Clinical prediction modelling in oral health: A review of study quality and empirical examples of model development

A thesis

Submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy

by

Mi Du

Supervisors:

Dr. Murthy Mittinty

and

Dr. Dandara Haag, Prof. John Lynch

School of Public Health

The University of Adelaide

August, 2021

Table of Contents

Abstract

Background

Substantial efforts have been made to improve the reproducibility and reliability of scientific findings in health research. These efforts include the development of guidelines for the design, conduct and reporting of preclinical studies (**ARRIVE**), clinical trials (**ROBINS-I**, **CONSORT**), observational studies (**STROBE**), and systematic reviews and meta-analyses (**PRISMA**). In recent years, the use of prediction modelling has increased in the health sciences. Clinical prediction models use information at the individual patient level to estimate the probability of a health outcome(s). Such models offer the potential to assist in clinical decision-making and to improve medical care. Guidelines such as **PROBAST** (Prediction model Risk Of Bias Assessment Tool) have been recently published to further inform the conduct of prediction modelling studies. Related guidelines for the reporting of these studies, such as **TRIPOD** (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) instrument, have also been developed.

Since the early 2000s, oral health prediction models have been used to predict the risk of various types of oral conditions, including dental caries, periodontal diseases and oral cancers. However, there is a lack of information on the methodological quality and reporting transparency of the published oral health prediction modelling studies. As a consequence, and due to the unknown quality and reliability of these studies, it remains unclear to what extent it is possible to generalise their findings and to replicate their derived models. Moreover, there remains a need to demonstrate the conduct of prediction modelling studies in oral health field following the contemporary guidelines. This doctoral project addresses these issues using two systematic reviews and two empirical analyses. This thesis is the first comprehensive and systematic project reviewing the study quality and demonstrating the use of registry data and longitudinal cohorts to develop clinical prediction models in oral health.

Aims

- To identify and examine the quality of existing prediction modelling studies in the major fields of oral health.

- To demonstrate the conduct and reporting of a prediction modelling study following current guidelines, incorporating machine learning algorithms and accounting for multiple sources of biases.

Methods

As one of the most prevalent oral conditions, chronic periodontitis was chosen as the exemplar pathology for the first part of this thesis. A systematic review was conducted to investigate the existing prediction models for the incidence and progression of this condition. Based upon this initial overview, a more comprehensive critical review was conducted to assess the methodological quality and completeness of reporting for prediction modelling studies in the field of oral health. The risk of bias in the existing literature was assessed using the **PROBAST** criteria, and the quality of study reporting was measured in accordance with the **TRIPOD** guidelines.

Following these two reviews, this research project demonstrated the conduct and reporting of a clinical prediction modelling study using two empirical examples. Two types of analyses that are commonly used for two different types of outcome data were adopted: survival analysis for censored outcomes and logistic regression analysis for binary outcomes. Models were developed to 1) predict the three- and five-year disease-specific survival of patients with oral and pharyngeal cancers, based on 21,154 cases collected by a large cancer registry program in the US, the Surveillance, Epidemiology and End Results (SEER) program, and 2) to predict the occurrence of acute and persistent pain following root canal treatment, based on the electronic dental records of 708 adult patients collected by the National Practice-Based Research Network. In these two case studies, all prediction models were developed in five steps: (i) framing the research question; (ii) data acquisition and pre-processing; (iii) model generation; (iv) model validation and performance evaluation; and (v) model presentation and reporting. In accordance with the **PROBAST** recommendations, the risk of bias during the modelling process was reduced in the following aspects:

- In the first case study, three types of biases were taken into account: (i) bias due to missing data was reduced by adopting compatible methods to conduct imputation; (ii) bias due to unmeasured predictors was tested by sensitivity analysis; and (iii) bias due to the initial choice of modelling approach was addressed by comparing tree-based machine learning algorithms (survival tree, random survival forest and conditional inference forest) with the traditional statistical model (Cox regression).
- In the second case study, the following strategies were employed: (i) missing data were addressed by multiple imputation with missing indicator methods; (ii) a multilevel logistic regression approach was adopted for model development in order to fit

the hierarchical structure of the data; (iii) model complexity was reduced using the Least Absolute Shrinkage and Selection Operator (LASSO) for predictor selection; and (iv) the models' predictive performance was evaluated comprehensively by using the Area Under the Precision Recall Curve (AUPRC) in addition to the Area Under the Receiver Operating Characteristic curve (AUROC); (v) finally, and most importantly, given the existing criticism in the research community concerning the gender-based and racial bias in risk prediction models, we compared the models' predictive performance built with different sets of predictors (including a clinical set, a sociodemographic set and a combination of both, the 'general' set).

Results

The first and second review studies indicated that, in the field of oral health, the popularity of multivariable prediction models has increased in recent years. Bias and variance are two components of the uncertainty (e.g., the mean squared error) in model estimation. However, the majority of the existing studies did not account for various sources of bias, such as measurement error and inappropriate handling of missing data. Moreover, non-transparent reporting and lack of reproducibility of the models were also identified in the existing oral health prediction modelling studies. These findings provided motivation to conduct two case studies aimed at demonstrating adherence to the contemporary guidelines and to best practice.

In the third study, comparable predictive capabilities between Cox regression and the non-parametric tree-based machine learning algorithms were observed for predicting the survival of patients with oral and pharyngeal cancers. For example, the C-index for a Cox model and a random survival forest in predicting three-year survival were 0.82 and 0.84, respectively. A novelty of this study was the development of an online calculator designed to provide an open and transparent estimation of patients' survival probability for up to five years after diagnosis. This calculator has clinical translational potential and could aid in patient stratification and treatment planning, at least in the context of ongoing research. In addition, the transparent reporting of this study was achieved by following the **TRIPOD** checklist and sharing all data and codes.

In the fourth study, LASSO regression suggested that pre-treatment clinical factors were important in the development of one-week and six-month postoperative pain following root canal treatment. Among all the developed multilevel logistic models, models with a clinical set of predictors yielded similar predictive performance to models with a general set of predictors, while the models with sociodemographic predictors showed the weakest predictive ability. For example, for predicting one-week postoperative pain, the AUROC

for models with clinical, sociodemographic and general predictors were 0.82, 0.68 and 0,84, respectively, and the AUPRC were 0.66, 0.40 and 0.72, respectively.

Conclusion

The significance of this research project is twofold. First, prediction models have been developed for potential clinical use in the context of various oral conditions. Second, this research represents the first attempt to standardise the conduct of this type of studies in oral health research. This thesis presents three conclusions: 1) Adherence to contemporary best practice guidelines such as **PROBAST** and **TRIPOD** is limited in the field of oral health research. In response, this PhD project disseminates these guidelines and leverages their advantages to develop effective prediction models for use in dentistry and oral health. 2) Use of appropriate procedures, accounting for and adapting to multiple sources of bias in model development, produces predictive tools of increased reliability and accuracy that hold the potential to be implemented in clinical practice. Therefore, for future prediction modelling research, it is important that data analysts work towards eliminating bias, regardless of the areas in which the models are employed. 3) Machine learning algorithms provide alternatives to traditional statistical models for clinical prediction purposes. Additionally, in the presence of clinical factors, sociodemographic characteristics contribute less to the improvement of models' predictive performance or to providing cogent explanations of the variance in the models, regardless of the modelling approach. Therefore, it is timely to reconsider the use of sociodemographic characteristics in clinical prediction modelling research. It is suggested that this is a proportionate and evidence-based strategy aimed at reducing biases in healthcare risk prediction that may be derived from gender and racial characteristics inherent in sociodemographic data sets.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Mi Du

Date: August 21, 2021

Acknowledgements

I am lucky to have an exceptional supervisory panel. I deeply appreciate the opportunity for the systematic research training they have offered me. It is memorable that during the difficult time due to COVID-19, they were always there to provide sufficient support to ensure my research goes smoothly. This thesis would not have reached the present form without the support and contribution of these wonderful individuals.

With immense pleasure and deep sense of gratitude, I wish to, first and foremost, express my sincere thanks to my principle supervisor Dr. Murthy Mittinty, for his continual guidance and motivation during my PhD studies. His initiative, intellect, and patience inspired the exploration of my interests and helped me navigate the challenging waters of academic research. Murthy has walked me through each stage of writing this thesis. Without his consistent encouragement and illuminating instruction, this research would not have been successfully completed. Murthy, thank you for your supervision and mentoring - your door has always been open throughout my candidature. I still remember, every time I thank you, you always say 'it is my job', to me, 'your job' is greatly appreciated!

I am extremely grateful to the head of *BetterStart* research group, Prof. John Lynch, for offering me an opportunity to carry out my research in the *BetterStart* research group. His persistent drive in pursuit of high-quality research kept me focused when my frustrations mounted. And his insightful feedbacks motivated me to sharpen my thinking and brought my work to a higher level. John, thank you for your dedication to crafting exemplary written, visual, and oral presentations of research. I take you as an idol in my research career!

I am also indebted to Dr. Dandara Haag for her kind words of support along with patience and understanding in these past three years. Dandara, thank you for your invaluable comments, suggestions and contributions throughout the PhD. I have learnt a lot from you!

My sincere thanks also goes to my previous supervisors in the Australian Research Center for Population Oral Health (ARCPOH) Prof. Marco Peres and Dr. Kostas Kapellas

for leading me to the world of clinical prediction modelling.

Thank you to the *BetterStart* group for providing an environment where researchers can flourish. A special thanks goes to Dr. Janet Grant for reading my documents and revising the language. I greatly appreciate it. Also, thank you Dr. Angela Gialamas and Dr. Rhiannon Pilkington for offering me opportunities to work as a teaching assistant and research assistant. With this group, I have developed my critical thinking and learnt to ask research questions with courage. I give my best wishes to everyone in *BetterStart*.

I was lucky to have supportive fellow PhD students and colleagues in the School of Public Health - Dr. Engida Yisma, Dr. Mumtaz Begum, Alexandra Procter, Dr. Razlyn Rahim, Jasmine Liu, Dr. Zhidong Liu, Dr. Micheal Tong, Dr. Jianjun Xiang and Dr. Blesson Varghese, as well as those at Adelaide Dental School - Dr. Rahul Nair, Dr Youngha Song, Dr. Sneha Sethi, Dr. Arash Ghanbarzadegan, and Dr. Pedro Santiago. Thank you for the walk, coffee, and chats. I must give them my best wishes.

My close friends in Adelaide - Xuyi Wang, Dr. Xiaolong You, Dr. Yasi Wang, Cong Xie, Menghe Liu, Dr. Sathvika Justin, Constanza Roa Vaca and whose names I may have missed out, I thank you for your support, time, laughter, stimulating discussions, encouragement, and for all the fun we had. A special thanks to Dr. Yuanqiu Mo for helping to fix the technical issues in LaTeX. I wish you all the bests in the future. Also, thank you to the Capstone Editing group for correcting the language for this thesis.

My family - you have been there all along. I extend my profound sense of gratitude to my parents Mr. Peicheng Du, Mrs. Haiyan Liu, and my sister Ms. Tiantian Du for your moral support and for all the sacrifices you have made during my PhD candidature. I look forward to celebrating and spending more time with you after submission.

And finally, I am very grateful for the excellent opportunity given to me through the provision of Adelaide University China Fee Scholarship.

A big thank you to everyone who has been a part of my PhD journey in one way or the other. This journey is indeed a life changing experience and I consider the skills I have gained to be an asset for my future endeavours.

Mi Du
Adelaide
August 21, 2021

To my beloved family

Publications contributing to this thesis

Published papers

1. **Du M.**, Haag, D. G., Lynch, J. W., Mittinty, M. N. Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on SEER database. *Cancers*. 2020;12(10), 2802. https://doi.org/10.3390/cancers12102802.

2. **Du M**, Haag D, Song Y, Lynch J, Mittinty M. Examining bias and reporting in oral health prediction modelling studies. *Journal of Dental Research*. 2020;99(4):374-387. https://doi.org/10.1177/0022034520903725.

3. **Du M**, Bo T, Kapellas K, Peres M. Prediction models for the incidence and progression of periodontitis: A systematic review. *Journal of Clinical Periodontology*. 2018;45:1408-1420. https://doi.org/10.1111/jcpe.13037.

Submitted paper(s)

1. **Du M**, Haag D.G., Lynch J.W., Mittinty M.N. Application of multilevel machine learning models for predicting pain following root canal treatment. *Journal of Dentistry*. (Submitted)

Presentations arising from this thesis

Oral presentations

1. **Du M**, Haag D, Song Y, Lynch J, Mittinty M. Examining bias and reporting transparency in prediction modelling studies in oral health research. 99^{th} General Session and Exhibition of the International Association for Dental Research (IADR), 18^{th} March 2020, Washington D.C., USA. (Invited but cancelled)

2. **Du M**, Haag D, Lynch J, Mittinty M. The application of machine learning algorithms in predicting survival of oral and pharyngeal cancers: Analyses based on SEER database. 2020 IADR Epi-forum, 16^{th} March 2020, Washington D.C., USA. (Invited but cancelled)

3. **Du M**, Haag D, Song Y, Lynch J, Mittinty M. Examining bias and reporting transparency in prediction modelling studies in oral health research. Australasian Epidemiological Association (AEA) Annual Scientific Meeting, 23-25^{th} October 2019, Brisbane Convention Exhibition centre, Brisbane, Australia.

4. **Du M**. Three Minutes Thesis (3MT) Competition, 24^{th} August 2018, AHMS building, Adelaide, Australia.

5. **Du M**, Bo T, Kapellas K, Peres M. Prediction models for the incidence and progression of periodontitis: a systematic review. Adelaide Dental School Research Day, 10^{th} August 2018, National Wine Centre, Adelaide, Australia.

6. **Du M**, Kapellas K, Peres M. Predicting the 12-year risks of incidence and progression of periodontitis in Australian adults: A nationwide cohort study (A PhD Research Proposal). ARCPOH Friday Seminar, AHMS building, 1^{st} June 2018, Adelaide, Australia.

Poster presentations

1. **Du M**, Haag D, Lynch J, Mittinty M. What are the survival chances for people like myself? 14^{th} Annual Florey Postgraduate Research Conference, 30^{th} September 2020 (virtual event).

2. **Du M**, Haag D, Song Y, Lynch J, Mittinty M. Examining bias and reporting transparency in prediction modelling studies in oral health research. 13^{th} Annual Florey Postgraduate Research Conference, 24^{th} September 2019, National Wine Centre, Adelaide, Australia.

3. **Du M**, Bo T, Kapellas K, Peres M. Prediction models for the incidence and progression of periodontitis: a systematic review. 12^{th} Annual Florey Postgraduate Research Conference, 25^{th} September 2018, National Wine Centre, Adelaide, Australia.

Awards arising from this thesis

1. Robinson Research Institute High Impact Paper Funding, The University of Adelaide, October 2020.
2. The top 10 % most downloaded papers in *Journal of Clinical Periodontology* 2018-2019. Wiley, April 2020.
3. The DR Stranks Travelling Scholarship, The University of Adelaide, November 2019.
4. Adelaide Dental School Prize, 12^{th} Annual Florey Postgraduate Research Conference, The University of Adelaide, September 2018.
5. Adelaide University China Fee Scholarship, The University of Adelaide & China Scholarship Council, October 2017.

List of Figures

List of Tables

List of Abbreviations

AJCC	American Joint Committee on Cancer
AUPRC	The Area Under the Precision-Recall Curve
AUROC	The Area Under the Receiver Operating Characteristic curve
ARRIVE	Animal Research: Reporting of In Vivo Experiments
BMI	The Body Mass Index
CCA	Complete Case Analysis
CDC/AAP	The Centers for Disease Control and Prevention - American Association of Periodontology
CHARMS	CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies
CONSORT	Consolidated Standards of Reporting Trials
DPBRN	The Dental Practice-Based Research Network
EFP	The European Federal of Periodontology
HIV	Human Immunodeficiency Virus
HPV	The Human Papilloma Virus
ICD	The International Disease Classification
LASSO	The Least Absolute Shrinkage and Selection Operator
MAR	Missing At Random
MCAR	Missing Completely At Random
MI	Multiple Imputation
MICE	Multiple Imputation with Chained Equation
MNAR	Missing Not At Random
MSE	The Mean Squared Error
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST	Prediction model Risk Of Bias ASsessment Tool
ROBINS-I	Risk Of Bias In Non-randomised Studies - of Interventions
SEER	The Surveillance, Epidemiology, and End Results program
STROBE	The Strengthening the Reporting of Observational studies in Epidemiology
TNM	Tumour size, lymph Node involved, Metastasis
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis
US	The United States of America

CHAPTER 1

Introduction

Preface

It has long been recognised that prediction models hold the potential to benefit clinical decision-making, health care policy making and to improve patients' health outcomes. Though prediction modelling represents an important element in the armamentarium of oral health promotion strategies, the clinical use of such models remains sparse due to their unknown reliability and reproducibility. Providing comprehensive insights into these aspects of clinical prediction modelling studies in the oral health context forms the basis of this PhD research.

Section 1.1 of this chapter introduces the background to the body of research underpinning this thesis. Section 1.2 summarises the current context and thereby identifies gaps in the existing body of knowledge. Section 1.3 provides the research questions and aims of the current research. Section 1.4 presents the original contributions of the project to the field. Finally, this chapter ends by providing an outline of the thesis structure in section 1.5.

1.1 Background

Oral diseases represent a major public health concern worldwide. In 2019, *The Lancet* published two articles (Peres et al., 2019; Watt et al., 2019) calling for radical actions to end the global neglect of oral health. The urgency and importance of this call were underpinned by the high prevalence and inequalities in oral diseases, and the adverse impacts they bring to the quality of life for afflicted individuals. Oral diseases include a variety of acute, aggressive and chronic oral conditions that affect the teeth, tongue, mouth and surrounding tissues. Globally, the top three most prevalent oral diseases are dental caries, periodontal diseases, and oral cancers (James et al., 2018). To tackle this global health challenge, there is a need for a range of oral health care strategies, including prevention, early detection and specific, appropriate treatment regimes. These activities benefit from prediction modelling research because prediction can inform the risk of developing a disease or estimate the outcomes of a course of treatment, thereby assisting in clinical decision-making. For example, prediction can help oral health professionals to identify populations who are at high risk of developing oral diseases and inform the appropriate response in framing early intervention opportunities.

Clinical prediction models (also known as risk prediction models, clinical prediction rules) have been known for more than 30 years in health and medical research. A typical example is QRISK (Anderson et al., 1991), a model used for predicting cardiovascular risk and assisting clinicians in identifying and formulating early interventions suitable for individual patients. In the field of oral health, clinical prediction models also play an important role and have garnered considerable interest from both researchers and oral health care practitioners. Using these data-driven predictive tools, oral health professionals seek to achieve better diagnoses and prognoses for a number of oral diseases, including dental caries (Abernathy et al., 1987), periodontal diseases, (Lai et al., 2015), tooth loss (Krois et al., 2019), fracture of dental ceramics (Ren and Zhang, 2014), and oral cancers (Kim et al., 2019).

Clinical prediction models can be categorised into diagnostic and prognostic variants (Steyerberg et al., 2019). A diagnostic model is developed to estimate the risk of having a disease. A prognostic model is used to predict the probability of the progression and prognosis of a disease (e.g., cancer survival). A prediction model allows questions such as: '*What is the risk of achieving a specific health outcome for a population at a specific time, given the data available at that time?*' to be answered. From this question, it can be seen that there are four key components of a clinical prediction model: a target population, a specific health outcome to be predicted, information characterising the population, and an

underpinning algorithmic mathematical or statistical model. These four elements include the following characteristics:

- Target population: The population of interest.
- Outcome: What are we predicting?
- Available information used: This usually includes an individual's demographics, clinical characteristics, laboratory test results, and other variables of individual members of the target population. Various sources of data can be used in prediction modelling research. These include longitudinal cohorts, cross-sectional and case-control studies, clinical trials and electronic medical records.
- Modelling approach: A mathematical/statistical/data dependent algorithm that describes how the model outcomes have been generated from available data. These approaches to modelling can be broadly classified into two categories: statistical models and machine learning algorithms.

Once a model is developed, its predictive outcomes must be communicated to other researchers and potential users, including clinicians and data custodians. A common way to present such a clinical prediction model is to report the models' characteristics (e.g., its variables' corresponding coefficients) along with its predictive ability (i.e., how well the model can predict the outcome). The performance/capability of a prediction model can be evaluated using a range of metrics, including the models' ability to distinguish the positive from the negative observations (referred to as 'discrimination'), and its ability to achieve a predicted probability as close as possible to the observed ones (referred to as 'calibration'). Therefore, the primary goal of a prediction modelling study is to optimise the performance metrics used in the evaluation as determined by the reduction in the squared difference between the predicted and the observed values (referred to as mean square error, MSE). The MSE can be described as the total of $Bias^2 + Variance$. By reducing the MSE, both the bias and the variance are minimised. Keeping the MSE low may, in the absence of any additional contributions from the unmeasured data, enhance the generalisability and applicability of prediction models and their outputs from one study to another.

Once the model and its predictive ability are presented, assessing the quality, reliability, and clinical impact of the model become important for other researchers and end users of the model (e.g., clinicians and policy makers). These assessments not only help to answer questions such as: '*To what extent can we trust the developed models?*', but also encourage model developers to consider a range of aspects which may improve the quality of their prediction modelling studies. These aspects include (but are not limited to): data quality, appropriate approaches used for capturing the relationship between the

variables, models' interpretability, models' ease of use and result interpretation, and the study reproducibility.

- Data quality concerns the consistency of collection and coding of the variables in multiple data sets. This includes defining data dictionaries, using consistent definitions and measurement systems for variable values. If the collected data has subtle biases (e.g., measurement error), then the models will incorporate these errors and generate biased estimates as a consequence. Additionally, data quality also relates to data completeness and addresses questions concerning the proportion of, and reasons for, missing information. Missing data occur when data values are not observed and recorded for any given variable in a sample. Multiple studies have shown that missing data may compromise the validity of study findings in clinical trials (Little et al., 2012), observational cohort studies (Howe et al., 2016), and studies using electronic health records (Petersen et al., 2019; Stiglic et al., 2019).

- Modelling approaches concern the model's complexity (e.g., how long does it take to train the model?), the model's interpretability (e.g., can every step of the model be understood?), the model's ability to reflect the data generating mechanism (e.g., what, if any, are the pre-defined assumptions relating to the relationships between variables?).

- The reproducibility of a study concerns the transparent reporting of the conduct of a study and how easily it may be replicated. These considerations relate to the various stages of modelling, including the selection procedure for participants, the definition of predictors and outcomes, the consistent measurement of variables and their values, consideration and accommodation of missing data, methods and codes used in statistical analyses, and the methods used for evaluating the models' performance evaluation.

As has been observed over recent decades, it is not unusual for clinical prediction models to outperform human clinical judgement alone (Ægisdóttir et al., 2006; Wiggins, 1981; Zellner et al., 2021). The capacity of prediction models to assist with clinical activities has led to a continual increase in the number of publications discussing prediction models (Steyerberg et al., 2019). As shown in Figure 1.1, in the 1960s, there were fewer than 10 publications listed on PubMed with the terms 'prediction model' or 'prognostic model', as compared to over 3500 identified in 2020.

1.2 Gaps in the existing research

Though the number of prediction models is increasing, the implementation of these models in clinical practice has remained stagnant. Potential reasons, reported in the literature

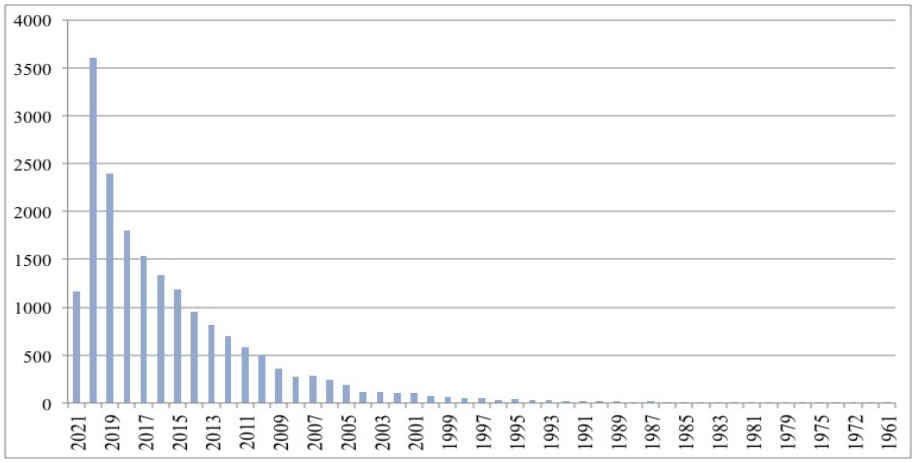


Figure 1.1: Number of publications per year listed in PubMed searched by title and abstract 'prediction model' OR 'prognostic model' OR 'diagnostic model' up to 2021 (searching was conducted on 18^{th} March 2021).

such as *The Lancet - Digital Health* (Futoma et al., 2020) and *Nature - Digital Medicine* (Sutton et al., 2020), include inter alia: (i) poor generalisability and predictive performance with respects to discrimination and calibration when used with new populations, (ii) the paucity of evidence to indicate models can improve patients' health outcomes and aid in clinical decision-making, and (iii) doubt surrounding the quality of the studies used to develop the models. This thesis aims to address the third concern. The first research gap identified for this PhD project is the dearth of evidence addressing the quality of the published prediction modelling studies in the field of oral health.

The second research gap is the need to ensure the quality of prediction modelling studies. In recent years, researchers have developed methodological guidelines such as **CHARMS** (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) (Moons et al., 2014) and **PROBAST** (Prediction model Risk Of Bias ASsessment Tool) for conducting prediction modelling studies (Moons et al., 2019). Additionally, reporting checklists such as **TRIPOD** (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) (Collins et al., 2015) have also been developed to guide researchers on the reporting of findings from this type of study. However, the use of these guidelines is sparse in oral health prediction modelling research. Implementation of the **PROBAST** strategy suggests there is a need to address the common types of biases arising from prediction modelling studies which include, for example, measurement error and anomalies arising from missing data.

1.3 Aims

The overall purpose of this research was to establish a better understanding of prediction modelling studies in oral health, thereby contributing to new knowledge that may be

useful in dental practice. This thesis presents the findings of the research conducted in four separate but interrelated studies.

Studies 1 and 2 comprise the first half of this thesis and formed the basis of the remaining chapters. Studies 3 and 4 form the second half of this thesis and demonstrate the incorporation of multiple sources of bias and the applications of machine learning methods in oral health prediction. The research aims and objectives are:

Aim of Study 1: To investigate the quality of prediction modelling studies in periodontology, one of the major disciplines of oral health.

Aim of Study 2: To identify the best available evidence addressing the overall quality of oral health prediction modelling research and to present optimal contemporary methodological standards and reporting guidelines for conducting prediction modelling studies.

- Objective 1: to examine the risk of bias of recent oral health prediction modelling studies, using the **PROBAST** instrument.
- Objective 2: to examine the completeness and transparency of reporting of these studies using the **TRIPOD** instrument.

Aim of Study 3: To account for some of the identified biases (missing data, unmeasured predictors) in the existing literature, and to demonstrate the conduct of a prediction modelling study designed to predict the survival probability for patients with oral and pharyngeal cancers.

- Objective 1: to develop prediction models and a web-based calculator for estimating the three- and five- year disease-specific survival of oral and pharyngeal cancer patients, using cancer registry data.
- Objective 2: to compare the predictive capability of Cox proportional hazard regression, a traditional statistical method, with machine learning methods (e.g., survival tree and random survival forest).

Aim of Study 4: To account for some of the potential biases in model development when using multilevel oral health data by conducting a predictive case study of predicting acute and persistent pain following root canal treatment.

- Objective 1: to use a prospective cohort data set (containing sociodemographic and clinical characteristics) to develop models for predicting the postoperative pain for patients who received root canal treatment.

- Objective 2: to demonstrate the conduct of predictor selection using multilevel machine learning algorithms.
- Objective 3: to demonstrate the use of precision-recall curves for assessing the performance of prediction models when the distribution of outcome data is imbalanced.
- Objective 4: to increase the research focus on effective strategies to incorporate sociodemographic variables when developing clinical prediction models.

1.4 Original contributions

The major contributions of this research work can be outlined as follows:

1. The best available evidence in prediction modelling research in the major fields of oral health has been identified and collated. Moreover, the quality of existing studies has been evaluated by applying two validated and accepted quality assessment tools to these studies.
2. This work has led to the publication of a number of relevant papers and presentations at (inter)national conferences. These focus attention on the need for researchers to improve the quality (including reliability and reproducibility) of oral health prediction modelling studies.
3. This research clearly identifies that multiple types of biases (e.g., inappropriate handling of missing data, unmeasured predictors, modelling uncertainty) can arise during the prediction modelling process. Using two empirical case studies, this research demonstrates how these biases can be avoided in prediction modelling and how the results from such analyses can be presented and reported.
4. Models have been developed for predicting the outcome of various oral diseases, such as the disease-specific survival of oral and pharyngeal cancers and pain following a common endodontic treatment. This thesis describes the application of machine learning algorithms for oral health prediction purposes and has provided a potentially useful translation of the research findings into clinical practice by developing a user-friendly, web-based application.

1.5 Thesis Structure

This thesis is a result of a PhD project funded by the Adelaide University China Fee Scholarship at the University of Adelaide, carried out by Mi Du under the supervision of Dr. Murthy Mittinty, Dr. Dandara Haag, and Prof. John Lynch.

This thesis has been structured in publication format and comprises eight chapters. Peer-reviewed papers published or submitted for publication have been included in relevant chapters. Additional chapters including the Introduction, Research Context, Methodology, Discussion and Conclusion have been provided to give the readers a clear description of the research undertaken.

The remainder of this thesis is organised as follows:

- Chapter 2 provides a comprehensive review of prediction epidemiology in oral health and reviews the commonly used practices and their limitations for the development of clinical prediction models.
- Chapter 3 describes the various research methods used in each of the four studies in this thesis, as well as the data sources used in the empirical analysis.
- Chapter 4 identifies the existing prediction models in periodontology, one of the major oral health areas. A systematic review published in the *Journal of Clinical Periodontology* is included in this chapter.
- Chapter 5 examines the overall quality (including methodological bias and reporting transparency) of prediction modelling studies in oral health. A review article published in the *Journal of Dental Research* is included in this chapter.
- Chapter 6 reports the use of multiple statistical and tree-based machine learning methods to develop prediction models for estimating the three- and five-year survival probability of patients with oral and pharyngeal cancers. An original research article published in *Cancers* is included in this chapter.
- Chapter 7 presents the use of multilevel machine learning models, incorporating sociodemographics and clinical information, to predict acute and persistent pain following root canal treatment. This chapter is presented in publication format submitted to the *Journal of Dental Research*.
- Chapter 8 presents a discussion of the main findings, strengths, limitations, and implications of the research, as well as an overall conclusion.
- This thesis ends with the inclusion of relevant appendices.

CHAPTER 2

Research Context

Preface

This chapter presents a summary of descriptive epidemiology in the context of major oral health conditions along with a detailed literature review of predictive epidemiology in the oral health field.

There are eight sections in this chapter. Section 2.1 presents a brief review of the major burden of the three most prevalent oral diseases – dental caries, periodontal diseases, and oral cancers. Sections 2.2 and 2.3 introduce the general concepts of clinical prediction modelling research, followed by their contributions to oral health promotion in Section 2.4. Section 2.5 reviews the step-by-step procedure for conducting a clinical prediction modelling study: from data preparation to model performance evaluation. Of special interest, with the advancement of machine learning methods, we also look at the current application and evolution of machine learning algorithms in oral health-related prediction research in Section 2.6. Section 2.7 provides guidelines on choosing between statistical models and machine learning algorithms for prediction purposes. We then investigate the potential limitations of the commonly used practices in clinical prediction modelling in Section 2.8. In alignment with the guidelines - **PROBAST** instrument, this section starts by understanding potential bias during clinical prediction modelling (e.g., selection bias, measurement error, overfitting/underfitting) and ends by discussing current efforts and possible solutions for reducing these biases.

KEY QUESTIONS

- What is the major burden of oral diseases worldwide?
- What role does prediction modelling research play in promoting oral health?
- How are clinical prediction models developed?
- What are the limitations of the existing methods for conducting clinical prediction modelling?
- Why does bias matter in prediction modelling?
- How does bias arise during prediction modelling?
- What are the common sources of bias in prediction modelling study?
- How can researchers reduce these biases?

2.1 Overview of oral diseases burden

Most oral diseases do not cause death but have been identified as a major public health burden worldwide due to their high prevalence and adverse impact on the quality of life. The top three most prevalent oral conditions are periodontal diseases, dental caries (tooth decay), and oral cancers. According to the estimates by the Global Burden of Disease study 2017 (James et al., 2018):

- About 2.3 billion people (1/3 of the global population) suffer from caries of permanent teeth. The plaque on the surface of a tooth can lead to dental caries and destroy the teeth over time. Risk factors for dental caries include bacteria deposits, poor oral hygiene, frequent exposure to sugar contained diets, developmental defects of tooth enamel, family history of caries, poor oral health habits (e.g., inadequate tooth brushing or dental care).

- Two common types of periodontal disease are gingivitis and periodontitis. Periodontal diseases affect the tissues that surround the teeth, therefore they are usually characterised by swollen or bleeding gums, causing teeth to become loose and even fall out. Gingivitis is a common oral condition in children and adolescents and it can be prevented by improving oral hygiene. Periodontitis has several forms, with aggressive periodontitis occurring more commonly in children and adolescents while chronic periodontitis is more prevalent in adults. Severe periodontitis is affecting almost 10% of the global population. The common risk factors for periodontal diseases include smoking, systemic conditions such as diabetes, stress, and human immunodeficiency virus (HIV) (Scannapieco and Gershovich, 2020).

- Oral and pharyngeal cancers represent the eighth most prevalent cancer worldwide among the male population and are one of the three most common cancers in some countries of Asia and the Pacific. According to the American Cancer Society, the estimated new cases and deaths due to oral oropharyngeal cancers are 54,010 and 10,850 in the US for 2021 (American Cancer Society, 2021). Based on the International Disease Classification (ICD) 10^{th} edition, oral and pharyngeal cancers (ICD C00-C14) usually include cancers of the lip, tongue, salivary glands, oropharynx, hypopharynx, nasopharynx, and other surrounding sites in the mouth. Risk factors for oral and pharyngeal cancers include smoking, tobacco consumption, alcohol consumption, viruses such as human papillomavirus (HPV), and recurrent oral inflammation (Shield et al., 2017).

Despite considerable progress being made to understand the causes of the major oral diseases, the burden of oral conditions worldwide persists and may have even become more severe (Kassebaum et al., 2017; Watt et al., 2019). To reverse the trend, clinical dentistry adopts a treatment/intervention-dominated approach to care. All of these interventions and clinical activities rely on evidence-based dentistry, which requires joint efforts from dental practitioners, epidemiologists, biostatisticians, ethicists, and others. These efforts have multiple aims, covering the causal exploration, disease prevention, early detection, treatment decision, and treatment effect evaluation. Among most of those aims, predictive modelling research can play important roles, such as identifying the risk factors for a disease, early detection of the occurrence of an adverse health outcome, and predicting the treatment effect of a specific therapy.

2.2 What are clinical prediction models?

Clinical prediction models (also named clinical prediction rules or risk scores) are mathematically derived tools that predict the health outcomes of interest and inform evidence-based clinical decisions (Kuhn et al., 2013). Since 2000, the number of studies describing the development (and the validation) of clinical prediction models has been increased. A well-known example for clinical prediction models is QRISK for cardiovascular risk (Anderson et al., 1991; Hippisley-Cox et al., 2008). Models such as this are being improved constantly to predict outcomes covering numerous disease areas, including cancers, obstetrics, diabetes respiration diseases, and oral conditions.

What is prediction modelling in the first place? Over decades, the discussion of explanation versus prediction has been actively pursued in the philosophy of science, however, since statistical modelling can be and is used for each of these goals, there has been confusion between causal explanation and empirical prediction. Statistical models are initially and mostly used for causal explanation and inference (the ultimate goal of scientific research). Models with high explanatory (causal) power are often assumed

to possess predictive power, which may not necessarily be the case. Depending on the focus of a study (explanatory or predictive), there are important differences in the design and data analysis methodology for explanatory and prediction modelling. The work by (Shmueli, 2010) '*To Explain or to Predict?*' provides a detailed discussion/guidance on this. Similar to Shumeli, Hernán (Dickerman and Hernán, 2020; Hernán et al., 2019) and van Geloven and colleagues (van Geloven et al., 2020) also highlight the difference between (factual/observed) prediction and counterfactual prediction (causal inference) in *European Journal of Epidemiology*. Based on these works, we briefly summarise the differences between explanatory and prediction models in Table 2.1. More considerations on factual/observed prediction *v.* counterfactual prediction (causal inference) can be found in Chapter 8.

Table 2.1: Distinction between explanatory and prediction models.

	Explanatory Models	Prediction Models
Goal	Establish causal relationships (estimating the distribution of an outcome under hypothetical interventions)	Estimate the probability of an event in binary or categorical outcome variable or estimate the conditional mean of a continuous outcome variable
Candidate variables	Expert-specified risk factors and confounders (common causes) identified through an explicit causal diagram	A larger set of potential predictors; Causal relationship not required
Variable selection	Theory informed and hypothesis-driven	Exploratory; Automated selection procedures
Choices of methods	Interpretable statistical models that adequately represent the underlying causal structures	Interpretable statistical models; Algorithmic models (e.g., machine learning algorithms)
Model estimates of interest	Size of β coefficients; Effect size measures (e.g., odds ratios)	Discrimination (e.g., AUROC [1], sensitivity, specificity); Calibration (e.g., Hosmer-Lemeshow test); Reclassification (e.g., net reclassification index); Clinical utility; Variation explained

Table 2.1 continued from previous page

Validation	Confirming causal relationships	Internal (e.g., split-sample, cross-validation); External
Challenges to validity	Confounding; Selection bias; Information bias	Overfitting; Generalisability

[1] AUROC: Area Under the Receiver Operating Characteristic curve

2.3 Types of clinical prediction modelling studies

The studies used to develop, validate, and test the effect of clinical prediction models are called clinical prediction modelling studies. It is common to classify multivariable prediction modelling studies into four groups: (i) predictor finding studies; (ii) model development studies; (iii) external validation studies; and (iv) model impact studies. Most clinical prediction modelling studies describe both the development and internal/external validation of new models. Relatively fewer studies investigate the evidence of models' impact (e.g., improving patients' outcomes) and widespread implementation of these developed models (clinical application) is rare. The four types of multivariable prediction modelling studies can be defined as follows, and this thesis covers the second and third categories:

1. *Predictor finding studies*:

These studies, alternatively known as predictor importance studies, allows identifying which variables (also namely predictors, characteristics, covariates, factors, features, markers, etc.) from a number of possible predictors contribute to the prediction of a health outcome (Altman and Lyman, 1998; Hayden et al., 2008; Moons et al., 2009). However, identifying the 'significant' predictors does not translate to its ability to discriminate 'cases' and 'non-cases'.

2. *Model development studies*:

The process of model development can be summarised as follows: 1) selecting predictors from a candidate set; 2) assigning the weights for each predictor (e.g., the β coefficients) in some kind of multivariable analysis; and then 3) combining them by a mathematical function to yield accurate forecasts when new observations are given. The main difference between the *Predictor finding studies* and *Model development studies* is that the latter builds a final multivariable predictive model using the predictors identified

by the former (Royston et al., 2009; Sterne et al., 2009).

3. *External validation studies*:

Once the model is developed using the above procedure, the next step is to validate this model in two different ways, internal and external, with and without model updating. The aim of these studies is to test the performance of the existing models using new data from new participants that were not used in the model development process (Altman et al., 2009; Bedogni, 2009; Efron, 2020; Janssen et al., 2008; Royston et al., 2009; Sterne et al., 2009). Let's see one particular situation of external validation without model updating. In this scenario, we apply the coefficient values from the developed model to the new data from external sources. For example, let's say we have developed an oral cancer screening model using data from New Zealand. Then using the same model, we select similar variables from the Australian cancer registry and use the data from the Australian Cancer registry to validate this model. If the model performance is similar to that of New Zealand, the model does not require any update. However, in some cases, when there was a poor predictive performance in the external validation cohort, there is a need to update the model based on validation data to improve the models' transportability in new data sets, by, for example, adding new predictors.

4. *Model impact studies*:

The aim of this type of prediction modelling study is to evaluate the impacts of using such a diagnostic or prognostic prediction model on the behaviour of patient or clinicians or patient health outcomes, compared to not using the model (Altman et al., 2009; Reilly and Evans, 2006; Toll et al., 2008).

2.4 The significance of prediction modelling research in promoting oral health

To better understand how oral health care can benefit from predictive modelling research, it is important to understand the various applications of predictive models in oral health care. This can be explained from three perspectives: clinical dentistry, public oral health, and oral epidemiology. Examples are shown in (Table 2.2).

- From clinical dentistry point of view

Reducing unpredictability and uncertainty is important in clinical activities.

Table 2.2: Illustrating potential applications of risk prediction models in dentistry and oral health.

Domain	Example question regarding risk	Potential decision made according to the prediction
Clinical dentistry	Is the loose tooth preservable?	Systematic periodontal therapy or tooth extraction
Public oral health	How many children are estimated to have deciduous dental caries in South Australia next year?	Quantity and location of dental services
Oral epidemiology	How many survivors from oral cancers do I expect in my clinical trial?	Recruit more or fewer participants

Everyone is a patient at one point in time, and we always expect the 'best' medical care and assume that all their decisions are evidence-based. However, that may not always be the case. Unpredictability and uncertainty exist in many medical activities. Therefore, forecasts are needed to improve our decisions regarding future events in the field of medicine. Usually, the lower their uncertainty and the more accurate that we can predict the future, the higher our confidence that our decisions may be correct.

To achieve this, prediction modelling research helps to estimate the uncertainty using statistical analysis. In determining the risk of having certain outcomes among a population, prediction modelling research takes various factors into account, such as individual's health behaviours, living environment, lifestyles, disease history, socioeconomic characteristics, and even genetic information. With these elements, a profile can be generated among the population group that predicts their health outcomes and lays the basis for addressing risk factors accordingly (Moons et al., 2012). In other words, prediction models quantify the average risk for a population or subgroup of a population of experiencing an adverse outcome either currently (diagnostic model) or in the future (prognostic model), by learning patterns from the training data and then applying to the test data. In the clinical dentistry sector, prediction modelling research allows dental practitioners to use predictive tools to make more accurate diagnoses, carry out the best treatment, and forecast the prognosis of oral diseases in order to make relevant supportive therapy.

By risk prediction modelling, each individual in a population can be assigned a probability representing their risk of having the adverse outcome. However, it does not answer questions like '*What is the **individual** risk of having the predicted outcome?*', but it answers questions like '*What is the **average** risk of having the predicted outcome for population like this individual?*' Though there are claims about personalised risk prediction following the -omic information such as the

sequencing of the human genome, we argue that individualised/personalised risk prediction is unlikely to be achievable. First, the predicted value is based on the mean estimation over a population or over many repetitive estimations for an individual (Rockhill et al., 2000). Therefore, decisions of health care made based on risk factors are general among a population who is similar to the specific individual rather than tailored. Second, at the individual level, the outcomes will vary because other risk factors besides genetics have not been suggested as a guide to change lifestyles (Marteau and Weinman, 2006). Therefore, the predicted risk might be modified by adding more predictors such as environmental factors. Detailed discussion on 'population-level risk prediction' versus 'individual-level risk prediction' can be found in Chapter 8.

- From a public health point of view

  Public oral health emphasises early detection, surveillance, and prevention of oral conditions. With prediction modelling research, scientists can estimate and monitor the prevalence, incidence, and changing patterns of oral conditions over a period of time among a specific population at a specific location.

  In addition, risk prediction modelling research will facilitate public health policy decisions. Population-based prediction modelling research usually uses a large amount of health data including geographic, sociodemographic, and systemic medical information (Dash et al., 2019). These population-based prediction modelling studies can generate health patterns of a community and provide suggestions/guidance to health policymakers on where to make interventions, such as 'dental fluoridation'.

- From an oral epidemiology point of view

  The discipline of Epidemiology investigates the distribution and determinants of diseases and focuses on causal inference (Frérot et al., 2018). In order to make an effective intervention, precision prediction and identification of high-risk populations should be prior made. Based on risk assessments, oral epidemiological studies aim to identify the risk factors and prevent oral diseases at the population level, then ultimately intervene to lower risk of oral conditions.

In summary, risk prediction modelling plays important role in dentistry, public oral health, and oral epidemiology. The results from a specific clinical prediction modelling study can be used to inform decision-making by patients, dental practitioners, health policymakers, and academics. In this thesis, we discuss the application of prediction modelling research in clinical dentistry (the diagnosis and prognosis of clinical oral conditions).

2.5 General practices for clinical prediction modelling

Given the significance of clinical prediction models in oral health, researchers have called for rigorous development and validation of these models (Kundu et al., 2017). People may think developing a prediction model is easy, using existing data (collected for various purposes) and a well-packaged statistical method. Obviously, this is an oversimplification, but in fact, prediction modelling is a complex process requiring sound clinical judgement and careful statistical analyses (Lee et al., 2016). Currently, there is no agreement on the best practices for developing, validating, and updating clinical prediction models, several studies and books (Hastie et al., 2009; James et al., 2013; Steyerberg et al., 2019) have published frameworks and recommendations for methodological approaches, with Harrell and Steyerberg being key leaders in this field. Overall, a prediction model can be developed by the following steps: (i) consideration of the research question; (ii) data collection (or inspection) and pre-processing; (iii) model generation; (iv) model validation and evaluation; and (v) model presentation. Figure 2.1 illustrates the general process of conducting a prediction modelling study. The following sections present detailed explanations of the process step by step.

