Importance of socioeconomic factors in predicting tooth loss among older adults in Japan: Evidence from a machine learning analysis

Upul Cooray (MSc)[a], Richard G Watt (PhD)[b], Georgios Tsakos (PhD)[b], Anja Heilmann (PhD)[b], Masanori Hariyama (PhD)[e], Takafumi Yamamoto (PhD)[a], Isuruni Kuruppuarachchige[f], Katsunori Kondo (PhD)[g,h], Ken Osaka (PhD)[a], Jun Aida (PhD)[c,d]

[a] Department of International and Community Oral Health, Tohoku University Graduate School of Dentistry, Sendai, Japan

[b] Department of Epidemiology and Public Health, University College London, London, United Kingdom

[c] Department of Oral Health Promotion, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan

[d] Division for Regional Community Development, Liaison Center for Innovative Dentistry, Graduate School of Dentistry, Tohoku University, Sendai, Japan

[e] Intelligent Integrated Systems Laboratory, Graduate School of Information Sciences, Tohoku University, Miyagi, Japan

[f] Department of Dental and Digital Forensics Tohoku University Graduate School of Dentistry, Sendai, Japan

[g] Center for Preventive Medical Sciences, Chiba University, Chiba, Japan

[h] Center for Gerontology and Social Science, National Center for Geriatrics and Gerontology, Obu, Japan

Corresponding Author:

Upul Cooray

Department of International and Community Oral Health, Graduate School of Dentistry, Tohoku University, Sendai, Japan

Address: 4-1, Seiryo-machi, Aoba-ku, Sendai, Miyagi, 980-8575, Japan

Tel: +81-2-2717-7639

E-mail: upul.cooray.15@ucl.ac.uk

Importance of socioeconomic factors in predicting tooth loss among older adults in Japan: Evidence from a machine learning analysis

Abstract

Prevalence of tooth loss has increased due to population aging. Tooth loss negatively affects the overall physical and social wellbeing of older adults. Understanding socio-demographic and other predictors associated with tooth loss which were measured within non-clinical settings can be useful in community-level prevention. Hence, we investigated important factors in predicting tooth loss among older adults over 6 years of follow-up using high-dimensional epidemiological data. Data was from participants of 2010 and 2016 waves of the Japan Gerontological Evaluation Study (JAGES). A total of 19,407 community-dwelling functionally independent older adults aged 65 and older were included in the analysis. Tooth loss was measured as moving from a higher number of teeth category at the baseline to a lower number of teeth category at the follow-up. Out of 119 potential predictors, age, sex, number of teeth, denture use, chewing difficulty, household income, employment, education, smoking, fruit and vegetable consumption, community participation, time since last health check-up, having a hobby, and feeling worthless were selected by Boruta algorithm. Within the 6-year follow-up, 3,013 individuals (15.5%) reported incidence of tooth loss. People who experienced tooth loss were older (72.9±5.2 vs 71.8±4.7), and predominantly men (18.3% vs 13.1%). Extreme gradient boosting (XGBoost) machine learning prediction model had a mean accuracy of 90.5% (±0.9%). A visual analysis of machine learning predictions revealed that the prediction of tooth loss was mainly driven by demographic (older age), baseline oral health (having 10-19 teeth, wearing dentures), and socioeconomic (lower household

income, manual occupations) variables. Predictors related to wide a range of determinants contribute towards tooth loss among older adults. In addition to oral health related and demographic factors, socioeconomic factors are important in predicting future tooth loss. Understanding the behaviour of these predictors can thus be useful in developing prevention strategies for tooth loss among older adults.

Introduction

Tooth loss can have a negative impact on older adults' quality of life, affecting function, nutrition, aesthetics, as well as psychological and social well-being. The prevalence of tooth loss has increased due to population aging (James et al., 2018). Tooth loss indicates an individual's experience of dental disease and its treatment throughout the life-course. Therefore, it is considered one of the most useful oral health indicators (Kassebaum et al., 2014).

A wide range of social determinants, such as socioeconomic, community-level, psychosocial, behavioural, and demographic factors are associated with tooth loss (Aida et al., 2009; Silva et al., 2019). However, existing prediction models of tooth loss have mainly used tooth-level clinical features, as studies focused on people with oral diseases or those who sought or received dental treatment (Ravidà et al., 2020; Schwendicke et al., 2018). Furthermore, there are statistical limitations in such conventional multivariable models as parametric techniques obviate the use of multidimensional data with complex associations, thus limiting the predictive capacity of models.

