.551	27.1%	80.7%	0.681	25.8%	92.0%
Support Vector Machine	0.574	64.4%	38.1%	0.720	93.2%	30.9%	0.734	90.3%	49.1%
Ensembled Model (neural network + logistic regression)	0.736	43.0%	88.6%	0.731	40.7%	88.3%	0.705	29.0%	84.5%

Table 4: Predictive performance of the trained machine-learning for overall tooth loss on the different cohorts included in this study.

Periodontitis-Related molar loss	Beijing Cohort (n = 681)			Frankfurt Cohort (n = 597)			London Cohort (n = 404)		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Neural Network	0.709	41.5%	87.9%	0.837	60.0%	83.0%	0.764	63.6%	77.6%
Naive Bayes	0.651	28.9%	88.3%	0.736	36.0%	88.8%	0.720	54.5%	77.9%
Random Forest	0.684	11.9%	98.4%	0.756	24.0%	89.5%	0.701	54.5%	73.0%
Logistic Regression	0.728	43.7%	84.4%	0.768	84.0%	54.4%	0.758	90.9%	38.2%
K-Nearest Neighbors	0.607	23.7%	90.5%	0.549	24.0%	81.6%	0.580	36.4%	77.4%
Gradient Boosting	0.583	5.2%	98.4%	0.534	4.0%	93.0%	0.731	18.2%	95.7%
Support Vector Machine	0.561	39.3%	68.7%	0.809	93.2%	30.9%	0.673	81.8%	48.3%
Ensembled Model (neural network + logistic regression)	0.709	36.3%	90.3%	0.837	56.0%	86.0%	0.764	63.60%	79.1%

Table 5: Predictive performance of the trained machine-learning for periodontitis-related molar loss on the different cohorts included in this study.