2.5.1 Consideration of the research question

Before developing a prediction model, some key background questions should be set up:

- What would be the intended use of the developed model?
- What outcomes need to be predicted, in whom and how?
- Why is the model needed? Is there any similar prediction model currently being used in practice?
- What would be the predictors?
- Is the study a model development or model validation study, or a combination of both?

Before we apply a developed prediction model in clinical practice, more questions should be asked:

- Is this model easy to use?
- Has this model been validated by more than one cohort?
- What are the differences between the cohorts used for model development and model validation?

- Is there any evidence that the patients' health outcomes can be improved by using this prediction model?

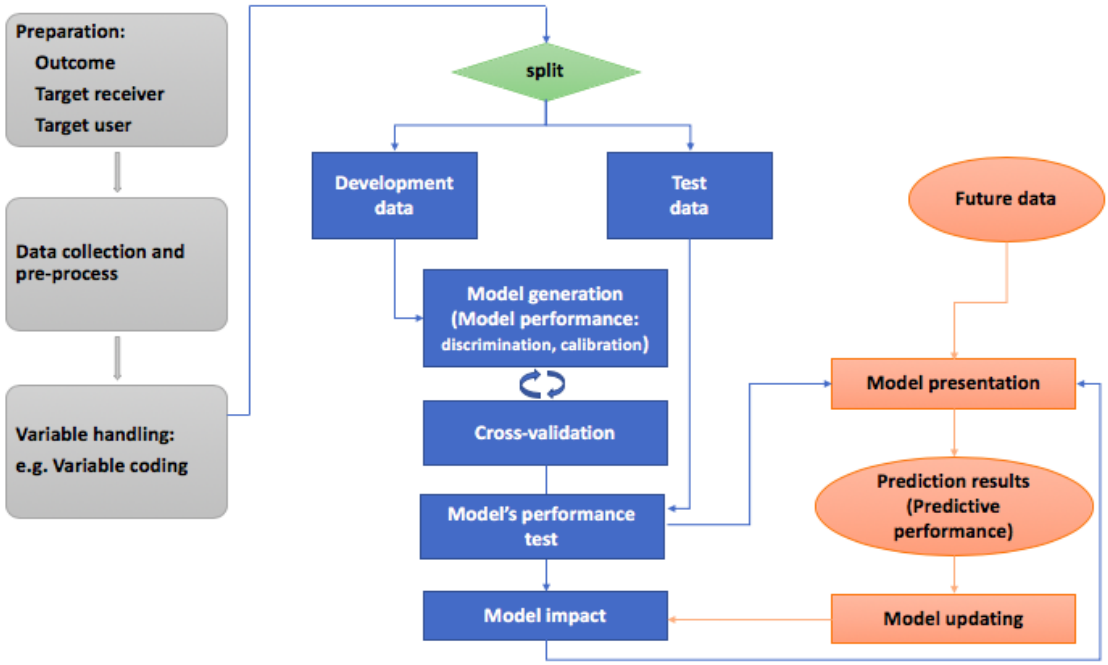


Figure 2.1: Flowchart of conducting a prediction modelling study.

In this figure, the **left** column is the preparation stage, at this stage, background questions should be set up: what the target outcome is to predict, who the target receiver is, and who the target user is. Following this, data collection and pre-processing are conducted, for example, cross-sectional data can be collected to predict prevalent events, longitudinal cohorts can be used to predict incident events. The **middle** part of this figure shows how to build the model. Usually, the data are split into two parts, the data to be used to develop the model are named development data, and the remaining data are named test data, which are to be used for testing the model's prediction performance. For a proposed model, a performance test is needed to understand the model's predictive ability. The commonly used performance measures include measures for model calibration and discrimination. Following this, prediction models can be presented as web-based calculators, or as applications for mobile phones. However, it is very common for a proposed model to perform well in the development cohort, but accuracy almost always decreases when the model is applied in other settings. There are several reasons, such as different baseline risk distributions, and different measurement systems. Therefore, for predicting a new subject with new data, if poor prediction is obtained, we can move back and adjust the model (model updating), till it performs well both in the original and future data. Following the model's performance test, another path is conducting model impact studies to determine whether the model would ultimately improve patients' health outcomes, in order to facilitate the model's clinical utility.

2.5.2 Data acquisition and pre-processing

As said by George Box in 1976 '*All models are wrong, but some are useful*', a common agreement is that there is no such thing as a perfect model and perfect data. Researchers could use different types of data sets (e.g., registry data, administrative data), depending on the research questions. In general, a clinical prediction modelling study has three key components: outcome data, predictor data, and a model that links the relationship between the two.

- Outcome data: The health outcome that we want to predict, for example, tooth loss.
- Predictor data: Data involved the models to make a prediction, for example, patients' clinical symptoms, demographical characteristics, etc.
- Modelling: A mathematical function that links the predictors and the outcome data and discovers the relationships between them, for example, logistic regression analysis, machine learning algorithms (e.g., neural network, decision trees, etc).

The outcome of a prediction model can be any event that might happen in the future, such as the probability of having a tooth loss, or the chance of surviving from cancers at a particular time point. It is noteworthy that the outcome measures and definitions are sometimes debatable. For example, there are various definitions of chronic periodontitis. Based on the definitions by the Centers for Disease Control and Prevention - American Association of Periodontology (CDC/AAP) (Eke et al., 2012; Page and Eke, 2007), mild periodontitis is defined as: there are $\geq$ 2 interproximal sites having $\geq$ 3mm clinical attachment loss and $\geq$ 2 interproximal sites having $\geq$ 4mm pocket probing depth or one site having $\geq$ 5mm probing depth. Moderate periodontitis is defined as: there are $\geq$ 2 interproximal sites having $\geq$ 4mm clinical attachment loss or $\geq$ 2 interproximal sites having probing depth $\geq$ 5mm; Severe periodontitis is defined as: there are $\geq$ 2 interproximal sites having $\geq$ 6mm clinical attachment loss and $\geq$ 1 interproximal site(s) having $\geq$ 5mm probing depth. However, according to the definitions by the European Federal of Periodontology (EFP) (Tonetti and Claffey, 2005), incipient periodontitis is defined as: there are $\geq$ 2 non-adjacent teeth having proximal attachment loss of $\geq$ 3mm; Severe periodontitis is defined as: there are $\geq$ 30% of teeth having proximal attachment loss of $\geq$ 5mm. Therefore, we raise researchers' attention on using consistent outcome definitions and measurement systems when developing a prediction model.

Handling of continuous and categorical variables With regards to data pre-processing, we can group variables into continuous or categorical variables. The type of the outcome variable(s) can be used to determine the methods for model generation. For example, continuous outcome variables, presented by numerical values, can be predicted by regression models (e.g., linear regression). Categorical outcome variables, described by

two or more classes, can be predicted by classification models (e.g., logistic regression). There are also studies aiming to not only predict/classify a particular outcome but also account for the time period for this outcome to happen. This is known as time-to-event analysis (e.g., survival analysis). For this type of analysis, Cox proportional hazards regression and Kaplan-Meier curves can be used (Efron, 1988; Walters, 2012). For predictor variables, researchers have more and more realised that we should not categorise continuous predictors at the initial stage of model development (Royston et al., 2006). However, though categorisation decreases the heterogeneity and individualisation in the population, at times interpretation of continuous values is not easy, categorisation might provide more meaningful interpretations. For instance, it may not be useful to use body mass index (BMI) as a continuous measure, BMI as a categorical variable is more useful as the pre-specified categories (underweight ($\leq$ 18.5), normal (18.5 - 24.9), overweight (25.0 - 29.9), and obese ($\geq$ 30) BMI categories) may convey more information or be easier to interpret. These subtleties need to be thought through with relevance to the study and communication of results from the study. Moreover, in oral health prediction research, when the investigated outcomes are chronic (e.g., dental caries, chronic periodontitis) and the target population is among adults, the 'risk' of getting certain outcome does not change too much by one year age differences, instead, it is quite stable within an age group, thus categorising age by five-year interval should not introduce much bias but be more interpretable. Therefore, handling predictors should be done in light of the target population and investigated outcome.

Handling of missing data The management of missing data is another important part of data pre-processing. Missing data is unavoidable in most data analyses, leading to loss of information and bias. Missing data occurs for various reasons, such as missingness during collection, not applicable, dropout, or '*unknown reason*'. Three common missingness mechanisms are: 1) '*missing completely at random*' (MCAR), 2) '*missing at random*' (MAR), and 3) '*missing not at random*' (MNAR) (Rubin, 1976). To address this issue, researchers may use singularly-imputed data, multiply-imputed data and complete case analysis (CCA), etc. (see Table 2.3 for definitions and examples). The selection guidance of an appropriate method to account for missing data can be found here (Hughes et al., 2019).

Table 2.3: Definitions and examples of missingness mechanisms.

Item	Definition	Example
Missing Completely At Random (MCAR)	When there is no systematic difference between the missing data and the observed data, it is MCAR.	A questionnaire might be lost in the post.
Missing At Random (MAR)	When the observed information can explain the systematic differences between the missing and the observed data, it is called MAR.	If a child does not attend an educational assessment because the child is (genuinely) ill, this might be predictable from other data we have about the child's health.
Missing Not At Random (MNAR)	When the observed information cannot explain the systematic differences between the missing data and the observed data, it is MNAR.	An adolescent took drugs the night before he/she might not attend a drug test, because they might not want the information to be passed on to their supervisors or parents.
Complete Case Analysis (CCA)	Analysis conducted among the individuals with complete information on the available predictors.	
Multiple Imputation (MI)	In MI, the missing values will be replaced by the 'imputed values', and multiple imputed data sets are created to account for the uncertainty due to imputation. Then the separate analysis is conducted using these data sets and Rubin's rules are used to combine the multiple results.	

2.5.3 Model generation

According to the discussion by Efron (Efron, 2020), in general, statistical tasks can be categorised into three: (i) predicting the new cases, (ii) estimating the regression parameters (e.g., coefficients), and (iii) assigning significance to each of the predictors. For prediction purposes, there are many ways to generate the model, we can choose either statistical methods or 'pure prediction algorithms' (e.g., machine learning algorithms). A statistical model is using statistical analysis to represent and/or infer any relationships between variables. Machine learning is using mathematical algorithms to learn the underlying patterns of the data without relying on rules-based programming or pre-defined assumptions in order to make predictions of the unknown (Murphy, 2012). A detailed review of their

application in oral health and the difference between statistical models and machine learning algorithms can be found in section 2.6.

2.5.4 Model performance evaluation

The study of performance measures has matured over the past two decades. For example, the binary classification tasks have two objectives: minimising the number of False Positives (e.g., where a disease is mistakenly diagnosed) as well as the number of False Negatives (e.g., where a diagnosis of the disease is missed). The binary classifiers can trade-off one type of misclassification against another. It is common to represent this trade-off by a Receiver Operating Characteristic (ROC) curve (see the following section). Efforts have also been made to put this ROC curve into a simple value that would allow for comparisons across different classifiers, known as the Area Under the ROC curve (AUROC). When two cases (a positive and a negative case) are selected at random, the AUROC reflects the probability of the classifier assigning a higher risk score to the positive case (Hanley and McNeil, 1982). Besides AUROC, many metrics are nowadays available, such as sensitivity, specificity, etc. These measures are detailed and reviewed elsewhere (Alba et al., 2017; Li and Wang, 2019; Steyerberg et al., 2010). In brief, a good prediction modelling study should not only assess model fit but also assess and evaluate the model's discrimination and calibration as well as compare the performance of different prediction models.

Model discrimination

Discrimination refers to the models' ability to classify cases *v.* non-cases (e.g., dead *v.* alive). A commonly used measure of discrimination is the AUROC, which sometimes refers to the concordance (C)-statistic, and their extensions. The AUROC is equal to the probability that an event is given a higher risk score than a non-event across all possible thresholds. When a threshold is set, the individuals whose risks estimates are lower or higher than the threshold are assigned different levels of risk. For example, if the threshold is 0.3, then the individuals whose estimated risk ≥ 0.3 are assigned to the higher risk group, and vice versa. Then the sensitivity and specificity for this threshold can be calculated. In the calculation of an AUROC, the thresholds vary from 0 to 1. Therefore, a ROC curve is plotted over a wide variety of sensitivity and specificity. An AUROC value of 0.5 indicates a random chance. However, when outcomes are infrequent, measures of overall accuracy can be misleading: if there are only 5% of patients experiencing a positive outcome, the model that predicts 0% positives obtain an accuracy of 95%, however, such a 95% accuracy does not mean a 'strong' predictive ability. To overcome this limitation, sensitivity and specificity were created as alternate properties for evaluating models' performance. However, when the proportion of positive and negative outcomes is imbalanced, alternate

measures were proposed and found to be more useful (Saito and Rehmsmeier, 2015). One alternative is the Area Under the Precision Recall Curve (AUPRC), which uses information on sensitivity (referred to as recall) and positive predictive values (referred to as precision). The AUPRC is more informative for evaluating the model's discrimination because it identifies the joint probability of being positive in both the predicted world and the observed world, meaning AUPRC focuses on the performance of a classifier on the positive (minority class) only. Therefore, AUPRC is more suitable for imbalanced data where the positive observations are less. Let's take binary outcome prediction as an example, the confusion matrix of the predicted and observed positives can be expressed by a 2×2 table.

Table 2.4: Confusion matrix of the binary outcomes

	Observed distribution of the binary outcome		Total
Predicted distribution of the outcome	1	0	
1	N_{11} (True positive)	N_{10} (False positive)	$N_{1.}$
0	N_{01} (False negative)	N_{00} (True negative)	$N_{0.}$
Total	$N_{.1}$	$N_{.0}$	N

The calculations of these measures are as follows:

$$Correctly\ classified\ = \frac{N_{11}+N_{00}}{N} \tag{2.1}$$

$$Positive\ predicted\ value\ (PPV) = Precision = \frac{N_{11}}{N_{1.}} \tag{2.2}$$

$$Sensitivity = Recall = True\ Positive\ Rate\ (TPR) = \frac{N_{11}}{N_{.1}} \tag{2.3}$$

$$Specificity = False\ Positive\ Rate\ (FPR) = \frac{N_{00}}{N_{.0}} \tag{2.4}$$

$$Accuracy = \frac{N_{00}+N_{11}}{N} \tag{2.5}$$

Then we have:

- **ROC: Plot of Sensitivity (TPR) *v.* 1-Specificity (FPR).**

$$AUROC = \int_0^1 Sensitivity(Specificity)d(Specificity) \tag{2.6}$$

- **PRC: Plot of Recall (TPR) *v.* Precision (PPV).**

To calculate the AUPRC, we can follow two steps: 1) Calculate precision-recall value from multiple confusion matrices for different thresholds. For example, if the threshold is 0.8, all the participants whose predicted probability greater than 0.8 are considered as cases. For $threshold = 0.8$, we can calculate the precision-recall values based on the true positives,

true negatives, false positives and false negatives. 2) Similarly, we can also calculate the precision-recall values for the other thresholds. Once we have these precision-recall values, we can calculate the AUPRC using an integral. Mathematically, it can be expressed as Equation 2.7 (Brodersen et al., 2010):

$$AUPRC = \int_0^1 Precision(recall)d(recall) \tag{2.7}$$

Model calibration

Model calibration measures the agreement between the predicted probabilities/risk *v.* the observed probabilities/risk. Let's say we have a cohort with 100 participants, among which 10 patients having a positive outcome. If our model predicts 10% of the population developing the outcome, then we believe that our model has 'good' calibration because the predicted probability matches the observed frequency.

A common way to evaluate the model calibration is called the Hosmer-Lemeshow (H-L) test (Hosmer et al., 2013). The H-L test measures the correspondence between the prediction and observation in two steps: first, dividing the ranges of the predicted and observed probability (0-1) into n subgroups (usually $n = 10$), second, comparing their difference by calculating a chi-square value and a p-value. Another way for measuring model calibration is to generate a calibration curve (Almeida et al., 2002). The calibration curve can be constructed in three steps: 1) ranking the predicted probabilities, 2) categorising these probabilities into n subgroups (usually $n = 10$) and 3) plotting the mean of the predicted probabilities *v.* the mean of the observed probabilities for each subgroup.

Interpreting performance

The performance of a particular model only should be interpreted with comparison to another (an existing) one. For example, a prediction model with an accuracy of 70% is useful if the accuracy of the currently-used models is just better than a random classifier (with an accuracy of 50%). In contrast, a model with an accuracy of 80% may not be useful if the existing models can predict the outcome correctly 90% of the time. Therefore, models' performance can only be meaningful when compared with others.

2.5.5 Model validation

For evaluating model performance, different metrics have been proposed in the above paragraphs. However, those measures do not provide guidance on the model adoption. Because there is a possibility that models overly adapted to the development data set may perform well in that particular data set but have poor predictive performance using new observations (Steyerberg and Vergouwe, 2014). Therefore, it is important to determine under what circumstances the model can be used, and how well is the model's

transportability (Altman and Royston, 2000). Model's transportability refers to the idea that the model's results from one population can be exchangeable with a different source of population. Models' reproducibility and transportability can be evaluated in internal and external validation studies, respectively. The typical life cycle of clinical prediction models thus includes multiple stages of model development and model validation. Validation refers to the process of testing whether the results from a prediction modelling study can be generalised to data that were not used in the previous study.

The importance of splitting data into a training and test data set

Talking about 'good' predictive performance, it is important to evaluate models' performance on both the training and the test data that was not seen by the model during the generation process. This ensures that the models are not overfitting or underfitting, which might result in poor performance in new observations. An example of a training and test data set can be found in Figure 2.2.

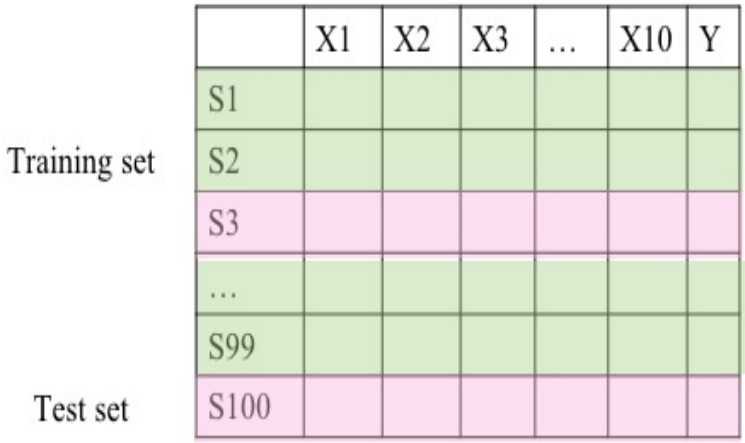


Figure 2.2: Example of training and test data sets.
Say we have a data frame with 100 observations (S1 to S100) and 10 predictors (X1 to X10) and outcome Y. We randomly pick a proportion (e.g., 80%) of the rows (green box) to training set and the rest 20% (pink box) go to test set. We train our model with training data set, while test data are only used for the evaluation of the model's performance.

Internal validation

Generally, we should distinguish the internal and external validation. Internal validation refers to the process where we split the data into a development data set (for model generation) and a test data set (for model validation). It tests whether the models' predictive ability remains at a similar level in the similar underlying population (not exactly the same as the specific population used to develop it). Several methods are available to do internal validation including split-sample validation, *k-fold* cross-validation, etc.

Split-sample validation referring to randomly splitting a single data set into two parts. For example, we use 70% of the available data for training and 30% for validation. The problem with split-sample validation is that we can never know whether the estimate is a realistic estimate of the model or due to a 'lucky' randomisation, therefore, it is important to try several split proportions (e.g., 9:1, 8:2, 7:3, 5:5, 4:6, etc.) with many iterations (e.g., setting different seeds to allow for repetitions).

Another approach for internal validation is *k-fold* cross-validation, where the data is stratified into k folds. Figure 2.3 box bar shows the example of 10 folds, the yellow box represents validation fold and the blue boxes are training folds. We use one fold as the test set once (the yellow box goes from the first to the last fold) to complete one iteration. Such iteration should be repeated many times (usually at least 50). There were two aims to use cross-validation, the first is to limit overfitting, the second is to tune the models. For example, when we develop a random forest, we can set different parameters for the model such as the number of trees. The tuning process would pick the forest with the best predictive ability. Finally, our models can be evaluated in the test sets. We illustrate the use of *k-fold* cross-validation for modelling training in Figure 2.3.

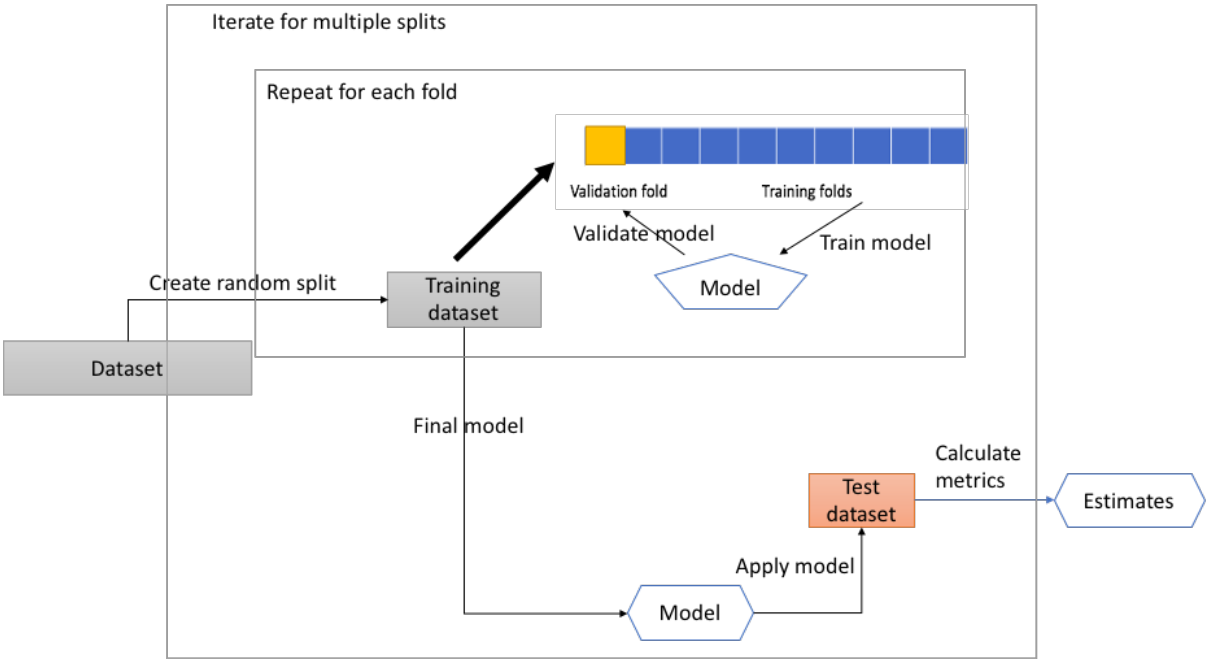


Figure 2.3: Schematic overview of 10-fold cross-validation.

Researchers have also argued that bootstrapping-validation is another effective version of internal validation, as it can avoid using the development data for estimating models' prediction performance (Gerds et al., 2008). Differing from cross-validation that splits the available data to create multiple data sets, bootstrapping works by performing sampling from the original data set with replacement, and assuming that the data that have not been chosen are the test data set. For this nature, bootstrapping can avoid the possibility

that a single model fit of the original data might overestimate the models' prediction performance. Therefore, bootstrapping can be viewed as a method that we fit the model once, then evaluate it many times on various resampled data sets, i.e., bootstrapping has dual purposes: estimating the variance of the parameters of interest and building ensemble models then testing the models' performance. Cross-validation is a method where we fit the model many times using each of real sub-data sets, each time evaluate it against the real data, i.e., cross-validation is more about testing the models' validity.

External validation

Having excellent model discrimination, high accuracy, and good model fit using data on hand does not ensure that the model will have 'good' performance when tested on new patient cohorts. We here take the Framingham risk score as an example. Framingham risk score is a commonly used method for assessing the risk of cardiovascular diseases. Though it obtains high accuracy in general, it is found to underestimate the risk of sub clinical atherosclerosis in some women (Michos et al., 2005). Thus this comes to a further test - evaluating the predictive performance of a model under external validation. External validation is different from what is given in Figure 2.2. It refers to the validation process where two different data sets are used. One is used for model generation and a new data set is used for validation. If we obtain 'good' performance metrics on multiple external data sets, this indicates 'good' transportability and generalisability of the model and further strengthens the acceptance of the model. External validation of a model requires open and transparent reporting of the original study, such as the patients' inclusion and exclusion criteria, data pre-processing steps (e.g., handling of missing data, etc.), and model performance metrics, etc.

2.6 Application of machine learning-based algorithms for prediction purposes in oral health

This section is an extension of the subsection 'model generation'. As mentioned above in subsection 2.5.3, the approaches used for predictive model generation are generally classified into two categories: statistical models and machine learning-based algorithms. The term 'machine learning' refers to the idea of applying algorithms to learn the structures in the observed data, therefore, allows for predicting the outcome using unseen data. Examples of machine learning-based algorithms include bagging, boosting, recursive partitioning, random forests, support vector machines, and neural networks, etc.

To understand the trends of the application of machine learning in oral health, electronic searching was carried out using the PubMed database. We searched keywords including 'machine learning', 'deep learning' and 'artificial intelligence'. Other items such as 'dentistry' and 'oral health' were used as conjunctive search terms. The distribution of

the number of articles can be found in Figure 2.4. We found that though the term 'machine

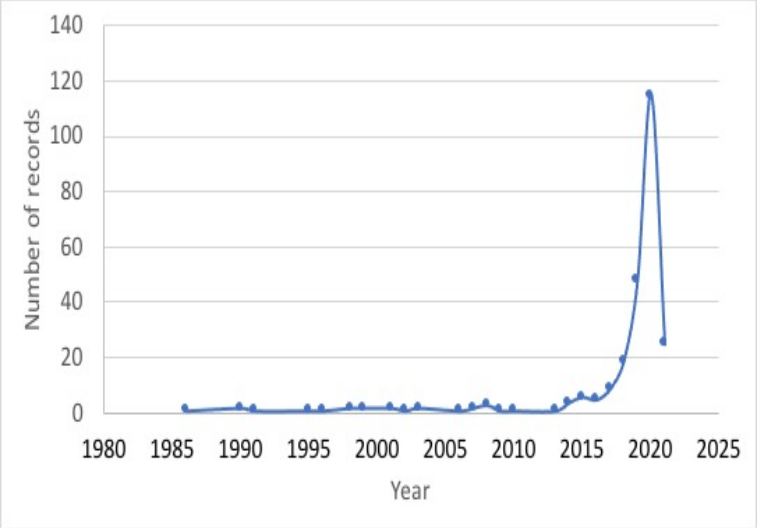


Figure 2.4: Number of articles on machine learning in the field of oral health (searched on PubMed on 5^{th} February 2021).

learning' was coined in the 1950s, its application in dentistry and oral health was not seen until the late 1990s. In recent years, its applications have evolved rapidly, and most of the existing machine learning algorithms were used to address various clinical issues in the fields of dental medicine. We here present its applications using recent examples:

- In endodontics, machine learning has been applied to detect dental caries. For example, using information relating to dental caries history and periodontal health from parents, a decision tree was 'grown' to detect dental caries in children (Dima et al., 2018). In addition, using radiology images, a study has shown the reliability of using machine learning to find the minor apical foramen in order to determine the working length for patients undergoing root canal treatment (Saghiri et al., 2012). The accurate evaluation of endodontists was 76%. The artificial neural network found correct anatomic positions in 96% of the teeth and was more accurate than endodontists' markings.

- In periodontics, machine learning algorithms were applied to better understand periodontal diseases. For example, a study has presented the classification of periodontitis using decision trees and neural networks (Ozden et al., 2015). Moreover, a large number of radiographs was analysed by the convolutional neural network models for the diagnosis of periodontally-compromised teeth (Lee et al., 2018b). Additionally, under the help of the microbial profiles of subgingival plaque, support vector machine models were applied to classify the aggressive and chronic periodontitis (Feres et al., 2018).

- In prosthodontics, a study by Chen et al. (Chen et al., 2016) was conducted to describe a clinical decision support system for the design of removable partial dentures. Moreover, a machine learning system was also used for predicting the colour of the teeth after the bleaching procedure (Thanathornwong et al., 2016). The system needed two input components: patients' data and pre-bleaching colour. Based on these information, the developed system can predict the post-treatment colour and colour change obtained with a whitening system.

- In the management of other oral conditions such as oral cancers, the last two decades have also seen the evolution of machine learning in these fields. In 1995, it was first proposed to use a machine learning-based approach to identify individuals who are at high risk of developing oral cancers (Speight et al., 1995). More recently, based on the clinical, pathological, and genomic markers, (Chang et al., 2013) applied machine learning methods for feature selection and prognosis of patients with oral cancers.

In summary, there are a variety of machine learning applications in dentistry and oral health, but we only present a small part using the above-mentioned examples. From these examples, we conclude that machine learning algorithms offer new supportive diagnostic and prognostic tools for various oral diseases with high accuracy and facilitate the development of a clinical decision support system. We see there is a significant uptake of machine learning in oral health research. An important driver for this uptake is that in research that involves the analysis of a vast amount of data (e.g., electronic health records, including clinical data, behavioural data, imagery data, etc.), machine learning-aided models are capable of analysing these data efficiently.

Health care data are rather heterogeneous as it includes data in various types such as demographic data, clinical data, behavioural data, -omics information, etc. Obviously, it is difficult for humans to infer the data and to make decisions. While data analysts are working towards providing clean, curated, and structured data, machine learning has been proposed for a better understanding of data, making the best use of the various types of data, and allowing to grasp their interactions.

2.7 The choice between statistical models and machine learning algorithms for health care prediction purposes

There have been several vague statements on the difference between statistical models and machine learning algorithms. Generally, the statistical models depend on statistical hypothesis testing, which tries to identify whether the observed patterns match our assumed data generating mechanism. Before we start collecting data and performing research, we

will need to formulate our hypothesis. The original aim of statistical methods is to make inferences on the predictors so as to intervene on the outcomes.

One of the advantages of machine learning is that many forms are non-parametric, data-driven, more flexible, and do not rely on prior assumptions. Meaning, there is no specific requirement for the hypothesis that the independent predictor X is associated with dependent outcome Y. Instead, the primary hypothesis for machine learning is that there is a pattern in the set of predictors that will identify the outcome, therefore they allow models to identify the underlying relations between variables. Second, statistical modelling aims to produce the simplest and explainable models that fit the data. So the predictors are generally assumed to be independent of one another. However, machine learning algorithms consider all possible interactions between variables.

Though there are many published comparisons/differences of statistical models and machine learning algorithms, these comparisons do not help researchers to choose an appropriate approach for their prediction modelling studies. Based on Harrel's suggestions (Harrell, 2020), we here describe some general guidelines in order to help researchers choose between these two approaches.

Statistical models may be a better choice when

- We want to find out the (causal) effects of a number of predictors on the outcome.
- Predictors mostly affect the outcome in an additive way, or the interactions between predictors are 'small' and/or can be pre-specified.
- Sample size is inadequate for the alpha level and analyses chosen (i.e., statistical models typically require 20 events per candidate predictor).
- One wants to isolate the effects of a specific predictor(s) such as treatment or a risk factor.
- The interpretability of the model is one of the major pursuits.

Machine learning algorithms may be more suitable when

- The algorithms can be trained on a number of replications.
- Obtaining 'high' prediction accuracy is the main purpose while describing the impact of any particular predictor (e.g., treatment) is less of a concern.
- The interaction between variables and/or non-additivity can be strong and cannot be isolated.
- The sample size is large (i.e., machine learning algorithms usually require > 200 events per candidate predictor).

- One does not pursue the interpretability of the model (i.e., the model is a 'black box').

2.8 Bias in prediction modelling research

Whenever we discuss model prediction, it is important to understand prediction errors (bias and variance). Like many other types of research, the low risk of bias has always been a pursuit for prediction modelling research. Understanding the nature of bias is important for understanding and interpreting a model's prediction ability. In this section, we talk about what bias is, how bias occurs, and why bias matters in prediction modelling studies.

2.8.1 What is bias

Generally speaking, in data science, bias refers to a deviation from expectation in the data. In the context of risk prediction, we have defined the risk of bias to occur when systematic errors of a model's estimates or predictive performance are led by the limitations in the study design, conduct or data analysis, etc. However, the error is often subtle and neglected. So, the first question is how does bias occur?

2.8.2 How bias occurs

By understanding the general steps to develop a clinical prediction model in the last paragraphs, it is not hard to find that there are many chances for bias to occur. To understand this, we mapped the relationship between the real world, the observed world, and the predicted world in Figure 2.5.

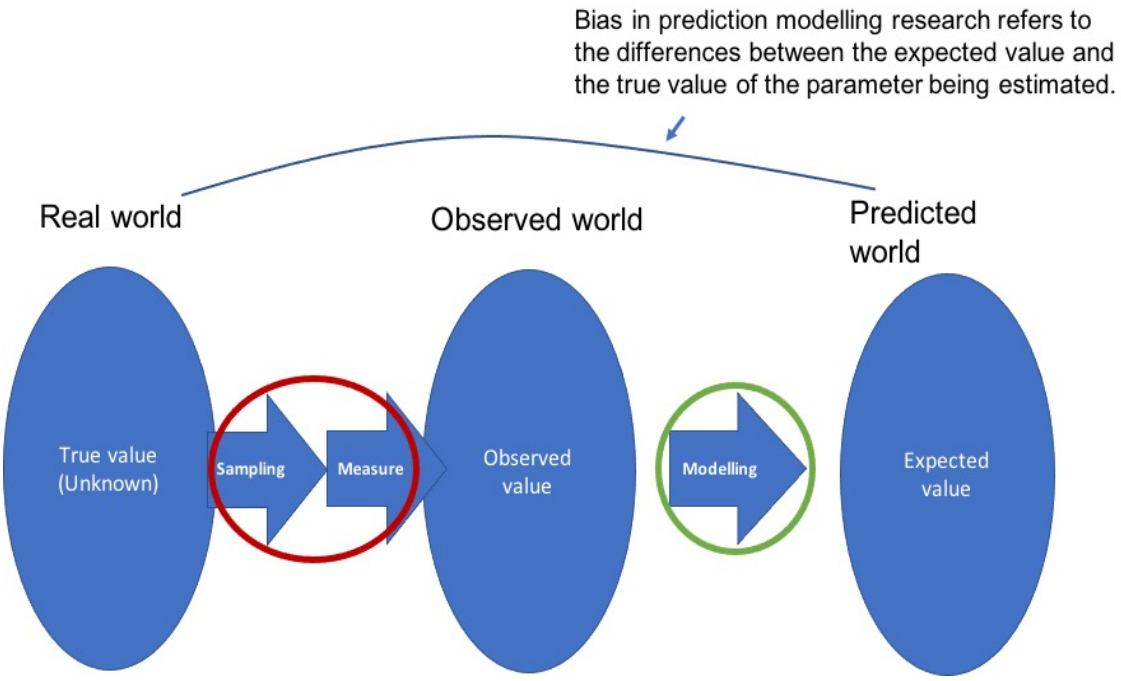


Figure 2.5: Overview of how bias arises in the process of developing a prediction model.

As shown in Figure 2.5, the real world contains the true values that we are interested in, however, these true values are never known. If we could know 'everything' about all the individuals in the real world and could store and identify all of this information, our data and model would have no bias. Therefore, to understand the true values, we usually collect data by sampling and measuring random observations from the real world, thus an observed world can be obtained. Next, we conduct modelling procedures to identify the data generating mechanism and develop the prediction models. Finally, we estimate the unobserved outcomes/parameters of interest using these models and we thereby obtain the predicted world.

Notably, in prediction modelling research, bias refers to the differences between the predicted value and the true value of the parameters being estimated (indicated by the blue curve). Therefore it is obvious that uncertainties that happen during sampling, measuring (red circle in Figure 2.5), and the modelling process (green circle in Figure 2.5) may bias our prediction. In short, bias occurs because of sampling, measurement, and estimation.

2.8.3 Why does bias matter in prediction modelling studies

Prediction models observe the real world through the data used for training. When the data used for model development are biased, models' prediction accuracy and estimates are compromised, because the model will not only learn those biases but will end up amplifying them (Mehrabi et al., 2019). Being aware of these risks allows us to better eliminate bias. Mathematically, let's define the true outcome and predictors as Y and X, and the predicted outcome value is $\hat{Y}$. Let's suppose that the data is generated from Equation 2.8:

$$Y = f(X) + \varepsilon \tag{2.8}$$

Where the random error, ε has a mean $E(\varepsilon) = 0$ and is independent of X. Variance of the error is $Var(\varepsilon) = \sigma^2$. Let's define the prediction error as the difference between the observed outcome (Y) and the predicted outcome ($\hat{Y}$). Now the mean square error (MSE) of an estimate $\hat{Y}$ of an observation Y is defined by Equation 2.9:

$$MSE_{\hat{Y}} = E[(\hat{Y} - Y)^2] \tag{2.9}$$

Note that MSE measures the average squared difference between the estimator $\hat{Y}$ and the true value Y, and MSE can also be expressed as the following equations:

$$\begin{aligned} MSE_{\hat{Y}} = E(\hat{Y} - Y)^2 &= E(\hat{Y}^2) + Y^2 - 2YE(\hat{Y}) && (2.10)\\ &= E(\hat{Y}^2) - [E(\hat{Y})]^2 + [E(\hat{Y})]^2 + Y^2 - 2YE(\hat{Y}) && (2.11)\\ &= Var(\hat{Y}) + [E(\hat{Y})]^2 + Y^2 - 2YE(\hat{Y}) && (2.12)\\ &= Var(\hat{Y}) + [E(\hat{Y}) - Y]^2 && (2.13) \end{aligned}$$

$$= Var + Bias^2 \tag{2.14}$$

Hence, bias and variance are the two sources of imprecision in prediction models.

2.8.4 Common sources of bias in prediction modelling research

Recently, substantial efforts have been made to improve the reproducibility and reliability of scientific findings in health research. These efforts include the development of guidelines for designing, conducting, and reporting preclinical studies (**ARRIVE**), clinical trials (**ROBINS-I**, **CONSORT**), observational studies (**STROBE**), and systematic reviews and meta-analyses (**PRISMA**). Given the use of prediction has increased in health sciences, guidelines such as **PROBAST** have been recently published for the conduct and risk of bias assessment for prediction modelling studies. As suggested by **PROBAST** (Moons et al., 2019), common types of bias could be classified into four domains, relating to Participants, Predictors, Outcome, and Analysis. We summarise the actions that may introduce bias in Table 2.5, and these actions are avoidable.

Table 2.5: Actions introducing bias in prediction modelling studies.

Domain	Actions should be avoided
Participants	• Selection bias, e.g., lack of specification of inclusion/exclusion criteria, lack of sampling and/or data collection methods.
Predictors	• Measurement error. Poor measurement of predictors is likely to degrade their predictive power. This usually happens when inconsistent measures were used across all the participants. • Categorising continuous variables. • Lack of consideration on unmeasured predictors. In external validation studies, this usually happens when omitting unavailable predictors.

Table 2.5 continued from previous page

Outcomes	• Unclear definition. • Measurement error. This usually happens when inconsistent measures were used across all the participants. • Incorporation of predictors. This usually happens when the predictors include part of the information captured by the outcome. For example, using bleeding to predict gingivitis.
Analysis	• Selecting predictors based on univariate analyses. • Lack of consideration on overfitting. This usually happens when lacking internal validation (e.g., *k-fold* cross-validation). • Lack of appropriate missing data handling. • Insufficient evaluation of the model's prediction performance. This usually happens when lacking either assessment of discrimination or calibration.

2.8.5 Efforts to improve the quality of clinical prediction modelling studies

Here we define that the quality of a clinical prediction modelling study can be influenced by two aspects: the methodology that is used to reduce biases in the study and completeness of reporting of the study. There has been lots of progress in recent years to improve the quality of prediction modelling studies, including the publication of several books, various series of papers in *PLOS Med* and *BMJ* (Hemingway et al., 2013; Hingorani et al., 2013; Riley et al., 2013; Steyerberg et al., 2013), the **CHARMS** checklist in 2014 (Moons et al., 2014), the **TRIPOD** statement in 2015 (Collins et al., 2015), and the **PROBAST** guidelines in 2019 (Moons et al., 2019). The efforts to reduce bias and improve the reporting completeness are discussed in the following paragraphs.

Efforts to reduce bias To standardise methodological principles, the earliest efforts were from the PROGnosis RESearch Strategy (PROGRESS) group, this group has proposed the methods that can be used in predictive models in particular using a series of publications (the so-called PROGRESS series). More recently, **CHARMS** checklist was designed for two purposes: 1) to provide explicit guidance to help reviewers and users of diagnostic or

prognostic models framing the right (review) questions, and 2) to provide a data extraction list with guidance on which items to be extracted from prediction modelling studies. **CHARMS** can be applied to all types of primary model development studies, for all types of target population, outcomes, predictors, and regardless of the used statistical methods. However, **CHARMS** does not evaluate the risk of bias in the studies. To address this gap, **PROBAST** was introduced in 2019 as an updated tool to assess the risk of bias in a multivariate prediction modelling study. We would not reiterate the details in **PROBAST** documents (Moons et al., 2019; Wolff et al., 2019), however, key guidance to reduce bias according to **PROBAST** are summarised in Table 2.6.

Table 2.6: Key suggestions to reduce bias in prediction modelling research.

Suggested practices	Cautions
Prioritising precision and transparency when working on a prediction task.	Approaches for solving a prediction problem can be different from causal inference.
Considerations on the methods for feature selection.	Using p values from bivariable comparison or step-wise procedures to select predictors might introduce bias and overfitting.
Transparently report the presence and handling of missing data. Approaches to handle missing data are suggested to be able to improve models' prediction capability.	CCA [1] may not always be a good practice to handle missing data.
Conduct external validation to test the generalisability of a model.	External validation should use exactly the same model developed (same predictors, same modelling approaches, etc.).
Seek a reference model/comparator other than 'no model' when evaluating and interpreting model performance.	Relying on the AUROC [2] alone may lead to an in-comprehensive understanding of the model's predictive performance.
Follow checklists such as **TRIPOD** [3] for the purpose of reporting completeness and transparency.	

[1]CCA: Complete Case Analysis. [2]AUROC: Area Under the Receiver Operator Characteristic curve. [3]TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.

Although the criteria of risk assessment for clinical prediction modelling studies are in infancy, when we use them to assess the previous literature, we can obtain the changes and progress during the past years and the investments on standardising the methodological issues in prediction modelling studies will surely increase the quality of models.

Efforts to improve transparent reporting With regards to reporting principles of a prediction modelling study, there were no defined reporting guidelines of predictive modelling studies for researchers to follow until the development of **TRIPOD** statement in 2015. The **TRIPOD** statement lists 22 items (covering title, abstract, introduction, methods, results, conclusion, and supplements) that are considered crucial to effectively report a clinical prediction modelling study. There are two groups working towards standardising the reporting of a clinical prediction model (Collins et al., 2015; Hemingway et al., 2013). These groups and their reports take important steps towards outlining ways that the results of this type of study can be effectively communicated in the research community so that these prediction tools can be validated and might improve decision-making. Moving forward, the dental and oral health journals should require that clinical prediction modelling studies conform to the standards outlined in the **TRIPOD** statement before publication.

CHAPTER 3

Methods

Preface

Chapter 3 comprises the methodologies used in the thesis. It consists of four sections: Section 3.1 gives a brief summary of the adopted methods in this thesis. Section 3.2 describes the methods for conducting systematic reviews (Chapters 4 and 5). These methods include framing a review question, registering a protocol, defining study selection criteria, formulating search strategy, critical appraisal for the studies included, and evidence synthesis and study reporting. Section 3.3 briefly describes the data sources and aspects of data management used in each of the empirical analyses included in later chapters (Chapters 6 and 7). Section 3.4 presents statistical methods used in each of the studies undertaken in relevant chapters.

3.1 Summary of the methods used in this thesis

The term, 'evidence-based medicine', was introduced to generate high-quality evidence in medical research and using that evidence to make better clinical decisions (Sackett, 1997). A milestone in 'evidence-based medicine' was the development of systems for classifying the 'level of evidence' (Hill et al., 1979). Based on the traditional hierarchy of evidence for therapeutic studies, systematic reviews are positioned at the top, followed by randomised controlled trials, then observational studies such as cohort studies and case–control studies, then case studies, then laboratory studies and 'expert opinions' (Greenhalgh, 1997). When coming to the field of prediction modelling studies, the highest level of evidence can be provided by the 'high-quality' prospective cohort studies with 'sufficient' predictive power or systematic reviews of these studies (Burns et al., 2011).

In this thesis, two types of research were conducted: systematic reviews and empirical analyses using observational cohorts. Table 3.1 summarises the methods used in each of the four studies included in the current thesis.

Table 3.1: Summary of the adopted methods in each of the included publications

Publication	Title	Methods adopted
Publication 1 (Chapter 4)	Prediction models for the incidence and progression of periodontitis: A systematic review	Systematic review + narrative report.
Publication 2 (Chapter 5)	Examining bias and reporting in oral health prediction modelling studies	Systematic review + narrative report.
Publication 3 (Chapter 6)	Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on SEER database	*Prediction modelling approaches:* 1. Cox proportional hazard regression 2. Random Forest for survival 3. Survival Tree *Handling of missing data*: 1. Substantive Model Compatible Imputation 2. Random Forest for survival
Publication 4 (Chapter 7)	Application of multilevel machine learning for predicting pain following root canal treatment	*Prediction modelling approaches:* Multilevel logistic regression *Variable selection approaches*: LASSO [1] *Handling of missing data*: MICE [2] + missing indicator

[1]LASSO: Least Absolute Shrinkage and Selection Operator; [2]MICE: Multiple Imputation with Chained Equations.