In contrast to conventional statistical methods, machine learning based models can make accurate predictions while considering a wider range of factors without strict assumptions regarding the predictors and the outcome (Breiman, 2001; Bzdok et al., 2017). Machine learning is an umbrella term used to describe a wide variety of models and strategies that focus on algorithmic modelling (Royal Society 2017). The use of machine learning methods has become popular in medical research fields such as genomics and diagnostic imaging (Sidey-Gibbons and Sidey-Gibbons, 2019). Although machine learning methods are capable of powerful predictions by detecting complex relationships in data, their application in epidemiological and public health research has been limited (Wiemken and Kelley, 2019). This could mainly be due to the perception of machine learning models as "black boxes"; that is, being not straightforward in explaining how a given predictor contributed to the prediction, hence not being suitable for obtaining actionable interpretations (Bi et al., 2019).

In a recent cross-sectional study, Elani et al. (2021) predicted tooth loss among adults (aged 20 years or older) in the United States using machine learning algorithms (Elani et al., 2021). However, none of the previous studies have used longitudinal data to predict tooth loss among older adults. Some studies have used machine learning to predict health outcomes such as mobility loss, fall risk, and mortality among older adults (Speiser et al., 2020; Stenholm et al., 2015). None of those studies used explainable machine learning methods to obtain interpretations for their prediction models.

The main oral conditions that lead to tooth loss among older adults (dental caries and periodontal disease) are largely preventable by community level interventions (Peres et al., 2019). Therefore, understanding of the predictors related to a wide variety of determinants could provide useful insights to optimise relevant prevention strategies and for policy interventions. In the current study, we investigated the possibility of using machine learning methods to identify the most important predictors of tooth loss, to predict the incidence of tooth loss, and to understand the behaviour of those predictors.

Methods

Study settings

The data for this study came from the Japan Gerontological Evaluation Study (JAGES) (Kondo et al., 2018). JAGES is an ongoing cohort study for over 65-year-old community-dwelling older adults living in Japan. We used two waves of data: the baseline in 2010 and the follow-up in 2016. In 2010, 95,827 postal survey questionnaires were randomly distributed among 16 municipalities and 62,418 people responded (response rate: 65.1%). In the 2016 follow-up, 54,529 of the baseline survey participants were successfully identified (n=7,889 were not identified due to invalid information at baseline). Among them, 7,744 were ineligible to participate in the 2016 survey because they were in care facilities, 6,148 had died, and 12,570 were lost to follow-up. Hence, a total of 28,067 baseline participants took part in the 2016 follow-up survey.

Only the respondents who were functionally independent at baseline (i.e., who could walk, take a bath or use a toilet without assistance) were included in the analysis. The participants who were edentate at baseline (n=2,717), had missing (n=1,558) or invalid information for the number of teeth variable (n=4,231) were also excluded. As a result, the final analytical sample included 19,407 individuals. The selection of the analytical sample is shown in Figure 1.

Ethical approval

The JAGES survey was approved by the ethics committee of the XXXX (No. XXX) and the ethics committee of XXXX University (No. XXX).

Outcome variable

In JAGES, the number of teeth was recorded as a categorical variable (i.e. >20 teeth/10-19 teeth/1-9 teeth/no teeth). Transition from a higher number of teeth category at baseline to a lower number of teeth category at the follow-up indicated "tooth loss". A binary outcome variable was created to indicate the tooth loss (0: no tooth loss, 1: tooth loss).

Variable transformations

Since low frequencies in categorical variable categories can have a negative impact on the robustness of machine learning models, some categorical variables were recoded to assign less frequent values to a more general category. Some variables were combined into single variables to create composite variables related to certain constructs (e.g. community participation) or aggregate variables of established measures (e.g. IADLs, body mass index). All variable transformations are reported in Supplementary Table 1.

Imputation of missing values

Variables with more than 30% missing information were dropped to minimise possible bias due to imputation. Remaining variables were imputed using the 'missForest' multivariate iterative random forest (RF) imputation algorithm with five iterations and 100 estimators to impute each variable (Kokla et al., 2019).

Variable selection and analyses

Selection of predictors

Selection of predictors for the final model was performed in two steps. First, as JAGES contains a large number of variables, we manually excluded theoretically irrelevant variables for tooth loss, informed by existing literature and the domain knowledge of the authors. JAGES contained a large number of variables. This process resulted in a pool of 119 potential predictors. Then, to further reduce the dimensionality of the data and to select only the most important variables for the final model we used the random forest based Boruta feature selection algorithm(Kursa and Rudnicki, 2010). The Boruta algorithm has been shown to be more robust and statistically grounded compared to other feature selection methods used in machine learning (Kursa, 2014).