3.2 Methods for systematic reviews on prediction modelling studies

In this thesis, available guidelines were followed regarding the steps in conducting systematic reviews of prediction modelling studies. These steps include framing a review question, literature searching (Geersing et al., 2012; Ingui and Rogers, 2001; Wong et al., 2003), study selection (Steyerberg et al., 2013), data extraction, critical appraisal of prediction modelling studies (Moons et al., 2014; Wolff et al., 2019), and evidence synthesis and reporting. A brief description of each step can be found in the following sections.

3.2.1 Registering a protocol and framing the review question

Usually, a systematic review starts with a protocol describing the review question, background, aims, study design, methodology (e.g., literature searching, study inclusion and exclusion criteria), and statistical analysis (if applicable) of the study. Two review questions that are of interest in this thesis are:

- Chapter 4: Are there any prediction modelling studies for the incidence and progression of periodontitis in adults?
- Chapter 5: Whether the oral health prediction modelling studies in the recent literature were conducted following the methodological and reporting recommendations?

According to the International Prospective Register of Systematic Reviews (known as PROSPERO), once the review questions are framed, systematic reviews should be registered at inception using public platforms, in order to avoid unplanned duplication. A typical PROSPERO protocol has 39 items, which defines various aspects of a systematic review, including the aim of the systematic review, the timeline for conducting the study, searching strategies, the criteria for selecting studies, the planned approaches for data analysis. In this thesis, we registered two protocols on PROSPERO, details can be found by ID numbers CRD42018085437 and CRD42019122274.

3.2.2 Formulating the search strategy

Following the recommendations by Moons et al. (Moons et al., 2014), the search strategies for our systematic reviews were framed according to the PICOTS system (population, intervention, comparator, outcome(s), timing, setting). The definitions and details of PICOTS items applied to our two systematic reviews are described in Tables 3.2 and 3.3.

Table 3.2: PICOTS system

- **P**opulation — The target population (model receivers) on which the prediction model(s) will be used.
- **I**ntervention (Model) — The prediction model(s) to be reviewed.
- **C**omparator — 'C' means comparison group, sometimes it is not applicable in systematic reviews for prediction modelling studies. In systematic reviews for intervention studies, the 'C' represents confounding.
- **O**utcome(s) — The outcome(s) that the model is to predict.
- **T**iming — The time period between the record of predictors and the occurrence of the outcome.
- **S**etting — Under what circumstances or settings is the prediction model to be used.

Table 3.3: Framing a systematic review search strategy by use of the PICOTS* system

	Publication 1 (Chapter 4)	Publication 2 (Chapter 5)
Population	Adults (aged 18 and over)	Adults (aged 18 and over)
Intervention (Model)	Models to predict the future incidence and progression of periodontitis	Models to predict the risk of any oral conditions
Outcome	Periodontitis incidence and progression	Any investigated oral conditions
Time span	Predictors measured after age 18, outcome measured at adequate time after the occurrence and progression of periodontitis	Predictors measured after age 18, outcome measured at adequate time after the occurrence of oral conditions
Settings of using the model	Generally healthy adults who are free of systematic conditions, e.g., intellectual disability, HIV, drug dependent and alcohol dependent	Relevant only to adults (aged 18 and over)

* The 'C' (Comparator) in the PICOTS system was not used in our studies.

3.2.3 Searching

Here we present a summary of electronic searching, the detailed search strategies can be found in the relevant chapters (Chapters 4 and 5). For the systematic review of prediction models for periodontitis (Publication 1), six databases (PubMed, Embase, DOSS, Scope, Web of Science, Proquest) were reviewed to identify the existing literature without time restrictions. Additionally, to avoid missing out on potential papers/reports, we reviewed the bibliography of the included full-text articles. In our second systematic review (Publication 2), assessing bias and reporting transparency in predictive modelling, our search was limited to 'high-impact' journals, including oral health, dentistry, epidemiology and biostatistics. The reason why we only searched these journals is that these journals are believed to provide the best available evidence in the field of oral health, and the findings from these studies are believed be generalised to a broader area of oral health prediction modelling studies.

3.2.4 Critical appraisal and information extraction

The quality of a systematic review depends not only on the methodology used in the review but also on the quality and reliability of the studies included. Therefore, critical appraisal (also referred to as the risk of bias assessment) is essential in any systematic review. In Chapters 4 and 5, studies' quality was critically appraised using tools such as **CHARMS** published in 2014 and **PROBAST** published in 2019 (the most updated tool by the time when the systematic review was conducted). **CHARMS** assesses the potential risk of bias in five domains: participant selection, predictor, outcome, attrition, and analysis for model development, while **PROBAST** assesses the study quality in four domains: participants, predictor, outcome, and analysis.

Following the critical appraisal, another important step in a systematic review is to extract information from the included studies. This information are to be used for evidence synthesis. In this thesis, **CHARMS** was also used for information extraction in Chapters 4 and 5. A standard **CHARMS** checklist outlined 11 items that need to be extracted from a study: setting, source of data, population characteristics, follow-up period, sample size, outcomes, predictors, missing value, variable selection, modelling approach, model presentation, interpretation and performance evaluation (e.g., discrimination, calibration).

3.2.5 Evidence synthesis across studies

Researchers should adopt an appropriate and 'tailored' approach to synthesise the extracted evidence across the reviewed studies. In Chapters 4 and 5, narrative reports and qualitative synthesis of findings were provided, focusing on the description of study characteristics and the presentation of study quality, respectively.

Usually, evidence from a systematic review can be synthesised into two types: narrative and quantitative. When the quantitative analysis is conducted, the estimates of

model performance measures (e.g., models' sensitivity, specificity) can be extracted and meta-analysed (Snell et al., 2016). However, when quantitative analysis is not conducted then studies need to qualitatively synthesise the information. Collating evidence from qualitative synthesis is challenging as distinguishing linguistic statements can be time-consuming and the interpretation of the qualitative evidence may vary across different investigators. We argue that the synthesis of the available evidence should be conducted in light of the review question and the specific context of studies. For example, in our first systematic review (Chapter 4), we aimed to identify the existing prediction models for the incidence and progression of periodontitis, therefore, we emphasised the identification and description of the commonly used predictors and modelling approaches. Moreover, the aim of our second systematic review (Chapter 5) was to examine the quality of the included studies. We therefore highlighted the findings from critical appraisal and adopted tables and Gantt charts to provide an understanding of the limitations and overall quality of those studies, in a visual-friendly manner.

3.2.6 Reporting and presentation

Communicating and presenting the results from systematic reviews (and meta-analysis) in sufficient details are important. The PRISMA statement (Moher et al., 2009) was originally developed for reporting systematic reviews (and meta-analysis) of intervention studies, with many items being suitable to report systematic reviews of prediction modelling studies. Therefore, these two systematic reviews were reported and presented following the PRISMA statement.

3.3 Data used in the empirical analysis in this thesis

Data for this thesis were drawn from two open-access databases: the Surveillance, Epidemiology, and End Results (SEER) program and the Dental Practice-Based Research Network (DPBRN) in the US. The US states covered by the SEER program and the DPBRN project are mapped in Figure 3.1).

3.3.1 Data used in Chapter 6: SEER program

The SEER program collects information on cancers prevalence, incidence, mortality and survival covering over 30% of the population in the US (Hankey et al., 1999). There are 18 registration stations across the country (states can be found in Figure 3.1). These registration stations collect data using SEER*STAT software and submit the information to the National Cancer Institute for data aggregation and statistical analysis. The data collected contains information on patients' sociodemographics (e.g., sex, age) and clinical characteristics (e.g., tumour site, tumour size, differentiation stage at diagnosis and treatment). In addition, SEER registries follow up with the participants for vital status

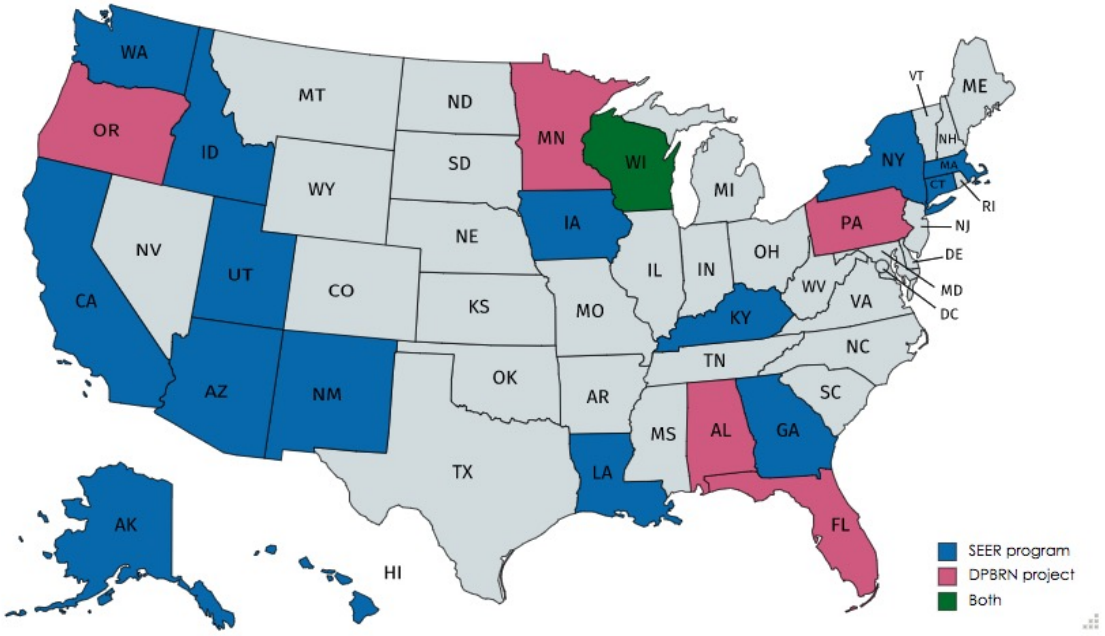


Figure 3.1: The US states from which the data for studies in the thesis were drawn. States coloured in blue are covered by the SEER program. States coloured in pink are covered by the DPBRN. State(s) coloured in green is covered by both SEER and DPBRN.

to provide survival information. Because of its broad coverage and comprehensive data collection, the SEER data were used in this thesis to serve as a basis for the survival prediction modelling using machine learning. SEER releases data submissions annually, containing new incidences and updated information for existing cases. Our study (Chapter 6, Publication 3) is based on the data from 1973 to 2015, released in November 2017. The Data-Use agreement for SEER 1973-2015 file can be found in Appendix A.

Outcome and predictors in Chapter 6 A summary of predictors and outcome variables and their descriptions can be found in Table 3.4. Among these variables, 'Survival months' and 'Death status' were used as outcome data.

Table 3.4: Description of the predictor and outcome variables used in SEER study

Variable	Description	Type
Age	Age at time of diagnosis	Continuous
Sex	Male or female	Binary
Race	White, Black, American Indian/Alaska native, Asian or pacific islander, Others	Categorical
Marital status	Marital status at diagnosis	Categorical
Grade	Tumour grading and differentiation	Categorical

Table 3.4 continued from previous page

Derived AJCC [1] Stage Group, 6^{th} ed (2004+)	Tumour stage - based on T, N, and M [2]	Categorical
Derived AJCC T, 6^{th} ed (2004+)	AJCC component describing tumour size	Categorical
Derived AJCC N, $6th$ ed (2004+)	AJCC component describing lymph node involvement	Categorical
Derived AJCC M, 6^{th} ed (2004+)	AJCC component describing tumour dissemination to other organs	Categorical
Histology	Histology type	Categorical
CS tumour size (2004+)	Information on tumour size	Categorical
No. of lymph nodes removed	Information on the involvement of lymph nodes	Categorical
Surgery	Whether or not the patient underwent a surgery	Binary
Survival months	Number of months that patient is alive from date of diagnosis	Count
Death status	Alive or death	Categorical

[1]AJCC: American Joint Committee on Cancer, [2]TNM: Tumour size, lymph Node involved, Metastasis

3.3.2 Data used in Chapter 7: the DPBRN endodontic study

In Chapter 7 (Publication 4), a data set based on patients' electronic dental records was used. It was derived from an endodontic study involving 62 dental practitioner-investigator from the DPBRN who recruited 708 patients - over 6 months - receiving initial root canal treatment. The DPBRN is research networking between a group of dental practices in the US that share expertise and work together on research questions.

Study design and participants in Chapter 7

We employed a prospective cohort where data were collected at four time points: pre-operation, intra-operation, one-week and six-month after operation. As shown in Figure 3.2, data collected at time points coloured in yellow were used as predictors and data collected at time points coloured in green were used as outcomes. Further, the one-week postoperative information was also used for predicting the six-month outcome.

Outcome and predictors in Chapter 7

Outcome The outcome measures of pre-operative, intra-operative and postoperative pain were assessed using the Graded Chronic Pain Scale (GCPS). (Smith et al., 1997; Von Korff

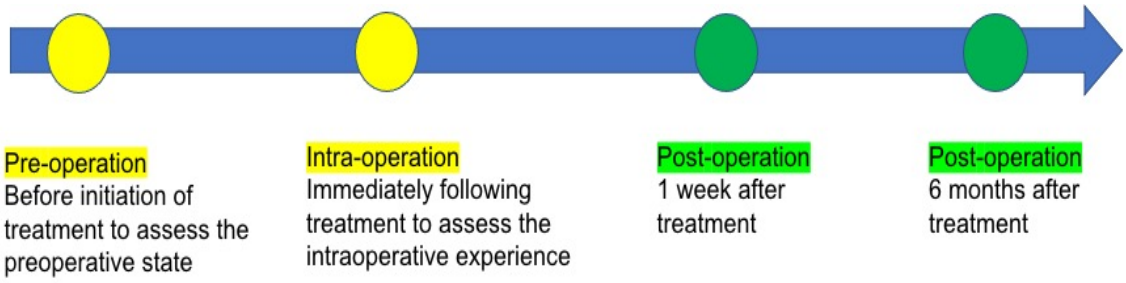


Figure 3.2: Time points for data collection in the root canal treatment cohort.

et al., 1992). The intensity of the one-week postoperative pain is defined by '*In the past one week, on average, how intense was your tooth pain rated on a 0 to 10 scale?*'. If the patient answered '≥ 7', then we defined the individual as a positive observation for the one-week cohort. The six-month outcome measure of persistent tooth pain is defined by '*On how many days in the last one month have you had tooth pain in the root canal treated tooth?*'. If the patient answered '≥ 1 day', then we defined the individual as a positive observation for the six-month cohort.

Predictors Based on the electronic health records, a total of 76 variables, obtained from questionnaires answered by patients and dentists, were used as predictors. These predictor variables can be grouped into the following categories:

- Patient sociodemographic characteristics (e.g., sex, age, insurance status)
- Pain measures (e.g., pre-operative pain intensity, the experience of intra-operative pain)
- Psychosocial variables (e.g., dental anxiety, dental fear)
- Medical characteristics (e.g., diabetes)
- Procedural characteristics (e.g., type of root canal, pulp status, presence of periapical periodontitis, number of appointments, use of rubber dam, procedural difficulty)

3.4 Analytic approaches

The adoption of prediction modelling approaches usually depends on the data type of the outcome variables. In this thesis, two common types of outcome data were included in Chapters 6 and 7, respectively: a time-to-event outcome and a binary outcome.

Overview of the analytical methods used in Chapters 6 and 7 To understand the methodology of a prediction modelling study, it is common to answer the following questions: *'What is predicted, in whom, for whom, and how?'*, i.e., What is the outcome(s)?

Who are the model users? Who is the target population? and How to conduct this study? Methods used in Chapters 6 and 7 are presented in Table 3.5.

Table 3.5: What is predicted, in whom, for whom, and how in Chapters 6 and 7?

Research scenario	Examples in Chapters 6 and 7
Specification of aim	**Chapter 6**: Predict the 1-5 year survival probability for patients with oral and pharyngeal cancers **Chapter 7**: Inform clinicians and patients about the risk of developing pain following root canal treatment.
Outcome (What is predicted?)	**Chapter 6**: Three- and five-year disease-specific survival of the patients with oral and pharyngeal cancers **Chapter 7**: One-week and six-month pain following a root canal treatment
Target population	**Chapter 6**: Adults patients diagnosed with oral and pharyngeal cancers **Chapter 7**: Adult patients undergoing root canal treatment
Model users (Who will the prediction benefit?)	**Chapters 6 and 7**: Clinicians, patients, and researchers who are interested
Methods used for predictor selection	**Chapter 6**: Predictors were collected based on previous literature **Chapter 7**: Multilevel LASSO [1]
Models used in analysis	**Chapter 6**: 1. Cox proportional hazard regression 2. Survival Tree 3. Random Survival Forest 4. Conditional Inference Forest **Chapter 7**: Multilevel logistic regression
How was missing data handled?	**Chapter 6**: 1.Substantive model compatible imputation 2. Random survival forest **Chapter 7**: Multiple imputation with missing indicator method
Other considerations	**Chapter 6**: Effects of unmeasured predictors on models' performance **Chapter 7**: Evaluation for models' prediction performance when the outcome is imbalanced (AUROC [2] and AUPRC [3])

[1]LASSO: the Least Absolute Shrinkage and Selection Operator; [2]AUROC: Area Under the Receiver Operating Characteristic curve; [3]AUPRC: Area Under the Precision Recall Curve

3.4.1 Chapter 6: Right-censored data and survival analysis

Right-censored data

In Chapter 6 (Publication 3), the outcome variable is time of death among patients with oral cancers. As illustrated in Figure 3.3, the observations in the data set can be distinguished into three types. The first type is that we observe the 'event' (i.e., death occurs before time t, as shown for Patients 1 and 2). The second type is that we observe the patient being alive at time t, all we know is that his/her death time is after his/her censored time t. The phenomenon is called *right censoring*, as the unknown event time is on the 'right' side of censored time t (as shown for Patients 3 and 4). The third type is that, if a patient drops out before time t, then we will never know his/her outcome status at t (as shown for Patients 5 and 6), this is called *left censoring*. In our study, we exclude the left-censored observations due to the unknown outcome status. The observations with '✓' in Figure 3.3 were included, the observations with '×' were excluded.

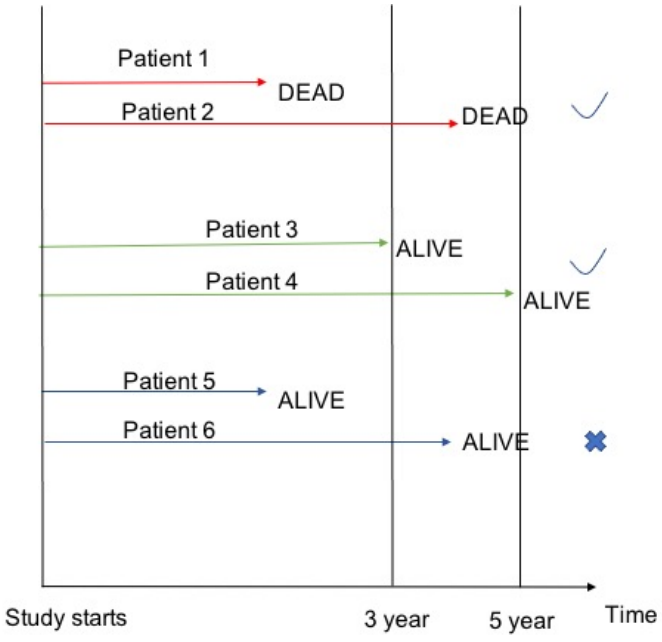


Figure 3.3: Illustration of the 'event', the left- and right-censored observations.

Survival prediction: Cox proportional hazard regression

For the usual time-to-event data setup, a commonly used analytical method is the Cox proportional hazard regression model (David et al., 1972). A Cox model computes the

impact of a given predictor(s) on the hazard (i.e., risk) of an event occurring (e.g., death). For the i^{th} individual, let's denote the predictors being $x(x_1, x_2, ...x_k)$ and the death/event time being y_i. In the Cox model, it is assumed that:

$$h(t|x) = h_0(t)exp(\sum_j x_j\beta_j) \tag{3.1}$$

where $h(t|x)$ is the hazard at time t given the predictor values of that individual, and h_0 is the baseline hazard function.

Though the Cox model has been widely adopted for survival analysis due to fast computation and straightforward interpretation, this method has some shortcomings. For example, Cox models rely on the pre-specified assumptions: the hazard of a patient is a linear function of the baseline hazard of the population and of the static predictor values over time of that patient. Additionally, Cox regression is unable to model nonlinearities and interactions between variables. To overcome these limitations, machine learning has provided alternate solutions, and are being increasingly used in survival prediction.

Survival prediction: tree-based machine learning algorithms

While methods such as support vector machine and neural networks can be used for survival prediction, real data experiments have shown that they require larger computational power and have similar efficiency when compared to the methods that do not have the same computational burden, such as tree-based methods (e.g., random forest) (Fouodo et al., 2018; Kvamme et al., 2019). For this reason, we were specifically interested in the use of 'tree-based machine learning algorithms'. In Chapter 6 (Publication 3), three tree-based machine learning approaches (survival tree, random survival forest and conditional inference forest) were adopted along with Cox regression. A detailed description of tree-based models can be found in Chapter 6.

A survival tree:

In the terminology of tree models, one *node* would *split* the data into two *children nodes* based on a pre-defined *splitting rule*. For the purpose of growing and pruning the tree, one needs to define a splitting statistic (e.g., log-rank) that handles the dependence of failure times and a performance metric (e.g., C-index) to evaluate the predictive ability of the tree. In the survival tree, the split rule was defined using the 'log-rank' test (Damato and Taktak, 2007). For the purpose of predicting a new observation, the *if-then* rules was followed using predictor values of that observation until coming to the terminal node. A step-by-step procedure for growing a survival tree is shown in Algorithm 1.

Algorithm 1 The development of a basic survival tree

Start

Create an initial survival tree with a root node k_0

Create a stack S of open nodes

while *S is not empty* **do**

  $k = k_0 + k_1$

  if *the stopping criterion is met for k;* **then**

    end

  else

    Find the splitting node that maximises the survival difference between the children nodes

    Partition data into two children nodes of k

  end

end

Calculate the tree's prediction performance

A random survival forest:

The implementation of a basic random survival forest follows three steps: 1) The basic survival trees are fully grown using a bootstrapped sample of the original data. 2) The tree nodes are split using the best splitting criterion among a number of candidate criteria. Usually, the one that maximises the survival difference between two children nodes is selected. 3) The prediction from a random survival forest is calculated as the average of the individual survival trees. A step-by-step procedure for growing a random survival forest is shown in Algorithm 2.

Algorithm 2 The development of a basic random survival forest

Start

Define the number of survival trees (*ntree*) to develop

for $i = 1$ *to ntree* **do**

Create a B bootstrap sample (usually two thirds of the original data set), one third is left as out-of-bag (OOB) data Develop a survival tree model using this sample

for *each split* **do**

Randomly select k ($< K$) of the candidate predictors

Select the predictor from the k predictors that can give the optimal estimates (e.g., highest homogeneity within children node)

for *each splitting point of the best k* **do**

Compare the survival curves of the two groups using one splitting rule among the multiple splitting rules

Select the best splitting rule that maximises the survival difference between two children nodes

Partition the data

(for determining splitting rule)

end

(for determining one split in one tree)

end

Using tree stopping criterion to test whether the development of a tree is completed.

Using OOB data, the prediction performance is calculated and recorded.

(for developing one tree)

end

The performance metrics of individual trees are then averaged to obtain the ensemble performance of this forest.

In Chapter 6, a handful of predictors was selected following the previous literature. However, in much of the real-world data settings, some data complexities need to be considered before modelling, such as the selection of predictors from a large set of candidate variables and the nesting of information. To demonstrate how to take these complexities into account in prediction modelling, a data set with a 'small' number of samples and a 'large' number of variables was used. Additionally, this data set has a two-level structure, which allows the use of multilevel modelling. Detailed explanations can be found in the following paragraphs.

3.4.2 Chapter 7: Multilevel data and multilevel models

Multilevel data

As shown in Figure 3.4, our data have a nested (clustered) structure: 708 patients were grouped by 62 dental practitioners across six states in the US. Data with such a hierarchical structure is called multilevel data.

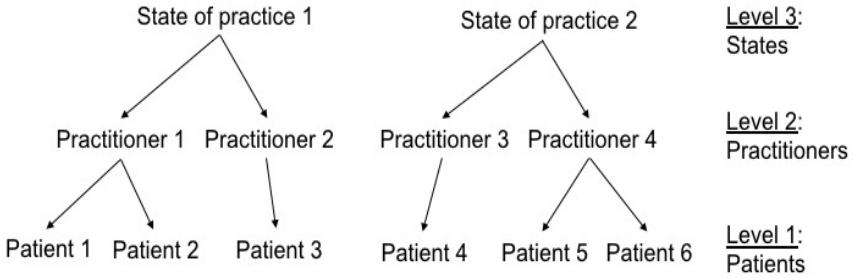


Figure 3.4: Schematic structure of the data set.

Multilevel models

When the data a two- (or multi-) level structure and also it is believed that the total variation is made up of the variation of level-1 and level-2, then one may consider using multilevel models (also refer to mixed effect models or random effect models) (Mason, 2001). Multilevel models can recognise the hierarchies in the data set by accounting for variance components at each level in the hierarchy. In our study, a two-level logistic regression model accounting for the clustering of patients' outcomes within dental practitioners was used. In this study, the practitioner residuals represent the 'random effect', referring to the unobserved practitioner characteristics that affect patients' outcomes. It is the unobserved characteristics that led to correlations between the outcomes for different patients treated by the same dental practitioner. More details of the two-level model can be found in Chapter 7 (Publication 4).

3.4.3 Variable selection in Chapter 7: the Least Absolute Shrinkage and Selection Operator (LASSO)

In Chapter 7, variable selection was an important criterion as the data set had a smaller sample size (number of participants = 708) and a large number of variables (number of variables = 152). The reasons for conducting variable selection are three:

1. Removing the irrelevant variables that do not add to the models' prediction performance can make the models easier to interpret.

2. Variable selection can speed up the training of an algorithm, especially for high-dimensional data sets.

3. If we do not select variables in this study there is a chance that the determinant(s) of the variance-covariance matrix equates to zero thus resulting in unstable effect estimates, commonly referred to as overfitting.

To select variables in a study one may use methods such as backward or forward selection in training and test data sets. However, these methods are not free from limitations (Greenland et al., 2016). Alternatively, several machine learning algorithms are available for variable selection, and in this study we used the LASSO introduced by (Tibshirani, 1996). The idea of conducting LASSO is to constrain the sum of the model parameters (e.g., the sum of variables' coefficients) to be less than a fixed value. To do so, the LASSO applies a shrinking function to penalise the variables' coefficients and shrink some of them towards zero. Then the variables with non-zero coefficients following the shrinking function are selected into the final model. Given that our data has a two-level structure, a two-level logistic regression was used for model development. To be compatible with the intended model, multilevel LASSO was therefore conducted to select predictors (Schelldorfer et al., 2011). This method is discussed in detail in Chapter 7.

3.5 Model performance measures

3.5.1 Measures for models' discrimination

Chapter 6: Discriminating measures for survival prediction models

There are a number of measures being constantly developed for evaluating the accuracy of survival predictions, of which some are reviewed by (Bøvelstad and Borgan, 2011), including the log-rank test, the Harrell's C-index (also known as C-index), and the Brier score. Following the literature which evaluates survival prediction models in medical research (Austin et al., 2017; Chen et al., 2012; Choodari-Oskooei et al., 2012; Graf et al., 1999; Nasejje et al., 2017; Pencina et al., 2012; Rahman et al., 2017; Schemper and Stare, 1996; Schmid and Potapov, 2012), C-index and the Integrated Brier Score (IBS) were used to measure models' discriminative ability in Chapter 6 (Publication 3). The definitions and calculation of C-index and IBS can be found in Chapter 6.

- In time-to-event analysis, say we have a pair of patients (i, j), and the i^{th} patient has an event (e.g., death) prior to the j^{th} patient. If the model predicts that the i^{th} patient having a higher risk score than the j^{th} patient, then this is a concordant pair. The C-index is calculated as the proportion of concordant pairs divided by the total number of possible pairs. It measures the probability of successfully predicting the sequence of events for a random pair of cases. A C-index of 0.5 or lower indicates the model predicting an outcome no better than random chance and a higher C-index indicates a higher predictive ability.

- The Brier score was first introduced in 1950 by (Brier, 1950) as a method of assessing the prediction error of weather forecasts. For binary prediction models, the Brier score is calculated as the mean squared prediction error. In the study by (Graf et al., 1999), the Brier score was adopted as a performance measure for predicting survival up to sometime t, and was given the name 'IBS'. For models based on time-to-event data, the Brier score is defined as: at a given time point t, the average squared distances between the predicted survival probability and the observed survival status. The IBS represents an overall estimate of the models' prediction error at all available time points. Similar to the C-index, the IBS is a number between 0 and 1, with 0 being the best value.

Chapter 7: Discriminating measures for binary classification models

In Chapter 7, two measures were used to evaluate the discriminative ability of binary classifiers: the AUROC and the AUPRC. We here define what these two measures are, describe why they are suitable for our study, and present how to interpret these measures.

What are AUROC and AUPRC?

- For a binary classification problem, when two cases (a positive and a negative case) are selected at random, the AUROC reflects the probability of the classifier assigning a higher risk score to the positive case (Hanley and McNeil, 1982). In a plot of a ROC, the x-axis is the sensitivity (recall) and the y-axis is 1-specificity (false positive rate). Therefore, AUROC is calculated as the average sensitivity, regarding all values of the specificity as equally likely (Hand, 2009).

- In a plot of a PRC, the x-axis is the sensitivity (recall) and the y-axis is the precision (positive predictive value). Therefore, AUPRC is calculated as the average positive predictive values, regarding all values of the sensitivity as equally likely (Saito and Rehmsmeier, 2015).

Besides AUROC, why was AUPRC used as an additional measure?

For binary classification tasks, the two outcome categories are often labelled as positives (e.g., diseased) and negatives (e.g., disease-free). Sometimes the interest of a study would be to get the prediction of both the outcomes correct (i.e., positives, '1' and negatives, '0'), else the interest can be in getting the prediction correct only for the positives. Moreover, when the outcome distribution is imbalanced (the negatives being more than the positives), then predicting the positives become more important. The AUPRC is used as a performance metric when the outcome data are imbalanced and the investigators care a lot about recognising the positives. If we are interested in how the model performs on both the

positive and the negative categories, then AUROC is a good choice. A typical example is for classifying images between dogs and cats. In this case, there is no distinguishing between the positives and negatives, we thus want the model to perform equally well in both the dogs and cats categories. However, if we are not interested in how well the model performance is on the negatives (i.e., the true negatives are less of a concern), but just focus on positive prediction (i.e., we want a high value of sensitivity and to have as many of the positives classified as positives as possible), then AUPRC is a better choice. For example, in our case for predicting pain following root canal treatment, the proportion of developing one-week and six-month pain were 24% and 11%, respectively, we considered it as a data set with imbalanced outcome distribution. And for this research question, we care less about how many of the negative predictions (non-pain cases) are correct, but we want to ensure none of the positive observations (pain cases) missed out, and that most of the positive predictions are correctly classified.

How to interpret AUROC and AUPRC?

There are similarities and difference between the interpretation of AUROC and AUPRC. While the higher values for both AUROC and AUPRC indicate the better models' performance, they differ in terms of what is considered as a 'good' baseline value. Differing to AUROC where a baseline value always being 0.5, the baseline value of AUPRC is equal to the proportion of positive observations (Saito and Rehmsmeier, 2015), calculated as $Positives/(Positives+Negatives)$. Thus, models built on different data sets have different AUPRC baselines. For example, a data set with 10% positives has an AUPRC baseline of 0.1, so if a model obtains an AUPRC of 0.3, then it means using the model is better than not using the model. However, for a data set with 80% positives as baseline, a model with an AUPRC of 0.7 maybe viewed as worse than not using the model.

3.5.2 Measures for models' calibration

Besides discrimination, the performance of prediction models also relates to the agreement between the predicted and the observed probabilities, and this is called models' calibration (Van Calster et al., 2016). In this thesis, calibration curves were adopted to present the models' calibration ability for both empirical examples. Specifically, in Chapter 7, we used a method to create a confidence belt for the calibration curve (Nattino et al., 2016). Compared to the standard calibration curve, the calibration belt can spot the range and direction of deviation from the ideal calibration, therefore providing suggestions for revising the model.

CHAPTER 4

Prediction models for the incidence and progression of periodontitis ——A systematic review

Preface

The primary aim of this chapter is to identify existing prediction modelling studies in the field of periodontology. Additionally, this chapter is included in this thesis to achieve the second aim: to demonstrate a systematic approach for collecting evidence on the prediction modelling research in the field of oral health. We were interested in periodontitis because it is one of the most prevalent oral diseases and prediction modelling research around this discipline has been largely investigated. This paper acts as our first step to gain insights into clinical prediction models in oral health.

This chapter contains the first of a series of four studies contributing to this thesis. Details of this publication are:

Du M., Bo T, Kapellas K, Peres M. Prediction models for the incidence and progression of periodontitis: A systematic review. Journal of Clinical Periodontology. 2018;45:1408-1420. https://doi.org/10.1111/jcpe.13037.

The accepted version of the published paper is reproduced as follows.

Statement of Authorship

Title of Paper	Prediction models for the incidence and progression of periodontitis: A systematic review
Publication Status	√ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Du M., Bo T, Kapellas K, Peres M. Prediction models for the incidence and progression of periodontitis: A systematic review. Journal of Clinical Periodontology. 2018;45:1408-1420.

Principal Author

Name of Principal Author (Candidate)	Mi Du		
Contribution to the Paper	MD contributed to conception, design, studies selections, data extraction, critical appraisal, results interpretation, drafted and critically revised the manuscript.		
Overall percentage (%)	90%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24-04-2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Tao Bo		
Contribution to the Paper	TB contributed to studies selections, and critically revised the manuscript.		
Signature		Date	28-04-2021

Name of Co-Author	Kostas Kapellas		
Contribution to the Paper	KK contributed to conception, design, critical appraisal, results interpretation, and critically revised the manuscript.		
Signature		Date	28/04/2021

Name of Co-Author	Marco Peres		
Contribution to the Paper	MP contributed to conception, design, critical appraisal, results interpretation, and critically revised the manuscript.		
Signature		Date	28/04/2021

4.1 Abstract

Aim: To comprehensively review, identify and critically assess the performance of models predicting the incidence and progression of periodontitis.

Methods: Electronic searches of the MEDLINE via PubMed, EMBASE, DOSS, Web of Science, Scopus and ProQuest databases, and hand searching of reference lists and citations were conducted. No date or language restrictions were used. The Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies checklist was followed when extracting data and appraising the selected studies.

Results: Of the 2,560 records, five studies with 12 prediction models and three risk assessment studies were included. The prediction models showed great heterogeneity precluding meta-analysis. Eight criteria were identified for periodontitis incidence and progression. Four models from one study examined the incidence, while others assessed progression. Age, smoking and diabetes status were common predictors used in modelling. Only two studies reported external validation. Predictive performance of the models (discrimination and calibration) was unable to be fully assessed or compared quantitatively. Nevertheless, most models had 'good' ability to discriminate between people at risk for periodontitis.

Conclusions: Existing predictive modelling approaches were identified. However, no studies followed the recommended methodology, and almost all models were characterized by a generally poor level of reporting.

4.2 Introduction

Periodontitis has become a significant global healthcare problem with increasing costs for both the individual and society (Tonetti et al., 2017). Periodontitis is one of the leading causes of tooth loss and ranks as the sixth most prevalent disease globally (Kassebaum et al., 2017). Early identification of people at risk of periodontitis and early treatment are important to retain teeth and improve oral health-related quality of life (Ramseier et al., 2017; Tan et al., 2016). In order to improve prevention, efforts in epidemiology have shifted from identifying new risk factors to developing viable algorithms to assess individuals at risk (Page and Beck, 1997). The American Academy of Periodontology (AAP) stated: *'the use of risk assessment will become a component of all dental and periodontal evaluation as well as part of all periodic dental and periodontal examination'* (American Academy of Periodontology, 2008). Criteria and measures of periodontitis and progression vary across the literature (Savage et al., 2009). The World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions recently agreed on a new classification and case definition system based on a multidimensional staging and grading system, which not only reflects the severity of disease, but also accounts for the risk of aggressiveness and progression (Tonetti et al., 2018). Previously, several risk assessment tools for periodontitis were based on a list of single factors developed to delineate different risk levels, and often risk portrayed the extent and severity of periodontal status (Lang et al., 2015). That is, people with fewer risk factors and presenting little periodontal destruction are assumed to be at 'low' risk for developing and worsening of the disease, whereas those deemed as 'high' risk are considered to have a high probability of future disease. Despite these tools fitting the definition of 'risk assessment' as defined by the AAP (American Academy of Periodontology, 2008), most are qualitative assessments, as they do not calculate the accurate probability required by prediction studies (Collins et al., 2014) .

Prediction in medicine includes both studies of the presence of disease (diagnosis) or an event in the future course of disease (prognosis). A prediction model contains more than two predictors and aims at converting observed values in individuals to absolute and objective probability, going beyond correlation coefficients and risks (Steyerberg and Vergouwe, 2014). Beside the goodness of fit, a prediction model should be evaluated in terms of discrimination, which is a model's ability to distinguish individuals with and without an outcome event; and calibration, which is the agreement between predicted and observed outcomes (Harrell et al., 1996; Steyerberg et al., 2010; Vergouwe et al., 2002). Prediction is not new in medicine, and there are a range of prediction models existing in the fields of cardiovascular diseases (Shariat et al., 2008), cancers (Altman, 2009; Shariat et al., 2008), stroke (Counsell and Dennis, 2001), diabetes (Collins et al.,

2011; Damen et al., 2016), and reproductive medicine (Leushuis et al., 2009). Compared to subjectively made predictions, prediction models provide more accurate and fewer variable estimates of risk (Kattan et al., 2013; Ross et al., 2002). However, systematic reviews evaluating the methodology and reporting of prediction modelling studies all conclude that these studies are deficient in study design, statistical approaches, and suffer from poor reporting (Bouwmeester et al., 2012; Collins et al., 2013; Jaja et al., 2013).

Several models aiming at predicting periodontitis prevalence, incidence, progression, and tooth loss have been developed, but their performance, validity, and clinical applicability raise concerns (Schwendicke et al., 2018). In addition, a universally accepted objective method/model for calculating the probability of prospective development or deterioration of periodontitis does not exist. A review of all existing prediction models is lacking, particularly in terms of their methodological quality, validity and clinical reliability.

Thus, the aim of this study was to comprehensively review, identify and critically appraise studies presenting prediction models for periodontitis incidence and progression. Specifically, the review question was: '*What models with clustered risk factors predict the incidence and progression of periodontitis in adults? Moreover, the existing modelling approaches are of special interests*'.

4.3 Methods

The protocol for this review was prepared according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). Registration of the protocol was on PROSPERO (No.CRD42018085437) on 13rd March 2018, prior to the formal commencement of this systematic review, and an update on 26th April 2018 was made to formally include the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) (Moons et al., 2014) criteria into the evaluation methodology. Structure of this systematic review was prepared based on Cochrane Handbook of Systematic Reviews of Intervention (5.1 version) (Higgins and Green, 2011) and Joanna Briggs Institute Reviewers' Manual (4th edition) (Aromataris and Munn, 2017).

4.3.1 Inclusion and exclusion criteria

The predictors include, but are not limited to: tooth-related factors (initial periodontal status), oral health-related factors (tooth brushing, interdental cleaning, pattern of dental visits), subject-related factors (smoking, diabetes, alcohol consumption, overweight/obesity), inherited factors (family history of periodontitis), psychological factors, and socioeconomic/demographic factors.

Studies were eligible if they met the following criteria: (1) Described the development, validation or assessment of a model that was constructed to predict the incidence or progression of periodontitis used in the general population; (2) Population-based cohort studies with samples selected probabilistically; (3) Targeted on adults aged 18 years or over; (4) The outcomes were periodontitis incidence or progression; (5) The model contained at least two risk factors as predictors.

Studies were excluded if they were: (1) Not original data, such as review article, meta-analysis, letter to editor, editorials or comments on prediction studies; (2) With participants younger than 18 at baseline; (3) Targeting on specific populations like those with chronic diseases, pregnant, intellectual disability, HIV, drug-dependent person and alcohol-dependent person; (4) Models including a single predictor, test, or marker; (5) With no probabilistic sample; (6) Not peer reviewed.

4.3.2 Search strategy

Six electronic databases (MEDLINE via PubMed, EMBASE, DOSS, Web of Science, Scopus, and ProQuest) were used for article searching and collection (supporting information included in Appendix B Supplement 1). The primary search strategy was constructed based on four domains ('periodontitis' AND ('prediction' OR 'risk factors' OR 'risk assessment') AND ('incidence' OR 'progression') AND general aspects for longitudinal studies). No time and language restrictions were used. Additional searches were performed by reviewing the bibliography and citation of the retrieved full-text articles.

4.3.3 Study selection

Two reviewers (MD, TB) independently scanned titles and abstracts in parallel and selected the articles that meet the inclusion criteria, then full texts of the selected articles were read. There was disagreement for three studies, where the referees (MP, KK) were called and an agreement was reached.

4.3.4 Data extraction

Two reviewers (MD, TB) collected key characteristics of the study and quantitative data related to results by pre-defined data-abstraction forms. The CHARMS (Moons et al., 2014) (Appendix B Supplement 2) was used for data extraction.

Study characteristics included data such as author/publication year, setting/context, type of study, participant characteristics (mean age at baseline, sex, study inclusion/exclusion information), sample size, outcomes to be predicted, periodontitis classification criteria or definition for the outcome, predictors. Quantitative data extraction focused on the two most common statistical measures of predictive performance: discrimination and calibration. In terms of model development, type of model, missing values, selection of candidate predictors and selection of final predictors were extracted. In terms of model

performance, the reported concordance (C) statistic, sensitivity, specificity, false-positive and false-negative proportions were extracted; other data related to model evaluation (external/internal validation), model presentation (formula/ score chart), and model interpretation were extracted. If a model presented several cut-off scores, the one that represented the highest sensitivity model was selected (Verstraete et al., 2015).

4.3.5 Risk of bias assessment

The CHARMS checklist for critical appraisal of prediction modelling studies was used to assess the risk of bias that may occur in participant selection, predictor, outcome, attrition, and analysis for model development. CHARMS defines the risk of bias as 'low', 'moderate' and 'high' level. Cohort studies were assessed by the Newcastle-Ottawa Scale (NOS) in three domains: selection, comparability, and outcome. Detailed criteria of the CHARMS checklist and NOS are provided in Appendix B Supplements 3 and 4.

4.3.6 Data synthesis and reporting

Due to study heterogeneity, it was deemed inappropriate to conduct a meta-analysis. Therefore, a narrative report of findings is provided.

4.4 Results

4.4.1 Studies searches and selection

2,560 titles and abstracts were identified, of which 41 papers were examined in full text, however, no grey literature was identified. Reasons for exclusion are shown in S5, common reasons included tooth- and site-level studies, and validation study of a risk assessment tool. Finally, seven studies were included in the analysis (Figure 4.1). Three papers reported risk assessment models, three papers proposed prediction models, and one reported both.

4.4.2 Characteristics of studies and data extraction

Overview

The characteristics of studies are shown in Tables 4.1 and 4.2. Three studies were from the USA, one each was from Brazil, Italy, Spain, Sweden, and Switzerland.

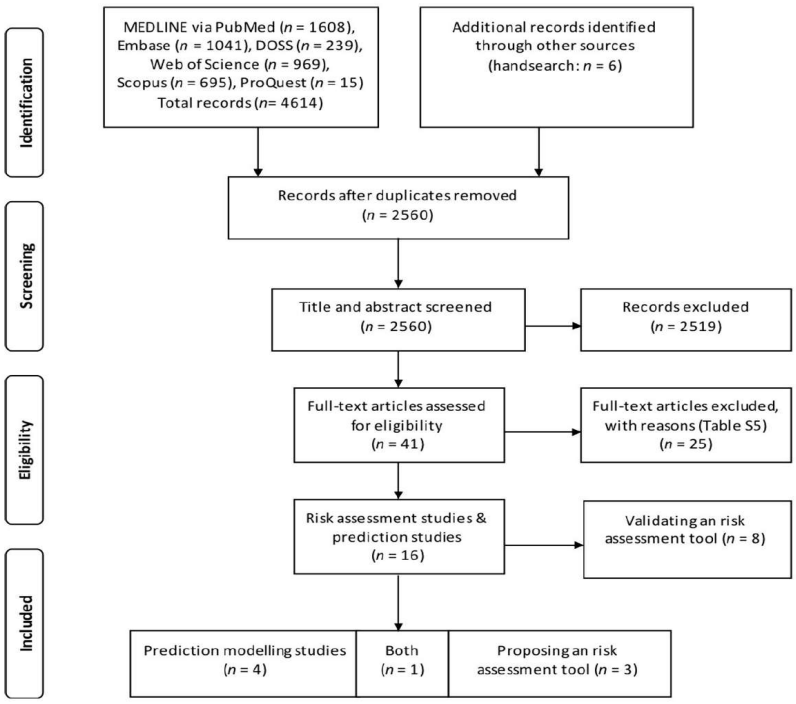


Figure 4.1: PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow diagram of the studies search and selection.

Table 4.1: Characteristics of various risk assessment tools.

Risk assessment tool		Aim	Variables (n)	Description	Validation or a application	Updating
Page 2002	Periodontal risk calculator (PRC)	To provide a risk score of a patient's susceptibility for the progression of periodontal disease.	9	Computer-based tool, Weighted mathematical algorithm. PRC assigns the individual a level of risk on a scale from 1 (lowest risk) to 5 (highest risk). Quantifies Disease level 1 - 100 Scale (1= health - 100 = severe disease) (www.previser.com)	Retrospective Cohort Study (Page et al., 2003) (USA); Retrospective Cohort Study (Martin et al., 2009) (USA); Retrospective Cohort Study (Martin et al., 2011) (USA); Retrospective Cohort Study (Trombelli et al., 2017) (Italy)	Later incorporated with disease severity score to form oral health information suite (OHIS). Periodontal assessment tool (PAT) is the periodontal assessment component. (Page, 2007)

Table 4.1 continued from previous page

Lang & Tonetti 2003	Periodontal risk assessment (PRA) (hexagonal risk diagram)	To classify patients as low, medium and high risk for the progression of periodontal disease.	6	Diagram with six parameters. Each vector represents one risk factor or indicator. All factors have to be evaluated together and hence, the area of relatively low risk is found within the centre circle of the polygon, while the area of high risk is found outside the periphery of the second ring in bold. Between the two rings in bold, there is the area of moderate risk.	Retrospective Cohort Study (Jansson and Norderyd, 2008) (Sweden); Retrospective Cohort Study (Eickholz et al., 2008) (Germany); Retrospective Cohort Study (Matuliene et al., 2010) (Switzerland); Prospective Cohort Study (Costa et al., 2012) (Brazil)	PRA/hexagonal risk diagram (Persson et al., 2003) ; Periodontal Risk Assessment Diagram Surface (PRAS) (Leininger et al., 2010); PPRDs (Periodontal pentagon risk diagrams) (Renvert and Persson, 2004); Modified PRA (Chandra, 2007)

Table 4.1 continued from previous page

Trombelli 2009	University of Ferrara (UniFe) Score	To compare the simplified risk score of a patient's susceptibility for the progression of periodontal disease with modified PRC.	5	Algebraic sum of the parameter scores is calculated and relates to a risk score between 1 and 5.		

Table 4.1 continued from previous page

Lindskog 2010	DRS a patient risk score (DRS denti-tion) or tooth risk score (DRS tooth).	To provide a patient level risk score and a tooth level risk if the patient level risk is found to be elevated.	22		A Web-based analytic tool that calculates chronic periodontitis risk for the dentition (Level I) and, if an elevated risk is found, prognosticates disease progression tooth by tooth (Level II).	

Table 4.2: Characteristics of prediction models

Study		Setting/ Context	Source of data	Participants characteristics	Follow-up (Year)	Outcomes to be predicted	Sample Size	Predictors (*n*)	Outcome definitions/ classifications
Author/ Year	Model								
Beck 1994	Risk model (Blacks)	USA, Community	Pro. cohort[1]	Blacks; age > 65 years	1.5 and 3	Incidence of CAL[2]	169	6	> 2 sites with > 3mm of CAL
	Risk model (Whites)	USA, Community	Pro. cohort	Whites; age > 65 years	1.5 and 3	Incidence of CAL	169	6	> 2 sites with > 3mm of CAL
	Prediction model	USA, Community	Pro. cohort	Whites; age > 65 years	1.5 and 3	Incidence of CAL	169	4	> 2 sites with > 3mm of CAL
Leite 2017	Model 1	Brazil, Birth cohort	Pro. cohort	52.3% Female; Age 24 years	7	Periodontitis occurrence	471	9	Mild: <10% of sites with CAL 3mm;
	Model 2	Brazil, Birth cohort	Pro. cohort	52.3% Female; Age 24 years	7	Periodontitis occurrence	471	7	EFP definition[3]
	Model 3	Brazil, Birth cohort	Pro. cohort	52.3% Female; Age 24 years	7	Periodontitis occurrence	471	9	CDC-AAP definition[4]

Table 4.2 continued from previous page

	Model 4	Brazil, Birth cohort	Pro. cohort	52.3% Female; Age 24 years	7	Periodontitis occurrence	471	5	1 sites CAL 4mm + BOP[5] in 1 site
Lindskog 2010	DRS dentition	Sweden, Dental clinics	Pro. cohort	55% Female; Mean age 47.9	Mean 3.8	Periodontitis progression, radiographic marginal BL[6].	183	11	Periodontitis progression was defined: (1) at any proximal surface (radiographic marginal BL, FI[7], or angular bony destruction); or (2) at any proximal, facial,or oral surface; or (3) increased in severity (radiographic marginal BL or FI) between baseline and follow-up.
	DRS tooth					Tooth loss	1408 teeth		
Martinez-Canut 2018	Molars	Spain, Dental practice	Retro. cohort[8]	68.8% Female; Age 40.3±9.07	20.2 ± 2.04	TLPD[9]	500	11	TLPR were defined (molars): BL > 50% associated to FI grade III and repeated abscesses.
	Non-molars	Spain, Dental practice	Retro. cohort	68.8% Female; Age 40.3 ± 9.07	20.2 ± 2.04	TLPD	500	11	TLPR were defined (non-molars): spontaneous exfoliation; and BL > 75% with mobility of grade III, which caused pain under function or spontaneously.

Table 4.2 continued from previous page

Morelli 2018	Index of periodontal classes (IPC)	USA, Community	Pro. cohort	54.5% Female; Mean age 62.4	10	Periodontitis progression and tooth loss	3985 development; 697 validation	6	Periodontitis progression was defined: 10% of sites exhibiting 3mm attachment loss in a 3-year period.

[1]Pro. cohort: prospective cohort study. [2]CAL: clinical attachment loss. [3]EFP definition: Sensitive: defined as presence of proximal attachment loss of 3mm in 2 non-adjacent teeth; or Case: proximal attachment loss of 5mm in 30% non-adjacent teeth. [4]CDC-AAP definition: Healthy: No mild, moderate, or severe periodontitis; Mild: 2 proximal sites with CAL 3mm AND 2 proximal sites with PD 4mm; Moderate: 2 proximal sites with CAL 4mm OR 2 proximal sites with PD 5mm, in different teeth; Severe: 2 proximal sites with CAL 6mm in different teeth AND 1 proximal sites with PD 5mm. [5]BOP: Bleeding on probing. [6]BL: bone loss. [7]FI: furcation involvement. [8]Retro. cohort: retrospective cohort study. [9]TLPD: Tooth loss due to periodontal disease.

For risk assessment tools, in 2002, Page et al. introduced the Periodontal Risk Calculator (PRC) (Page et al., 2002), and in 2007, they added disease status score to the risk score (PreViser) (Page, 2007), which became a component of the Oral Health Information Suite. Based on 11 parameters: age, smoking, diabetes, history of periodontal surgery, periodontal probing depth (PPD), bleeding on probing (BOP), furcation involvement (FI), root restorations or sub-gingival calculus, radiographic bone height and vertical bone lesions, the risk score and disease severity score can be calculated to establish both risk assessment as well as disease severity. The Periodontal Risk Assessment (PRA) model, proposed by Lang et al in 2003, is a multifactorial graphic composed of six vectors representing six systemic and clinical factors: systemic and genetic aspects, diabetes, cardiovascular disease, smoking, percentage of sites with BOP, prevalence of residual pockets > 5mm (residual pocket > 4mm), number of tooth loss, loss of periodontal support in relation to the patient's age. In contrast to PRC, the targeted population are patients during the supportive periodontal treatment. Moreover, PRA was designed to classify patients as either low-, moderate- or high-risk profile. Some validation and updating studies were conducted related to PRC and PRA (Table 4.1). The simplified risk assessment model (UniFe) proposed by Trombelli et al. (Trombelli et al., 2009) included the five key parameters: smoking, diabetes, BOP, number of sites with PD > 5mm, and radiographic bone loss-to-age ratio. Patients are assigned to five risk categories: score 1 (low), 2 (low-medium), 3 (medium), 4 (medium-high), and 5 (high). In general, all of these risk assessment tools share several common attributes. They all aimed at assigning patients into one of those risk categories, but did not provide the predicted probability of getting the outcome. For prediction modelling studies, they varied in derivation cohort, follow-up period, type of predictors, and the definition of outcome (Table 4.2). Of the five prediction modelling studies, one acted as a validation study, while the other four were model development studies, without (n = 3) or with (n = 1) external validation. Twelve prediction models were identified, of which three were based on tooth level and nine were on individual level. Two models were explicitly developed only for white or black populations in the US.

Study design and participant sampling

While prediction model development and validation can be achieved from nested case-control or case-cohort data sets, prospective cohort studies are the preferred option (Moons et al., 2009). Four studies used prospective cohort data while one utilised data from a retrospective cohort study. Participant recruitment was well-described in all five studies. Three studies involved participants from clinical settings with an average age of 59.5 years, and two used cohort data sets from general populations. Of the cohort studies, one was a birth cohort, which recruited their sample at age 24 years while the other investigated

people aged 65 years and over longitudinally. Follow-up times ranged from 3 to 20 years. Study samples sizes ranged from 183 to 3985 participants.

Outcome

The outcome definition of periodontitis varied across studies, and ultimately eight criteria were identified (Table 4.2). Four models from one study predicted the occurrence of periodontitis using four alternate case definitions; three models from one study predicted the incidence of attachment loss of 3mm or more, the other three studies with five models predicted periodontitis progression, of which tooth loss was mostly selected as one of the outcomes of progression.

4.4.3 Development, presentation, and performance of the prediction models

Candidate predictors and selection of final predictors

Figure 4.2 and Appendix B Supplement 6 list > 30 variables included across the prediction models. Most incorporated the well-established periodontitis risk factors age, smoking and diabetes. Other common predictors included oral examination parameters such as BOP, clinical attachment loss (CAL), and degree of tooth loss. The number of predictors in the studies varied between four and 11 and were selected based on clinical knowledge or literature. Two studies applied predictor selection in multivariable analyses: one reported criterion for predictor selection was p-value = 0.05 in univariate analysis (Beck, 1994), whilst the other used variables in nodes of the first two leaves from the decision tree analysis (DTA) (Leite et al., 2017). Both reported using automated variable selection (forward and/or backward selection and stepwise) regression procedures to decide the final variables.

Missing values

Three studies handled missing data by excluding subjects lost to follow-up from the analyses, while the other two studies did not provide a description of how missing data was managed.

Development and performance of the models

Models within a particular study used the same model generation method however, none of the five studies used the same statistical methods. At least one measure of predictive performance was reported for all the 12 models. Discrimination was reported in the form of C-statistics, Area under the receiver operating characteristic curve (AUROC), sensitivity + specificity, positive/negative predictive value (PV+, PV-), but none of these models reported calibration. Detailed information is listed in Tables 4.3 and 4.4.

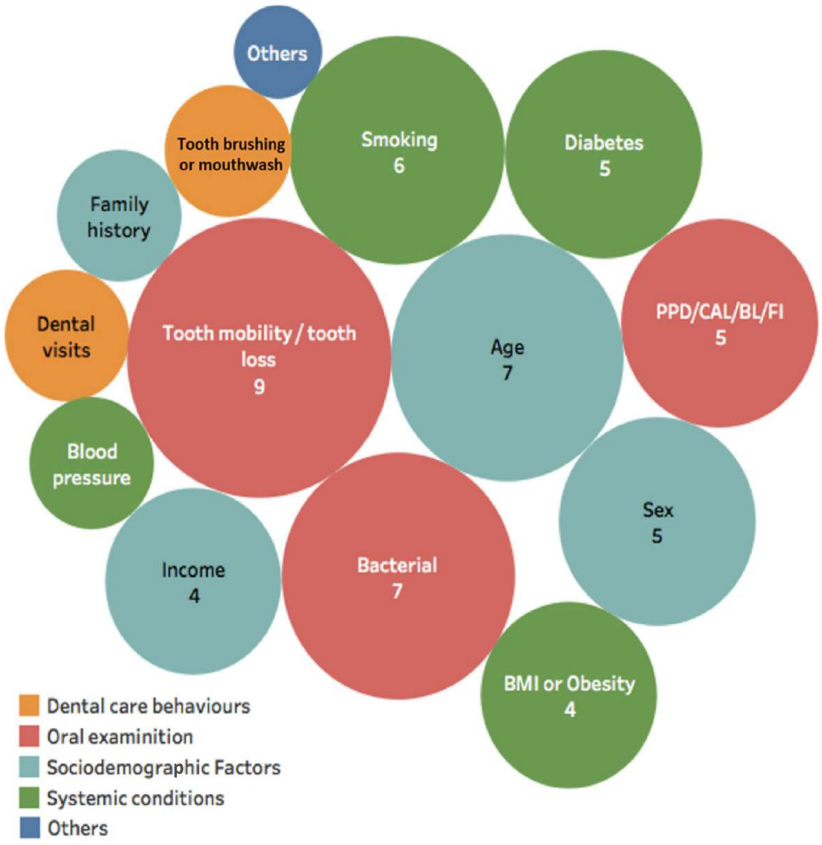


Figure 4.2: Frequency of identified risk predictors in the final prediction models. BL: bone loss; BMI: body mass index; CAL: clinical attachment loss; FI: furcation involvement; PPD: periodontal probing depth. Others includes education, race, gingival recession, alcohol, bruxism, patient awareness, therapist's experience, stress-related factor. Bacterial includes plaque, calculus, and/or presence of periodontal pathogens. Frequency for "*Tooth brushing or mouthwash*", "*Family history*", "*Dental visits*" and "*Blood pressure*" is 2, frequency for "*Others*" is 1.

Table 4.3: Model development, presentation, and interpretation.

Study		Missing value	Variable selection	Model development	Model presentation	Model interpretation
Author/ Year	model					
Beck 1994	risk model (Blacks)	The subjects lost follow-up were excluded.	Variables with p-value <0.02 in Cochran -Mantel-Haenszel (CMH) test was included. A forward and backward selection (p-value = 0.05) were used to decide the final variables.	Ordinal logistic regression model.	Beta for each variable were shown in table 7 and figure 2.	The model does poorly in discrimination both times, but can more accurately predict the people who would experience attachment loss when combined two follow-up time points.
	risk model (Whites)	The subjects lost follow-up were excluded.	Variables with p-value <0.02 in Cochran -Mantel-Haenszel (CMH) test was included. A forward and backward selection (p-value = 0.05) were used to decide the final variables.	Ordinal logistic regression model.	Beta for each variable were shown in table 9 and figure 3.	Model performance was poor at neither ruling in nor ruling out the outcome, no better than chance.

Table 4.3 continued from previous page

	Prediction model	The subjects lost follow-up were excluded.	Variables with p-value <0.02 in Cochran -Mantel-Haenszel (CMH) test was included. A forward and backward selection (p-value = 0.05) were used to decide the final variables.	Ordinal logistic regression model.	Beta for each variable were shown in figure 4.	Prediction models with both risk factors and predictors could increase model's sensitivity while accompanied by a decrease in specificity.
Leite 2017	Model 1,2,3,4	The subjects lost follow-up were excluded.	Candidate variables were collected at Birth or at 23 and 24 years of age (OHS-06). Oral health variables (Table 1) and grouped periodontal data at OHS-06 were included for final selection. Decision tree analysis and backward stepwise method (p-value <0.20) were used to decide the final variables.	Multivariable logistic regression models.	Logistic regression coefficients for each variable were shown in Table 5 and supplementary Tables 5-11.	Combining periodontal information, sociodemographic information, and general health history, multivariable logistic regression models can predict development of periodontitis in young adults. Choice of classification might have an impact on accuracy to predict periodontitis occurrence.

Table 4.3 continued from previous page

Lindskog 2010	DRS dentition & DRS tooth	Missing items were not imputed in any way, and subjects with missing data were excluded.	Predictors were included based on published literature.	Not provided	Web-based analytic tool (www.dentosystem.se)	Validation of this risk tool showed that it can select of risk patients (Level I) for periodontitis progression and risk tooth (Level II) for tooth loss. But clinical use of this product remains to be determined.
Martinez-Canut 2018	Molar and Non-molar	Not provided	Variables are the ones that are most consistently found to be associated with TLPD[1] in the literature.	Generalized linear mixed regression model.	Web-based algorithm (www.perioproject.es)	This model can predict TLPD with high accuracy and it can be useful for survival time in different TLPD samples.

Table 4.3 continued from previous page

Morelli 2018	Index of periodontal classes (IPC)	Not provided	Predictors were included based on published literature.	Latent Class Analysis.	7*7 table	First, each participant was assigned to one of the seven PPCs (periodontal profile classes); then, each tooth was classified to one of the seven TPCs (tooth profile classes). The IPR (Index of periodontal risk) was calculated as the mean predicted probability for 10-year tooth loss across all teeth present for each individual. This model acted as not only a periodontal/tooth profile classes, but also a useful system for patient stratification that is predictive for disease progression and tooth loss.

[1]TLPD: Tooth loss due to periodontitis.

Table 4.4: Model Performance and evaluation.

Study		R^2	Predictive performance (Discrimination)						Evaluation
Author/ Year	Model		AUROC [1]	Sen [2]	Spec [3]	PV+ [4]	PV- [5]	Accu. [6]	
Beck 1994	Risk model (Blacks)			0.54	0.34				No
	Risk model (Whites)			0.49	0.82				No
	Prediction model			0.85	0.57				No
Leite 2017	Model 1		0.75	0.62	0.88				No
	Model 2		0.65	0.54	0.75				No
	Model 3		0.64	0.70	0.58				No
	Model 4		0.58	0.41	0.75				No
Lindskog 2010	DRS dentition	0.53		0.86	0.71	0.76	0.83	0.79	External validation
	DRS tooth	0.77		0.66	0.64	0.73	0.55	0.65	External validation
Martinez -Canut 2018	Molars	0.31	0.93	0.39	0.98	0.72	0.94		External validation
	Non-molars	0.24	0.97	0.43	0.99	0.6	0.98		External validation
Morelli 2018	Index of periodontal classes (IPC)		0.72						External Validation (C-statistics = 0.72, 0.75)

[1]AUROC: Area under the receiver operating characteristic curve, [2]Sen: Sensitivity, [3]Spec: Specificity, [4]PV+: Positive prediction value, [5]PV-: Negative prediction value, [6]Accu.: Accuracy.

4.4.4 Quality and risk of bias assessment

The risk of bias of prediction modelling studies is shown in Figure 4.3 and Appendix B Supplement 7. Of all selected prediction studies, participant-related bias occurred in two, no predictor-related and outcome-related bias occurred, however, all studies have a bias in sample flow-related and statistical analysis-related area. Newcastle-Ottawa Scale scoring is presented in supporting information Appendix B Supplement 8. Among five cohort studies, two showed '*Good*' quality, and three showed '*Poor*' quality. No study was excluded because of quality or bias.

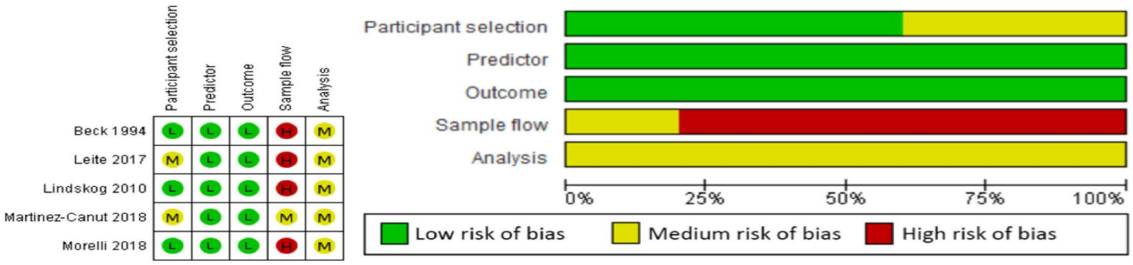


Figure 4.3: Bias assessment of the prediction modelling studies according to CHARMS. L: low risk of bias; M: Medium risk of bias; H: high risk of bias.

Model presentation and interpretation

The presentation of the prediction models varied from narrative and graphical to detailed, quantitative formulas (Table 4.3). Among five studies, one used a 7×7 table (Morelli et al., 2018), two prediction models are available online as web-based tools (Lindskog et al., 2010; Martinez-Canut et al., 2018), while the other two did not report the presentation.

Beck (Beck, 1994) used an ordinal logistic regression model to predict whether the individual will develop CAL > 3mm, with a low sensitivity of 56% and a specificity of 74%. The study distinguished prediction (where the model included risk predictors, such as degree of tooth loss), with risk (whereby the model only contained risk factors). Leite et al. (Leite et al., 2017) established four multivariable logistic regression models to predict the occurrence of periodontitis. Choice of classification influenced the prediction accuracy with sensitivity varying between 41.4% to 69.8%, and specificity variability ranging from

58.2% to 88.5%. Lindskog et al. (Lindskog et al., 2010) validated a web-based algorithm for predicting periodontitis progression, radiographic marginal bone loss, and tooth loss, with accuracy of 79%, sensitivity of 86%, specificity of 71%, PV+ of 76%, and PV- of 83% at the patient-level; and with accuracy of 65%, sensitivity of 66%, specificity of 64%, PV+ of 73%, and PV- of 55% at the tooth-level. Generalized linear mixed regression was used by Martinez-Canut et al. (Martinez-Canut et al., 2018) to predict the dichotomous event: tooth loss due to periodontal disease (TLPD+ or TLPD-). The predictive models for molars and non-molars achieved an AUROC of 93% and 97%, a sensitivity of 39% and 43%, a specificity of 98% and 99%, a positive predictive value of 72% and 60%, and a negative predictive value of 94% and 98% respectively. In the model by Morelli et al. (Morelli et al., 2018), the latent class analysis was used to calculate the Index of Periodontal Risk (IPR) score and classify individuals or teeth into different classifications. This IPR score can predict the 10-year tooth loss with C-statistics of 0.72. No study conducted internal validation. External validation was conducted in Martinez-Canut's and Morelli's model, but only Morelli's model reported the model performance with C-statistics of 0.72 and 0.75 for 3-year attachment loss and 5-year tooth loss, respectively.

4.5 Discussion

4.5.1 Summary of main findings and quality of the evidence

In this systematic review, we distinguished the risk prediction, estimating the probability of an event, from risk assessment, stratifying the risk levels, and we provided an overview of the currently available prediction models of periodontitis incidence and progression based on clustering of risk factors and predictors. Fifteen models from seven studies were identified, of which three were risk assessment tools, and 12 were prediction models. Despite concerted efforts to develop and improve periodontitis prediction modelling, the overall results have not been ideal. Risk assessment tools were designed either qualitatively, classifying patients into low, medium, and high-risk level, or quantitatively, converting the disease status and risk status into risk scores. More review articles about the status of periodontal risk assessment are available in the literature (Heitz-Mayfield, 2005; Kye et al., 2012; Lang et al., 2015). The 12 prediction models varied in derivation cohort, type of predictors, outcome, statistical approaches, model performance, and model presentations. Each model had its merits and limitations. Some are able to calculate the exact probability of periodontitis development or tooth loss while others are more suited to predicting periodontal health. Nonetheless, it is difficult to determine which models stood out when it comes to predictive performance and clinical usefulness. Most studies were of 'good' quality concerning participant selection, predictors, and outcomes. However, prediction models were deemed to be of 'high' risk of bias due to poor handling with missing data, statistical methods for model development,

and the lack of validation.

Variables covered the common risk factors and indicators for periodontitis (Albandar, 2002), the most frequent were age, smoking, diabetes, and sex. Some other clinical parameters such as number of teeth lost and plaque-related factors also were included in risk prediction studies. All models selected candidate variables based on the published literature, but two studies selecting final predictors based on statistical significance, may lead to a model fitting the data too closely (Bedogni, 2009; Collins et al., 2015). Eight identified criteria for periodontitis incidence and progression showed the absence of consensus, resulting in variation across data collection of longitudinal studies (Leite et al., 2017). Two previous prominent definitions from AAP/CDC and EFP were not given priorities in existing prediction models, and the change of PPD, CAL, bone loss, or tooth loss have been used to investigate the disease progression. The prediction models commonly reported performance on discrimination but cannot be compared quantitatively, because they reported performance in the original derivation cohort only. However, the reported performance could give an indication about the maximum potential predictive performance in other populations, because accuracy usually decreases in future external populations (Collins et al., 2015; Steyerberg et al., 2013).

Overall, risk assessment/scoring systems cannot replace the need of prediction models. Generalizability of the prediction models in other settings with different population and socioeconomic compositions remains unclear. Periodontitis incidence and progression has been differently defined for epidemiological purposes (Holtfreter et al., 2015), complicating prediction. Assessment of methodological quality revealed that improvements are clearly needed, both in conducting studies and eventual reporting, in-line with conclusions made previously (Bouwmeester et al., 2012). The recently published transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement, may lead to improvements in conducting and reporting of future studies (Collins et al., 2015). We showed that nearly half of the items considered essential in the TRIPOD statement were either not or inadequately reported (Appendix B Supplements 9 and 10).

4.5.2 Strengths and potential limitations

The strengths of this study include the very first perspective of prediction modelling studies for periodontitis, comprehensive literature search, adherence to rigorous systematic review methodology, and adherence to recommended guidelines for reporting. However, this study is not free of limitations. The broad inclusion criteria could allow the inclusion of studies not meant to concern accurate probability, but only qualitatively evaluating the risk levels. Outcomes included tooth loss though it was not a part of this systematic review since periodontitis progression and tooth loss are somewhat linked. Finally, high study

heterogeneity precluded overall comparisons of prediction models despite most containing the same performance measures.

4.5.3 Implications for future research

In order to develop high-quality prediction models for periodontitis, a standard methodology for model generation should be adopted (Moons et al., 2012). Several strategies may be adopted to improve the performance and usability of the available and further models. First, a comprehensive external validation of existing models in an independent eligible population will allow a model-to-model comparison and identify the strength and weakness of each model. Second, the predictive performance of a model might improve by accounting for new and better predictors such as gingival crevicular fluid and salivary markers (e.g., interleukin-1, and matrix metalloproteinase-8) (Giannobile et al., 2009; Gursoy et al., 2011) or microbiology information. Models should predict specific events accurately and be relatively easy to use. If a prediction model provides inaccurate estimates of future-event development, it may mislead healthcare professionals and provide inappropriate management of patients. In contrast, if a model has good predictive ability but is hard to apply (e.g., with complicated examinations or questions), time-consuming, or costly, it will not be commonly used (Pencina et al., 2008). Therefore, an ideal prediction model should achieve a balance between predictability and simplicity. Third, different modelling methods can be employed such as machine-learning approaches, tree-based algorithms, and neural networks (Goldstein et al., 2017; Peissig et al., 2014). In the field of healthcare, machine learning algorithms have outperformed conventional regression models in predictive ability in identifying patients at high risk for developing disease (Singal et al., 2013). Lastly, using consistent outcome definitions will enhance not only the reporting, but also the generalizability and comparability of models. The novel classification of periodontal manifestations and conditions of AAP and EFP is strongly recommended for further studies (Tonetti et al., 2018).

4.6 Conclusion

This is the first systematic review of prediction modelling studies for periodontitis incidence and progression. It revealed several methodological and reporting shortcomings of published prediction models and indicated further research is required. Logistic regression, generalized linear mixed regression, and latent class analysis were used as predictive modelling approaches. Existing models covered the most likely predictors for periodontitis development in adults and achieved acceptable predictive accuracy, but the absence of consensus regarding periodontitis measurement and classification complicates modelling. In the future predictive modelling studies, it is essential to conduct cross-validation and/or out-of-sample estimation and/or validation with a second data set, both for variable selection, and for measures of prediction accuracy, in order to avoid overfitting. More efforts

can be made in taking into account new predictors, such as microbiological information, biological markers, and machine learning approaches, such as decision tree, or neural network.

4.7 Acknowledgments

The authors acknowledge Vikki Langton, a research librarian from the University of Adelaide for her assistance and guidance for the development of the database search strategies.

CHAPTER 5

Examining bias and reporting in oral health prediction modelling studies

Preface

In Chapter 4, we identified that the quality of most prediction models in the discipline of periodontology is poor. Based on these findings, Chapter 5 takes a further step to obtain more information on the methodological quality and reporting transparency of prediction modelling studies in oral health, regardless of their investigated outcomes. In this chapter, we review the recent prediction modelling studies published in major dental, epidemiological and biostatistical journals. We find that oral health prediction modelling studies suffer from various potential biases, such as selection bias, measurement errors. These biases lead to a lack of reproducibility and non-transparent reporting of the existing models. To solve these issues, suggestions and achievable steps are provided in this chapter to improve the reliability of future oral health prediction modelling research. This is the first study of this nature in oral health field.

This chapter contains the second of a series of four studies contributing to this thesis. Details of the publication are:

Du M, Haag D, Song Y, Lynch J, Mittinty M. Examining bias and reporting in oral health prediction modelling studies. Journal of Dental Research. 2020;99(4):374-387. https://doi.org/10.1177/0022034520903725.

The accepted version of the published study is reproduced as follows.

Statement of Authorship

Title of Paper	Examining bias and reporting in oral health prediction modelling studies
Publication Status	√ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Du M, Haag D, Song Y, Lynch J, Mittinty M. Examining bias and reporting in oral health prediction modelling studies. Journal of Dental Research. 2020;99(4):374-387.

Principal Author

Name of Principal Author (Candidate)	Mi Du		
Contribution to the Paper	MD contributed to conception, design, studies selections, data extraction, critical appraisal, results interpretation, drafted and critically revised the manuscript.		
Overall percentage (%)	90%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	24-04-2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dandara Haag		
Contribution to the Paper	DH contributed to conception, design, critical appraisal, results interpretation, and critically revised the manuscript.		
Signature		Date	3/05/21

Name of Co-Author	Youngha Song		
Contribution to the Paper	YS contributed to conception, studies selections, and critically revised the manuscript.		
Signature		Date	24-04-2021

Name of Co-Author	John Lynch		
Contribution to the Paper	JL contributed to conception, design, critical appraisal, results interpretation, and critically revised the manuscript.		
Signature		Date	3/05/21

Name of Co-Author	Murthy Mittinty		
Contribution to the Paper	MM contributed to conception, design, critical appraisal, results interpretation, and critically revised the manuscript.		
Signature		Date	3/05/21

Abstract

Background and aims: Recent efforts to improve the reliability and efficiency of scientific research has caught the attention of researchers conducting prediction modelling studies. Use of prediction models in oral health has become more common over the past decades for predicting the risk of diseases and treatment outcomes. Risk of bias (ROB) and insufficient reporting present challenges to the reproducibility and implementation of these models. A recent tool for bias assessment—PROBAST (Prediction model Risk Of Bias Assessment Tool) and a reporting guideline—TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) have been proposed to guide researchers in the development and reporting of prediction modelling studies, but their application has been limited.

Methods: Following the standards proposed in these tools and a systematic review approach, a literature search was carried out on PubMed to identify oral health prediction modelling studies published in dental, epidemiological and biostatistical journals. ROB and transparency of reporting was assessed using PROBAST and TRIPOD.

Results: Among 2,881 papers identified, 34 studies containing 58 models were included. The most investigated outcomes were periodontal diseases (42%) and oral cancers (30%). 75% of the studies were susceptible to at least four (out of 20) sources of bias,including measurement error in predictors (n = 12) and/or outcome (n = 7), omitting samples with missing data (n = 10), selecting variables based on univariate analyses (n = 9), overfitting (n = 13), and lack of model performance assessment (n = 24). Based on TRIPOD, at least five (out of 31) items were inadequately reported in 95% of the studies. These items included sampling approaches (n = 15), participant eligibility criteria (n = 6), and model-building procedures (n = 16).

Conclusion: There was a general lack of transparent reporting and identification of bias across the studies. Application of the recommendations proposed in PROBAST and TRIPOD can benefit future research and improve the reproducibility and applicability of prediction models in oral health.

5.1 Introduction

In general, over 85% of the health research is unreliable due to the lack of reproducibility (Chalmers et al., 2014). In order to improve the transparency and reliability of health research, collaborative efforts (e.g., EQUATOR network: http://www.equator-network.org/) have been made to produce measures for quality assessment of various types of studies. These measures include guidelines for reporting preclinical studies (ARRIVE) (Kilkenny et al., 2012), clinical trials (CONSORT) (Schulz et al., 2011), observational studies (STROBE) (von Elm et al., 2014), and systematic reviews and meta-analyses (PRISMA) (Moher et al., 2009). Additionally, bias assessment tools have also been developed, such as Cochrane risk-of-bias (for randomised studies), ROBINS-I (Risk of Bias in Nonrandomized Studies of Interventions), and QUIPS (Quality In Prognosis Studies). In health research, prediction modelling studies are used to identify population at high risk of a particular health condition, and determine the benefit from a care management plan

The primary goal of prediction models is to improve the accuracy with which a particular outcome can be predicted given a set of observed predictors. However, a difference between the predicted and the observed outcome always exists due to random or systematic error or both (Rothman, 2008). Systematic error, which also refers to bias, is the difference between the expected and the true value of the parameters of interest in the population. Random error occurs due to variations in a given population and is largely influenced by sample size (Rothman, 2008).

One key aspect of the quality of prediction modelling studies is related to the potential of bias in data collection and modelling processes, which involves misclassification, selection of participants, missing information and unmeasured covariates (Lash et al., 2014). Therefore, inaccurate prediction due to systematic errors can be avoided or at least reduced if appropriate methodology concerning data collection and analytical procedures are used.

Additional issues concerning the quality of prediction modelling studies include transparency and completeness of reporting of a prediction modelling study. Insufficient reporting of data sources and cleaning, as well as model development processes hinders study replication and inhibits the assessment of the reliability of such studies.

Systematic reviews have shown that prediction models have been developed to assess the risk of multiple oral conditions among various populations. These conditions include dental caries (Senneby et al., 2015), periodontal diseases (Lang et al., 2015), and oral cancers (Sharma and Om, 2013). Most of these systematic reviews are interested in estimating the effect size or to understand the importance of certain predictors, rather than assessing the overall quality of the included prediction modelling studies. Therefore, the present study aims to assess the ROB (using PROBATS) and reporting transparency (using TRIPOD) of recent oral health prediction modelling studies.

5.2 Methods

5.2.1 Protocol and registration

This study was prepared in accordance with the PRISMA guidelines (Moher et al., 2009). The review protocol was registered on PROSPERO (No. CRD42019122274). This is not a typical systematic review, since we limited our literature search to certain journals and publication times, however strictly adhere to the guidelines of conducting and reporting a systematic review.

5.2.2 Inclusion and exclusion criteria

Studies were included if they met the following criteria: 1) diagnostic and prognostic multivariable prediction modelling studies; 2) for prognostic purposes, study designs included observational cohorts, case-control, and RCTs. For diagnostic purposes, cross-sectional designs were also included; 3) outcomes were oral health-related and clearly defined, including but not limited to diagnosis (incidence, occurrence, prevalence) and prognosis (outcome of treatment, progression, survival).

Studies were excluded if they: 1) were not original research (e.g., review articles, letters, editorials); 2) were not human research (e.g., cell- and molecular-level); 3) were not multivariable modelling studies; 4) did not include oral health outcomes (e.g., study investigating periodontitis as a predictor for type 2 diabetes).

5.2.3 Literature search and study selection

To reflect the 'good-quality' research and best available evidence in oral health prediction modelling studies, 14 dental journals, eight epidemiology journals and seven biostatistical journals were fully searched on PubMed from 2016/01/01 to 2018/12/31, following the development of TRIPOD in 2015 (ranking based on Thomson Reuters Journal Citation Reports, Google scholar, and Scopus). Figure 5.1 shows the flowchart of study selection. Two authors (MD and YS) screened all the titles and abstracts based on inclusion and exclusion criteria in parallel, then full-text reading of all selected articles was performed for eligibility assessment. In cases of disagreement, a third referee (DH) was involved. Detailed search strategy and journals selection criteria are provided in Appendix C Supplements 1, 2, and 3.

5.2.4 Data extraction

Key information extracted from each eligible study include: setting, study design, population characteristics, follow-up period, sample size, outcomes, predictors, missing data, variable selection, modelling approach, model presentation, interpretation, performance (discrimination and calibration) and evaluation.

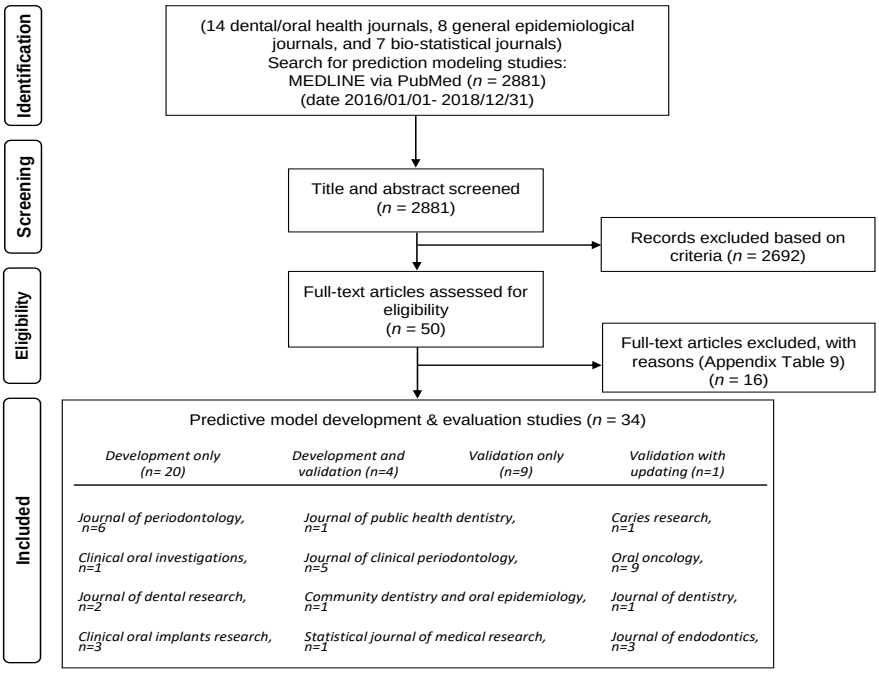


Figure 5.1: PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow diagram of the studies search and selection.

5.2.5 Application of PROBAST and TRIPOD

A pilot study on applying each PROBAST item was conducted and agreement was achieved among MD, DH, JL, and MM. Two authors (MD and DH) applied PROBAST across all selected studies and a third referee (MM or JL) was involved in cases of doubt. Following PROBAST recommendations (Moons et al., 2019), we classified studies into high, low and unclear ROB, and recorded the answer of 'Yes (Y)', 'Probably Yes (PY)', 'No (N)', 'Probably No (PN)', or 'No Information (NI)' for 20 signalling questions (Appendix C Supplement 4) along with supportive background (Appendix C Supplement 5). TRIPOD was applied by one author (MD) with experience on this tool (Du et al., 2018) and doubts were resolved with the involvement of a second reviewer (DH or JL or MM). A score for each study was then calculated in order to quantify the completeness and transparency of reporting (maximum 31 items with a score of 1 for item met and a score of 0 if the item was not met). Detailed PROBAST and TRIPOD criteria are provided in Appendix C Supplements 6, 7, and 8.

5.3 Results

5.3.1 Literature search, study selection and general characteristics of the included prediction modelling studies

Out of the 2,881 searched articles, 50 were included following title and abstract screening. After reading the full-text, 34 studies were included in the review (exclusion reasons are provided in Appendix C Supplement 9). There were 24 studies emphasising model development for risk assessment, and ten studies focusing on model validation with (n = 1) or without (n = 9) updating. Study characteristics are summarized and presented in Tables 5.1 and 5.2. Details of the investigated outcomes, predictors used, model deviation methods and model presentations are provided in Appendix C Supplements 10 and 11.

Table 5.1: Characteristics of 34 prediction modelling studies.

Outcomes	Study (Author/ Year)	Setting/ Context	Study design	Participants characteristics (% of females, years)	Follow-up (years)	Sample Size	Predictors (*n*)
Model development studies							
Periodontal disease	Chatzopoulos et al. 2016	Greece ç	Cro.sec	53.9%, 50.5 ± 15.1	NA	535	6
	Eke et al. 2016	USA §	Cro.sec	45.8%, 30-79	NA	7,066	5
	Kuboniwa et al. 2016	Japan ¶	Cro.sec	78.9%, 39.2 ± 11.6	NA	19	8
	Lee et al. 2018	South Korea §	Cro.sec	61.8 % (non-cases); 52.6% (cases)	NA	34,950	8
	Leite et al. 2017	Brazil §	Pro.	52.3%; 24	7	471	5-9
Tooth loss	Martinez-Canut et al. 2018	Spain ç	Retro.	68.8%; 40.3 ± 9.07	20.2 ± 2.04	500	11
	Meisel et al. 2018	Germany §	Pro.	51.8%, 20-81	5 and 10	2,776 (5-year) 2,016 (10-year)	9
Tooth loss and periodontitis progression	Morelli et al. 2018	USA §	Pro.	54.5%; mean 62.4	10	3,985 development; 697 validation	6

Table 5.1 continued from previous page

Treatment of periodontal disease	Gul et al. 2017	UK ¶	Pro.	50.6%, 49.7 ± 8.9	0.5	77	6
Implant-related (Implantitis)	Canullo et al. 2016	Italy ¶	Retro.	57%, 63.83 ± 9.68 (non-cases); 68%, 61.24 ± 9.91 (cases)	NP	56 (125 implantitis, 207 healthy)	9
Implant-related (Implantitis)	Canullo et al. 2017	Spain ¶	Cro.sec	53.2%, mean 48.4 (non-cases); 58.1%, mean 54.1 (cases)	NA	47	23
Implant-related (Implant bone level)	Papantonopoulos et al. 2017	Greece ç	Cro.sec	51.4%, 61.9 ± 11.1	Over 2	72	6
Mucositis (among NPC)	Orlandi et al. 2018	Italy ¶	Retro.	30.3%, median 49 [18-81]	Median 50[43–61] days	132	3
Oral cancer	Rao et al. 2016	India ¶	Case-control	45.2%, 16-55(57.4%), > 55(43.6%) (non-cases); 20.6%, 18-55(52.8%), > 55(47.5%) (cases)	NP	452	9
Root caries	Ritter et al. 2016	USA ç	RCT(control arm)	60%, 52.42 ± 12.51	3	155	6-11
Caries transition	Pak et al. 2019	USA §	Pro.	52.3%, 0-5	NP	1,020	15

Table 5.1 continued from previous page

Intraoperative pain (among endodontic patients)	Kayaoglu et al. 2016	Turkey ç	Cro.sec	56.2%, 39 ± 15.2	NA	1,435	6
Periapical cyst	Pitcher et al. 2017	USA ç	Cro.sec	56.6%, NP	NA	113	5
OC survival (among HPV-positive TSCC and BOTSCC patients) Þ	Bersani et al. 2017	Sweden ¶	Pro.	25%, median 60[30-90] (training); 22%, median 60[30-84] (validation)	3	258	11
OC survival (among surgically treated T4 buccal mucosa cancer patients) õ	Bobdey et al. 2018	India ¶	Retro.	17.6%, median 50[23-76] (training); 24.2%, median 51[24-79] (validation)	Median 21 [2-68] months	205	4
OC survival (among NPC) Þ õ	Xu et al. 2017	China ¶	Retro.	24.5%, 18-38(27.3%), 39-45(23.4%), 46-53(25%), > 54(24.3%)	Median 56.6 [51.5-63.1] months	1,230	6
OC survival Þ	Zhang et al. 2017	South Korea ¶	Retro.	37.5%, median 54 [13-89]	Median 11.3 [4.6-23.2]	160	4
OC recurrence (among laryngeal glottis cancer)	Jover-Espla et al. 2018	Spain ¶	Pro.	3.7%, 62.9 ± 10.9	3.4 ± 3.0	189	4

Table 5.1 continued from previous page

OC survival (among NPC) Þ	Peng et al. 2018	China ¶	Retro.	26.4%, < 44(50%)	Median 43.8 months	7,413	3
Model validation studies with or without updating							
Periodontal disease	Carra et al. 2018	France ¶	Cro.sec	40.1%, 46.1 ± 12.6	NA	232	12 questionnaires
	Heaton et al. 2017	USA §	Cro.sec	100%, mean 59	NA	77	8 questionnaires
	Su et al. 2017	Taiwan §	Cro.sec	58.2%, mean 45	NA	4,061	Calibrated CPI score
Tooth loss	Martinez-Canut and Llobell 2018	Spain ç	Retro.	66%, 42.3 ± 6.95	24.7 ± 2.4	100	11
	Schwendi cke et al. 2018	Germany ç	Retro.	57.1%, 47.6 ±1 0.1	Over 9	301	Mar-19
Odontogenic or temporomandibular disorder pain	Alonso et al. 2017	UK ç	Cro.sec	52.9%, 49 ± 12 (Odontogenic pain);86.5%, 45 ± 18 (TMD pain)	NA	71	6 and 14 questionnaires
Root caries	Hayes et al. 2017	Ireland ¶	Pro.	NP, > 65	1 and 2	334	14
Oromandibular dystonia	Yoshida 2019	Japan ¶	Retro.	53.9%, 53.4 ± 17.4	NP	553	10 questionnaires

Table 5.1 continued from previous page

OC survival (among NPC) Þ	OuYang et al. 2017	China ¶	Pro.	27.9%, 18-29(7.3%), 30-39(24.5%), 40-49(35.9%), 50-59(22.4%), > 60(10%)	Median 59 [3-109] months	920	7, 9
OC survival õ	Prince et al. 2016	USA ¶	Pro.	43%, < 40(6%), 40-49(13%), 50-59(27%), 60-69(23%), 70-79(18%), > 80(12%)	Median 53 months	492	7-8

ç: Clinic setting; ¶: Hospital-based; §: General population; Cro.sec: Cross-sectional study; Pro.: Prospective cohort study; Retro.: Retrospective cohort study. NA: Not applicable; NP: Not provided. OC: Oral cancer; Þ: Progression-free survival, õ: Overall survival; NPC: Nasopharyngeal carcinoma; HPV: Human papilloma virus; TSCC: Tonsillar-dominated oropharyngeal squamous cell carcinoma; BOTSCC: TSCC base of tongue.