The final selected predictors were age, sex (male/female), number of teeth at baseline (>=20 teeth/10-19 teeth/<10 teeth), denture use (yes/no), chewing difficulty (yes/no), annual household income in Japanese yen (<1 million/1 to <1.5 million/1.5 to <2 million/2 to <3 million /3 to <5 million/ >=5 million), years of education (<10/10-12/>=13), occupation category (manual/sales or services/clerical/managerial/specialist), smoking status (never smoked/stopped >=5years ago/stopped <5years ago/current smoker), frequency of fruits and vegetables consumption (once a week or less/2-3 times a week/4-6 times a week/ once a day/ twice or more a day), time since last health check-up (within a year/within 2-3 years/>4years ago/never), leisure activities (yes/no), feeling useless (yes/no), and frequency of community participation (never/few times a year/1-2 times a month/once a week/>once a week). Table 1 reports characteristics of the study population by outcome variable.

Minority class oversampling

In our study sample, the incidence of tooth loss was 15.5% (minority class) as opposed to 84.5% (majority class) for no tooth loss. This large variance between classes can lead to a poor performing machine learning model due to few instances of the outcome (Wiemken and Kelley, 2019). Oversampling of the minority class is commonly used procedure with imbalanced data. Therefore, random oversampling of the minority class was performed to obtain class balance (Khaldy, 2018).

Prediction model

We used the extreme gradient boosting (XGBoost) classification algorithm (Chen and Guestrin, 2016) to model the relationships between the selected predictors and incidence of tooth loss. XGBoost was selected as the main algorithm for the following reasons: 1) its superior performance over logistic regression and random forest algorithms in our preliminary analysis compare performance (Figure 2), 2) its ability to directly handle encoded categorical variables, 3) having a better compatibility with the features of the shapley additive explanations (SHAP) framework. In addition, XGBoost algorithm is well-known for its high efficiency and accuracy (Huang et al., 2018; Sagi and Rokach, 2018). Random forest classification models were used as baseline tree-based model to compare the performance of XGBoost models.

k- fold cross-validation

k-Fold cross validation was to assess the performance of machine learning models, which is performed by splitting data into *k* number of groups; each unique group is held out as test data while the remaining *k-1* groups are used as training data (Hastie et al., 2009). We nested a separate k-fold hyperparameter optimisation procedure within the training data split of the model evaluation procedure. 10-fold cross validation was used for model evaluation (Figure 3; outer split) and 5-fold cross validation was used for nested hyperparameter optimisation (Figure 3; inner split). Cross validation procedure was repeated 10 times, thus evaluating 100 (10x10) models to obtain mean performance scores (Table 2). Implementation of repeated nested cross validation is illustrated in Figure 3. Results of all XGBoost model evaluations are reported in Supplementary Table 2.

Performance metrics

Model performance was evaluated using accuracy score, F1 score (i.e., weighted average of the precision and sensitivity obtained using the formula: *2 (precision x sensitivity)/ (precision + sensitivity)*), and Area Under the Receiver Operating Characteristic Curve (ROC_AUC).

Interpretability

We used the SHAP framework to get insight into the behaviour of the XGBoost prediction model. SHAP values can provide both local (each prediction) and global (overall) explanations regarding the behaviour of predictors in predicting the outcome. This enables visualisation of how a given value (of a continuous variables) or a specific group (of a categorical variable) in a predictor contributes to outcome prediction. A detailed description of SHAP value computation can be found in Lundberg and Lee 2017 (https://github.com/slundberg/shap) (Lundberg and Lee, 2017).

Sensitivity analyses

In order to assess the sensitivity of the minority class oversampling procedure, we conducted several model performance analyses using different oversampling ratios (1.0, 0.75, 0.5, none). Model performances related to all oversampling ratios are reported in Table 2. All analyses were conducted using Jupyter Notebook with python 3.8 kernel and Stata MP 16.1 (StataCorp LLC). XGBoost version 1.3.3 and SHAP version 0.38.1 were used. The reporting of this study conforms to STROBE guidelines.