Table 5.2: Model development, presentation, performance and interpretation

Study (Author/Year)	Missing value	Variable selection	Model development	Model presentation	Model interpretation	Model performance measurement		Model validation
						Discrimination	**Calibration**	
Model development studies								
Chatzopoulos et al, 2016	Complete-cases	Literature search, automated selection based on Sen/Spec	Logistic regression	Four item questionnaire	A two-domain self-report measure combining two self-report items with age and sex has good sensitivity and specificity for periodontitis screening in a white, university-based population.	Sen, Spec, C-statistics, PV+, PV-	NP	Internal 1,000 bootstrap

Table 5.2 continued from previous page

Eke et al, 2016	Unavailable predictors were assigned by estimates in other dataset	NP	Logistic regression	NP	Periodontitis prevalence models generated from NHANES can be used at state and local levels.	NP	NP	Internal validity tests were performed by comparing summary prevalence estimates. External validated in other 2 data sets.
Kuboniwa et al, 2016	NP	Variable importance in the projection score >1	Logistic regression	Logistic regression algorithm (Table 1)	Potential salivary metabolites that may be useful for reflecting the severity of periodontal inflammation for monitoring disease activity in periodontitis patients.	AUROC	NP	NP
Lee et al, 2018	NP	NP	Decision tree, NN, Regression models	Decision tree structure (Figure 3)	Decision tree model is an effective data mining technique for identifying the complex risk factors for PD.	AUROC, RMSE, Misclassification rate	NP	Internal split-sample validation

Table 5.2 continued from previous page

Leite et al, 2017	Complete-cases	Decision tree analysis and backward stepwise method (p-value < 0.20)	Logistic regression	Logistic regression algorithms (Table 5, supplementary Tables 5-11)	Combining periodontal information, sociodemographic information, and general health history, multivariable logistic regression models can predict development of periodontitis in young adults. Choice of definition might have an impact on accuracy to predict periodontitis occurrence.	AUROC, Sen, Spec	NP	NP
Martinez-Canut, 2018 A	NP	Literature search	Generalized linear mixed regression.	Web-based (www. perioproject. es)	This model can predict TLPD with high accuracy and it can be useful for survival time in different TLPD samples.	Sen, Spec, PV+,PV-, Accu.	NP	External validation
Meisel et al, 2018	NP	NP	Logistic regression	Logistic regression algorithm (Tables 2 and 3)	Self-reported oral health provides reliable predictions of tooth loss comparable to those assessed by clinical diagnostics.	Sen, Spec	NP	NP

Table 5.2 continued from previous page

Canullo et al, 2016	NP	Univariate logistic regression (p-value $<$ 0.05)	Decision tree	Decision tree structure (Figure 2)	This decision tree technique seems to be a promising tool for diagnostics of peri-implantitis subtypes.	Accu.	NP	10-fold cross validation
Canullo et al, 2017	NP	NP	Decision tree, Logistic regression, SVM, Artificial NN	Decision tree structure (Figure 2)	Predictive models developed in the present study discriminated some qualitative and quantitative characteristics of microbiological profile associated with peri-implantitis and were estimated as highly accurate for microbiologically based implant diagnostics.	Accu.	NP	10-fold cross-validation
Papantonopoulos et al, 2017	NP	Principal component analysis	SVM	Figure 5	Prediction of individual implant mean bone levels could be achieved by using ensemble selection and SVM with six variables.	RMSE, Sen, Spec	NP	10-fold cross-validation

Table 5.2 continued from previous page

Orlandi et al, 2018	NP	LASSO and univariate logistic analyses	Logistic regression	Logistic regression algorithm (Table 3)	A predictive model combining clinical and dosimetric items to identify NPC patients at risk for developing severe oral and oropharyngeal mucositis during radio- and chemo-therapy was built.	AUROC	Calibration plot and H-L test	Internal 1,000 bootstrap resampling
Rao et al, 2016	Complete-cases	Literature search	Regression models	Regression algorithm (Table 2)	A risk score model to screen for individuals with high risk of oral cancer with satisfactory predictive ability was developed in the Indian population.	AUROC, Sen, Spec, PV+, PV-	H-L test	Internal 200 bootstrap resampling

Table 5.2 continued from previous page

Ritter et al, 2016	Conditional mean imputation	NP	Logistic regression	Logistic regression algorithms (Table 3)	Five prediction models were built and their performance were compared. Model M2 (which included the number of years and number of root surfaces at risk, RCI, gender, race, age, and tobacco use) demonstrated the best prediction performance in predicting incidence of root caries.	AUROC, Brier score, Sen, Spec, PV+, PV-	NP	Internal 500 bootstrap resampling
Kayaoglu et al. 2016	NP	Stepwise forward-entry process (p-value <0.05)	Logistic regression	Web-based (http:// websitem. gazi. edu.tr/site/ guvenk/ files).	A model including the variables age, dental arc, tooth type, pulpal diagnosis, presence of pain within the previous 24 hours, and type of anesthetic solution was highly predictive for forecasting incidence of intraoperative pain.	Correct classification rate, Sen, Spec, PV+, PV-	H-L test	15% internal split sample validation

Table 5.2 continued from previous page

Pitcher et al, 2016	NP	Univariate Logistic regression	Decision tree	Decision tree structure (Figure 4)	Binary decision tree classifier determined that if the CBCT volume of the lesion was > $247mm^3$, there was 80% probability of a cyst. If volume was < $247mm^3$ and root displacement was present, cyst probability was 60%.	Spec, Accu.	NP	NP
Yoshida, 2018	NP	Experts experience. Cronbach's $\alpha(>0.8)$	NA	10 item questionnaires (Table 1)	The present questionnaire is a simple diagnostic tool that is useful for tentative differentiation of oromandibular dystonia from temporomandibular disorders.	Pearson's correlation was used to analyze test-retest reliability	NP	NP
Gul et al, 2017	Complete-cases	Literature search, Backward stepwise technique	Logistic regression	Logistic regression algorithm (Table 4)	The levels of three GCF enzymes plus two bacterial species at a site comprises a unique biomarker profile or fingerprint that is useful for predicting the outcome of periodontal treatment.	AUROC, Sen, Spec	NP	External validation

Table 5.2 continued from previous page

Morelli et al, 2018	NP	Literature search	Latent Class Analysis	7×7 scoring table	First, each participant was assigned to one of the seven periodontal profile classes; then, each tooth was classified to one of the seven tooth profile classes. Index of periodontal class was calculated as the mean predicted probability for 10-year tooth loss across all teeth present for each individual. This model acted as not only a periodontal/tooth profile classes, but also a useful system for patient stratification that is predictive for disease progression and tooth loss.	AUROC	NP	External Validation
Bersani et al, 2017	Missing data was imputed by patient median	LASSO regression	LASSO regression	LASSO regression algorithm (Figure 3)	A lasso model with four variables (CD8+ TIL counts, age, T-stage and E2 expression) could predict progression-free survival.	AUROC	NP	10-fold cross-validation

Table 5.2 continued from previous page

Bobdey et al, 2018	Complete-cases	Literature search, Stepdown reduction (p-value <0.05)	Cox proportional hazards regression	A nomogram (Figure 1)	The ability of nomograms taking into account more variables than the conventional TNM staging system would more accurately predict 3-year overall survival of buccal mucosa cancer than the currently used TNM system.	AUROC, Youden Index, Accu. Sen, Spec, PV+, PV-, PLR, NLR	NP	Internal 1000 bootstrap resampling and external validation
Xu et al, 2017	NP	Backward stepdown selection using AIC	Recursive partitioning analysis	Nomograms (Figure 2)	Prognostic nomograms based on the 8^{th} edition of the UICC/ AJCC staging system, plasma EBV DNA and other predictors have good prognostic accuracy in patients with NPC and the subgroup of patients with NPC.	C-index	Calibration plot	Internal 1000 bootstrap resampling and subgroup validation
Zhang et al, 2017	Complete-cases	Univariate Cox regression	Cox proportional hazards regression	A nomogram (Figure 3)	The proposed nomogram combined biomarkers could be useful for the accurate and individual prediction of the probability of 5-, 10- and 15-year progression free survival.	Accu. C-index, Sen, Spec, PV+, PV-	NP	NP

Table 5.2 continued from previous page

Jover-Espla et al, 2018	No missing data	Based on models C statistics	Cox proportional hazards regression	An application on Google Play, named Glottic cancer recurrence	A points system and a mobile application to obtain the probability of recurrence of laryngeal glottic cancer within five years from diagnosis was constructed.	C-statistic	Calibration plot	Internal 1000 bootstrap resampling
Peng et al, 2018	NP	Backward elimination (criterion = 0.1)	Cox proportional hazards regression	Scoring system (page 260) and a flow chart (Figure 5)	A novel prediction model to discriminate NPC patients who would differently benefit from the additional induction chemotherapy to concurrent chemoradiotherapy was built. And it consequently help with decision-making in clinical practice.	NP	NP	NP
Model validation studies with or without updating								
Carra et al, 2018	NP	NA	Logistic regression	Questionnaires	The PESS based on five self-report Items and age and smoking demonstrated a good accuracy and moderate sensitivity and specificity.	C-statistics, AUROC, Sen, Spec	NP	Questionnaire updating study

Table 5.2 continued from previous page

Heaton et al, 2017	Complete-cases	NA	NA	NA	Combinations of questionnaire items improved the predictive ability with respect to severe disease beyond that of individual questionnaire items.	AUROC, Sen, Spec, PV+, PV-	NP	External validation study
Su et al, 2017	Complete-cases	NA	Bayesian hierarchal logistic regression	NA	An improvement in the updated prediction model was demonstrated for periodontal disease as measured by the calibrated CPI derived from a large epidemiologic survey.	AUROC, Sen, Spec	NP	Model updating study
Martinez-Canut et al, 2018 B	NP	NA	NA	NA	The TLPD rate increased as the risk of TLPD increased while the percentage of TLPD increased as the survival time decreased.	NP	NP	External validation study

Table 5.2 continued from previous page

Schwendicke et al, 2018	Unavailable predictors excluded	NA	NA	NA	Most of the investigated multivariable tooth loss prediction models was limited when applied (usually after some modifications) to the specific cohort of patients.	AUROC	NP	External validation study
Alonso et al, 2017	Complete-cases	NA	NA	NA	DePaQ and TMD screener were “acceptable” in identifying patients who had the pain condition in question (ie, sensitivity), whereas the point estimate for appropriately identifying patients who did not have the pain condition when they did not have it (ie, specificity) was “nonacceptable” for both.	Sen, Spec	NP	External validation study
Hayes et al 2017	Complete-cases	NA	NA	NA	Cariogram may be clinically useful in determining future root caries risk in independently living older adults.	AUROC, Sen, Spec, PV+, PV-	NP	External validation study

Table 5.2 continued from previous page

OuYang et al, 2017	NP	NA	NA	NA	Tang's nomogram performed better in risk stratification.	AUROC	Calibration plot	External validation study
Prince et al, 2016	Complete-cases	NA	NA	NA	Five prognostic calculators designed to predict individual outcomes of oral cancer differed significantly in their assessments of risk. Most were well calibrated and had modest discriminatory ability.	AUROC, C-index	Calibration plot	External validation study

AUROC: Area under the receiver operating characteristic curve; Sen: Sensitivity; Spec: Specificity; PV+: Positive prediction value; PV-: Negative prediction value, Accu.: Accuracy; RMSE: Root-mean-squared error; PLR: Positive likelihood ratio; NLR: Negative likelihood ratio; LASSO: Least absolute shrinkage and selection operator; Hosmer-Lemeshow (H-L) test; NPC: Nasopharyngeal carcinoma; AIC: Akaike information criterion; BIC: Bayesian information criterion; NA: Not applicable; NP: Not provided.

5.3.2 Identification of main sources of bias in the included prediction modelling studies

Figure 5.2A shows the proportion of Y, PY, N, PN and NI for each PROBAST item. Figure 5.2B presents the proportion of studies with potential biases according to four domains of PROBAST (Participants, Predictors, Outcomes, Analysis), and Figure 5.2C presents the 'biased' domain(s) identified in each study. Overall, 76% of the studies were susceptible to at least four (out of 20) sources of bias (e.g., measurement error, missing data), 40% and 30% of the studies were identified with at least five and six sources of bias, respectively.

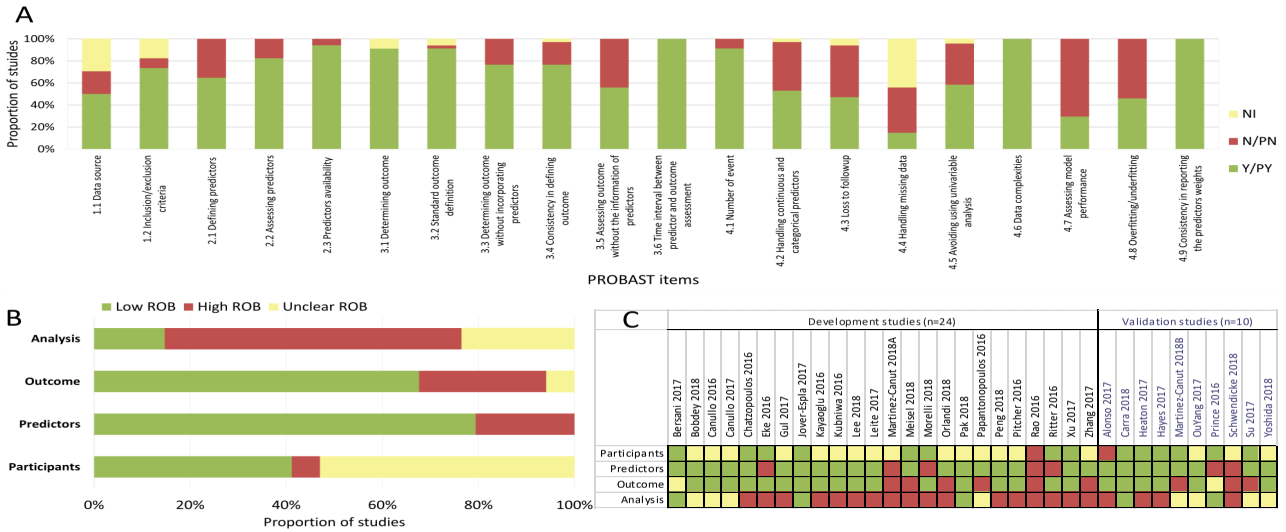


Figure 5.2: Bias assessment of 34 studies (24 model development studies and 10 model validation studies) based on PROBAST. (A) Proportion of studies being answered Y/PY (Yes/Probably Yes), N/PN (No/Probably No), and NP (Not Provided) for each PROBAST item. † Development studies only. (B) Overall rate of risk of bias (ROB) based on PROBAST. Low ROB (green); High ROB (red); Unclear ROB (yellow). Bias related to analysis refer to categorising continuous variables (n = 16, 47%), selecting variables based on univariate analyses (n = 9, 26%), inappropriate handling of missing data (n = 29, 85%), overfitting (n = 13/24, 54%), and lack of model performance measure (n = 24, 71%). Bias related to outcome refer to unclear definition (n = 3, 9%), measurement error (n = 7 , 21%), and incorporation bias (n = 8, 24%). Bias related to predictor refer to measurement error (n = 12, 35%), lack of blinding to the outcome when assessing predictors (n = 6, 18%), and unmeasured predictors (n = 2, 6%). Bias related to participants refer to unclear sampling (n = 14, 41%), lack of inclusion and exclusion criteria (n = 6, 18%), and convenience sample (n = 1, 3%). (C) ROB in each study.

Bias related to participants

Selection bias may be present when the association between the predictors and the outcome is different between those who were part of the study versus those who should have been theoretically eligible but were not part of the study (Rothman, 2008).

Overall, six of the studies showed unclear ROB due to lack of specification of inclusion/exclusion criteria. Among the 22 studies which utilised existing data sets (e.g., electronical health records, disease registries), 14 did not specify the sampling and/or data collection methods nor referenced previous studies that did so, making it impossible for us to judge the potential for selection bias in these studies. According to PROBAST, one study (Alonso et al., 2017) showed high ROB due to convenience sampling, as convenience samples may not be generalized to the target population. Another study (Rao et al., 2016) showed high ROB because the baseline risk/outcome frequency was not adjusted when using a non-nested case-control design.

Bias related to predictors

The identified potential bias related to predictors include (i) measurement error (n = 12), (ii) lack of blinding to the outcome when assessing predictors (n = 6), and (iii) unmeasured predictors (n = 2). Overall, 27 studies showed low ROB and 7 studies showed high ROB in the predictor section.

Twelve studies used data from multiple medical/dental centres (Martinez-Canut et al., 2018; Rao et al., 2016; Ritter et al., 2016) , different periods (n = 7) and different studies (Eke et al., 2016; Martinez-Canut and Llobell, 2018). Measurement error for predictors is likely to exist in these studies if the predictors were assessed in various ways across the centres or time periods. Another seven studies clearly specified that different examiners/observers were calibrated to minimize measurement errors.

Lack of blinding of predictors assessors to outcome information may increase the predictor-outcome association and lead to a biased prediction (Moons and Grobbee, 2002). While this does not represent a problem in prediction modelling studies using data from RCTs and prospective cohort studies, lack of blinding to the outcome could not be achieved in prediction modelling studies using data from retrospective cohorts and cross-sectional designs, as the predictors and outcomes were assessed at a similar time (n = 6).

Though the included prediction models cover multiple well-known risk factors, we should be aware that unmeasured predictors always exist, e.g., genetic information. For external validation studies, omitting unavailable predictors that presented in the derived model (Prince et al., 2016; Schwendicke et al., 2018) can result in validation of another model rather than the intended model.

Bias related to outcomes

Bias related to outcomes refer to factors that can influence the outcome assessment, such as (i) unclear definition, (ii) measurement error and (iii) incorporation bias, which usually occurs when the predictors share the same or part of the information captured by the outcome, and is usually observed when the outcome is a manifest/latent variable. Overall, 9 studies showed high ROB related to outcomes, with some of them being suspected to multiple biases. Three out of the 34 identified studies did not use a pre-specified or clear outcome definition (Martinez-Canut et al., 2018; Prince et al., 2016; Su et al., 2017).

Similar to predictor measurement error, inconsistent outcome measurement (e.g., registry data, multiple-centre study) was likely to exist due to variations in assessments among professionals from multiple health care facilities (n = 2) (Martinez-Canut and Llobell, 2018; Rao et al., 2016) and across different time periods (n = 7). Seven studies stated that the agreement or consistency between various examiners was verified and achieved, thus reducing the chance of outcome measurement error.

Incorporation bias was observed in eight studies. One example can be found in the study by (Su et al., 2017), where the objective was to verify the accuracy of the Community Periodontal Index (CPI) in predicting periodontitis. However, the outcome under study (CPI score) is based, among other factors, on the periodontal pocket depth, which is the most important criteria for the definition of periodontitis.

Bias related to analysis

Overall, 23 studies were identified to be high ROB on the analysis due to categorization of continuous variables (n = 16), omission of samples with missing data (n = 10), selection of predictors based on univariate analyses (n = 9 out of 24), lack of over-fitting consideration (n = 13 out of 24), and insufficient model performance assessment (n = 24).

Twelve out of the 16 studies which categorized continuous variables classified age into 5-year age groups (n = 7) or larger categories (n = 5). Three studies categorised pocket probing depth and/or level of bone loss (Martinez-Canut et al., 2018; Martinez-Canut and Llobell, 2018; Schwendicke et al., 2018), and two studies categorised laboratory test results such as low-density lipoproteine (OuYang et al., 2017; Peng et al., 2018).

Half of the studies (n = 17) did not provide information on missing data handling. Among the remaining 17 studies, 13 studies conducted 'complete-case' analyses (ten and three studies had missing information proportion greater and less than 5%, respectively), two studies used mean/median imputation (Bersani et al., 2017; Lee et al., 2018a), one used conditional mean imputation (Ritter et al., 2016) to assign missing data and one study reported there was no missing data (Jover-Espla et al., 2018).

Over a third of the model development studies (n = 9) adopted 'univariate analyses' with various criterion (e.g., p-value < 0.05) to select predictors. Seven studies used variable

selection methods such as backward elimination and stepwise forward-entry. The predictor selection process was unclear in six model development studies.

Our review found a good consideration of data complexity (e.g., censoring) in survival predictions (n = 7), where all the studies have applied time-to-event analysis (e.g., Cox regression). Among the 24 model development studies, five conducted external validation with an independent dataset, and 13 studies conducted an internal validation of developed models. Out of these 13 studies, two used a split-sample procedure (Kayaoglu et al., 2016; Lee et al., 2018a), five used cross validation (Bersani et al., 2017; Canullo et al., 2017, 2016; Papantonopoulos et al., 2017; Ritter et al., 2016), and six adopted bootstrap resampling. Bootstrapping, however, is a form of estimating variance using a computational approach as opposed to an analytical approach, and should not be considered as a way of internal validation.

Insufficient model performance assessment (e.g., discrimination and calibration) was observed in the majority of the studies (n = 24). While 28 studies assessed the discriminative ability of the prediction models, only seven studies assessed their calibration, out of which five presented it graphically and two performed Hosmer-Lemeshow test. The Area under the receiver operating characteristics curve (n = 19) was the most-adopted discrimination measurement, followed by specificity (n = 17), sensitivity (n = 16), positive/negative predictive value (n = 9), and accuracy (n = 8).

5.3.3 Transparent reporting of the included prediction modelling studies

Reporting of individual TRIPOD items are presented in Figure 5.3, and reporting of TRIPOD for each study is shown in Figure 5.4. Overall, at least five (out of 31) items were inadequately reported in 95% of the studies.

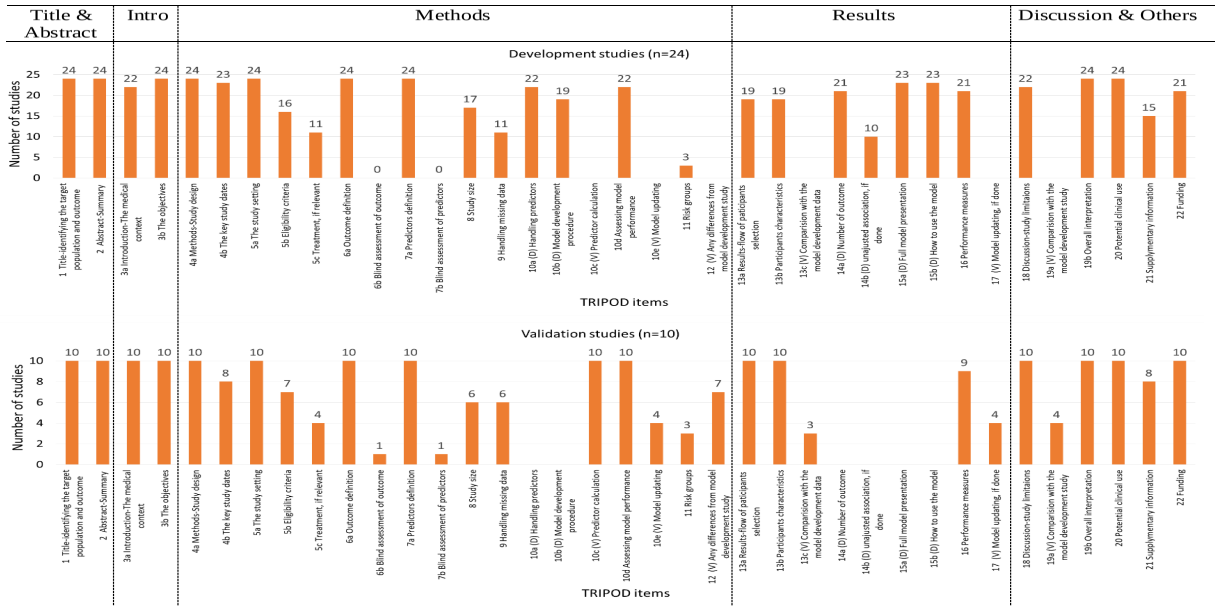


Figure 5.3: Numbers of studies that reported each TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) item.

$_{(D)}$ Development study only; $_{(V)}$ Validation study only.

The TRIPOD items most often reported included title (item 1), abstract (item 2), and introduction (item 3) across all studies. Methods, results and discussion items are reported separately for model development and validation below:

Model development studies

TRIPOD scores varied from to 20 to 28 (out of 31) in the 24 studies.

A) Methods: Although all the studies reported the data source (e.g., study design, key study dates etc.) (item 4), lack of reporting on data collection procedures (e.g., sampling approaches, data entry, data cleaning), inclusion and exclusion criteria (item 5b), and treatment (item 5c) were identified in more than half of the studies. No study reported blind assessment of predictors (item 6b) and outcomes (item 7b). 30% of the studies did not report how the sample size was arrived at (item 8). In addition, missing data handling (item 9) was not reported in 60% of studies (n = 14).

B) Results and Discussion: 63% of the development studies did not report unadjusted associations between each candidate predictor and outcome (item 14b). Other aspects regarding description of participants (item 13), model development procedure (item 14a), models' predictive performance (item 16), study limitations (item 18), model interpretation (item 19), and study implications (item 20) were reported in at least 80% of the studies.

Model validation studies

Ten model validation studies yield a range of 20 to 27 TRIPOD scores (total 31).

A) Methods: Similar to model development studies, only 10% of the validation studies reported blind assessment of predictors (item 6b) and outcomes (item 7b). Information regarding model updating (e.g., recalibration) (item 10e) was reported in 40% of the studies. 60% of the studies reported information on sample size calculation (item 8) and missing data handling (item 9). Only 70% of the validation studies provided a comparison of the distribution of important characteristics of the population (e.g., age distribution, sex) between the model development and validation studies (item 12).

B) Results and Discussion: Insufficient reporting of the updated model metrics (item 17) lead to a lack of discussion on comparison of model performance between development and validation cohorts (item 19a) in 60% of the studies. Similar to model development studies, many aspects regarding description of participants (item 13), model predictive performance (item 16), study limitation (item 18), model interpretation (item 19), and study implication (item 20) were reported in at least 90% of the studies.

5.4 Discussion

Bias and non-transparent reporting were identified across all the included studies. Three quarters of the studies were susceptible to at least four (out of 20) sources of bias, including

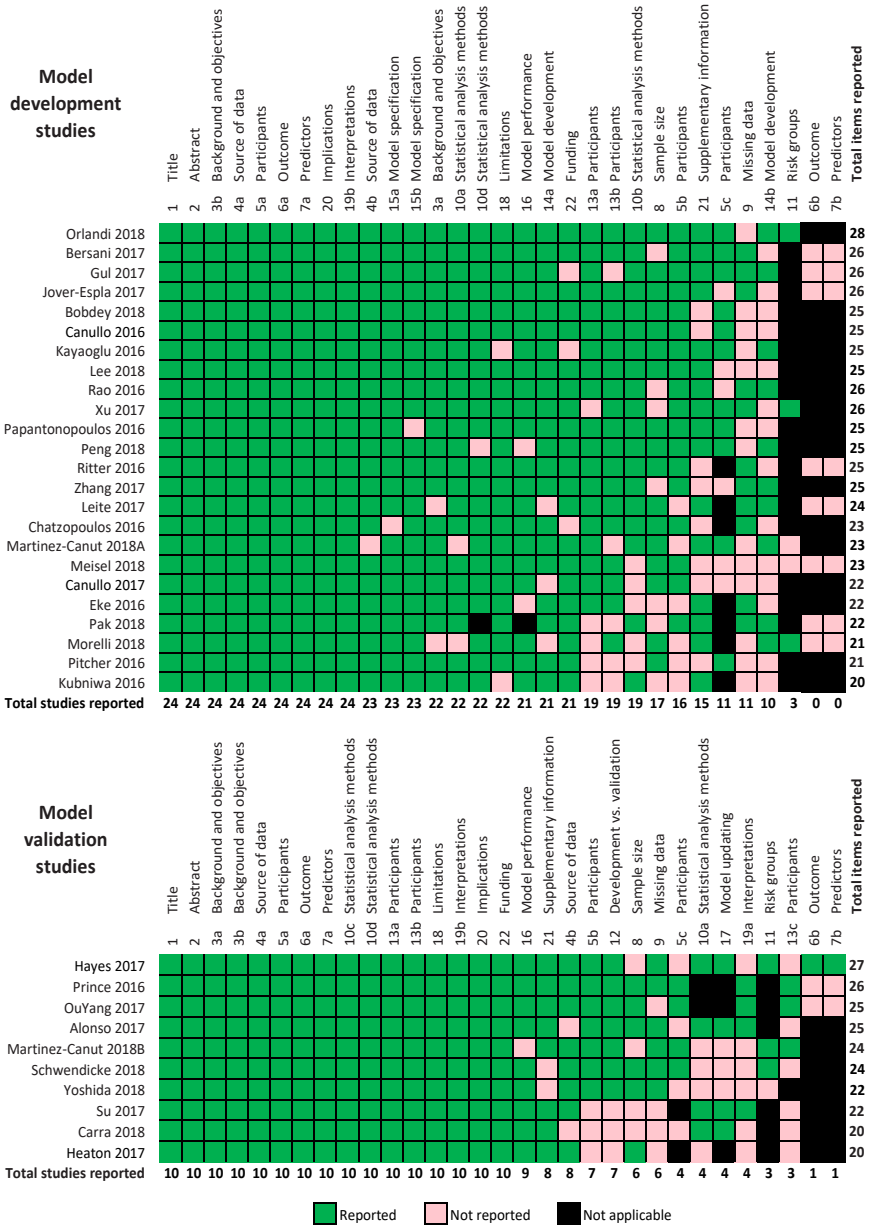


Figure 5.4: Completeness of TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) checklist for 34 prediction modelling studies.

Reported (green), Not reported (pink), Not applicable (black). Right columns are the least reported items, and the bottom studies represent the 'least-reported' studies.

measurement error in predictors and/or outcome, omitting samples with missing data, selecting variables based on univariate analyses, overfitting, and lack of model performance assessment. Similarly, 95% of the studies presented inadequate reporting in at least five (out of 31) TRIPOD items, especially regarding sampling approaches, eligibility criteria, and model-building procedures.

5.4.1 Summary of main findings and quality of the evidence

- To allow selection bias assessment, studies using data from existing cohorts, data registries, experiments and surveys are required to provide detailed information regarding sampling process, data collection and cleaning procedures. In case data are from case-control studies, the prediction model should adjust for the baseline risk (Moons et al., 2019) to reduce the variability of the starting difference among the two groups as well allows for re-calibration when this model to be externally validated in another population (Ahmed et al., 2014). When using medical records, information on record collection, data entry, and data cleaning should be provided.

- Measurement error has the potential to influence the overall performance of a risk prediction model (Rosella et al., 2012) therefore consistent assessment criteria should be followed when measuring both predictors and outcomes. When multiple data sources are used, it would be a good practice if model developers explained transparently/clearly the methods used for assessing the predictors and outcomes.

- The most commonly categorised variable in the included studies was age (n = 12). The degree of loss of information caused by categorising continuous variables may vary according to the population and outcomes under study. For example, among adults, the risk of periodontal diseases (Eke et al., 2015) and dental caries (Slade, 2007) is fairly stable within 5-15 year age groups, however the risk of dental caries varies greatly from one year to another among children (Gradella et al., 2011). Therefore, categorising age into age groups among children when predicting dental caries could lead to bias whereas it may not be a concern among adults and when outcomes, such as periodontal diseases are studied.

- Missing data occurs both in longitudinal and cross sectional studies. The missing data problem at the item level needs to be tackled from three aspects: the missing data mechanisms, the proportion of missing data, and patterns of missing data.

  A common approach to handle missing data in oral health prediction modelling studies is complete-case analysis, while the use of multiple imputation (MI) is sparse. Complete cases analysis are unbiased if the missing mechanism is missing completely at random (MCAR). Usually this is not the case in real life, hence missing data can be imputed using MI if the missing data

are missing at random (MAR) to retain the sample size and reduce biased prediction due to complete-case analysis (Madley-Dowd et al., 2019). Bias caused by missing data when they are missing not at random (MNAR) can be addressed by sensitivity analyses examining the effect of different assumptions about the missing data mechanism (Sterne et al., 2009).

Regarding the proportion of missing data, there is no set cut off from the literature regarding an acceptable percentage of missing data. However, following (Schafer, 1999) and (Little and Rubin, 2002), missing information less than 5% is inconsequential, and complete-case analysis and single imputation can be used along with estimating variance due to single imputation.

- Univariate analyses are discouraged to be used for predictor selection (Rothman, 2008), because (i) if two predictors are highly correlated with each other and with the response, then the univariate analysis will identify both as 'significant', and (ii) some variables are only 'significant' when we adjust other covariates. We suggest researchers select predictors based on the clinical knowledge (but not relying on *p*-values) and on the availability of such information when the models intend to be used.

- Lack of internal validation may lead to over-fitting because quantifying the predictive performance of a model on the same data from which the model was developed tends to give optimistic estimates of performance, that is, the model is over adapted to the development data set. We thus suggest that any form of internal validation such as *k-fold* cross-validation be applied before moving to external validation, which is necessary before implementing prediction models in clinical practices.

- As suggested by PROBAST item 4.7, assessment of both discrimination and calibration is essential to make models predictive ability (model performance) known.

These abovementioned suggestions are not an exhaustive list. To avoid potential threats to reproducible prediction modelling studies (Appendix C Supplement 12), more efforts could be made, such as making the data used in the research accessible (if no confidential agreements exist), providing statistical codes used in analysis, and pre-registering RCTs.

5.4.2 Strengths and potential limitations

To our knowledge, this is the first study that has comprehensively appraised the methodological and reporting quality of multivariable prediction modelling studies in oral health research. A main strength of this study is the adoption of PROBAST and TRIPOD guidelines as benchmarks, although an 'acceptable TRIPOD score' still is lacking, and some PROBAST criteria are not tailored for oral health prediction research.

Limitations

The current review is limited to studies published in pre-specified journals over the past three years, leading us to underestimate the number of studies with high ROB and poor reporting.

It must also be said that the tools that were used are not free from limitations. For example, PROBAST emphasizes MI for handling any missing data (e.g., MCAR, MAR, MNAR); this is incomplete. Even though PROBAST is very instructive in regard to the method of MI, methods such as single imputation can be used when the proportion of missing information is less than 5% with a detailed description of variance estimation methods. Moreover, when the missing data are MNAR, methods such as pattern mixture models can be more appropriate than the MI.

5.5 Conclusion

The majority of the prediction modelling studies identified in this review fell short in addressing bias due to missing data (85%), overfitting/underfitting (54%), categorising continuous variables (46%), univariate variable selection (41%) and measurement error (35%). None of the included study rigorously followed the TRIPOD recommendations for study reporting, especially in reporting model performance metrics (70%) and missing data handling (47%). To improve the overall quality of prediction modelling studies, PROBAST criteria need to be taken into account to address the potential of bias during modelling procedure, and TRIPOD checklist is encouraged to be provided to reviewers and readers for the assessment of the study reporting transparency. Our suggestions and proposed achievable steps towards improving the reproducibility of prediction modelling studies can be wider adopted in oral health research.

5.6 Acknowledgments

MD contributed to conception, design, studies selections, data extraction, critical appraisal, results interpretation, drafted and critically revised the manuscript. DH contributed to conception, design, critical appraisal, results interpretation, and critically revised the manuscript. YS contributed to studies selections, and critically revised the manuscript. JL contributed to conception, design, and critically revised the manuscript. MM contributed to conception, design, critical appraisal, results interpretation, and critically revised the manuscript. All authors gave their final approval and agree to be accountable for all aspects of the work. MD is supported by Adelaide University China Fee Scholarship. We acknowledge the research librarian at the University of Adelaide Ms.Vikki Langton for the help in literature search. The authors declare that there are no conflicts of interest in this study.

CHAPTER 6

Comparison of the Tree-Based Machine Learning Algorithms to Cox Regression in Predicting the Survival of Oral and Pharyngeal Cancers: Analyses Based on SEER Database

Preface

Chapter 5 identified the existence of multiple sources of bias in the current oral health prediction modelling studies. In this chapter, we conduct a case study to address some of the identified bias, as well as to demonstrate the adherence to the reporting guidelines of clinical prediction modelling study. We choose to predict the survival of oral and pharyngeal cancers (OPCs), for two reasons: 1) OPCs are the only fatal oral disease and predicting the survival of patients with OPCs has the potential to benefit patients' lives; 2) time-to-event data is a common type of outcome data in health research, and we use this chapter as an example to demonstrate how to develop models for survival outcomes. Following the growing trend in the application of machine learning methods in cancer research, we present the use of tree-based machine learning algorithms and compare their predictive performance to the standard Cox proportional hazard regression. We have used a real-world cancer registry data set, which includes various prognosis factors and is subjected to different forms of biases such as missing data and unmeasured predictors. Moreover, it is criticised to handle missingness in the time-to-event data using multiple imputation based on standard multivariate normal distribution. For this reason, we demonstrate the use of a compatible method – substantive model compatible fully conditional specification (SMC-FCS) approach for handling missing data for Cox regression. SMC-FCS allows for compatibility between the imputation model and the final prediction model. Additionally, a web-based calculator is presented in this chapter for estimating the 3- and 5-year survival probability of OPCs patients.

This chapter contains the third of a series of four studies contributing to this thesis. Following the requirements of the journal, we describe the results of this study prior to the methods, and this chapter is organized in a different order compared to previous chapters: Abstract, Introduction, Results, Discussion, and Methods. Details for this publication are:

Du M., Haag, D. G., Lynch, J. W., Mittinty, M. N. Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on SEER database. Cancers. 2020;12(10), 2802. https://doi.org/10.3390/cancers12102802.

The accepted version of the published paper is reproduced as follows.

Statement of Authorship

Title of Paper	Comparison of the Tree-Based Machine Learning Algorithms to Cox Regression in Predicting the Survival of Oral and Pharyngeal Cancers: Analyses Based on SEER Database
Publication Status	√ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Du M., Haag, D. G., Lynch, J. W., Mittinty, M. N. Comparison of the Tree-Based Machine Learning Algorithms to Cox Regression in Predicting the Survival of Oral and Pharyngeal Cancers: Analyses Based on SEER Database. Cancers. 2020;12(10), 2802.

Principal Author

Name of Principal Author (Candidate)	Mi Du		
Contribution to the Paper	MD contributed to conception, design, data preparation, data analysis, results interpretation, drafted and critically revised the manuscript.		
Overall percentage (%)	90%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	03-05-2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dandara G. Haag		
Contribution to the Paper	DGH contributed to conception, design, results interpretation and critically revised the manuscript.		
Signature		Date	3/05/21

Name of Co-Author	John W. Lynch		
Contribution to the Paper	JWL contributed to conception, design, results interpretation and critically revised the manuscript.		
Signature		Date	3/05/21

Name of Co-Author	Murthy N. Mittinty		
Contribution to the Paper	MNM contributed to conception, design, results interpretation, drafted and critically revised the manuscript.		
Signature		Date	3/05/21

Abstract

Aims: This study aims to demonstrate the use of the tree-based machine learning algorithms to predict the 3- and 5-year disease-specific survival of oral and pharyngeal cancers (OPCs) and compare their performance with the traditional Cox regression.

Methods: A total of 21,154 individuals diagnosed with OPCs between 2004 and 2009 were obtained from the Surveillance, Epidemiology, and End Results (SEER) database. Three tree-based machine learning algorithms (survival tree (ST), random forest (RF) and conditional inference forest (CF)), together with a reference technique (Cox proportional hazard models (Cox)), were used to develop the survival prediction models. To handle the missing values in predictors, we applied the substantive model compatible version of the fully conditional specification imputation approach to the Cox model, whereas we used RF to impute missing data for the ST, RF and CF models. For internal validation, we used *10-fold* cross-validation with 50 iterations in the model development data sets. Following this, model performance was evaluated using the C-index, integrated Brier score (IBS) and calibration curves in the test data sets.

Results: For predicting the 3-year survival of OPCs with the complete cases, the C-index in the development sets were 0.77 (0.77, 0.77), 0.70 (0.70, 0.70), 0.83 (0.83, 0.84) and 0.83 (0.83, 0.86) for Cox, ST, RF and CF, respectively. Similar results were observed in the 5-year survival prediction models, with C-index for Cox, ST, RF and CF being 0.76 (0.76, 0.76), 0.69 (0.69, 0.70), 0.83 (0.83, 0,83) and 0.85 (0.84, 0.86), respectively, in development data sets. The prediction error curves based on IBS showed a similar pattern for these models. The predictive performance remained unchanged in the analyses with imputed data. Additionally, a free web-based calculator was developed for potential clinical use.

Conclusion: Compared to Cox regression, ST had a lower and RF and CF had a higher predictive accuracy in predicting the 3- and 5-year OPCs survival using SEER data. The RF and CF algorithms provide non-parametric alternatives to Cox regression to be of clinical use for estimating the survival probability of OPCs patients.

6.1 Introduction

Globally, oral and pharyngeal cancers (OPCs) are ranked as the ninth most prevalent type of cancers (Bray et al., 2018). As the only life-threatening diseases in oral health, OPCs have an estimated incidence of 834,860 new cases worldwide in 2018 (Bray et al., 2018), and have shown an increasing incidence trend over the past two decades (Du et al., 2019). Despite the advances in multiple types of therapies for OPCs, such as tumour removal surgery, chemo(radio)therapy and molecular-targeted therapy (Kioi, 2017), the current 5-year overall survival rate of OPCs remains 64.8% in the United States. In response to the need for improving medical care delivery in the oral health field, there are clinical decision support tools being developed to aid the early detection, diagnosis, treatment and prognosis of oral diseases, including OPCs (Patton, 2017). These clinical decision support tools are all developed based on clinical prediction modelling research, which aims to yield the most accurate outcome prediction by capturing patterns in the available data (known as data-generating mechanisms) and minimizing the difference between the predicted and observed outcome (known as bias). However, following the up-to-date bias assessment criteria (PROBAST - Prediction model Risk Of Bias ASsessment Tool) (Wolff et al., 2019) and reporting guidelines (TRIPOD - Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) (Collins et al., 2015), the overall quality of oral health prediction modelling studies was found to be less than optimal due to the presence of multiple sources of bias (e.g., measurement error, unmeasured predictors) and lack of reporting transparency (Du et al., 2020).

The traditional method of survival prediction for OPCs has been building nomograms using Cox proportional hazard (Cox) regression analysis based on available clinical and sociodemographic predictors (Chen and Asch, 2017; Wang et al., 2018). Such models are generally based on the assumptions that each predictor is linearly associated with OPCs survival outcomes. Thus, there is a possibility that these models may oversimplify complex relationships, which potentially include both non-linear associations, non-linear interactions and effect modification (Breslow, 1975; Chen and Asch, 2017). To overcome this limitation, the evolution of machine learning provides an alternative to (semi)parametric modelling by relaxing the hypothesis of the data-generating mechanism and considering all possible interactions and effect modification between variables (Ryo and Rillig, 2017). Among the commonly used machine learning algorithms, tree-based methods (e.g., decision tree, random forest) are well-known for the ease of use, interpretability and the nature of preventing overfitting (Duda et al., 2012).

To date, despite machine learning algorithms being used for predicting OPCs prognosis (Kim et al., 2019; Tseng et al., 2015), they rarely accommodate the potential systematic bias arising from data collection (e.g., missing data, measurement error) and modelling process (e.g., unmeasured predictors). Moreover, very few outputs from prediction mod-

elling research have been implemented to assist clinical practice. Therefore, this study was performed to contribute to the clinical decision support system in the field of OPCs by 1) developing and validating various models to predict the 3- and 5-year disease-specific survival of OPCs; 2) comparing the predictive accuracy of the tree-based machine learning algorithms and the standard parametric Cox method; and 3) developing a web-based calculator to estimate the individual survival probability of OPCs patients. Additionally, this study demonstrated the conduct and reporting of a clinical prediction modelling study, following up-to-date guidelines (PROBAST and TRIPOD). The significance of this study not only lies in the development of prediction models and an online calculator for OPCs survival, but also includes a call for action to improve the quality (reduce bias) of prediction modelling studies in the field of oral health. Specifically, this study demonstrates how biases due to missing data and unmeasured predictors can be incorporated into predictive modelling.

6.2 Results

6.2.1 Patient selection

A total of 54,955 primary OPC patients diagnosed between 2004 and 2009 were collected from the Surveillance, Epidemiology, and End Results (SEER) database; 27,569 records were excluded based on inclusion and exclusion selection criteria. Patients with survival months of less than 1 month were excluded (n = 771). Left censored samples for 3-year (n = 157) and 5-year (n = 260) cohorts were also excluded. For complete case analysis, patients who had missing values (n = 13,455, 49.49%) on any of the examined variables, including race (n = 327, 1.2%), marital status (n = 1745, 6.42%), tumour size (n = 7654, 28.15%), lymph node involvement (n = 2131, 7.84%), tumour (T), node (N), and metastasis (M) categories (n = 515, 1.89%), clinical stage (n = 4729, 17.39%), differentiated grades (n = 5803, 21.35%), surgery history (n = 262, 0.96%) and ICD (International Classification of Disease) classification (n = 437, 1.61%) were excluded. Due to the different number of left censored samples for 3- and 5-year cohorts, we ended up with 21,154 patients (21,000 for 5-year survival) and 11,888 patients (11,870 for 5-year survival) for final imputation and complete case analysis, respectively. Figure 6.1 shows the flowchart of patient selection.

6.2.2 Characteristics of the patients

Here we only describe the characteristics of patients without missing information in Tables 6.1 and 6.2. Detailed characteristics of patients in the imputed data sets can be found in Appendix D Supplements 1 and 2.