Results

Among 19,407 people in the analytical sample, 3,013 individuals (15.5%) reported a category with fewer number of teeth in 2016 compared to the baseline (2010). Table 1 shows the incidence of tooth loss for all the predictors used in the study. People who experienced incidence of tooth loss were older (72.9±5.2 vs 71.8±4.7), and predominantly men (18.3% vs 13.1%). Tooth loss incidence was higher among individuals who at baseline were in the 10-19 teeth category, were denture wearers, reported chewing difficulties, and were in a lower socioeconomic position.

Our feature selection algorithm selected 14 relevant predictors of tooth loss that can be categorised into seven broader determinants of oral health. These were two demographic predictors (age, sex), three oral health related predictors (number of teeth at baseline, denture use, chewing difficulty), three socioeconomic predictors (household income, employment category, years of formal education), two behavioural predictors (smoking, fruit and vegetable consumption), two psychological predictors (having a hobby, feeling worthless), a predictor related to community participation, and a predictor related to health service use (time since last health check-up).

Table 2 summarises the performance of all the machine learning models evaluated in this study. XGBoost models outperformed RF models in all three oversampling scenarios. However, RF performed slightly better with imbalanced data. XGBoost with balanced outcome classes had the best performance metrics with mean accuracy score of 90.5%±0.9%, F1 score of 91.0%±0.9, and ROC_AUC of 90.5%±0.9. Model performance appeared to decline when the oversampling ratio of the minority class was reduced.

Figure 4 visualises the behaviour of predictors within the XGBoost model in predicting tooth loss, using SHAP values. Figure 4(A) shows the average magnitude of SHAP values indicating the overall importance of each predictor within the model. Our model identified baseline age and the number of teeth as the primary drivers of tooth loss in older adults, in accordance with fundamental intuition. Figure 4(B) summarises the model behaviour for each local prediction (each dot represents an individual prediction), hence revealing the direction of effects at different levels of each predictor, such as, higher values of age (red) being related to higher risk of tooth loss and lower values of age (blue) being related to lower risk, and mid-level values of number of teeth (purple= 10-19 teeth category) being associated with higher risk of tooth loss compared to the other two categories (red: 1-9 teeth, blue: >=20 teeth). Figure 4(B) also revealed that having a denture, belonging to the manual occupation category, having a lower household income, fewer years of education, and chewing difficulties were associated with a higher risk of tooth loss being predicted by the machine learning model.

The effect of community activities, fruit and vegetable consumption, smoking status, leisure activities, and feeling worthless, was not apparent in a single dimension, as the XGBoost model captures complex interactions between variables. We analysed SHAP interaction values to capture some of the predictor interactions (reported in Figure 4(C) to Figure 4(G)). Figure 4(C) revealed that the risk of tooth loss among number of teeth categories varied based on denture use, i.e., wearing a denture while being in the >=20 teeth category was associated with an increased risk of tooth loss, while denture use reduced the risk of tooth loss for those in the lower number of teeth categories. Figure 4(D) showed that the effect of smoking status on tooth loss prediction was mainly driven by men except in the never smoked category. Similarly, interaction effects between 'participation in community activities and years of education', 'chewing difficulty and denture use', and 'fruit and vegetable consumption and number of teeth' are shown in Figures 4(E), 4(F), and 4(G) respectively.

Discussion

In this study, we explored predictors of the tooth loss among older adults over 6 years, using high-dimensional epidemiological data that contained more than 100 variables. Feature selection algorithm selected 14 important predictors that belonged to seven broader determinants associated with tooth loss among older adults. Although the predictor selection was done in a highly data driven manner, all selected predictors were theoretically associated with tooth loss, but have not been studied together in the same analysis. The machine learning model had over 90% accuracy score, F1 score and ROC_AUC in predicting tooth loss (Table 2). The SHAP value analysis provided useful insights regarding behaviour of the machine learning model in predicting tooth loss. Such as, the overall importance of individual predictor, direction of their effect in predicting tooth loss, and interactions between predictors (Figure 4).

Tooth loss at older age can be considered as an accumulative effect of life course oral health conditions. Therefore, with aging the risk of tooth loss should naturally increase. Our prediction model had behaved accordance with this basic intuition. Men had higher risk of tooth loss compared to women. Interaction analysis suggested that the sex difference was largely affected by smoking status, as the majority of the smokers were men. However, based on previous Japanese studies, oral hygiene behaviours and attendance to dental care services might also affect the observed sex difference in tooth loss among Japanese older adults (Cooray et al., 2020; Fukai et al., 1999). In addition to non-modifiable demographic factors (age, sex) and oral health condition at the baseline (number of teeth, denture use), socioeconomic factors such as lower household income, manual labour occupations, and lower education increased the risk of tooth loss. This highlights the importance of targeted preventive approaches for older people who belong to lower socioeconomic positions. Furthermore, the predictors identified in this study has shown to be associated many other health outcomes (Braveman and Gottlieb, 2014; Kojima et al., 2018; Montano, 2021). Therefore, tooth loss prevention could potentially be a consequence of reduced health inequalities and improved living conditions of older adults.