Of the 11,888 (11,807 for 5-year survival) patients, the overall disease-specific survival rates for all patients in the 3- and 5-year cohorts were 65.0% and 60.1%, respectively. Mean survival times were 66.7 (SD = 44.1) and 66.8 (SD = 44.3) months for 3- and 5-year

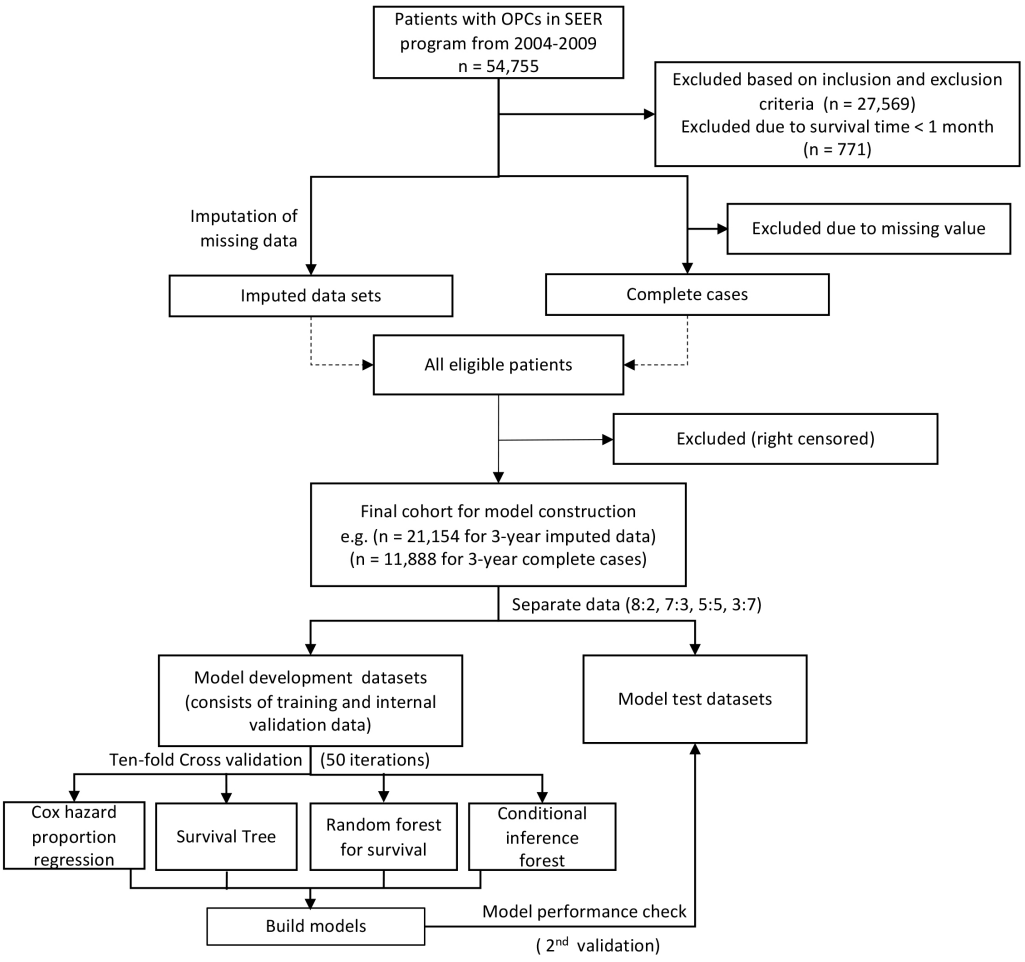


Figure 6.1: Flowchart of study design and patient selection.

cohorts. In the 3-year cohort, patients had a mean age of 59.0 (SD = 12.3) years. Of the 11,888 eligible patients, 8774 (73.8%) were male, 10,020 (84.3%) were white and 6997 (58.9%) were married at diagnosis. Overall, 7293 (61.3%) tumours arose from the oral cavity (C00-06) and tumours were poorly differentiated or undifferentiated. T3-T4 tumours accounted for 31.1% of all tumours and positive neck lymph nodes and distant metastases accounted for 44.8% and 2.9%, respectively; 7189 (60.5%) patients underwent surgery. The 5-year cohort consisted of 11,807 patients, of which 73.9% were male, 84.3% were white and 58.9% were married at diagnosis. Similar to the 3-year cohort, over half of the tumours (61.2%) arose from the oral cavity, while 39.6% tumours were poorly differentiated or undifferentiated. T3-T4 tumours accounted for 31.2% of all tumours, lymph nodes were removed in 44.8% cases and distant metastases occurred in 2.9% tumours; 60.4% patients received surgery.

Table 6.1: Demographic characteristics of patients with oral and pharyngeal cancers in SEER (Surveillance, Epidemiology, and End Results) cohorts.

Characteristics	3-year cohort (n = 11,888)	5-year cohort (n = 11,807)
Death Status		
Alive	7731 (65.0%)	7092 (60.1%)
Dead	4157 (35.0%)	4715 (39.9%)
Survival months		
Mean (SD)	66.7 (44.1)	66.8 (44.3)
Median [min, max]	77.0 [2.00, 143]	78.0 [2.00, 143]
Age(Years)		
Mean (SD)	59.0 (12.3)	59.0 (12.3)
Median [min, max]	58.0 [18.0, 103]	58.0 [18.0, 103]
Sex		
Female	3114 (26.2%)	3086 (26.1%)
Male	8774 (73.8%)	8721 (73.9%)
Race		
American Indian/ Alaska native	55 (0.5%)	54 (0.5%)
Asian or Pacific Islander	762 (6.4%)	750 (6.4%)
Black	1051 (8.8%)	1046 (8.9%)
White	10020 (84.3%)	9957 (84.3%)
Marital status		
Divorced	1532 (12.9%)	1526 (12.9%)
Married (including common law)	6997 (58.9%)	6951 (58.9%)
Separated	135 (1.1%)	133 (1.1%)
Single (never married)	2218 (18.7%)	2198 (18.6%)
Widowed	1006 (8.5%)	999 (8.5%)

Table 6.2: Tumour-related characteristics of patients with oral and pharyngeal cancers in SEER cohorts.

Characteristics	3-year cohort (n = 11,888)	5-year cohort (n = 11,807)
Differentiation grade		
Well differentiated; grade I	1535 (12.9%)	1521 (12.9%)
Moderately differentiated; grade II	5667 (47.7%)	5626 (47.6%)
Poorly differentiated; grade III	4434 (37.3%)	4410 (37.4%)
Undifferentiated; anaplastic; grade IV	252 (2.1%)	250 (2.1%)
T category		
T1	3956 (33.3%)	3917 (33.2%)

Table 6.2 continued from previous page

T2	4204 (35.4%)	4171 (35.3%)
T3	1562 (13.1%)	1556 (13.2%)
T4	2132 (17.9%)	2129 (18.0%)
TX	34 (0.3%)	34 (0.3%)
N category		
N0	4659 (39.2%)	4615 (39.1%)
N1	2416 (20.3%)	2404 (20.4%)
N2	4386 (36.9%)	4362 (36.9%)
N3	391 (3.3%)	390 (3.3%)
NX	36 (0.3%)	36 (0.3%)
M category		
M0	11,447 (96.3%)	11367 (96.3%)
M1	348 (2.9%)	347 (2.9%)
MX	93 (0.8%)	93 (0.8%)
Stage		
I	2183 (18.4%)	2160 (18.3%)
II	1540 (13.0%)	1522 (12.9%)
III	2356 (19.8%)	2341 (19.8%)
IV	5809 (48.9%)	5784 (49.0%)
Lymph nodes removed		
None	6557 (55.2%)	6515 (55.2%)
Yes	5331 (44.8%)	5292 (44.8%)
Tumour size		
0∼1cm	1420 (11.9%)	1404 (11.9%)
1∼2cm	2806 (23.6%)	2783 (23.6%)
2∼3cm	3120 (26.2%)	3104 (26.3%)
3∼4cm	2097 (17.6%)	2081 (17.6%)
4∼5cm	1369 (11.5%)	1363 (11.5%)
5∼6cm	561 (4.7%)	559 (4.7%)
6∼7cm	253 (2.1%)	252 (2.1%)
7∼8cm	128 (1.1%)	127 (1.1%)
8∼9cm	55 (0.5%)	55 (0.5%)
9∼10cm	41 (0.3%)	41 (0.3%)
>10cm	38 (0.3%)	38 (0.3%)
Surgical therapy		
Surgery not performed	4699 (39.5%)	4673 (39.6%)

Table 6.2 continued from previous page

Surgery performed	7189 (60.5%)	7134 (60.4%)
Tumour sites (ICD code)		
Lip (C00)	540 (4.5%)	536 (4.5%)
Base of tongue (C01)	2192 (18.4%)	2177 (18.4%)
Other parts of tongue (C02)	2400 (20.2%)	2378 (20.1%)
Gum (C03)	412 (3.5%)	410 (3.5%)
Floor of mouth (C04)	786 (6.6%)	783 (6.6%)
Palate (C05)	318 (2.7%)	314 (2.7%)
Other oral cavity (C06)	645 (5.4%)	639 (5.4%)
Parotid gland (C07)	253 (2.1%)	253 (2.1%)
Other salivary glands (C08)	39 (0.3%)	38 (0.3%)
Tonsil (C09)	2858 (24.0%)	2840 (24.1%)
Oropharynx (C10)	363 (3.1%)	362 (3.1%)
Nasopharynx (C11)	408 (3.4%)	404 (3.4%)
Pyriform sinus (C12)	391 (3.3%)	391 (3.3%)
Hypopharynx (C13)	283 (2.4%)	282 (2.4%)

6.2.3 Model specification

In this study, a commonly used Cox regression was chosen as the reference model and three tree-based machine learning algorithms (survival tree (ST), random forest (RF) for survival and conditional inference forest (CF)) were used to develop prediction models. These tree-based models were applied because our data set was a mix of continuous and categorical variables, among which a large proportion had polychotomous values (i.e., more than two levels), and a major advantage of tree-based models is that they can handle this type of data by allowing for multiple splits of a selected node. Additionally, tree-based models can rank the importance of variables based on their location depth in the tree structure whereas other popular machine learning algorithms (e.g., neural networks) focus on outcome prediction with less consideration of the variables' contribution. Here we describe the prediction models for 3-year OPC survival for the complete cases. According to the hazard ratios returned by Cox regression, all the included predictors were identified as having prognostic value for predicting OPC survival (majority of the hazard ratios with 95% confidence intervals not crossing 1) (Appendix D Supplement 3). The two most important predictors that determine the 3- and 5-year survival of an individual patient were tumour site (hazard ratios for ICD being C02—tongue excluding base of tongue,

Table 6.3: C-index (Median (IQR) in the development and test datasets for various models for predicting 3- and 5-year disease-specific survival of oral and pharyngeal cancers based on the complete case analysis.

Modelling approaches	Development data set	Test data set
Three-year survival cohort		
Cox	0.77 (0.77, 0.77)	0.76 (0.76, 0.77)
Survival tree (ST)	0.70 (0.70, 0.70)	0.70 (0.69, 0.71)
Random forest for survival (RF)	0.83 (0.83, 0.84)	0.77 (0.76, 0.77)
Conditional inference forest (CF)	0.83 (0.83, 0.86)	0.76 (0.75, 0.76)
Five-year survival cohort		
Cox	0.76 (0.76, 0.76)	0.76 (0.76, 0.76)
Survival tree (ST)	0.69 (0.69, 0.70)	0.69 (0.68, 0.70)
Random forest for survival (RF)	0.83 (0.83, 0.83)	0.76 (0.76, 0.76)
Conditional inference forest (CF)	0.85 (0.84, 0.86)	0.75 (0.75, 0.76)

C04—floor of mouth, C08—other salivary glands, were > 3) and tumour metastasis (hazard ratio = 2.56). The final ST was with two parameters: the number of the minimum number of observations that must exist in a node ("*minsplit*") = 11 and maximum length from root node to leaf node ("*maxdepth*") = 18. The constructed RF model which returned the highest value of C-index included the following parameters: number of trees ("*ntree*") = 1217, number of variables tested in any split ("*mtry*") = 11, the size of random split points for each "*mtry*" candidate ("*nsplit*") = 3 and split rule/formula ("*splitrule*") = "*logrank*".

6.2.4 Model performance

The predictive performance of the models was measured using Harrell's C-index (C-index) (Tseng et al., 2015), integrated Brier score (IBS) (Mogensen, 2012) and calibration plots. Values of C-index for each model are presented in Table 6.3. In the main text, we present results from complete case analysis for the 8:2 split of the model development and test datasets. Results for the imputed datasets and any other splits can be found in Appendix D Supplements 4–7.

For 3-year survival prediction, in the complete case analysis (n = 11,888), the C-indexes in development data sets were 0.77 (0.77, 0.77), 0.70 (0.70, 0.70), 0.83 (0.83, 0.84) and 0.83 (0.83, 0.86) for Cox, ST, RF and CF respectively. Similar results were found in 5-year survival cohort: CF yielded the highest values of C-index of 0.85 (0.84, 0.86), followed by RF (0.83 (0.83, 0.83)), Cox (0.76 (0.76, 0.76)) and ST (0.69 (0.69, 0.70)). The values of C-indexes in the imputed data were similar to the complete case analysis. As shown by the over-time C-index in Figure 6.2, RF and CF constantly exhibited the best C-index throughout the investigated period across all settings. However, this comes with a computational burden: for example, in the complete case analysis of 3-year cohort, the average training times for Cox, ST, RF and CF for each iteration were 2.78s, 0.61s, 348.95s and 2513.72s, respectively. In addition to the over-time C-index plots, Figure 6.3

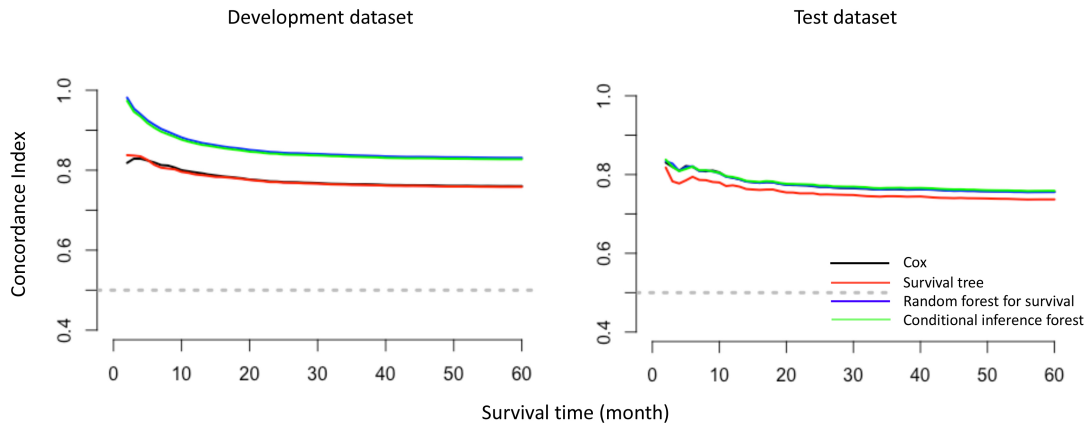


Figure 6.2: Overtime C-index for predicting disease-specific survival of oral and pharyngeal cancers with various models (Cox regression, survival tree (ST), random forest for survival (RF) and conditional inference forest (CF)) based on the complete case analysis.

presents the prediction error curves for each model based on IBS. We found that all models performed better than the default benchmark Kaplan–Meier model. The RF and CF models (blue and green curves) had approximately the same values. Compared with the Cox model, ST showed a higher prediction error while RF and CF showed lower prediction errors in the test data sets. For the completeness and comparability of prediction models with the binary family we plotted the time-dependent receiver operator curves (ROC) based on the cumulative sensitivity and dynamic specificity for Cox models (Kamarudin et al., 2017). The values of area under the ROC at specific time points (1- to 5-years) were consistent with the results of over-time C-index plots. Detailed explanation and results can be found in Appendix D Supplement 8. In terms of calibration, the calibration curves (Figure 6.4) displayed by all the algorithms in the test data sets appeared close to each other, despite the weaker calibration exhibited by RF and CF in development data sets.

6.2.5 Model presentation and development of an online survival calculator

In attempt to contribute to clinical decision-making, we have developed a web-based OPCs survival probability calculator based on a Cox regression model (this online tool was designed for research only and should not be used clinically until externally validated). This online tool was developed as a ShinyApp software using R package *shiny* (https://dumizai.shinyapps.io/apptest2/). A snapshot of the online calculator can be found in Appendix D Supplement 9. The potential users of this calculator are OPCs clinicians, patients and interested researchers. This calculator can 1) present the 3- and 5-year overall survival curves of OPCs cohorts in SEER database; 2) interact with the users to provide survival curves stratified by each predictor; and 3) interact with the users to provide survival probability of an individual OPCs patient for 1, 2, 3, 4 and 5 years after diagnosis.

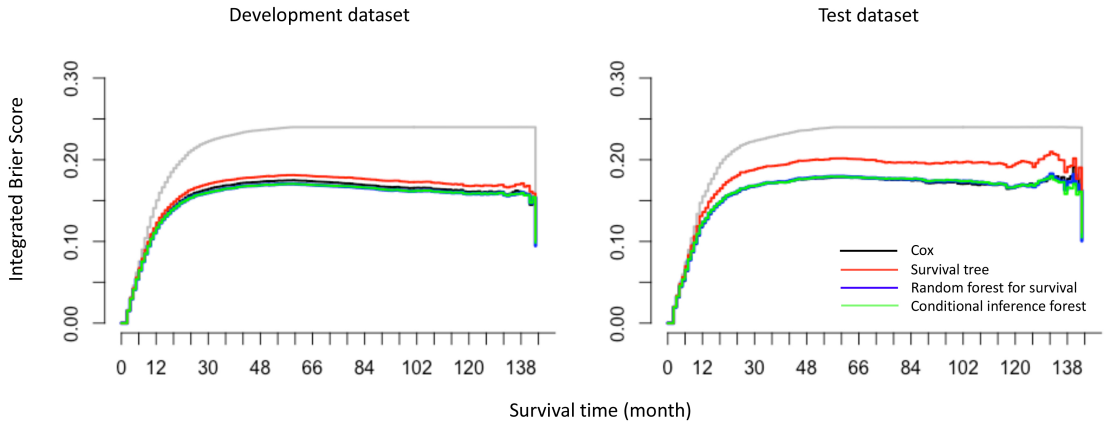


Figure 6.3: The prediction error curves for models (Cox regression, survival tree (ST), random forest for survival (RF) and conditional inference forest (CF)) in predicting disease-specific survival of oral and pharyngeal cancers based on the integrated Brier score (IBS).

The maximum (observed) survival month was 143 for OPC patients in the SEER database. Different models are presented by different colours, where the grey curve represents a default benchmark Kaplan–Meier model. All curves start at time 0 where all subjects are alive and all predictions equal to 1.

6.3 Discussion

By comparing the prediction performance of three tree-based machine learning algorithms (ST, RF and CF) to the reference method (Cox), our findings suggest that Cox regression performed robustly as a conventional method for OPCs survival prediction; despite this, we observed an increase in C-index for RF and CF and a decrease for ST. To facilitate the translation of our developed model into clinical practice, we developed a web-based survival probability calculator to allow better visualization and ease-of-use for clinicians. It can dynamically predict the disease-specific survival probability of OPCs patients at various time points and help identify patients at high risk of OPC-specific death.

Different C-indexes obtained by Cox and tree-based models led to our further thinking on the reasons for these differences. In general, conventional statistical models (e.g., Cox) attempt to fit the data to an investigator-specified model, whereas machine learning algorithms allow the data to dictate the form of the model. In our study, Cox regression examines each predictor's effect by testing the proportional hazards assumptions, ST focuses on data partitioning by maximizing the between-node differences, while RF and CF focus on overall prediction accuracy by reducing the prediction error. Traditional Cox regression is popular and well-studied, results are easily interpretable and it remains the most convenient solution for most survival problems. However, it is a (semi)parametric model and only works when the number of predictors is less than the number of events (Vittinghoff and McCulloch, 2007). ST, built with "*rpart*" package (Therneau et al., 2015), employed a log-rank test statistic (a statistic for comparing the survival curves of two

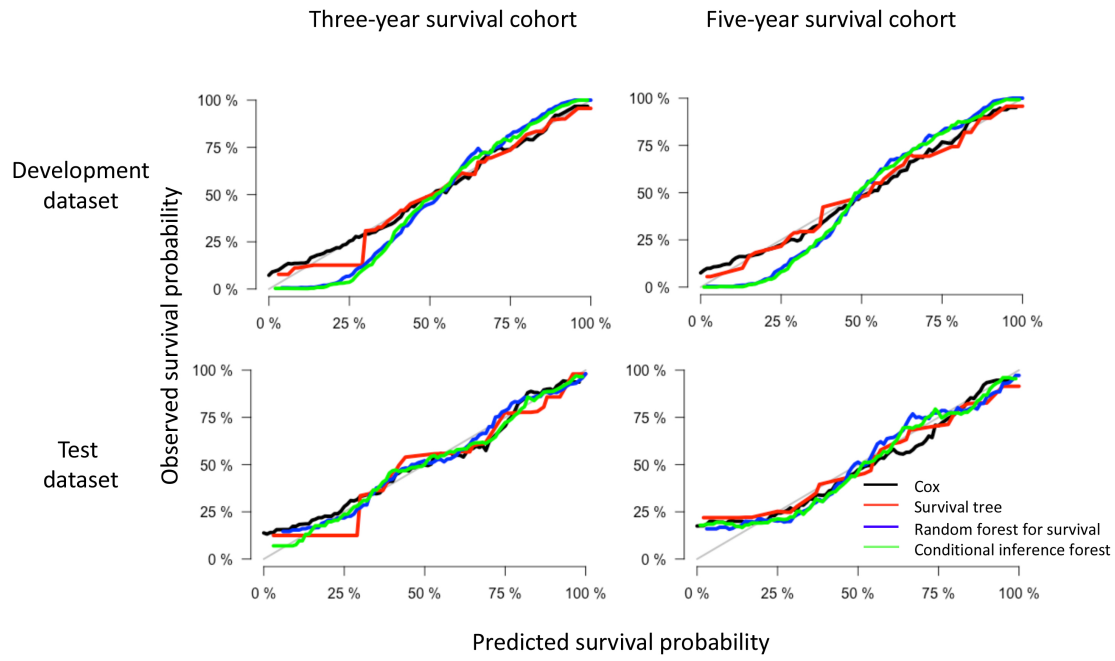


Figure 6.4: Example of calibration plots for predicting 3- and 5-year disease-specific survival of oral and pharyngeal cancers with various models (Cox regression, survival tree (ST), random forest for survival (RF) and conditional inference forest (CF)).

The 45-degree straight line represents the perfect match between the observed (y-axis) and predicted (x-axis) survival probabilities. A closer distance between two curves indicates higher accuracy.

samples) to maximize the between-subgroup heterogeneity. However, the large variance in the survival time in our study (2 to 143 months) may add complexity when distinguishing between subgroups. RF, implemented by the "*randomForestSRC*" package (Ishwaran et al., 2020) and CF, implemented by the "*partykit*" package (Mogensen, 2012), are ensemble methods which aggregate a large number of trees. By combining thousands of trees and testing multiple node splitting rules, forest took into account all possible link functions between the outcome and predictors, as well as all possible interactions between variables. Therefore, it approximates the data-generating mechanism that is in the observed data and gains the closest predicted value to the observed value (Breiman, 1996). In short, in the traditional statistical sense, we selected one model (Cox), set up the model's parameters and evaluated its accuracy. Obviously, the initial choice of algorithm would limit the flexibility of the model, while machine learning algorithms (as alternatives for prediction) have no concerns of non-proportionalities, multicollinearity or nonlinearity; thus, they may reduce the prediction bias (systematic prediction error) stemming from modelling uncertainty.

Interestingly, the predictive capabilities represented by the C-index in test data sets were similar between RF, CF and Cox for predicting OPCs survival based on the

SEER database, which suggests that the superiority of machine learning is not always seen but is seen only in situations when the conventional methods meet their limits. These situations include 1) data sets with large numbers of predictors and relatively small sample size; 2) datasets with large numbers of uncorrelated candidate predictors (Ishwaran et al., 2008) (high dimension data sets such as "omics" data); 3) datasets with complex confounding factors, interactions and nonlinearities (e.g., no clear theory or hypothesis of the proportional hazard assumption available) (Breiman, 1996); and 4) survival data sets with high censoring rate, where the predictors were responsible for censoring (Zhou and McArdle, 2015). Therefore, there are several possible explanations to the comparable predictive performance between RF, CF and Cox. The first reason is the small number of predictors used in this study. Though no threshold number of predictors was available, more variables may enhance machine learning algorithms to outperform the conventional statistical models. Secondly, the predictors in the SEER program were collected based on prior clinical knowledge and many of the variables were mostly linearly correlated. Thus, there is a possibility that although RF and CF are non-parametric methods, they captured the underlying structure of data which aligned with the hypothesis that we made in the parametric model (Cox).

In summary, for prediction purposes, it is appropriate to consider a Cox model first for a given survival data set with a smaller number of predictors. Meanwhile, machine learning algorithms can also be adopted in combination with conventional methods, so that one may obtain extra information (e.g., non-linear interactions) in the data that are not grasped by Cox models. It is also noteworthy that there is a computational burden when using RF and it cannot match the speed of the computation of Cox regression methods.

The strengths of our study include that according to PROBAST, there are a variety of systematic biases that can be present in prediction modelling studies, one of which is the bias due to missing information. In this study, we address this issue using imputed data. However, when imputing missing data, one might introduce bias by using incompatible imputation models. For example, with the imputation model being Cox regression and the final prediction model being RF, this creates incompatibility in the inference as well as estimation. We show how this can be avoided by adopting the Substantive Model Compatible Fully Conditional Specification (SMC-FCS) and RF to impute missing data for the Cox regression and tree-based models, respectively. Another bias which can occur is due to pre-specified choice of data analysis model. We address this issue by comparing both parametric and non-parametric models. We also further discuss the possible reasons that may lead to the differences of C-index between Cox and tree-based models and therefore guide users to choose different models accordingly. Finally, the web-based prediction tool for OPCs survival represents an attempt to translate research outputs into clinical practices. This prognostic tool not only informs clinicians and patients of the possible outcome of

OPCs survival, but also provides suggestions to clinicians in decision-making, such as treatment determinations.

There are several limitations in this study. The main limitation of this study is the lack of records of other well-known predictors for OPCs survival in the SEER program. For example, for patients with oropharyngeal squamous cell carcinoma, individuals with human papillomavirus (HPV) positive status are likely to have better prognosis than their HPV-negative counterparts (Norregaard et al., 2018). Therefore, it is worth noting that including HPV information and other prognostic factors such as smoking, alcohol consumption and chemotherapy (Sakamoto et al., 2016) are likely to modify the models' predictive ability. Nevertheless, sensitivity analysis was performed using the R package "*obsSens*" (Snow and Snow, 2015) to test the impact of unmeasured predictors on our estimates for Cox models. In the sensitivity analysis, we added another hypothetical unmeasured predictor with a different effect size (a range of 0.1–2-fold to our included predictors) on the outcome. We then examined how this added unmeasured factor impacted the estimated hazard ratios of the existing predictors. Our results showed that the conclusion might change only when the unmeasured predictor(s) had an (combined) effect on the outcome 2-fold higher than the existing predictor. This suggests that the lack of an unmeasured predictor may not inflate the effect of the existing predictors on survival outcome, and we can trust the hazard ratios and use them for new predictions. Detailed methodology and results can be found in the Supplementary Materials (Appendix D Supplement 10). In future research, new methods are needed for incorporating sensitivity analysis for the computed estimates using the tree-based non-parametric methods. Another major limitation stemmed from the lack of validation in an external cohort; nevertheless, the replicability of the models should be sufficient with a *10-fold* cross-validation method. Moreover, following the reviewers suggestions and also the guidance document of handling missing data in SEER, we presented the results from a complete case analysis in the main document. Additionally, given in the guidance document by SEER there are occasions where data can be imputed under the missing at random assumption, it is for this reason we have presented the results from imputed data in the web supplements. The results from imputed data need to be interpreted with caution. These data were imputed using the missing at random mechanism, which implies that missingness can be fully accounted for by variables that have complete information. However, this assumption might not be appropriate for all variables in the SEER data as it is linked data from multiple sources with multiple types of missingness.

Therefore, future directions derived from our study include: 1) more comprehensive models with better predictive performance can potentially be developed by adding more predictors; 2) external validation of models on another data set (independent of the model development data set) is required for assessing the models predictive capability; 3) apart from the tree-based models, multiple types of machine learning algorithms (e.g., support vector machine, neural networks) could be used for clinical prediction purposes; and 4)

more research is needed to accommodate multiple sources of bias while developing a prediction model (e.g., measurement error/misclassification). We have not conducted sensitivity analysis around all the bias suggested by PROBAST, as it remains unclear how to incorporate these biases in machine learning algorithms.

6.4 Methods

6.4.1 Data source and study population

Data for this study were obtained from the SEER database (approval number: 15617-Nov2017), a population-based cancer registry in the National Cancer Institute in the United States (https://seer.cancer.gov/). The University of Adelaide Human Research Ethics Committee waived the provision of ethics approval for this de-identified secondary data analysis.

The five criteria for patient inclusion were listed as follows. The first criterion was that patients were diagnosed with OPCs as the "Primary and only cancer diagnosis" during 2004 to 2009. The inception year of 2004 was chosen to allow capture of definitions of T, N and M categories as published in the sixth edition of the AJCC Cancer Staging Manual. The last year of 2009 was chosen to guarantee the completion of 5-year follow-up for each patient. The second criterion was defined by the International Classification of Disease 10th revision (ICD-10) as cancer of lips, tongue, gum, floor of mouth, palate, cheek mucosa, vestibule of the mouth and retromolar area (C00-C06), salivary glands (C07-C08), oropharynx (C09-C10), nasopharynx (C11), hypopharynx (C12-C13) (Bray et al., 2018). The remaining three criteria were that patients had a histologically confirmed diagnosis and histological examination of squamous cell neoplasms; an active follow-up with complete dates and a known outcome with "Alive or dead due to cancer". Patients were excluded if they had an unknown survival time, had multiple primary cancers, or were diagnosed by autopsy or death certificate (i.e., unknown date of diagnosis).

6.4.2 Predictors and outcome

Thirteen predictor variables included in this study were age at diagnosis, sex, race, marital status at diagnosis, AJCC TNM category, overall tumour stage, histologic differentiation grade, tumour site, tumour size, whether surgical therapy was undertaken and whether lymph node was removed. All variables were categorical except age (continuous). Specifically, in SEER program, surgical therapy was recorded as "Surgery performed", "Not recommended", "Recommended but not performed, patient refused", "Recommended but not performed, unknown reason", "Recommended, but unknown performed" and "Unknown". We have categorized this predictor as "Surgery performed", "Surgery not performed" and "Unknown" in this study. Tumours arising from different sites (e.g., oral cavity, pharynx, salivary glands) were treated separately in our study. Detailed information

on the categorization of each predictor are shown in Tables 6.1 and 6.2. The primary outcome of interest was 3- and 5-year disease-specific survival, calculated from the date of diagnosis to the date of death due to OPCs. "Survival months" and "Death status" as outcome variables were extracted.

6.4.3 Model development

For this study, patients were randomly assigned to a development or test data set using a sample of 8:2 proportion, where the development data set consisted of training and validation data. Proportion ratios of 7:3, 5:5 and 3:7 were also used to assess the models robustness. Cox regression, survival tree (ST), random forest (RF) and conditional inference forest (CF) were used to develop prediction models. All models were built on a development data set using *10-fold* cross-validation with 50 iterations. Samples in each iteration were randomly drawn from observed data using a different seed. After the final models were obtained, we evaluated, compared and reported the models' predictive performance in test data sets. To explain the *10-fold* cross-validation, 90% of the development data were used for training and the remaining 10% were used for validation. It was obvious that variation in the models estimates existed due to the different partitions of the data to form training and validation data sets. We adopted *10-fold* cross-validation to reduce this variance by averaging over 10 different partitions, so the performance estimates were less sensitive to the random partitioning of the data. For machine learning models, all 13 predictor variables were used as inputs. The outputs were not different from the Cox regression model, which were the estimated 3- and 5-year survival probability for patients with OPCs since diagnosis. Therefore, this was a regression problem based on time-to-event (censored) data.

Cox regression model

The Cox regression model can be expressed by the hazard function denoted by $h(t)$, which can be defined as Equation (6.1):

$$h(t) = h_0(t)exp(\alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_k X_k) \tag{6.1}$$

where t represents the survival time, $h_0(t)$ is the baseline hazard when all of the predictors are equal to zero. The coefficients (α_1, α_2,..., α_k) measure the effect size of predictors (X_1,X_2,...,X_k).

Survival tree model

A single ST can group observations according to their survival behaviour based on their predictors. To grow a tree, at each node every candidate predictor is tried for node splitting.

Within a set of predictors, the one with a split point that maximizes the survival differences between children nodes is finally chosen as the parent node. The growth of a decision tree is continued until the tree meets the stopping criteria, which refers to all terminal nodes containing only a minimal number of unique events which prevents further node splits (pre-defined by "*minsplit*"). Additionally, pruning procedures are used to reduce the size of the tree (pruned hyper parameters for ST can be found in Appendix D Supplement 11). A step-by-step practical procedure of growing a ST is described in Appendix D Supplement 12.

As for the splitting rule, the log-rank statistic can be applied (Segal, 1988). The log-rank statistic maximizes the dissimilarity between children nodes using the following Equation (6.2):

$$L(X,C) = \frac{\sum_{i=1}^{N}(d_{(i,1)} - Y_{(i,1)}(d_i/Y_i))}{\sqrt{\sum_{i=1}^{N} Y_{(i,1)}/Y_i(1-Y_{(i,1)}/Y_i)((Y_i-d_i)/(Y_i-1))d_i}} \tag{6.2}$$

where *L(X, C)* is the log-rank measure of node separation, X is the predictor, C is the splitting point, N is the number of individuals in the parent node, i is the i^{th} observation, d_i is the number of deaths at time ti in the children nodes, $d_{i,1}$ is the number of death in children node 1, therefore $d_i = \sum_{i=1}^{t} d_{(i,t)}$, for example, when $t = 2$, then $d_i = d_{(i,1)} + d_{(i,2)}$. Y_i are the individuals at risk at time t_i and $Y_{i,1}$ is the number of individuals of Y_i in children node 1, therefore $Y_i = \sum_{i=1}^{t} Y_{i,t}$

Random forest model

RF is an non-parametric ensemble method that introduces two forms of randomization into the tree growing process: bootstrap sampling from the data and selection of a limited number of predictors to construct the tree (known as "*mtry*") (Ishwaran et al., 2008). When growing a tree, a random B bootstrap sample that includes two thirds of the development data (in-bag data) is used. Based on the B samples, a node splitting process is applied. The node splitting process works as follows. At each node, according to a splitting criterion (pre-defined by "*splitrule*"), the predictor (among all candidate predictors, pre-defined by "*mtry*") with a split point (pre-defined by "*nsplit*") that maximizes the survival differences between children nodes is used for node splitting. The process of tree growing iterates "*ntree*" times to obtain a forest for final prediction. For predicting the survival of a new subject (S^{new}) at time t in the m^{th} tree, it eventually falls into a terminal node. The final prediction (Equation 6.3) can be obtained by calculating node level estimation and then averaging overall trees:

$$\hat{S}^{new}(t) = \frac{1}{M}\sum_{m=1}^{M} S_m^{new}(t) \tag{6.3}$$

When tuning the forest, the remaining one third of the development data (out-of-bag data) was used to avoid overfitting and to select the models' hyper parameters (Appendix D Supplement 11) which returned the highest prediction accuracy. Details of developing a RF are described in Appendix D Supplement 12.

Random forest based on conditional inference trees — Conditional inference forest (CF)

As we have stated above, the standard split criterion for a single ST is the log-rank test statistic, which favors splitting variables with many possible split points. Conditional inference trees can avoid this bias by using separate algorithms for selecting the best split-node from that of selecting the best splitting point (Wright et al., 2017). Specifically, the optimal split-node is obtained by testing the association of all the available predictors to the time-to-event outcome using a linear rank test based on the log-rank transformation (log-rank score). Following this, a standard binary split is done for the selected node. CF is an ensemble model with multiple conditional inference trees. We applied CF because there were more polytomous predictors than dichotomous predictors (the number of polytomous and dichotomous predictors was 9 and 3, respectively) in our data set and CF has been shown as superior in predictive performance to RF on time-to-event data sets with polytomous predictors (Nasejje et al., 2017).

6.4.4 Missing data

When using multiple imputation (MI) methods, it is important that there is compatibility between the imputation model and the analysis model (Allison, 2012). Moreover, it is criticised to handle missingness in the time-to-event data using multiple imputation based on standard multivariate normal distribution (Bartlett et al., 2015; White and Royston, 2009). Therefore, we applied different MI methods according to these three prediction modelling approaches. For Cox regression, the SMC-FCS approach for MI (Bartlett et al., 2015) was used to impute missing data. Compared to the traditional FCS MI method (also known as multivariate imputation by chained equations), SMC-FCS can not only specify an appropriate regression method for imputing each predictor X (depending on the type of X, e.g., linear regression for continuous variables, logistic regression for binary variables), but also ensures that each missing value in a partially observed predictor X is imputed from a model (Equation 6.4) that is compatible with the assumed substantive model (Bartlett et al., 2015):

$$f(X_n|Y,X_{-n},Z) \propto f(Y|X,Z)f(X_n|X_{-n},Z) \tag{6.4}$$

where X_n refers to n^{th} predictor with missing data, Y is the outcome and X_{-n} refers to the remaining predictors other than the X_n predictor. The types of regression methods used to impute each predictor when using SMC-FCS in this study can be found in Appendix D Supplement 13. For ST, RF and CF models, RF algorithm was used to impute missing

data because parametric imputation methods (e.g., FCS) may bias our estimation due to model incompatibility. The underlying two-step imputation strategy of RF is as follows. Step 1: for each splitting node, missing data are replaced ("imputed") with values drawn randomly from the non-missing in-bag data; Step 2: after splitting, the imputed data in the children nodes are reset to missing (then proceed as Step 1 until terminal nodes are reached). Therefore, RF imputation is carried out as a tree is being grown and all the missing values are imputed at the end of one iteration. To accommodate variance due to imputation, missing data were imputed for five sets in our study. We then performed modelling on each imputation and the estimates of models prediction performance over the five imputed data sets were combined using Rubin's rules.

6.4.5 Model validation and performance evaluation

Validations were conducted internally (in development data sets) and externally (in test data sets). The discriminative performance of prediction models was evaluated using an overall C-index (Harrell et al., 1982), as well as a C-index which indicates the models' C-index at each point of survival time. For example, for the i^{th} patient, let's say the event time is T_i, censoring time is D_i and the predicted risk score from a model is η_i, and let $\tilde{T}_i = min(T_i, D_i)$ denote the censored time or the latest observed time and $\xi_i = I(T_i < D_i)$ denote event indicator for right censoring. Then the C-index is an estimate of the probability that, in a randomly selected pair of cases (*i, j*), the sequence of events are successfully predicted (Pencina and D'Agostino, 2004):

$$Concordance probability = pr(\eta_i > \eta_j | T_i < T_j) \tag{6.5}$$

and the C-index is defined as Equation (6.6):

$$C - index = \frac{\sum_{i \neq j} I(\eta_i > \eta_j) I(\tilde{T}_i < \tilde{T}_j) \xi_i}{\sum_{i \neq j} I(\tilde{T}_i > \tilde{T}_j) \xi_i} \tag{6.6}$$

A C-index of 0.5 or lower finds the model is predicting an outcome no better than random chance and a higher C-index corresponds to a model with higher prediction accuracy. Additionally, calibration plots, graphs consisting of two types of curves, a 45-degree straight line (reference line, indicating perfect calibration) and irregular curves (calibration curves for each model), were constructed to determine whether the predicted survival probability and observed survival probability were in concordance. In addition to C-index, the integrated Brier score (IBS) (Graf et al., 1999) was also applied to assess models' prediction performance. For binary prediction models, the Brier score is the mean squared prediction error. For models based on time-to-event data, the Brier score for a single subject is defined as: at a given time point t, the squared difference between observed event status (e.g., death) and a model based prediction of survival time t. The IBS represents a

cumulative Brier score over time and is written as Equation (6.7):

$$IBS(t) = \frac{1}{t}\int_0^t E[(I(T > t) - S(tZ))^2]dF_Z(Z) \tag{6.7}$$

where $I(T > t) \in [0,1]$ is the individual survival status at time t and $S(t|Z)$ is the predicted survival probabilities from the model with covariates Z. The IBS is also known as prediction error rate, where a value of 0.5 or higher indicates the model's predictive performance is no better than a chance and a lower IBS corresponds to higher prediction ability.

6.4.6 Study reporting and software

For reporting the findings of this study, we followed the TRIPOD statement (Appendix D Supplement 14). All data were extracted from SEER*Stat 8.3.5, statistical analyses were performed using STATA (Version 15, StataCorp LP, College Station, TX, USA) and R (Version 3.6.4, R Foundation for Statistical Computing, Vienna, Austria). R package "*smcfcs*" and "*randomForestSRC*" were used for missing data imputation. For model development, Cox, ST, RF and CF were fitted using R package "*mlr*" incorporated in packages "*survival*", "*rpart*", "*randomForestSRC*" and "*partykit*" respectively.

6.5 Conclusion

Based on a cohort from the SEER database, various models were used for predicting 3- and 5-year OPCs survival, where RF and CF had a higher and ST had a lower predictive capability than the reference approach (Cox regression). Moreover, a web-based calculator was developed to predict the OPCs survival probability to potentially assist clinical decision-making. Even though no major differences in the predictive performance were seen between the imputation results and the complete case analysis, we recommend using imputation as it allows a check if there was any information loss due to missing observations. Additionally, since we are unaware of the true data-generating mechanism, it is good practice to apply multiple prediction models to check if they all lead to the same answer. This not only increases the confidence in the estimates but also increases the consistency in the estimation.

6.6 Acknowledgments

The authors would like to thank the SEER program for providing open access to the database. The authors thank Janet Grant, a postdoctoral researcher at *BetterStart* research group, The University of Adelaide, for correcting the language.

CHAPTER 7

Application of multilevel machine learning models for predicting pain following root canal treatment

Preface

In Chapter 6, we used a real-world cancer registry data to demonstrate how to develop prediction models for a time-to-event outcome. The analysis in that chapter was carried out assuming a pre-specified data generating mechanism. In this chapter, we use a prospective patient cohort to demonstrate how to develop prediction models for a binary outcome with a multilevel data structure. The outcome we predicted was the pain following a common treatment in dentistry – root canal treatment. We also aim to address different types of bias such as missing data and predictor selection. Moreover, this chapter presents the novel application of precision-recall curves for measuring the performance of prediction models when the distribution of outcome data is imbalanced (the negatives being more than the positives). This is a common phenomenon in real-world health care data. To the best of our knowledge, we are the first to recognise the issue of imbalanced data and build prediction models on such type of data in oral health.

This chapter contains the last of a series of four studies contributing to this thesis. Details for the submitted paper are:

Du M., Haag, D. G., Lynch, J. W., Mittinty, M. N. Application of multilevel machine learning models for predicting pain following root canal treatment. Journal of Dentistry. (Submitted).

The submitted version of the paper is reproduced as follows.

Statement of Authorship

Title of Paper	Application of multilevel machine learning models for predicting the pain following root canal treatment
Publication Status	☐ Published ☐ Accepted for Publication √ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Du M., Haag, D. G., Lynch, J. W., Mittinty, M. N. Application of multilevel machine learning models for predicting the pain following root canal treatment. Journal of Dentistry.

Principal Author

Name of Principal Author (Candidate)	Mi Du		
Contribution to the Paper	MD contributed to conception, design, data preparation, data analysis, results interpretation, drafted and critically revised the manuscript.		
Overall percentage (%)	90%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	03-05-2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dandara G. Haag		
Contribution to the Paper	DGH contributed to conception, design, results interpretation and critically revised the manuscript.		
Signature		Date	3/05/21

Name of Co-Author	John W. Lynch		
Contribution to the Paper	JWL contributed to conception, design, results interpretation and critically revised the manuscript.		
Signature		Date	3/05/21

Name of Co-Author	Murthy N. Mittinty		
Contribution to the Paper	MNM contributed to conception, design, results interpretation, drafted and critically revised the manuscript.		
Signature		Date	3/05/21

Abstract

Aims: This study developed predictive models for one-week acute and six-month persistent pain following root canal treatment. Additionally, we aimed to study the gain in predictive efficacy of models containing clinical factors only, over models containing sociodemographic characteristics.

Methods: We conducted a secondary data analysis of 708 patients who received root canal treatment. Three sets of predictors were used: 1) *combined* set, containing all predictors in the data set; 2) *clinical* set and 3) *sociodemographic* set. Missing data were handled by multiple imputation using the missing indicator method. The multilevel least absolute selection and shrinkage operator (LASSO) regression was used to select predictors into the final multilevel logistic models. Three measures, the area under the receiver operating characteristic curve (AUROC) and precision-recall curve (AUPRC), and calibration curves, were used to assess the predictive performance of the models.

Results: The selected-in factors in the final models, using LASSO regression, are related to pre- and intra-treatment clinical symptoms and pain experience. Predictive performance of the models remained the same with the inclusion (exclusion) of the socio-demographic factors. For predicting one-week outcome, the model built with *combined* set of predictors yielded the highest AUROC and AUPRC of 0.85 and 0.72, followed by the models built with clinical factors (AUROC = 0.82, AUPRC = 0.66). The lowest predictive ability was found in models with only *sociodemographic* characteristics (AUROC = 0.68, AUPRC = 0.40). Similar patterns were observed in predicting six-month outcome, where the AUROC for models with *combined*, *clinical* and *sociodemographic* sets of predictors were 0.85, 0.89, and 0.66, respectively, and the AUPRC were 0.48, 0.53 and 0.22, respectively.