A couple of previous studies have used machine learning to predict tooth loss. Krois et al. (2019) used logistic regression, recursive partitioning, random forest, and extreme gradient boosting machine learning algorithms to predict tooth loss among patients with periodontitis using only tooth-level variables (Krois et al., 2019). Furthermore, their study was mainly focused on suggesting different validation strategies for tooth loss predictions. Elani et al. (2021) used cross-sectional data to assess the performance of multiple machine learning models in predicting tooth loss related outcomes using a variety of socioeconomic, self-reported dental care, and general health related predictors. Although the demographics of our sample population is different from the sample used in Elani et al. (adults vs older adults, the United States vs Japan), socioeconomic factors such as income, education, and employment were found to be important predictors of tooth loss in both studies. To the best of our knowledge, the current study is the first to present an explainable machine learning model for an oral health outcome in older adults using longitudinal data.

There are some potential limitations of our study. Tooth loss measure was based on self-reported number of teeth. Although self-reported number of teeth has been validated and used in many epidemiological studies (Peres et al., 2021), self-reports of number of teeth, especially among older adults, could lead to random errors. However, JAGES participants reported number of teeth as categories which might be less error prone compared to counting individual teeth. As the number of natural teeth can change only in one direction, we identified and excluded all invalid responses. On the other hand, having number of teeth only as a categorical variable presented another potential limitation as we could not detect tooth loss within the same categories or the exact number of teeth an individual would have lost when moving to a lower category. Hence, we acknowledge that incidence of tooth loss could potentially be underestimated in this study. However, we believe that the number of teeth categories measured in this study represent wider variations of overall oral health and oral functions for older adults (i.e.; between 1-9 teeth, 10-19 teeth, 20 or more teeth categories). Hence, it could be argued that detecting tooth loss at categorical level carries more weight from a public health and policy perspective.

The strengths of this study hinge on three aspects. First, the Boruta feature selection algorithm selected a set of highly relevant predictors that represent a wide range of oral health determinants. This suggests that the predictor selection procedure alone could be useful to identify new predictors related to the outcome. However, we acknowledge that the predictors selected by a data driven procedure might not be generalisable as the algorithm behaviour is dependent on the data. Therefore, only the theoretically plausible 119 variables were used to select most important predictors. Second, the XGBoost model predicted the incident tooth loss with a satisfactory level of accuracy. More importantly, the model had behaved in a manner that is agreeable with basic intuition and existing evidence, for example, risk of tooth loss increases with age and lower socioeconomic position, and is higher among men (Buchwald et al., 2013; Meisel et al., 2014). Third, the use of SHAP framework enabled visual interpretations of machine learning predictions without having to use statistical jargon. In this study we focused more on the interpretability as it is considered one of the main barriers to integrate machine learning methods in health outcome research (Schwendicke et al., 2020).

Accurate prediction regarding a health outcome is useful as long as we have actionable insights to prevent or achieve the intended outcome. Also, such insights must be easily communicated to clinicians in clinical settings, and to policy makers, social workers, and the general public in public health settings. In this study, we predicted future tooth loss among older adults using epidemiological data, and were able to extract actionable interpretations from a complex machine learning prediction algorithm. Hence, we believe this study has important practical clinical and policy implications. Furthermore, future gerontological research should try to adapt explainable machine learning methods in order to extract actionable evidence from powerful prediction algorithms.

Conclusions and Implications

Factors related to multiple domains contribute towards the tooth loss among older adults. In addition to oral health related and demographic factors, socioeconomic factors are important in predicting tooth loss. Therefore, understanding of the behaviour of these predictors can be useful in formulating prevention strategies for tooth loss among alder adults.

.

The authors declare no potential conflicts of interest with respect to the authorship and/or publication.

References

Aida, J., Hanibuchi, T., Nakade, M., Hirai, H., Osaka, K., Kondo, K., 2009. The different effects of vertical social capital and horizontal social capital on dental status: A multilevel analysis. Soc. Sci. Med. 69, 512–518. https://doi.org/10.1016/j.socscimed.2009.06.003

Bi, Q., Goodman, K.E., Kaminsky, J., Lessler, J., 2019. What is machine learning? A primer for the epidemiologist. Am. J. Epidemiol. 188, 2222–2239. https://doi.org/10.1093/aje/kwz189

Braveman, P., Gottlieb, L., 2014. The social determinants of health: It's time to consider the causes of the causes. Public Health Rep. 129, 19–31. https://doi.org/10.1177/00333549141291s206

Breiman, L., 2001. Statistical Modeling: The Two Cultures. Stat. Sci. 16, 199–231.

https://doi.org/10.1214/ss/1009213726

Buchwald, S., Kocher, T., Biffar, R., Harb, A., Holtfreter, B., Meisel, P., 2013. Tooth loss and periodontitis by socio-economic status and inflammation in a longitudinal population-based study. J. Clin. Periodontol. 40, 203–211. https://doi.org/10.1111/jcpe.12056

Bzdok, D., Krzywinski, M., Altman, N., 2017. Points of Significance: Machine learning: A primer. Nat. Methods. https://doi.org/10.1038/nmeth.4526

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cooray, U., Aida, J., Watt, R.G., Heilmann, A., Kato, H., Kiuchi, S., Kondo, K., Osaka, K., 2020. Effect of Copayment on Dental Visits : A Regression Discontinuity Analysis. J. Dent. Res. https://doi.org/10.1177/0022034520946022

Elani, H.W., Batista, A.F.M., Thomson, W.M., Kawachi, I., Filho, A.D.P.C., 2021. Predictors of tooth loss: A machine learning approach. PLoS One 16, e0252873. https://doi.org/10.1371/JOURNAL.PONE.0252873

Fukai, K., Takaesu, Y., Maki, Y., 1999. Gender Differences in Oral Health Behavior and General Health Habits in an Adult Population. Bull. Tokyo Dent. Coll. 40, 187–193. https://doi.org/10.2209/tdcpublication.40.187

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd editio. ed. Springer Science and Business Media LLC.

Huang, C., Murugiah, K., Mahajan, S., Li, S.-X., Dhruva, S.S., Haimovich, J.S., Wang, Y., Schulz, W.L., Testani, J.M., Wilson, F.P., Mena, C.I., Masoudi, F.A., Rumsfeld, J.S., Spertus, J.A., Mortazavi, B.J., Krumholz, H.M., 2018. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: A retrospective cohort study. PLOS Med. 15, e1002703. https://doi.org/10.1371/journal.pmed.1002703

James, S.L., Abate, D., Abate, K.H., Abay, S.M., Abbafati, C., Abbasi, N. et al. , 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. Lancet 392, 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7

Kassebaum, N.J., Bernabé, E., Dahiya, M., Bhandari, B., Murray, C.J.L., Marcenes, W., 2014. Global Burden of Severe Tooth Loss: A Systematic Review and Meta-analysis. J. Dent. Res. 93, 20S-28S. https://doi.org/10.1177/0022034514537828

Khaldy, M. Al, 2018. Resampling Imbalanced Class and the Effectiveness of Feature Selection Methods for Heart Failure Dataset. Int. Robot. Autom. J. 4. https://doi.org/10.15406/iratj.2018.04.00090

Kojima, G., Avgerinou, C., Iliffe, S., Jivraj, S., Sekiguchi, K., Walters, K., 2018. Fruit and Vegetable Consumption and Frailty: A Systematic Review. J. Nutr. Heal. Aging. https://doi.org/10.1007/s12603-018-1069-6

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., Hanhineva, K., 2019. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study. BMC Bioinformatics 20, 492. https://doi.org/10.1186/s12859-019-3110-0

Kondo, K., Rosenberg, M., WHO, 2018. Advancing universal health coverage through knowledge translation for healthy ageing: lessons learnt from the Japan gerontological evaluation study. World Health Organization.

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, 1088–1095. https://doi.org/10.1177/0022034519864889

Kursa, M.B., 2014. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics 15. https://doi.org/10.1186/1471-2105-15-8

Kursa, M.B., Rudnicki, W.R., 2010. Feature Selection with the Boruta Package. J. Stat. Software; Vol 1, Issue 11 36, 1–13. https://doi.org/10.18637/jss.v036.i11

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Guyon, I., Luxburg, U. V, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774.