Conclusion: Models containing pre-treatment clinical factors adequately predict the development of one-week and six-month pain following root canal treatment. Adding sociodemographic characteristics to the models with clinical factors did not change the models' predictive performance or the proportion of explained variance.

Clinical significance: This study presents the use of pre- and intra-operative pain experience to predict the post-operative pain following RCT. Clinicians could use this information to better inform patients about pain outcome and possibly use different treatment strategies to manage their patients.

7.1 Introduction

Root canal treatment is an invasive procedure that dentists perform to remove the dead or infected pulp from the pulp chamber and then replace with a rubbery material to prevent bacteria from entering again. Sometimes patients experience pain after a root canal treatment due to various reasons, such as post-procedure inflammation, in-adequate/over-adequate fillings and tooth cracks. Pain following a root canal treatment can be either acute or persistent. The prevalence of acute pain was reported to vary from 5% to 14% (Pak and White, 2011), and persistent pain was estimated to have an occurrence of 3% to 12% (Polycarpou et al., 2005). As known, clinical prediction models usually combine available data on risk factors into a single index, which conveys some information about the likelihood of a certain healthcare outcome (Moons et al., 2009). Pre-operative identification of patients who are at high-risk of pain has the potential to help dentists and improve post-operative care of patients subject to root canal treatment.

The predictors for pain following root canal treatment include patient-level characteristics (e.g., sex, age), tooth-level characteristics (e.g., tooth anatomy, type of root canal), and treatment characteristics (e.g., instrumentation). Using these risk factors, numerous models are being developed to predict the post-operative pain of root canal treatment (Arias et al., 2013; Kayaoglu et al., 2016; Law et al., 2015). However, there are several limitations to current research. The first is that the existing studies have not compared the predictive ability of the predictors from across different domains of variables (e.g., clinical, sociodemographic domains). Moreover, sociodemographic characteristics summarise a combination of variables related to socio-economic position, cultural background, biology, and the impact of racism on health. Using sociodemographic characteristics for prediction in clinical settings may seem to push these variables from social paradigm to biological and legitimise the notions of racial/sexual essentialism (Paulus and Kent, 2017). Another limitation in much of the current literature revolves around the statistical modelling approach. Frequently, the study data used in models are clustered within dental practice or clinician (Bouwmeester et al., 2012). For example, patients treated by one clinician may have some similarities when compared to the patients treated by another clinician, thus making the patients treated by a particular physician clustered (dependent). However, the use of techniques that account for 'clustering' was limited in previous oral health prediction modelling research (Moerbeek et al., 2003; Twisk, 2006).

In this study, we employed a multisite prospective cohort and applied multilevel regression analysis for predicting the post-operative pain following root canal treatment. We compared the predictive performance of models which incorporated different domains of predictors (sociodemographic characteristics, clinical factors and the combination of both).

7.2 Methods

7.2.1 Data source

We used a publicly available data collected by the National Dental Practice-Based Research Network in the US. Ethics approval was not required by the University of Adelaide for secondary analysis of the de-identified open-access data. The study population consisted of 708 patients who received root canal treatment undertaken by 62 dental practitioners across six states in the US. Patient eligibility criteria and details of data collection can be found elsewhere (Law et al., 2014). We used records collected at four time points: before the initiation of treatment (pre-operative), immediately after treatment (intra-operative), one week after tooth obturation, and six months after the initiation of treatment (post-operative).

7.2.2 Predictor variables

Available predictors include: (1) Sociodemographic characteristics of patients and practitioners (e.g., sex, race, ethnicity); (2) Pain characteristics (e.g., pre- and intra-operative measures of pain intensity); (3) Systemic medical characteristics of patients (e.g., diabetes); and (4) Procedural characteristics (e.g., tooth anatomy, use of rubber dam). We organized the available information into three sets of predictors. The first set was a *combined* set, comprising both clinical and sociodemographic predictors (number of predictors = 76), the second set was a *clinical* set, consisting of only the clinical factors (number of predictors = 67, and the third set was a *sociodemographic* set, consisting of only the social and demographic characteristics (number of predictors = 9) from the patients and practitioners.

7.2.3 Outcome variables

There were two outcome variables used in this study. The first outcome was one-week acute pain following root canal treatment, i.e. severe pain (a rating of 7 on a scale of 0-10) occurring in or around the location of a tooth that received root canal treatment within the past week (Walton and Fouad, 1992). The other outcome was six-month persistent pain following root canal treatment, i.e., pain lasts for six months after the initiation of root canal treatment (Von Korff et al., 1992). This outcome was defined by an answer of 1 day(s) for the question '*How many days in the past month have you had pain in the area that was treated with a root canal?*'. In this study, pre- and intra-operative information were used for predicting one-week outcome while the six-month outcome was predicted by the information from pre-, intra- and one-week post-operative.

7.2.4 Handling of missing data

When missingness is informative, use of indicators of missing combined with multiple imputation can minimize bias compared with the complete case analysis (Choi et al., 2019; Fletcher Mercaldo and Blume, 2020; Sperrin and Martin, 2020). In the missing indicator method, we first created the dummy variable (0, 1) that indicated whether or not the value of that variable was missing. Then we incorporated this method with multiple imputation with chained equation, aiming to optimise prediction accuracy by making use of missing pattern information (Sperrin and Martin, 2020). To account for variance due to imputation, analysis was conducted using five imputed datasets. Using Rubin's rules, the estimates of prediction performance were combined over these five imputed datasets.

7.2.5 Model development

In this study, generalized linear multilevel regression analysis incorporated with 10-fold cross validation was used to develop prediction models. Our data have a two-level hierarchical structure with 708 patients at level 1, nested within 62 dental practitioners at level 2 (a schematic structure of our data can be found in Appendix E Supplement 1). A flowchart of statistical analyses can be found in Appendix E Supplement 2. Briefly, a two-level model was used. Level 1 was called fixed-effect level, and it took into account the patient-based characteristics. Level 2 was the random-effect level, and it took into account the practitioner-based characteristics. X_{ij} refers to the level 1 predictor of i^{th} patient of cluster j (e.g., patient's age), β_j is the effect size parameter, which differs for each practitioner, ε_{ij} refers to the random error of prediction for the level 1 equation. So, for the i^{th} patient of cluster j, the level 1 model is written as Equation (7.1):

$$Y_{ij} = \alpha + \beta_j X_{ij} + \varepsilon_{ij} \tag{7.1}$$

W_j refers to the level 2 predictor (e.g., practitioner's sex), γ_1 refers to the regression slope parameter of the level 2 predictor, and u_j is the error term from level 2, therefore $\beta_j = \gamma_0 + \gamma_1 W_j + u_j$ (Equation 2). For each of the outcomes, we fit three multilevel models. The first was the full model (Models 1 & 4, for one-week and six-month outcome, respectively) which contained the combined set of predictors. The second models (Model 2 & 5) contained only clinical factors from patients and practitioners. The third models (Model 3 & 6) included only sociodemographic characteristics.

Variable selection

We used the least absolute selection and shrinkage operator (LASSO) (Tibshirani, 1996) for variable selection. The LASSO minimizes the squared error between the observed and

predicted values, and additionally imposes a penalty if coefficients are not zero (Ahrens et al., 2020). To implement LASSO, we estimate the Equation (7.2) on the available data.

$$\hat{\beta}_{mLASSO}(\lambda) = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \lambda \sum_{x=1}^{m} |\hat{\beta}_x| \tag{7.2}$$

Where Y_i and $\hat{Y}_i$ are the observed and predicted outcome value for the patient i, n is the number of patients, $\hat{\beta}_x$ are the corresponding coefficients for predictor X, m is the number of predictors and λ is the shrinkage penalty parameter or tuning parameter. Given that our data were clustered, we fit multilevel LASSO models (Equation 4) introduced by (Schelldorfer et al., 2011). Besides the penalty function as described in Equation (7.2), the multilevel LASSO uses additional terms to account for the variance components (consists of between cluster variance and within cluster variance) that are parts of the multilevel model in Equation (7.1). Specifically, the multilevel LASSO aims to minimize:

$$\hat{\beta}_{mLASSO}(\lambda) = \frac{1}{2} ln|V| + \frac{1}{2}(Y_i - \widehat{Y}_i)' V^{-1} (Y_i - \widehat{Y}_i) + \lambda \sum_{x=1}^{m} |\hat{\beta}_x| \tag{7.3}$$

where V is the covariance matrix from the multilevel model, and $\hat{\beta}_x$ represents the mixed parameters of the level-1 and level-2 coefficients.

7.2.6 Models' discrimination, calibration and goodness of fit

Models' discrimination

We used two measures for models' discrimination: the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision Recall Curve (AUPRC). AUROC uses two aspects of information, sensitivity (true positives/true positives + false negatives) and 1-specificity (true negatives/true negatives + false positives), from the confusion matrix (A cross tabulation of the observed and predicted outcome for a given threshold). AUROC is then defined as the average sensitivity, regarding all values of the specificity as equally likely (Hand, 2009). However, when the outcome data are imbalanced, estimate of AUROC is severely impacted by the true negatives. To overcome this, alternative measures such as AUPRC have been proposed (Saito and Rehmsmeier, 2015). AUPRC also uses two aspects of information, sensitivity (referred to as recall) and positive predicted values (referred to as precision, true positives/true positives + false positives), from the confusion matrix. AUPRC is then defined as the average positive predictive values, regarding all values of the sensitivity as equally likely. In our study, we are building classifiers to detect pain following root canal treatment, therefore identifying all 'individuals with pain' is more important than predicting the 'non-pain cases'. Thus, true negatives are less of a concern, hence AUPRC is a more meaningful measure.

Model calibration

To measure models' calibration, we created a confidence belt for the calibration curve (Nattino et al., 2016). Compared to the standard calibration curve, the calibration belt can spot the range and direction of deviation from the ideal calibration, and it provides suggestions for revising the model.

Model goodness of fit

Models' goodness of fit was evaluated by Pseudo-R^2, Akaike information criterion (AIC) and Bayesian information criterion (BIC) value. A higher Pseudo-R^2, a lower BIC and/or AIC value indicates a better fit.

7.2.7 Statical software

The programming language R Version 3.6.2 was used for data pre-processing and analysis. Imputation was conducted using '*MICE*' package. LASSO analysis was performed with '*glmmLasso*' package, the '*lme4*' package was used to produce the models. The ROC, PRC and calibration belts were plotted using '*pROC*', '*ROCR*' and '*givitiR*' packages, respectively.

7.2.8 Data availability and study reporting

Data and data dictionary can be freely downloaded: https://www.nationaldentalpbrn.org/study-results/. R codes used for the analysis can be found in Appendix E Supplement 3. We reported our studying following the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (Collins et al., 2015).

7.3 Results

7.3.1 Characteristics of the participants

Table 7.1 reported the pre-operative (baseline) characteristics for patients. In this data set, 707 patients were included in our study who completed tooth obturation. Their mean age was 47.8 years and 408 (57.7%) were females. Of the 707 participants, 411 (58.1%) patients provided one-week follow-up data and 651 (92.1%) provided six-month follow-up data. We only included participants whose outcome data were available. At one-week post-treatment, 100 patients (24.3%) reported a score of 7 or higher for the worst pain following the first week of root canal treatment. At six-month follow-up, persistent pain presented in 70 patients (10.8%).

Table 7.1: Characteristics of the patients and outcome distribution. Characteristics are displayed at patient-, tooth- and practitioner-level.

Patient-related characteristics	
Age in years	
Mean (SD)	47.8 (13.0)
Median [Min, Max]	49.0 [19.0, 70.0]
Missing	15 (2.1%)
Gender	
Male	289 (40.9%)
Female	408 (57.7%)
Missing	10 (1.4%)
Ethnicity	
Hispanic	28 (4.0%)
Not Hispanic	661 (93.5%)
Missing	18 (2.5%)
Race	
White	631 (89.3%)
Black	37 (5.2%)
Asian	13 (1.8%)
Other	15 (2.1%)
Missing	11 (1.6%)
Dental insurance/3rd party coverage	
No	131 (18.5%)
Yes	570 (80.6%)
Missing	6 (0.8%)
Annual income	
<10K	26 (3.7%)
10K-29K	79 (11.2%)
30K-49K	139 (19.7%)
>50K	429 (60.7%)
Missing	34 (4.8%)
Highest level of education	
<High school	14 (2.0%)
High school	115 (16.3%)
Some college	202 (28.6%)
College	253 (35.8%)
Advanced/grad	103 (14.6%)
Missing	20 (2.8%)

Practitioner-related characteristics	
Practitioner gender	
Male	572 (80.9%)
Female	135 (19.1%)
Practitioner specialty	
General dentist	294 (41.6%)
Endodontists	413 (58.4%)
Practitioner ethnicity	
Hispanic	65 (9.2%)
Not Hispanic	596 (84.3%)
Missing	46 (6.5%)
Decade dentist graduated	
1970-79	158 (22.3%)
1980-89	255 (36.1%)
1990-99	161 (22.8%)
2000+	104 (14.7%)
Missing	29 (4.1%)
State of practice	
Alabama (AL)	98 (13.9%)
Florida (FL)	130 (18.4%)
Minnesota (MN)	377 (53.3%)
Saskatchewan (SK)	53 (7.5%)
Oregon (OR)	36 (5.1%)
Wisconsin (WI)	13 (1.8%)
Tooth-related characteristics	
Maxillary tooth	
Yes	416 (58.8%)
No	291 (41.2%)
Tooth site	
Molar	434 (61.4%)
Premolar	197 (27.9%)
Priorior	76 (10.7%)
Outcome distribution	
One-week acute pain (sample size = 411)	
Yes	100 (24.3%)
No	311 (75.7%)
Six-month persistent pain (sample size = 651)	
Yes	70 (10.8%)

No	581 (89.2%)

7.3.2 Selected-in variables by LASSO

According to the LASSO regression, there were 7-16 non-zero coefficient predictors entering the final models (Table 7.2). We found that most of the included clinical factors are related to pre-operative pain, such as the initial pain quality (e.g., 'pain getting worse by stress') and intra-operative pain experience (e.g., 'how numb felt during treatment'). Besides this, a large proportion (6 out of 9) of sociodemographic characteristics were selected in, including the patient's race, dental insurance status, income, practitioner's sex and ethnicity. For predicting six-month outcome, variables of pain experience at one-week follow-up were also identified important by LASSO regression, such as 'present pain rating at one-week follow-up'.

Table 7.2: Models specification and performance comparison

Outcome	Model	Selected-in variables by LASSO[1]	Performance (discrimination and goodness of fit)
One-week pain (1: pain; 0: non-pain)	Model 1 (*Combined* set)	*N*=16 in total Practitioner-related: sex, ethnicity, state of practice Patient-related: patients' education, patient's race, pain gets worse by stress, present time pain rating, # days pain kept from activities past wk[6], pain interfere work past wk, pain interfere social/fam act past wk, bleeding in pulp chamber, not negotiable canal, swelling, intensity of pain during treatment, how numb felt during treatment, nitrous oxide	AUROC[2] = 0.85 [0.81, 0.89] AUPRC[3] = 0.72 AIC[4] = 442.92, BIC[5] = 688.06 Pseudo-R^2 (fixed effects) = 0.95 -R^2 (total effects) = 0.95
	Model 2 (*Clinical* set)	*N*= 7 in total Practitioner-related: state of practice Patient-related: pain gets worse by stress, present time pain rating, # days pain kept from activities past wk, pain interfere work past wk, pain interfere social/fam act past wk, bleeding in pulp chamber	AUROC = 0.82 [0.77, 0.87] AUPRC = 0.66 AIC = 434.98, BIC = 603.77 Pseudo-R^2 (fixed effects) = 0.86 Pseudo-R^2 (total effects) = 0.86

Table 7.2 continued from previous page

Outcome	Model (variable)	Selected-in variables by LASSO	Performance (discrimination and goodness of fit)
	Model 3 (*Sociodemographic* set)	*N*= 9 in total Practitioner-related: sex, ethnicity, graduation year Patient-related: sex, race, ethnicity, income, education, dental insurance/3rd party coverage	AUROC = 0.68 [0.62, 0.74] AUPRC = 0.40 AIC = 471.39, BIC = 547.74 Pseudo-R^2 (fixed effects) = 0.07 Pseudo-R^2 (total effects) = 0.16
Six-month pain (1: pain; 0: non-pain)	Model 4 (*Combined* set)	*N*=9 in total Practitioner-related: graduation year, state of practice Patient-related: patients' race, income, patients' expectation of the treatment outcome, pain gets worse by cold, present time pain rating, days pain kept from activities past wk, location of the greatest probing depth, pain interfering daily activities in the first wk after treatment, pain rating at one-wk follow up	AUROC = 0.85 [0.80, 0.89] AUPRC = 0.48 AIC = 427.21, BIC = 646.66 Pseudo-R^2 (fixed effects) = 0.81 Pseudo-R^2 (total effects) = 0.81

Table 7.2 continued from previous page

Outcome	Model (variable)	Selected-in variables by LASSO	Performance (discrimination and goodness of fit)
	Model 5 (*Clinical* set)	*N*= 12 in total Practitioner-related: practitioners' graduation year, state of practice Patient-related: patients' expectation of the treatment outcome, pain gets worse by cold, present time pain rating, pain intensity past wk, days pain kept from activities past wk, location of the greatest probing depth, pain interfering daily activities in the first wk after treatment, pain interfere work in the first wk after treatment, pain rating at one-wk follow up, intensity of worst pain in the first wk after treatment	AUROC = 0.89 [0.85, 0.92] AUPRC = 0.53 AIC = 444.22, BIC = 757.72 Pseudo-R^2 (fixed effects) = 0.90 Pseudo-R^2 (total) = 0.90
	Model 6 (*Sociodemographic* set)	*N*= 9 in total Practitioner-related: sex, ethnicity, graduation year Patient-related: sex, race, ethnicity, income, education, dental insurance/3rd party coverage	AUROC = 0.66 [0.59, 0.72] AUPRC = 0.22 AIC = 461.23, BIC = 537.37 Pseudo-R^2 (fixed effects) = 0.06 Pseudo-R^2 (total) = 0.07

[1]LASSO: Least Absolute Shrinkage and Selection Operator. [2]AUROC: Area Under Receiver Operating Characteristics curve. [3]AUPRC: Area Under the Precision-Recall Curve. [4]AIC: Akaike information criterion. [5]BIC: Bayesian information criterion. [6]wk: week. All generalized linear mixed models were built with R *glmm()* function. *Combined* set contains sociodemographic characteristics + clinical factors. *Clinical* set contains only clinical factors and *sociodemographic* set contains only sociodemographic characteristics.

7.3.3 Models specification and performance measures

Model performance measures are specified in Table 7.2, Figures 7.1 and 7.2, and Appendix E Supplement 4. In terms of discrimination, models yield a wide range of AUROC from 0.66-0.89. Figure 7.1 presents a visual discriminative ability of all models used. We found that models with sociodemographic set of predictors underperformed others, and minor difference were found between the combined set and clinical set alone. For example, for predicting one-week outcome, the AUROC for Model 1-combined, Model 2-clinical, and Model 3-sociodemographic are 0.85 [95% CI 0.81, 0.89], 0.82 [0.77, 0.87] and 0.68 [0.62, 0.74], respectively. Similar pattern was found in six-month outcome prediction (AUROC for Model 4-*combined*, Model 5-*clinical* and Model 6-*sociodemographic* are 0.85 [0.80, 0.89], 0.89 [0.85, 0.92] and 0.66 [0.59, 0.72], respectively). In terms of AUPRC, the models with combined and clinical sets of predictors yielded similar AUPRC, while the models with only sociodemographic characteristics showed the lowest classification ability. As shown in Figure 7.1, the AUPRC for Model 1 to 6 are 0.72, 0.66, 0.40, 0.48, 0.53 and 0.22, respectively. Results of other measures such as Sensitivity, Specificity can be found in Appendix E Supplement 4.

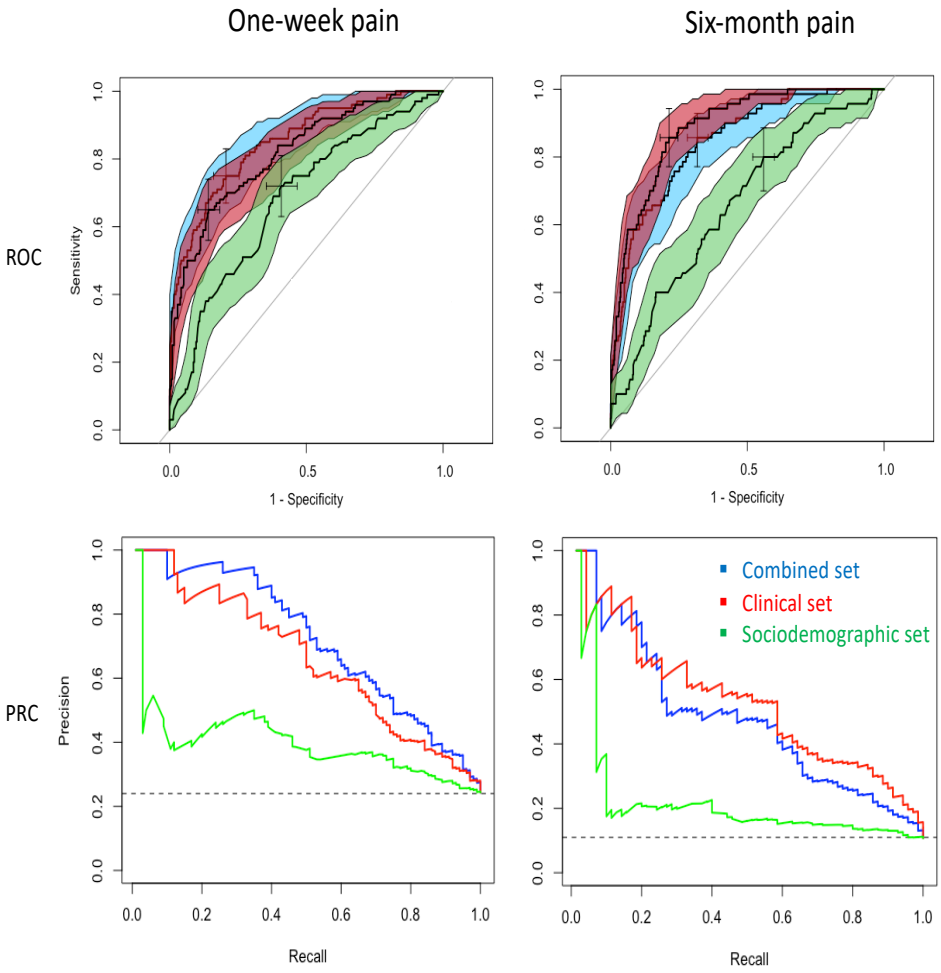


Figure 7.1: Models' area under the receiver operating characteristic curve (AUROC) and precision-recall curve (AUPRC) for predicting one-week and six-month post-operative pain in patients undergoing root canal treatment.

Curves in different colours represent models with different sets of predictors. Blue: *combined* set, containing sociodemographic characteristics + clinical factors. Red: *clinical* set, containing only clinical factors. Green: *sociodemographic* set, containing only sociodemographic characteristics. In ROC, the grey diagonal line represents a random classifier with an AUROC of 0.5. In PRC, the grey dashed line represents the baseline. Due to the different distribution of positives ('1') and negatives ('0') for the one-week (positives: negatives = 76% : 24%) and six-month outcome (positives: negatives = 89% : 11%), the random classifiers have AUPRC of 0.24 and 0.11, respectively. Therefore, all of the developed models have better AUPRC than random classifiers.

In terms of models' calibration, as shown in Figure 7.2, models built on combined set and clinical set of predictors calibrated well, whereas models built on sociodemographic characteristics did not calibrate, meaning the models may under/over-estimate the risk of developing pain, and even give extreme (too close to 0 and 1) risk estimates.

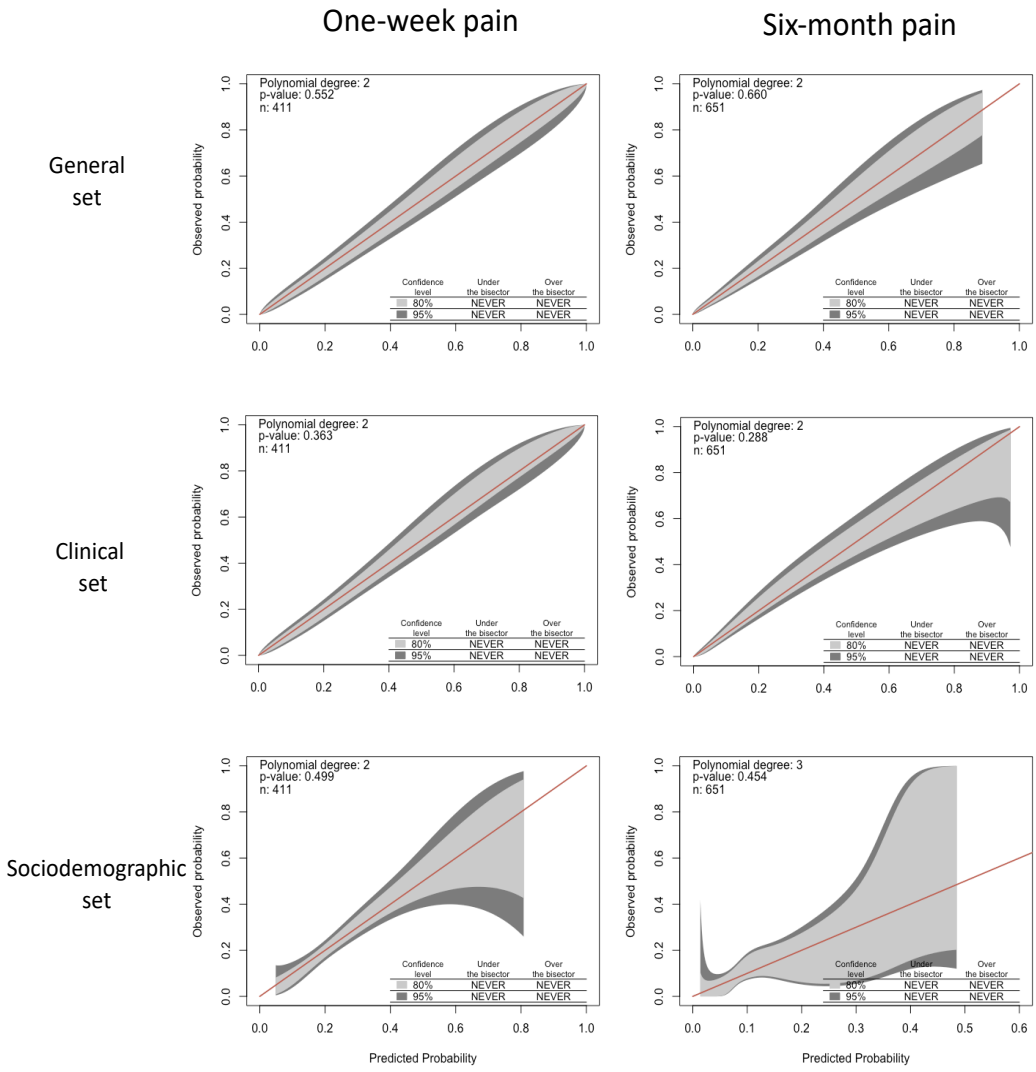


Figure 7.2: Models' calibration belts.

In terms of goodness of fit, there were three main findings. First, the Pseudo-R^2 for models including clinical set of predictors (Model 1, 2, 4 and 5) were 1.5 times higher than the models with sociodemographic set of predictors (Model 3 and 6). Second, similar proportions of the explained variance were observed among the models containing clinical factors, suggesting that minor differences in the included clinical factors does not change the models' goodness of fit. The third finding was that, when models were developed with the involvement of clinical factors, Pseudo-R^2 for fixed effects were equal to the Pseudo-R^2 for total effects, suggesting that majority of the explained variance was from level 1 (patient-level) characteristics, rather than the level 2 (practitioner-level) characteristics. This can be explained by the variance due to fixed and random effects in Appendix E Supplement 5.

Finally, for completeness of reporting following TRIPOD, the estimated regression coefficients and odds ratios are reported in Appendix E Supplement 6. TRIPOD checklist can be found in Appendix E Supplement 7.

7.4 Discussion

Our first major finding was that among all the available predictors, the intensity of pre-operative pain and the experience of intra-operative pain were important to the subsequent development of post-operative tooth pain regardless of the investigated time period. This is consistent with the previous literature in predicting post-operative pain in other health areas, such as breast surgery, limb amputation and thoracotomy (Kehlet et al., 2006). This finding also supports the importance of clinical factors for healthcare risk assessment.

Our second finding was that majority (6 out of 9) of the sociodemographic characteristics were identified important to predict the outcome. This leads to our further considerations on how to deal with variables from different domains in clinical prediction: when we mix the predictors across domains, we integrate modifiable clinical variables and unmodifiable sociodemographic variables that we may not intervene on. For example, in our data, the variable "patient's race" was one of the most relevant predictors for post-operative pain. However, it is known that race-based differences in health risk may be, in part, due to unmeasured differences between black and white patients that are related to the differential lived experiences of blacks versus whites in the US. Given the final goal of risk prediction is to assist on making intervention (i.e., changing a risk factor to a protective factor), it might make little sense to include these unmodifiable variables in clinical prediction models.

Our third finding was that in the presence of clinical factors, the sociodemographic characteristics contributed less to either improving the models' predictive performance or explaining the variance in the model. Though there is a lack of consensus in how to deal with sociodemographic variables in clinical prediction (Paulus and Kent, 2017), we argue that with the inclusion of sociodemographic characteristics (e.g., race/ethnicity), clinical prediction can create algorithmic bias, that is enhancing the inequality of healthcare attainment. Other research has argued that different races experience different levels of pain tolerance (Edwards et al., 2001; Wandner et al., 2012). When a patient is more tolerant to pain, he/she might be underscored and predicted as 'low' risk of developing pain, thus, less medical care would be allocated to him/her. As a result, individuals predicted to be at lower risk of pain because of their race/ethnicity may not be beneficiaries of enhanced pain care. In contrast, when a racial group is less tolerant to pain, this population are more likely to be predicted as 'high' risk and might be overtreated in order to prevent the development of post-operative pain.

Strengths

The first strength relates to the methodology used in this study: 1) We applied multilevel LASSO regressions to identify the importance of clinical and sociodemographic predictors in clinical prediction. 2) Use of multilevel regression models that take the hierarchies of data structure into consideration might increase both the amount of predictive information and models' overall predictive ability (Gelman, 2006). 3) Use of AUPRC along with AUROC to evaluate the models' performance increases our confidence in the findings. We here take six-month models as an example to interpret AUROC and AUPRC. As shown in Table 7.2, the proportion of positive cases in six-month cohort is 24%, meaning the baseline AUPRC for six-month models is 0.24, then achieving an AUPRC of 0.48 and 0.53 in Model 4 and 5 (higher than the baseline) is 'good' while an AUPRC of 0.22 in Model 6 (lower than the baseline) is 'bad'. For this reason, our findings can be interpreted as: No matter whether we look at all participants or we focus on the positive class only, the models with clinical set of predictors perform similar to models with combined set of predictors while models with sociodemographic set of predictors yield the lowest classification ability, even worse than not using the model. Second, the presented models are intended to have direct clinical relevance, as the predictor variables are easy to obtain from daily dental practices, making the models easy to validate. The third strength is that we are trying to develop models that rectify, rather than amplify sexist and racial discrimination by leaving out sociodemographic factors and maintaining comparable predictive ability. This attempt draws attention on the potential racial/sexual/ethnic bias in clinical prediction when involving sociodemographic variables into model development, particularly in oral health field. Hispanics living in the US represent an increasing diversity of national-origin groups with a relative lack of detailed epidemiological data on the incidence and prevalence of common and important diseases (Escarce et al., 2006). When 'Hispanic' identity was missing, the imputation could be challenging as 'Hispanic' identity is multidimensional and multifaceted and will align with multiracial identity (Parker et al., 2015). Therefore, the results need to be interpreted cautiously.

Limitations and future research

First, our models were developed and internally validated based on a small cohort, these models should be validated in other populations and tested in clinical settings before using in clinical practice. Second, there are a variety of biases which we did touch upon in our prediction, such as measurement error. We assumed there was no measurement error in the data which might not be the case. Third, when the original data has a two(multi)-level structure, the imputation model can be either hierarchical or non-hierarchical. The ideal way for imputation is to account for variance from both level-1 and level-2 using hierarchical imputation methods, such as the method described in the '*miceadds*' package

(Robitzsch et al., 2017). However, the hierarchical imputation was not conducted in this study for two reasons: i) the practitioner-level variance contributes less to the total variance; and ii) the '*miceadds*' packages does not support for imputing nominal variables. Fourth, in this study, the AUROC was used as the loss function to select best λ as in equation 7.2 and 7.3, however, since the data set has an imbalanced outcome distribution, the AUPRC may be a better replacement to AUROC. However, creating a new statistical method that incorporates AUPRC with LASSO extends beyond the study, therefore, further research is needed to address these methodological issues. Finally, more discussion is needed among a variety of stakeholders, in terms of developing a consensus about the best practice of coping with sociodemographic characteristics in clinical prediction modelling research.

7.5 Conclusion

Multiple multilevel models were built to predict the one-week acute pain and six-month persistent pain following root canal treatment. The severity and experience of pre-operative and intra-operative pain were discovered important to the subsequent development of post-operative pain. This provides information for managing pain expectations for both the patients and practitioners performing root canal treatment. Our results indicate that clinical factors play more important role in clinical prediction than sociodemographic characteristics. Using clinical factors over sociodemographic characteristics may reduce the sexual/racial biases when predicting risk.

7.6 Acknowledgments

The authors thank the National Dental Practice-Based Research Network for providing open access to the dataset. MD conceptualized the study, extracted data, performed analysis, interpreted results, drafted the manuscript and critically revised the manuscript. DGH, JWL and MNM conceptualized the study, interpreted results, drafted the manuscript and critically revised the manuscript. All authors gave their final approval and agree to be accountable for all aspects of the work. All authors declare no conflict of interest.

CHAPTER 8

General discussion and conclusion

Preface

Previous chapters have discussed the usefulness, strengths, and limitations pertaining to each individual study. This chapter brings together the evidence obtained throughout this PhD project and provides an overall discussion and conclusion. Section 8.1 presents an executive summary and key findings from this project. Drawing from these findings, further considerations are presented in Section 8.2 regarding a number of issues in prediction modelling research, including (factual) prediction *v.* counterfactual prediction (causal inference), individual-level *v.* population-level risk prediction, and the challenges of applying machine learning to oral health data. Sections 8.3 and 8.4 highlight the strengths and limitations of the design and the conduct of this project. Section 8.5 outlines the implications for future research and clinical practice. Finally, Section 8.6 provides an overall summarising conclusion.

8.1 Executive summary and key findings from the project

In recent years, eminent journals such as *Nature*, *The Lancet* and *The BMJ* have published a number of articles calling for global action to improve the quality and reproducibility of research in science, including the health and medical sciences (Baker, 2016; Begley and Ioannidis, 2015; Beran et al., 2019; Grol et al., 2002; McNutt, 2014; Munafò, 2019). In the field of prediction modelling, poor study quality and a lack of study reproducibility also constitute major concerns (Hayden et al., 2006, 2013; Moons et al., 2014). In this context, I commenced my PhD and structured the thesis with two systematic reviews and two empirical analyses, revolving around the research quality and reproducibility of oral health prediction models. The four studies included in this thesis address various aspects of the identification, quality assessment, development, internal validation and performance evaluation of clinical prediction models in various areas of oral health.

- The first two studies (systematic reviews in Chapters 4 and 5) were conducted to achieve two aims:
  1) to identify the prediction models in the major fields of oral health;
  2) to assess study quality in terms of addressing bias and reporting completeness.

  From these two studies, it was found that, the major oral health outcomes predicted were periodontal diseases, tooth loss, dental caries, and oral cancers. Our studies have shown that the majority of the existing studies do not incorporate various sources of potential bias, such as measurement error or the effective management of missing data. Moreover, opaque reporting and lack of reproducibility were also identified in the existing oral health prediction modelling studies. For example, it was found that 70% of the studies did not report the models' calibration and 47% of the studies did not report the methods for handling missing data. These findings provided motivations to conduct two further studies (Chapters 6 and 7) with the aim of demonstrating adherence to the recent guidelines in prediction modelling studies.

- Following the recommendations offered by **PROBAST**, the study in Chapter 6 was conducted to predict a time-to-event outcome - the survival of patients with oral and pharyngeal cancers. Using data from the SEER program, three tree-based machine learning methods (survival tree, random survival forest and conditional inference forest) and a traditional statistical model (Cox regression) were used to predict the three-year and five-year disease-specific survival of patients with oral cancers. Additionally, this study demonstrated how missing data can be handled in both machine learning and the traditional statistical models, using methods such as the substantive model compatible version of the fully conditional specification

imputation approach. Finally, C-index and IBS (Integrated Brier Score) were used to evaluate the models' predictive performance.

From this study, we observed comparable predictive capabilities between Cox regression and the non-parametric tree-based machine learning algorithms. For example, the C-index for a Cox model and a random survival forest in predicting three-year survival were 0.82 and 0.84, respectively. A novel application of this study was the development of an online calculator using ShinyApp software in R to initiate the open estimation of the survival probability for up to five years for patients. This calculator has clinical translational potential and may be useful in patient stratification and treatment planning, perhaps initially in the research context.

- Chapter 6 provides a simplistic view of data by considering a pre-specified set of predictors. However, in many real-world data settings, some data complexities need to be addressed prior to modelling. These complexities include the nesting of information and the selection of predictors. Chapter 7 demonstrates how to accommodate these issues using an example of developing clinical prediction models for the pain following root canal treatment. Data used in this study were from a multisite prospective patient cohort who underwent root canal treatment. This analysis comprises 76 predictors in the analysis, including the patients' sociodemographic descriptors (e.g., age, sex) and clinical symptoms (e.g., pre-treatment pain experience). The data set was hierarchically structured, with 708 patients being nested within 62 dental practitioners. Multiple imputation with missing indicator methods were applied to handle missing data. To fit the hierarchical structure of the data, multilevel logistic regression was used for model development. To reduce the number of predictors in the final models, the LASSO regression was used. In addition, since the data had an imbalanced outcome distribution (the negatives being more than the positives), measures such as the AUPRC were implemented to evaluate how well the positives are predicted by the models.

  From this study, we found that the pre-treatment clinical factors were identified as being important to the development of postoperative pain following root canal treatment, and that demographic characteristics did not add much to the models' predictive performance. Among all the developed multilevel logistic regression models, the models with a clinical set of predictors yielded similar predictive performance to models with the combined (clinical and sociodemographic) set of predictors, while the models with sociodemographic predictors alone showed the weakest predictive ability. For example, for the prediction of one-week postoperative pain, the AUROC for models with clinical, sociodemographic, and combined set of predictors were 0.82, 0.68 and 0,84 respectively, and the AUPRC were 0.66, 0.40 and 0.72, respec-

tively. AUROC and AUPRC convey different information and hence results from both these analyses are presented.

- Additionally, the transparent reporting and reproducibility of the empirical studies were achieved by following the **TRIPOD** checklist and sharing all the data and codes used via GitHub repositories (Chapters 6 and 7).

8.2 Issues to consider for future prediction modelling research

Though our findings show a promising future for the application of clinical prediction models in oral health, multiple issues remain to be considered.

8.2.1 Prediction *v.* causation and intervention

Discriminating between prediction and causation

Prediction relies on the correlation between the predictors and outcomes, however, the correlation between these two variables does not guarantee causation (Obermeyer and Emanuel, 2016). It is therefore critical to understand the difference between causation and correlation. In Chapter 2, the differences between explanatory models and prediction models were discussed. Their use in two different types of research can be described as:

- **Counterfactual prediction** (causal inference) research requires **causation** between the predictor(s) and the outcome(s), and relies on causal **explanatory models**. This is useful if the effect of interventions is the focus of the investigation.
- **Factual prediction** requires **correlation** between the predictor(s) and the outcome(s), and uses **prediction models**.

The nature of (factual) prediction is to link the predictors to the outcomes regardless of the data generating (causal) mechanism that produces the outcome in the real world. Therefore, (factual) prediction may contain both causal and non-causal factors that predict the outcome. In doing so, the focus of (factual) prediction lies in the identification of the data generating mechanism while using little or no knowledge of the conceptual model underlying it. Counterfactual prediction relies on posing a pre-specified data generating (causal) mechanism to predict the potential outcomes as if the world had been different to that which has been observed (Dickerman and Hernán, 2020; Hernán et al., 2019).

This distinction between the two types of prediction is perhaps best explained by using an example. To predict the five-year risk of tooth loss (due to periodontal disease), counterfactual prediction is used to answer questions such as: '*Given the baseline characteristics of the patients, what would be the five-year risk of having tooth loss if all individuals received periodontal therapy v. no individual received periodontal therapy?*' It

is the 'what if' approach that drives the counterfactual prediction and it measures how the prediction of a particular outcome changes when the value for one or more predictors is changed (Hernán and Robins, 2016; Wachter et al., 2017). In contrast, if the question is not a 'what if' question (e.g., 'What is the five-year risk of tooth loss among the individuals receiving the periodontal therapy?'), then a factual prediction is formulated.

The need for a shift from prediction to causation and intervention

While prediction allows for the estimation of the probability of an event, it cannot provide specific guidance on which interventions are appropriate to reduce the patients' risk. This is because we are often unaware of what variable(s) affect/lead to the predicted outcome. The nature of the research questions answered in this thesis is: '*What is the risk of having an adverse outcome for a patient at a specific time given the available data at that time?*' The subsequent logical step is to answer: '*Does a certain risk factor* ***cause*** *the outcome? Would the adverse outcome be changed if an intervention targeting the risk factor was implemented?*' To answer such questions, there must be a shift away from prediction modelling (factual prediction) towards causal inference (counterfactual prediction or the potential outcomes approaches). This shift in focus may assist in understanding the impacts of the prediction tools on clinical decision-making, as well as evaluate evaluating the effectiveness of interventions in improving patients' health outcomes. However, causal estimation has its own limitations, the major being the unmeasured confounding.

8.2.2 Population-level risk is not individual-level risk

As discussed in Chapter 2, definitive statements such as: '*My risk of having tooth loss is 10% in the next 10 years*' are problematic. Such statements are limited to forms similar to: '*If we consider a sample of 100 people exactly like 'myself', on average, 10 of them are predicted to have tooth loss in 10 years*.' This is because if a factor, X, is reported to increase the risk of developing a disease in a population, it does not mean that an individual, personally, will have a higher probability of developing the disease from exposure to factor X.

In 2003 the Human Genome Project was completed, following which an increasing number of researchers argue that 'individualised risk prediction' may be a possibility based on the information contained within -omics (Chen and Snyder, 2013). However, in a similar fashion to the commonly used predictors in health research such as sociodemographic characteristics, environmental factors and lifestyles, -omics data are unable to make predictions regarding an individual patient's likelihood of developing a disease or how a specific patient will respond to any given treatment. Because people may have either a protective or predisposing genotype with respect to a disease, there remains a genetically based risk difference to disease development between these two subgroups. Although it is

possible to identify the individuals with a genotype predisposition to disease development by mapping their entire genome, this information itself is not sufficient to identify specific individuals who will develop the disease. Achieving risk prediction at the level of the individual, therefore, remains challenging, even when in possession of -omics data. Additionally, -omics data are not routinely recorded in daily practice, meaning their inclusion as predictive variables in routine risk modelling approaches becomes problematic.

8.2.3 Racial and gender-based bias due to risk prediction

As discussed in Chapter 7, there has been a growing concern that predictive algorithms may reproduce racial and gender disparities via the data used to train them (Barocas and Selbst, 2016; Chouldechova and Roth, 2018). The 'high-impact' journal *Science* has called for specific attention regarding the racism in risk prediction tools (Benjamin, 2019). Empirical examples also support these concerns. One typical example of racial bias due to risk prediction can be found in the study by (Obermeyer and Emanuel, 2016). In that study, in order to inform the allocation of medical resources, the authors developed models aimed at predicting the health risk among millions of people. One of the predictors used was 'health need', represented by the 'cost on health care'. However, there was a problem with using 'cost' as a proxy of health need, because Black people tend to spend less on health care due to their relatively 'lower' economic status. But in fact, Black people with the same 'cost on health care' as White people tend to be much sicker and need much more health care. When the algorithm used 'cost' to predict the health risk, it falsely concluded that Black people are much healthier than White people. As a result, health care providers tend to allocate fewer resources to Black people, leading to inequality in health care provision.

In this example, bias was introduced because the same sick Black person may not expend an equivalent sum of money on health care as a result of a lower income. To overcome this bias, the findings from Chapter 7 (Publication 4) provide a potential solution. This may be achieved by removing the use of sociodemographic characteristics in model development when large clinical information is available.

8.2.4 Common challenges in applying machine learning for prediction analysis in oral health

Recent advances in machine learning present an exciting opportunity to improve oral health care, with potential applications ranging across different disciplines in oral health. However, challenges remain in applying machine learning algorithms to the analysis of oral health data.