Meisel, P., Holtfreter, B., Völzke, H., Kocher, T., 2014. Sex differences of tooth loss and obesity on systemic markers of inflammation. J. Dent. Res. 93, 774–779. https://doi.org/10.1177/0022034514535604

Montano, D., 2021. Socioeconomic status, well-being and mortality: a comprehensive life course analysis of panel data, Germany, 1984-2016. Arch. Public Heal. 79, 40. https://doi.org/10.1186/s13690-021-00559-7

Peres, M.A., Antunes, J.L.F., Watt, R.G., 2021. Oral Epidemiology, in: Oral Epidemiology. Springer International Publishing, pp. 223–233. https://doi.org/10.1007/978-3-030-50123-5

Peres, M.A., Macpherson, L.M.D., Weyant, R.J., Daly, B., Venturelli, R., Mathur, M.R., Listl, S., Celeste, R.K., Guarnizo-Herreño, C.C., Kearns, C., Benzian, H., Allison, P., Watt, R.G., 2019. Oral diseases: a global public health challenge. Lancet. https://doi.org/10.1016/S0140-6736(19)31146-8

Ravidà, A., Troiano, G., Qazi, M., Saleh, M.H.A., Lo Russo, L., Greenwell, H., Giannobile, W. V., Wang, H., 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, 1362–1370. https://doi.org/10.1111/jcpe.13362

Royal Society (G. B.), 2017. Machine learning : the power and promise of computers that learn by example, Report by the Royal Society.

Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. https://doi.org/10.1002/widm.1249

Schwendicke, F., Samek, W., Krois, J., 2020. Artificial Intelligence in Dentistry: Chances and Challenges. J. Dent. Res. 99, 769–774. https://doi.org/10.1177/0022034520915714

Schwendicke, F., Schmietendorf, E., Plaumann, A., Sälzer, S., Dörfer, C.E., Graetz, C., 2018. Validation of multivariable models for predicting tooth loss in periodontitis patients. J. Clin. Periodontol. 45, 701–710. https://doi.org/10.1111/jcpe.12900

Sidey-Gibbons, J.A.M., Sidey-Gibbons, C.J., 2019. Machine learning in medicine: a practical introduction. BMC Med. Res. Methodol. 19, 1–18. https://doi.org/10.1186/s12874-019-0681-4

Silva, M.F., Batista, M.J., da Luz Rosário de Sousa, M., 2019. Risk factors for tooth loss in adults: A population-based prospective cohort study. PLoS One 14, 1–14. https://doi.org/10.1371/journal.pone.0219240

Speiser, J.L., Callahan, K.E., Houston, D.K., Fanning, J., Gill, T.M., Guralnik, J.M., Newman, A.B., Pahor, M., Rejeski, W.J., Miller, M.E., 2020. Machine Learning in Aging: An Example of Developing Prediction Models for Serious Fall Injury in Older Adults. Journals Gerontol. Ser. A. https://doi.org/10.1093/gerona/glaa138

Stenholm, S., Shardell, M., Bandinelli, S., Guralnik, J.M., Ferrucci, L., 2015. Physiological Factors Contributing to Mobility Loss over 9 Years of Follow-Up - Results from the InCHIANTI Study. Journals Gerontol. - Ser. A Biol. Sci. Med. Sci. 70, 591–597. https://doi.org/10.1093/gerona/glv004

Wiemken, T.L., Kelley, R.R., 2019. Machine learning in epidemiology and health outcomes research. Annu. Rev. Public Health 41, 21–36. https://doi.org/10.1146/annurev-publhealth-040119-094437

Figure captions:

Figure 1. The flow of participant in the 6-year cohort to obtain the analytical sample.

Figure 2. A comparison of the preliminary performance of logistic regression, random forest classification, and XGBoost classification algorithms. All categorical variables were encoded with one hot encoding and standardised scaling was applied to continuous variables. Models were evaluated using 10-fold cross-validation. Confidence intervals for AUC are shown in grey shading.

Figure 3. A simplified illustration of repeated nested k-fold cross-validation procedure used to evaluate prediction models in this study.

Figure 4. A visual explanation of predictor behaviour based on the XGBoost prediction model. **A,** bar chart average global feature importance (separately for men and women) based on SHAP value magnitude. **B,** each dot represents an individual prediction**,** dot's position on the x-axis shows the impact that predictor has on the model's prediction for that individual. When multiple dots land at the same x position, they pile up to show density. The colour of the dot represents the level of the predictor related to that individual (ref: colour bar on the right). **C,** shows the interaction effect between the number of teeth and denture use on model predictions. Denture use among individuals in >=20 teeth category appeared to be increasing the risk of tooth loss. **D,** the interaction between smoking status and sex, indicating that smoking status effect higher tooth loss risk among men (ref: A & B). **E,** the effect of years of education is different for lower higher community activity categories compared to higher ones. **F,** individuals with chewing difficulty and dentures showed a higher risk of tooth loss and dentures posed a lower risk among no chewing difficulty individuals. **G,** shows different effects of the number of teeth on different fruit & veg consumption categories.

Tables

Table 1. Characteristics of the sample based on selected predictors and the tooth loss (N= 19,407)

Characteristics	Outcome (Incidence tooth loss) No (n= 16,394)	Outcome (Incidence tooth loss) Yes (n=3,013)
Age (mean±SD)	71.8±4.7	72.9±5.2
Sex		
Female	86.9%	13.1%
Male	81.7%	18.3%
Dental status		
>=20 teeth	88.3%	11.7%
10-19 teeth	75.9%	24.1%
<10 teeth	84.3%	15.7%
Denture use		

No	88.3%	11.7%
Yes	80.9%	19.1%
Chewing difficulty compared to 6 months ago		
No	87.8%	12.2%
Yes	80.3%	19.7%
Equivalent annual household income (yen)		
<1 million	82.2%	17.8%
1 to <1.5 million	82.0%	18.0%
1.5 to <2 million	84.8%	15.2%
2 to <3 million	85.3%	14.7%
3 to <5 million	85.3%	14.7%
>=5 million	86.2%	13.8%
Years of formal education		
<10 years	82.6%	17.4%
10-12 years	85.9%	14.1%
13 years or more	85.9%	14.1%
Longest occupation category		
Manual occupations	81.7%	18.3%
Sales/Services	84.7%	15.3%
Clerical work	87.6%	12.4%
Manager	85.2%	14.8%
Professional/Specialist	84.7%	15.3%
Smoking status		
Never smoked	86.2%	13.8%
Stopped >=5 years ago	83.7%	16.3%
Stopped <5 years ago	78.6%	21.4%
Current smoker	77.2%	22.8%
Fruits and vegetables consumption		
Once a week or less	77.8%	22.2%
2-3 times a week	81.0%	19.0%
4-6 times a week	80.4%	19.6%
Once a day	84.0%	16.0%
Twice or more a day	86.0%	14.0%
Time since the last health check-up		
within a year	85.4%	14.6%
within 2-3 years	83.7%	16.3%
> 4 years ago	83.7%	16.3%
never had	80.5%	19.5%
Engage in any leisure activities		
No	82.2%	17.8%
Yes	85.6%	14.4%
Having thoughts that you are not useful?		
No	85.34%	14.6%

Yes	81.5 %	18.5%
Participation in community groups/clubs		
Never	82.8%	17.2%
Few times a year	83.0%	17.0%
1-2 times a month	84.6%	15.4%
Once a week	86.7%	13.3%
> Once a week	86.1%	13.9%

Table 2. Mean performance metrics and class sizes of machine learning models evaluated in this study.

Classifier	Oversampling (N-$_{Majority}$ $=16394$)		Model performance metrics		
	Ratio (N_m/N_M)	N-minority	*Accuracy	*F1 Score	*ROC_AUC
XGBoost					
	1.0	16,394	90.5% (0.9%)	91.0% (0.9%)	90.5% (0.9%)
	0.75	12,295	89.2% (0.8%)	88.1% (0.9%)	89.7% (0.9%)
	0.5	8,197	88.1% (0.7%)	83.8% (0.9%)	89.2% (0.7%)
	None	3,013	73.2% (1.1%)	21.4% (1.8%)	53.0% (1.2%)
Random forest					
	1.0	16,394	82.7% (0.7%)	83.3% (0.7%)	82.7% (0.7%)
	0.75	12,295	81.9% (0.7%)	78.8% (0.9%)	81.4% (0.7%)
	0.5	8,197	79.9% (1.0%)	72.6% (1.1%)	79.9% (0.9%)
	None	3,013	71.6% (0.9%)	25.3% (1.9%)	55.0% (1.3%)

N-minority (Nm)= Number of people with tooth loss
N-Majority (NM)= Number of people with no tooth loss
*mean performance values of 100 independent models evaluated using repeated nested cross-validations are reported in the table.