Hierarchical structure of data in oral health research

Differing from other medical data, data in oral health are characterised by their multilevel structure. For example, for the assessment of periodontal parameters such as pocket depth,

data are evaluated at a contextual level more specific than the affected tooth. As shown in Figure 8.1, at least four, hierarchical, levels of lesion site, affected tooth, individual, and community characteristics (e.g., dental practitioner) must be accounted for. To handle such

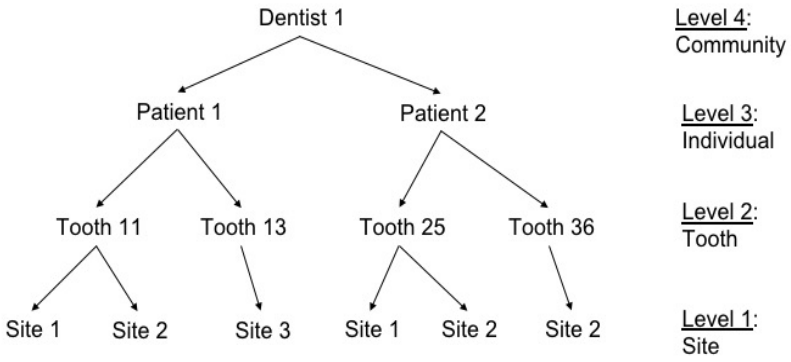


Figure 8.1: Hierarchical structure of data in oral health research

type of data, multilevel models were first applied in 1991 for assessing the craniofacial growth curves in orthodontics (Beek et al., 1991). Similarly, in Chapter 7, the data set used has a two-level structure with patients being nested within the overarching practitioner taxon. Therefore, a two-level logistic regression model was adopted. However, the use of multilevel machine learning models is still limited in predictive oral health research. Further developments in this field of research are crucial to enable the analysis of hierarchical oral health data using machine learning.

Data with an imbalanced outcome distribution

As shown in Chapter 7, the distribution of the real-world oral health outcome data is usually imbalanced (there exist both minority and majority categories for positive and negative observations). Usually, the population at high risk of adverse outcomes tend to be in the minority category. A recent study has found that when the outcome distribution is imbalanced, misclassification of the binary outcome is more likely to occur in the minority than in the majority category (Banerjee et al., 2018). Therefore, for this type of data, evaluating the models' ability to detect the minority class is more important than predicting both minority and majority categories. To address this, in Chapter 7, the imbalance of the outcome was accounted for by applying an appropriate performance measure - the AUPRC (Saito and Rehmsmeier, 2015). Studies have also shown that the imbalance of outcome distribution presents a barrier to the generalisability of results from a particular algorithm and a particular population to other populations (Kim, 2007; Zhang et al., 2010). Therefore, further investigation regarding machine learning is required to effectively analyse data with imbalanced outcome distribution.

Interpretability and transparency of machine learning models

In clinical settings, machine learning algorithms raise concerns due to a lack of transparency (Rudin and Radin, 2019). More specifically, if the connection between the model's input (the predictors) and output (the predicted outcome) is not knowable (hidden inside a 'black box'), then the question '*How can the models be trusted?*' naturally follows (Baselli et al., 2020; Miller, 2019; Zihni et al., 2020). In statistical models (e.g., linear and logistic regression models), interpretation of the statistical functions and how variables are interrelated to construct the final model are readily appreciated. However, when questions such as '*Is the model interpretable at every step, or with regards to its sub-components*?' as asked, some machine learning models provide no answer because they do not decompose the analytical process into steps that can be agreed upon (Ahmad et al., 2018). Improving process interpretability, therefore, presents an important challenge for machine learning models.

In summary, the application of machine learning in oral health research faces various challenges. Some of them cannot be addressed solely through the evolution of algorithms, and there is a need for effective interdisciplinary collaboration between experts in various fields. In doing so, it is possible to ask important questions, concerning appropriate study design, standardised data collection, and suitable data analysis approaches.

8.3 Strengths and contributions

The primary strength of this thesis is that to the best of the our knowledge, this is the first PhD project that comprehensively assesses the quality of oral health prediction modelling studies since the development of **PROBAST** and **TRIPOD** instruments. Following this, Chapter 5 provides recommendations to address multiple sources of bias that might arise during prediction modelling. More importantly, different techniques have been adopted to demonstrate, in Chapters 6 and 7, how some of these biases can be incorporated into prediction modelling.

The second strength of this thesis lies in the use of the selected data sets. There are two major advantages for using the data sets from SEER and DPBRN: 1) The two cohorts were recruited from large and diverse populations of dental practitioners and patients. This allows for rapid accrual of participants. 2) Both cohorts represent multiple geographic and ethnic context within the US. This allows for the potential generalisability of the results to a broader target population of interest.

The third strength of this thesis relates to its strict adherence to the most contemporary methodological standards outlined by the **PROBAST** group for clinical prediction

modelling, as well as the **TRIPOD** reporting guidelines. The methodology used in this thesis reflects the efforts to account for missing data (see examples in Chapters 6 and 7), unmeasured predictors (see an example in Chapter 6), and selection of predictors into the final models (see an example in Chapter 7), in both the machine learning as well as the statistical models.

The fourth strength of this thesis is that the discussion on the common issues of prediction modelling and machine learning extends beyond the field of oral health research. These considerations can be generalised to other disciplines in broader health and medical science.

Finally, from a clinical dentistry and oral health perspective, the models developed in this thesis hold the potential to assist in real-life dental practice. For example, an attempt has been made to translate the products of this research into clinical use by developing an online calculator for informing estimates of the risk of death due to oral and pharyngeal cancers.

8.4 Limitations

The limitations of the four studies conducted as parts of this project are discussed in each individual chapter. The following three points briefly discuss the overall limitations of the conduct of this PhD project.

- In the empirical analysis (Chapters 6 and 7), each prediction models was limited to the specific population upon which it was developed or validated. As a result, external validation utilising a different time, population, and setting may demonstrate a decline in model performance. Further validation and possibly re-calibration to other populations is important. It should also be assured that the prediction models have appropriate influences on clinical decision-making, which usually requires an impact analysis by means of randomised controlled trials. Neither of these requirements has been fulfilled by this thesis and must be addressed in future studies.

- Some of the empirical analyses were limited by the pre-specified statistical packages in R. For example, in Chapter 7, when the LASSO was conducted for variable selection using the R package '*glmmLASSO*', AUROC was adopted as a loss function. Since the data set has an imbalanced outcome distribution, the AUPRC may be a better replacement to AUROC. However, creating a new statistical method that incorporates AUPRC with LASSO extends beyond the thesis, for this reason, we limited our analysis to the available statistical software, and further research is needed to address this methodological issue.

- A further concern regarding the empirical prediction analysis is the quality of the data (e.g., the existence of measurement errors). Some of the data used in this thesis were collected using questionnaires; thus, the measurement of predictors may vary between investigators and studies. However, this thesis did not incorporate methods to address bias due to measurement error. Nevertheless, and as noted in the individual chapters, the study findings should be interpreted with caution due to the presence of measurement errors in self-reported and clinical record data sets.

- Measuring Hispanic identify has proved challenging for research, as the Hispanic individuals living in the US are cultural- and ethnical-diverse populations. The imputation of such a fuzzy concept can be spurious, this is in part because grouping together diverse racial backgrounds and nationalities into a single group masks the difference among them (Song, 2009). Therefore, we need more geographical, historical, epidemiological, and socio-political understanding of this group before aggregating them as Hispanic. Alternatively, using the family's country of origin (such as "Mexican", "Cuban", "Dominican") over pan-ethnic terms can be an option to describe their identity (Taylor et al., 2012).

8.5 Implications and future directions

In the context of these study limitations, there remain a number of implications for future research.

- At the time when we were conducting the systematic reviews, we only had access to PRISMA 2009 guidelines. It is for this reason in this thesis we followed PRISMA 2009 guidelines. However, in March 2021, the PRISMA 2020 statement was published (Page et al., 2021). For those who are conducting systematic reviews are advised to follow the up-to-date guidelines for transparent and complete reporting, thus facilitating evidence-based decision making.

- **PROBAST** could be updated to account for sources of bias such as measurement error. Additionally, some of the **PROBAST** items should be considered and revised where necessary. For example, **PROBAST** item 4.4 suggests that multiple imputation is the only valid method to handle missing data regardless of the nature and the extent of the missing data. However, this may not always be the case. Recent evidence in the published literature provides practical guidelines regarding the effective management and reporting of missing data in observational studies (Harrell Jr, 2015; Hughes et al., 2019; Lee et al., 2021).

- Future research can validate, update, and test the clinical impacts of the models developed in this thesis. For example, based on the prognostic models of oral and

pharyngeal cancers in Chapter 6, further research could be conducted to explore new prognostic factors that may improve the accuracy of survival prediction. Moreover, risk prediction strategies, in combination with medically relevant images (e.g., radiology images and histopathology slides) and 'advanced' learning algorithms (e.g., deep learning), can also be explored to provide information on the diagnosis and prognosis of oral diseases.

8.6 Concluding remarks

The significance of this research project is twofold. First, this research develops and presents prediction models that may be of clinical use for various oral conditions. Second, this project represents an attempt to standardise the conduct of this type of study in oral health research. The main conclusions from this thesis are:

- Various clinical prediction models were identified in the major fields of oral health, and their quality was examined using the latest guidelines. This research has found that 'high' quality prediction models remain scarce in oral health. Future prediction modelling studies should follow the appropriate methodological standards, including: i) pre-specification of predictors using clinical expertise and the best available knowledge in the literature, ii) consideration of missing data management and variable selection, iii) use of appropriate models to manage the data complexity (e.g., competing risk and multilevel data structures), iv) consideration of the internal validity and external validity of the developed models, and v) improvement in the completeness of study reporting.

- Chapter 5 provides suggestions for incorporating multiple sources of bias in prediction modelling. These suggestions were applied in Chapters 6 and 7 to develop reliable prediction tools for the diagnosis and prognosis of oral conditions. These tools hold the potential to benefit clinical practice. However, multi-centre studies are required to further refine these models and to confirm their predictive performance and reliability. For future prediction modelling research, it is important that data analysts work towards minimising bias, regardless of the areas where the models are employed.

- Machine learning algorithms provide alternatives to statistical models for oral health prediction purposes and have been successfully applied through the empirical analyses of Chapters 6 and 7. The use of machine learning may assist in the incorporation of data complexities such as hierarchical data structures and the selection of predictors from a large number of variables.

- Clinical prediction modelling research has the potential to benefit health care activities. However, studies should be conducted with cautions. For example, when using clinical information data sets of great size and complexity (e.g., data derived from electronic health records) for model development, consideration could be made to avoid the inclusion of sociodemographic characteristics in modelling, aiming to exclude unhelpful extraneous data, thereby reducing the risks of potential gender-based and racial bias in healthcare outcome prediction.

Appendices to Chapter 3

Last Name: DU
SEER ID: 15617-Nov2017
Request Type: Internet Access

SURVEILLANCE, EPIDEMIOLOGY, AND END RESULTS PROGRAM
Data-Use Agreement for the SEER 1973-2015 Research Data File

It is of utmost importance to protect the identities of cancer patients. Every effort has been made to exclude identifying information on individual patients from the computer files. Certain demographic information - such as sex, race, etc. - has been included for research purposes. All research results must be presented or published in a manner that ensures that no individual can be identified. In addition, there must be no attempt either to identify individuals from any computer file or to link with a computer file containing patient identifiers.

In order for the Surveillance, Epidemiology, and End Results Program to provide access to its Research Data File to you, it is necessary that you agree to the following provisions.

1. I will not use - or permit others to use - the data in any way other than for statistical reporting and analysis for research purposes. I must notify the SEER Program if I discover that there has been any other use of the data.
2. I will not present or publish data in which an individual patient can be identified. I will not publish any information on an individual patient, including any information generated on an individual case by the case listing session of SEER*Stat. In addition, I will avoid publication of statistics for very small groups.
3. I will not attempt either to link - or permit others to link - the data with individual level records in another database.
4. I will not attempt to learn the identity of any patient whose cancer data is contained in the supplied file(s).
5. If I inadvertently discover the identity of any patient, then (a) I will make no use of this knowledge, (b) I will notify the SEER Program of the incident, and (c) I will inform no one else of the discovered identity.
6. I will not either release - or permit others to release - the data - in full or in part - to any person except with the written approval of the SEER Program. In particular, all members of a research team who have access to the data must sign this data-use agreement.
7. I will use appropriate safeguards to prevent use or disclosure of the information other than as provided for by this data-use agreement. If accessing the data from a centralized location on a time sharing computer system or LAN with SEER*Stat or another statistical package, I will not share my logon name or password with any other individuals. I will also not allow any other individuals to use my computer account after I have logged on with my logon name and password.
8. For all software provided by the SEER Program, I will not copy it, distribute it, reverse engineer it, profit from its sale or use, or incorporate it in any other software system.
9. I will cite the source of information in all publications. The appropriate citation is associated with the data file used. (Please see either Suggested Citations on the SEER*Stat Help menu or the Readme.txt associated with the ASCII text version of the SEER data.)

My signature indicates that I agree to comply with the above stated provisions.

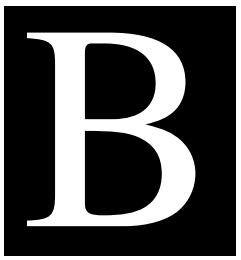
Signature

Date

Please print, sign, and date the agreement. Send the form to The SEER Program:

- **By fax to 301-680-9571**
- **Or, e-mail a scanned form to seerfax@imsweb.com**

Last Name: DU | SEER ID: 15617-Nov2017 | Request Type: Internet Access



Appendices to Chapter 4

Chapter 4 Supplement 1. Search strategy

Database	Search strategy
MEDLINE via PubMed	((((("periodontitis/epidemiology"[mh:noexp] OR "periodontitis"[tiab] OR "chronic periodontitis/epidemiology"[mh:noexp] OR "gingivitis"[mh:noexp] OR "tooth loss"[mh:noexp] OR "tooth loss"[tiab] OR "tooth mobility"[mh:noexp] OR "tooth mobility" OR "clinical attachment loss"[tiab] OR "alveolar bone loss"[tiab] OR "probing depth"[tiab])) AND ("forecasting"[mh] OR forecast*[tiab] OR predict*[tiab] OR "incidence"[mh] OR "incidence"[tiab] OR "progression"[tiab] OR "disease progression/epidemiology"[mh] OR "risk assessment/epidemiology"[mh] OR "risk assessment"[tiab] OR "risk factors"[mh] OR "risk factors"[tiab])) AND ("Adult"[mh] OR adult*[tiab] OR "population"[mh] OR "population"[tiab] OR "Adolescent"[mh] OR "adolescent*"[tiab])) AND ("cohort studies"[mh] OR "cohort"[tiab] OR "longitudinal studies"[mh] OR "longitudinal"[tiab] OR "prospective studies"[mh] OR "prospective"[tiab] OR "follow-up studies"[mh] OR "follow up"[tiab] OR "retrospective studies"[mh] OR "retrospective"[tiab])
Embase	#1 AND #2 AND #3 AND #4 #1: periodontitis/de OR "chronic periodontitis"/de OR gingivitis/exp OR "tooth loss":ti,ab OR "tooth mobility":ti,ab #2: "prediction and forecasting"/de OR prediction/de OR "predictive value"/de OR forecasting/de OR Incidence/de OR incidence:ti,ab OR progression:ti,ab OR "risk assessment"/de OR risk:ti,ab OR "risk factor"/de #3: Adult/de OR aged/exp OR adult*:ti,ab OR population/de OR "aged people":ti,ab OR "aged person":ti,ab OR elderly:ti,ab #4: "longitudinal study"/de OR longitudinal:ti,ab OR "prospective study"/de OR prospective:ti,ab OR "retrospective study"/de OR retrospective:ti,ab OR "follow up":ti,ab
DOSS (dentistry and oral sciences source)	(DE periodontitis OR TI periodontitis OR AB periodontitis OR DE gingivitis OR TI gingivitis OR AB gingivitis OR TI "tooth loss" OR AB "tooth loss" OR DE "tooth loss" OR DE "tooth mobility" OR TI "tooth mobility" OR AB "tooth mobility") AND (DE forecasting OR TI forecast* OR AB forecast* OR TI predict* OR AB predict* OR DE "disease incidence" OR TI incidence OR AB incidence OR TI progression OR AB progression OR DE "disease progression" OR DE "risk assessment" OR TI risk OR AB risk OR DE "disease risk factors") AND (DE Adults OR TI adult* AB adult* OR TI population OR AB population) AND (DE "cohort analysis" OR TI cohort OR AB cohort OR DE "longitudinal method" OR TI longitudinal OR AB longitudinal OR DE "prospective studies" OR TI prospective OR AB prospective OR DE "retrospective studies" OR TI retrospective OR AB retrospective OR "follow-up studies (Medicine)")
Scopus	((TITLE-ABS-KEY (periodontitis OR {chronic periodontitis} OR gingivitis OR {tooth loss} OR {tooth mobility}) AND TITLE-ABS-KEY (predict* OR forecast* OR incidence OR progression OR {risk assessment} OR {risk factor}) AND TITLE-ABS-KEY ({cohort study} OR {cohort studies} OR {longitudinal study} OR {longitudinal studies} OR {prospective study} OR {prospective studies} OR {retrospective study} OR {retrospective studies} OR {follow up} OR {follow-up}))) AND ((TITLE (adult OR population) OR ABS (adult OR population)))
Web of Science	TS= (periodontitis OR "chronic periodontitis" OR gingivitis OR "tooth loss" OR "tooth mobility") *AND* TS= (Predict* OR forecast* OR Incidence OR progression OR "risk assessment" OR "risk factor") *AND* TS= (Adult OR population) *AND* TS= ("cohort study" OR "cohort studies" OR "longitudinal study" OR "longitudinal studies" OR "prospective study" OR "prospective studies" OR "retrospective study" OR "retrospective studies" OR "follow up" OR "follow-up")
ProQuest	all(periodontitis OR "chronic periodontitis" OR gingivitis OR "tooth loss" OR "tooth mobility") AND all(Predict* OR forecast* OR Incidence OR progression OR "risk assessment" OR "risk factor") AND all(Adult OR population) AND all("cohort study" OR "cohort studies" OR "longitudinal study" OR "longitudinal studies" OR "prospective study" OR "prospective studies" OR "retrospective study" OR "retrospective studies" OR "follow up" OR "follow-up")

Chapter 4 Supplement 2. CHARMS checklist

Domain	Key items	Reported on page #
Source of data	Source of data (e.g., cohort, case-control, randomized trial participants, or registry data)	
Participants	Participant eligibility and recruitment method (e.g., consecutive participants, location, number of centers, setting, inclusion and exclusion criteria)	
	Participant description	
	Details of treatments received, if relevant	
	Study dates	
Outcome(s) to be predicted	Definition and method for measurement of outcome	
	Was the same outcome definition (and method for measurement) used in all patients?	
	Type of outcome (e.g., single or combined endpoints)	
	Was the outcome assessed without knowledge of the candidate predictors (i.e., blinded)?	
	Were candidate predictors part of the outcome (e.g., in panel or consensus diagnosis)?	
	Time of outcome occurrence or summary of duration of follow-up	
Candidate predictors (or index tests)	Number and type of predictors (e.g., demographics, patient history, physical examination, additional testing, disease characteristics)	
	Definition and method for measurement of candidate predictors	
	Timing of predictor measurement (e.g., at patient presentation, at diagnosis, at treatment initiation)	
	Were predictors assessed blinded for outcome, and for each other (if relevant)?	
	Handling of predictors in the modelling (e.g., continuous, linear, non-linear transformations or	
Sample size	Number of participants and number of outcomes/events	
	Number of outcomes/events in relation to the number of candidate predictors (Events Per Variable)	
Missing data	Number of participants with any missing value (include predictors and outcomes)	
	Number of participants with missing data for each predictor	
	Handling of missing data (e.g., complete-case analysis, imputation, or other methods)	
Model development	Modelling method (e.g., logistic, survival, neural network, or machine learning techniques)	
	Modelling assumptions satisfied	
	Method for selection of predictors **for inclusion** in multivariable modelling (e.g., all candidate predictors, pre-selection based on unadjusted association with the outcome)	
	Method for selection of predictors **during multivariable modelling** (e.g., full model approach, backward or forward selection) and criteria used (e.g., p-value, Akaike Information Criterion)	
	Shrinkage of predictor weights or regression coefficients (e.g., no shrinkage, uniform shrinkage, penalized estimation)	
Model performance	Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination (C-statistic, D-statistic, log-rank) measures with confidence intervals	
	Classification measures (e.g., sensitivity, specificity, predictive values, net reclassification improvement) and whether a-priori cut points were used	
Model Evaluation	Method used for testing model performance: development dataset only (random split of data, resampling methods e.g. bootstrap or cross-validation, none) or separate external validation (e.g. temporal, geographical, different setting, different investigators)	
	In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibrated, predictor effects adjusted, or new predictors added)	
Results	Final and other multivariable models (e.g., basic, extended, simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, model performance measures (with standard errors or confidence intervals)	
	Any alternative presentation of the final prediction models, e.g., sum score, nomogram, score chart, predictions for specific risk subgroups with performance	
	Comparison of the distribution of predictors (including missing data) for development and validation datasets	
Interpretation and discussion	Interpretation of presented models (confirmatory, i.e., model useful for practice versus exploratory, i.e., more research needed)	
	Comparison with other studies, discussion of generalizability, strengths and limitations.	

Source: Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, et al. (2014) Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med 11(10): e1001744. doi:10.1371/journal.pmed.1001744. assessment scale (Cohort study).

Chapter 4 Supplement 3. Criteria for scoring of risk of bias using the CHARMS checklist

Potential bias	Items to be considered for potential bias
Participant selection	Low risk of bias if - selection bias was unlikely, - study avoided inappropriate inclusions or exclusions, - in- and exclusion criteria were adequately described - participants were enrolled at a similar presentation of their disease - differences were accounted for by including appropriate predictors in the analysis
	Moderate risk of bias sample if - not satisfying one of the above or - no adequate description of recruitment of study sample - no adequate description of the sample for key predictors
	High risk of bias if both items were not adequately described
Predictor assessment	Low risk of bias if - predictor definitions were the same for all participants, - predictor measurement was blinded to outcome data - all predictors were available at the time the model is intended to be used - predictors were measured with valid and reproducible methods such that misclassification was limited and if - predictors were assessed in a similar way for all study participants
	Moderate risk of bias sample if one of the criteria was not satisfied
	High risk of bias if predictor assessment was not adequately described
Outcome assessment	Low risk of bias if outcome was pre-specified, measured with sufficient validity and reproducibility, measured in a similar way for all study participants and if the outcome was assessed independent from assessment of predictors. Note: for easy to obtain predictors such as gender, it is not possible to assess outcome independent of predictor information
	Moderate risk of bias if method for xxx
	High risk of bias if method for assessment of outcome was not adequately described
Attrition	Low risk of bias if - there was no loss-to-follow-up - there were no important differences on key characteristics between included participants and those who were lost-to-follow-up or missing
	Moderate risk of bias if - loss-to-follow-up was lower than 20% and there were no important differences on key characteristics between included participants and those who were lost-to-follow-up or missing OR: - loss-to-follow-up was higher than 20% but missing data and loss-to-follow-up were imputed adequately or there were no important differences on key characteristics between included participants and those who were lost-to-follow-up or missing
	High risk of bias if - loss-to-follow-up was higher than 20% and/or - there were important differences on key characteristics between included participants and those who were lost-to-

	follow-up or missing -loss-to-follow-up was not described
Analysis (time interval between predictor and outcome was reasonable, part of eligibility)	Low risk of bias if - relevant aspects of analysis were described allowing to judge the quality of the analysis to be adequate - # outcome events per candidate predictor reasonable - missing data handled appropriately or no differences - predictors included independent of p-value - overfitting and optimism accounted for - weights assigned according to regression coefficient - calibration and discrimination assessed - recalibrated or described that it was not needed
	Moderate risk of bias if: - relevant aspects of analysis were described allowing to judge the quality of the analysis to be adequate and part or none of the model evaluation items were reported
	High risk of bias if - not satisfying any of the aspects under low risk of bias

Source: Smit HA, Pinart M, Antó JM, et al. Childhood asthma prediction models: a systematic review. Lancet Respir Med 2015;3(12):973-84. doi: 10.1016/S2213-2600(15)00428-2.

Chapter 4 Supplement 4. Newcastle- Ottawa quality assessment scale (Cohort study)

Note: A study can be awarded a maximum of one star for each numbered item within the Selection and Outcome categories. A maximum of two stars can be given for Comparability

Selection

1) Representativeness of the exposed cohort
   a) truly representative of the average _____ (describe) in the community ✵
   b) somewhat representative of the average _____ in the community ✵
   c) selected group of users eg nurses, volunteers
   d) no description of the derivation of the cohort
2) Selection of the non exposed cohort
   a) drawn from the same community as the exposed cohort ✵
   b) drawn from a different source
   c) no description of the derivation of the non exposed cohort
3) Ascertainment of exposure
   a) secure record (eg surgical records) ✵
   b) structured interview ✵
   c) written self report
   d) no description
4) Demonstration that outcome of interest was not present at start of study
   a) yes ✵
   b) no

Comparability

1) Comparability of cohorts on the basis of the design or analysis
   a) study controls for _____ (select the most important factor) ✵
   b) study controls for any additional factor ✵ (This criteria could be modified to indicate specific control for a second important factor.)

Outcome

1) Assessment of outcome
   a) independent blind assessment ✵
   b) record linkage ✵
   c) self report
   d) no description
2) Was follow-up long enough for outcomes to occur
   a) yes (select an adequate follow up period for outcome of interest) ✵
   b) no
3) Adequacy of follow up of cohorts
   a) complete follow up - all subjects accounted for ✵
   b) subjects lost to follow up unlikely to introduce bias - small number lost - > ____ % (select an adequate %) follow up, or description provided of those lost) ✵
   c) follow up rate < ____% (select an adequate %) and no description of those lost
   d) no statement

Chapter 4 Supplement 5. Full texts exclusion reasons

Study	Reasons
Chen 2010	Older adults with physical disabilities, cognitive impairment or mental disorders included.
Chen 2011	Introduced a multidimensional risk assessment tool for longitudinal tooth loss.
Costa 2011	Aimed at the influence of compliance in the progression of periodontitis and tooth loss.
Costa 2012	Validation study. Aimed at investigating the association of PRA model with the recurrence of periodontitis and tooth loss.
Eichnolz 2008	Validation study. The PRA score at the start of SPT was significantly associated with tooth loss.
Genco 1998	Aimed at the association between stress and periodontal diseases, and presenting models explaining the mechanisms.
Gillbert 2007	Prevalence prediction. Aimed at the relationship between self-reported status and outcome.
Gilthorpe 2001	Aimed at introducing multilevel model and providing a new analytical way to longitudinal periodontal research.
Gilthorpe 2003	Presented a comprehensive multilevel model that describes the underlying progression of periodontal disease.
Gregg 2007	Use baseline self-reported information predict longitudinal prevalence.
Guarnizao 2014	Cross-sectional study.
Heaton 2017	Used self-reported information predict prevalence.
Hyun 2014	Screening model.
Jansson 2002	Aimed at evaluating the influence of potential risk predictors/risk factors on the longitudinal marginal bone loss and tooth loss. No prediction model mentioned.
Jansson 2008	Validation study. Aimed at evaluating PRA in periodontitis patients during SPT.
LaMonte 2014	Completion of self-reported periodontal disease does not align with either the baseline or the follow-up.
Leininger 2010	Updating study. Modified PRA to PRAS, and investigated the association between baseline periodontal risk assessment diagram surface (PRAS) and the outcomes.
Lindskog 2010 (Inflammatory)	The predictive value for using inflammatory test predict periodontitis. The association between inflammatory tests with periodontitis outcome and some risk factors.
Martin 2010	Validation study. Baseline disease score and risk score were significantly associated with tooth loss.
Martin 2009	Validation study. Baseline disease score and risk score were significantly associated with tooth loss.
Matuliene 2010	Validation study. Association between PRA risk level and periodontitis recurrence.
Mcleod 1998	Aimed at evaluating the predictability of given prognosis on tooth loss. Retrospectively test the predictability of treatment prognosis by tooth loss measurement.
Mdala 2014	Targeting on site-level. Use clinical attachment loss or pocket depth classify the progression sites. Compare two models.
Nieri 2002	Indicate the predictors on patients, tooth, and sites level, the association between predictors and outcomes. There were some formulas, but no model predictive performance.
Page 2002	Duplicate report. Developed a risk calculator, and then found the association between baseline score and outcome.

Page 2003	Development and Validation study. Proposed PRC. Aimed at the association between baseline score and outcome.
Page 2007	Updated PRC score. Adding disease severity score.
Peres 2012	Aimed at testing the accuracy of three partial protocols in estimating the prevalence of periodontal outcomes.
Reddy 2000	Targeting on site level. Use baseline clinical measurement such as plaque, gingival inflammation, attachment loss, and probing depth predict the regression, measured by digital subtraction radiography. No clustering of risk factors.
Renvert 2014	Cross-sectional study. Use patient-based data (smoking habit, bleeding on probing, plaque score, and pocket depth) predict alveolar bone loss.
Stoykova 2014	The association between each variables and outcomes.
Teles 2016	Does not construct a model to predict disease. Applied linear mixed models on longitudinal data. Targeting on site-level instead of patient level. Used sex, age, baseline CAL to classify progression sites.
Teles 2017	Description of periodontitis progression pattern.
Tu 2004	Aimed at the application of multilevel model in periodontal research.
Zhan 2014	Population prevalence prediction.

Chapter 4 Supplement 6. Variables included in prediction models

Study		Sociodemographic Factors	Systemic conditions	Dental care behaviours	Oral examination
Author /Year	model				
Beck 1994	risk model (Blacks)	Self-feeling of mouth appearing worse;	Take drugs resulting in soft tissue reactions due to heat disease, diabetes, ulcers and anxiety;	Regularly use mouthwash or rinse;	BANA of 3 or 4; Baseline P.g. + at 2+%; P.g. + 18 months.
	risk model (Whites)	Age; Perceived lack self-care; Don't see friends and relatives	Take drugs resulting in abnormal homeostasis;	visit dentist episodically;	P.g. +
	Predicti on model	Age; Perceived lack self-care;		visit dentist episodically;	less than 12 teeth
Leite 2017	Model 1	Family income at birth; Sex; Number of people living in the house;	Smoking; Diastolic blood pressure		Proportion of teeth with pocket, bleeding or calculus; Number of anterior teeth lost; Number of remaining teeth; Number of DMFT;
	Model 2	Sex; Family income at birth;	BMI		Number of posterior teeth lost; Proportion of teeth with pocket, bleeding or calculus; Number of remaining teeth; Number of DMFT
	Model 3	Sex; Education; Family income at birth;	BMI	Frequency of tooth brushing	Number of posterior teeth lost; Proportion of teeth with pocket, bleeding or calculus; Number of sound teeth; Number of anterior teeth lost;
	Model 4	Sex; Family income at birth;	Diastolic blood pressure;		Proportion of teeth with calculus; Number of DMFT
Lindsko g 2010	DRS dentitio n & DRS tooth	Age; Family history of chronic periodontitis,	Smoking; Systemic disease* and related diagnoses; Result of skin provocation test to assess the patient's inflammatory reactivity; nutritional deficiency; obesity; alcohol abuse; stress-related factors	Patient cooperation and disease awareness; The therapist's experience with periodontal care;	Bacterial plaque (oral hygiene); Endodontic pathology; FI; Angular bony destruction; Radiographic marginal bone loss; PPD; BOP; Marginal dental restorations; Increased tooth mobility; Missing teeth; Abutment teeth;
Martine z-Canut 2018	Molars & Non-molars	Age	Smoking;	Bruxism	Severe periodontitis; baseline number of teeth; type of tooth; FI; PPD; bone loss; mobility; C/R ratio
Morelli 2018	Index of periodo ntal classes (IPC)	Age; Sex; Race;	Smoking; Diabetes;		University of North Carolina-Periodontal profile class (UNC-PPC): IAL, direct attachment level, interproximal PPD, direct PPD, interproximal gingival recession, direct GR, BOP, GI, PI, decayed coronal surface, filled coronal surface, decayed root surface, filled root surface, presence/absence of full prosthetic crowns.

* including diabetes, immunopathies and hematologic disorders, hereditary disorders relevant to formation and maintenance of connective tissue and bone, granulomatous disease, osteoporosis, renal disorders, inflammatory vascular disease, Sjo¨gren syndrome, and rheumatism.

Chapter 4 Supplement 7. Critical appraisal of the 5 selected prediction modelling studies based on the CHARMS

Study (Author/Year)	Risk of bias				
	Participant selection	Predictor	Outcome	Sample flow/attrition	Analysis
Beck 1994	L	L	L	H	M
Leite 2017	M	L	L	H	M
Lindskog 2010	L	L	L	H	M
Martinez-Canut 2018	M	L	L	M	M
Morelli 2018	L	L	L	H	M

CHARMS checklist Criteria are listed in the appendix. CHARMS=Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies. L=low risk of bias. M=medium risk of bias. H=high risk of bias.

Chapter 4 Supplement 8. Critical appraisal of the 5 selected prediction modelling studies based on the NOS

Study (Author/ Year)	Selection				Comparability	Outcome			Thresholds
	Representativeness of the exposed cohort	Selection of the non-exposed cohort	Ascertainment of exposure	Demonstration that outcome was not present at the start of study	Based on design and analysis	Assessment of outcome	Follow-up long enough for outcomes to occur	Adequacy of follow-up cohort	(Good, Fair, Poor)
Beck 1994	★	★	★★	★	-	★★	★	-	**Poor**
Leite 2017	★	★	★★	★	★	★★	★	-	**Good**
Lindskog 2010	★	★	★	★	-	★★	★	★	**Poor**
Martinez-Canut 2018	★	★	★	★	-	★	★	-	**Poor**
Morelli 2018	★	★	★	-	-	★	★	-	**Good**

NOS: Newcastle- Ottawa quality

Chapter 4 Supplement 9. TRIPOD checklist for 5 selected studies

Section/Topic		Page (in selected journals)				
		Beck 1994	Leight 2017	Lindskog 2010	Martinez-Canut 2018	Morelli 2018
Title and abstract						
Title	1	468	731	584	46	148
Abstract	2	468	731	584	46	148
Introduction						
Background and objectives	3a	471	NP	NP	47	NP
	3b	471	732	585	47	149
Methods						
Source of data	4a	472	732-733	585	47	149
	4b	472	733	585	NP	149
Participants	5a	472	733	585	47	149
	5b	NP	NP	585	NP	NP
	5c	n/a	n/a	n/a	47	n/a
Outcome	6a	474	733, S. Table 2	586-587	50	150
	6b	NP	NP	NP	NP	NP
Predictors	7a	473	S. table 1	586	47-48	150
	7b	NP	NP	NP	NP	NP
Sample size	8	472	733	585	47	149
Missing data	9	472	NP	587	NP	NP
Statistical analysis methods	10a	473-474	733	n/a	NP	NP
	10b	475,477	733	n/a	48	NP
	10c	n/a	n/a	585	n/a	n/a
	10d	474	736	587	48,50	150
	10e	n/a	n/a	NP	n/a	n/a
Risk groups	11	473	NP	NP	NP	NP
Development vs. validation	12	n/a	n/a	NP	50	n/a
Results						
Participants	13a	472	737	NP	47	NP
	13b	474	734-735,737	587	NP	150
	13c	n/a	n/a	NP	NP	153
Model development	14a	474	NP	n/a	48	NP
	14b	NP	S. Table 5-11	589,591	49	151
Model specification	15a	NP	739	n/a	48	151

	15b	476-477	737,739	n/a	48	151-152
Model performance	16	476-477	738	587-590	50	152-153
Model updating	17	n/a	n/a	NP	n/a	n/a
Discussion						
Limitations	18	478	741	NP	52-53	155
Interpretations	19a	n/a	n/a	591	NP	n/a
	19b	478	742	593	53	155
Implications	20	NP	741	592	53-54	NP
Other information						
Supplementary information	21	n/a	733-740	n/a	55	156
Funding	22	478	742	592	54	155

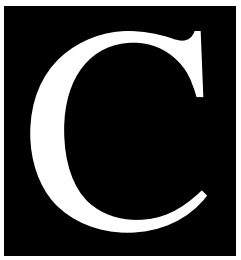
NP: Not provided; n/a: not applicable; S.: supplement information

Chapter 4 Supplement 10. Criteria for TRIPOD checklist (Checklist of Items to Include When Reporting a Study Developing or Validating a Multivariable Prediction Model for Diagnosis or Prognosis)

Section/Topic	Item	Development or Validation?	Checklist Item
Title and abstract			
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.
Introduction			
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model, or both.
Methods			
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation datasets, if applicable.
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.
	5b	D;V	Describe eligibility criteria for participants.
	5c	D;V	Give details of treatments received, if relevant.
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	D;V	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured.
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	D;V	Explain how the study size was arrived at.
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10c	V	For validation, describe how the predictions were calculated.
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.
Risk groups	11	D;V	Provide details on how risk groups were created, if done.
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.
Results			
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.

	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors, and outcome).
Model development	14a	D	Specify the number of participants and outcome events in each analysis.
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	D	Explain how to use the prediction model.
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.
Model updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).
Discussion			
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).
Interpretations	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.
Other information			
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and datasets.
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.

Source: Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55-63.



Appendices to Chapter 5

Chapter 5 Supplement 1. Journals identification approaches

Database	Rankings
Thomson Reuters	Incites journal citation report, Browse by category, Journals by rank, Choose 'DENTISTRY, ORAL SURGERY & MEDICINE' or 'PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH' https://jcr.incites.thomsonreuters.com/JCRJournalHomeAction.action
Scopus	Scopus, Source, 'Subject category', Choose 'Epidemiology' or 'Dentistry', Sort by 'Cite Score' https://www.scopus.com/sources.uri?DGCID=Scopus_blog_post_check2015
Google Scholar	Google Scholar, Metrics, Categories='Health & medicine sciences', Subcategories='Epidemiology' or 'Dentistry' https://scholar.google.com.au/citations?view_op=top_venues&hl=en&vq=med

Chapter 5 Supplement 2. List of searched journals

General dentistry (n=4)	Clinical dentistry (n=7)	Oral health and epidemiology (n=3)	General epidemiology (n=8)	Bio-statistical journals (n=7)
Journal of dental research	Oral oncology	Journal of public health dentistry	International Journal of Epidemiology	Biometrical Journal
Journal of dentistry	Journal of clinical periodontology	Clinical oral investigations	Journal of Clinical Epidemiology	Biostatistics
European journal of oral sciences	Journal of periodontology	Community dentistry and oral epidemiology	Epidemiology	Biometrics
International journal of oral science	Journal of endodontics		European Journal of Epidemiology	Biometrika
	Clinical oral implants research		American Journal of Epidemiology	Statistical methods in medical research
	Journal of prosthodontic research		Clinical epidemiology	Statistics in medicine
	Caries research		Annals of epidemiology	The International Journal of Biostatistics
			Cancer epidemiology, biomarkers & prevention	

Note: In three databases mentioned in Appendix table 1, biostatistics was not independent from general statistics field, thus we included seven statistical journals related to (bio)medicine and epidemiology by asking advice from an experienced biostatistician (MM).

Chapter 5 Supplement 3. Search strategy

Database	Search strategy
#1 MEDLINE via PubMed (on 17th January, 2019)	("International journal of epidemiology"[Journal] OR "European journal of epidemiology"[Journal] OR "American journal of epidemiology"[Journal] OR "Epidemiology"[Journal] OR "Journal of clinical epidemiology"[Journal] OR "Clinical epidemiology"[Journal] OR "Annals of epidemiology"[Journal] OR "Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology"[Journal] OR "International journal of oral science"[Journal] OR "Journal of public health dentistry"[Journal] OR "European journal of oral sciences"[Journal] OR "Community dentistry and oral epidemiology"[Journal] OR "clinical oral investigations"[Journal] OR "Journal of endodontics"[Journal] OR "Clinical oral implants research"[Journal] OR "Journal of prosthodontic research"[Journal] OR "Journal of dentistry"[Journal] OR "Journal of clinical periodontology"[Journal] OR "Journal of periodontology"[Journal] OR "Journal of dental research"[Journal] OR "Oral oncology"[Journal] OR "Caries research"[Journal]) AND (“forecasting”[mh] OR forecast*[tiab] OR predict*[tiab] OR diagnos*[tiab] OR prognos*[tiab]) AND ("2016/01/01"[Date - Publication] : "2018/12/31"[Date - Publication])
#2 MEDLINE via PubMed (on 1st April, 2019)	("Biometrical journal. Biometrische Zeitschrift"[Journal] OR "Biostatistics (Oxford, England)"[Journal] OR "Biometrics"[Journal] OR "Biometrika"[Journal] OR "Statistics in medicine"[Journal] OR "Statistical methods in medical research"[Journal] OR "The international journal of biostatistics"[Journal]) AND (count[all] OR dental[tiab] OR dentist*[tiab] OR dentistry[mh] OR oral[All] OR mouth[All] OR tooth[All] OR teeth[All]) AND (“forecasting”[mh] OR forecast*[tiab] OR predict*[tiab] OR diagnos*[tiab] OR prognos*[tiab]) AND ("2016/01/01"[Date - Publication] : "2018/12/31"[Date - Publication])

Note: search strategy for this systematic review consists of #1 and #2. In order to search prediction models in statistical journals, we have to restrict key words to dental data/count data/dental outcome, thus search strategy is different from other journals.

Chapter 5 Supplement 4. Answers for each signalling question in PROBAST for 34 studies

PROBAST = Prediction model Risk Of Bias ASsessment Tool.

(Web page)

https://docs.google.com/spreadsheets/d/e/2PACX-1vQxVqW3SfOfReQL_jlKEay5R96hH975f4IIrL9RsiIVPzWo8o0WL6VoglkTOg87lIEK1RpjicIkQuYf/pubhtml

Chapter 5 Supplement 5. Reasons for being answered "N/PN (No/Probably No)" for signalling question in PROBAST for 34 studies

PROBAST = Prediction model Risk Of Bias ASsessment Tool.

(Web page)

https://docs.google.com/spreadsheets/d/e/2PACX-1vQxVqW3SfOfReQL_jlKEay5R96hH975f4IIrL9RsiIVPzWo8o0WL6VoglkTOg87lIEK1RpjicIkQuYf/pubhtml#

Chapter 5 Supplement 6. Criteria for TRIPOD checklist (Checklist of Items to Include When Reporting a Study Developing or Validating a Multivariable Prediction Model for Diagnosis or Prognosis)

Section/Topic	Item	Development or Validation?	Checklist Item
Title and abstract			
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.
Introduction			
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model, or both.
Methods			
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation datasets, if applicable.
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.
	5b	D;V	Describe eligibility criteria for participants.
	5c	D;V	Give details of treatments received, if relevant.
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	D;V	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured.
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	D;V	Explain how the study size was arrived at.
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10c	V	For validation, describe how the predictions were calculated.
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.
Risk groups	11	D;V	Provide details on how risk groups were created, if done.
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.
Results			
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.

	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors, and outcome).
Model development	14a	D	Specify the number of participants and outcome events in each analysis.
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	D	Explain how to use the prediction model.
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.
Model updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).
Discussion			
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).
Interpretations	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.
Other information			
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and datasets.
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.

Source: Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55-63.

Chapter 5 Supplement 7. Signalling questions for PROBAST

1. Participants	2. Predictors	3. Outcome	4. Analysis
Signalling questions			
1.1. Were appropriate data sources used, e.g., cohort, RCT, or nested case–control study data?	2.1. Were predictors defined and assessed in a similar way for all participants?	3.1. Was the outcome determined appropriately?	4.1. Were there a reasonable number of participants with the outcome?
1.2. Were all inclusions and exclusions of participants appropriate?	2.2. Were predictor assessments made without knowledge of outcome data?	3.2. Was a prespecified or standard outcome definition used?	4.2. Were continuous and categorical predictors handled appropriately?
	2.3. Are all predictors available at the time the model is intended to be used?	3.3. Were predictors excluded from the outcome definition?	4.3. Were all enrolled participants included in the analysis?
		3.4. Was the outcome defined and determined in a similar way for all participants?	4.4. Were participants with missing data handled appropriately?
		3.5. Was the outcome determined without knowledge of predictor information?	4.5. Was selection of predictors based on univariable analysis avoided?†
		3.6. Was the time interval between predictor assessment and outcome determination appropriate?	4.6. Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?
			4.7. Were relevant model performance measures evaluated appropriately?
			4.8. Were model overfitting, underfitting, and optimism in model performance accounted for?†
			4.9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?†

PROBAST = Prediction model Risk Of Bias ASsessment Tool. † Development studies only.

Source: Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS et al. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 170(1):51-58.

Chapter 5 Supplement 8: Guidance notes for rating risk of bias using PROBAST

https://www.probast.org/translations/

Source: 1) Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS et al. (2019). PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med 170(1):W1-W33.

2) Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS et al. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 170(1):51-58.

Chapter 5 Supplement 9. Full texts exclusion reasons

Teles et al, 2016	No multivariate model was constructed. Applied linear mixed models on longitudinal data. Targeting on site-level instead of patient level. Used sex, age, baseline CAL to classify progression sites.
Wostmann et al, 2017	Created an assessment tool (chewing efficiency test) for dental treatment needs, targeting on non-dental staff.
Nasrin et al, 2017	Proposed a new staging system and then applied to an existing dataset.
Silveira et al, 2016	No model established, aimed to identify the association between risk factors and dental outcomes.
Jäger et al, 2016	Outcome was demanding of dentists and dental device.
Abreu-Placeres et al, 2018	Outcome was caries management for dentists.
Baelum et al, 2016	Different classification tool were compared in terms of if the same predictors were identified when different outcomes were used. Models were built based on different definition of outcome measurements.
Castilho et al, 2016	Models were on tooth level. Model were built to detecting occlusal caries on permanent molars.
Joda et al, 2017	Validate a diagnostic tool to classify fixed implant restorations by assigning 44 dentists applying the tool on 10 cases, and then compare the consistency.
Håkansson et al, 2017	Not distinguish head and neck cancer from oral cavity and pharyngeal cancer (OCPC).
Ashizawa et a,l 2017	Not distinguish head and neck cancer from OCPC.
Rasmussen et al, 2017	Not distinguish head and neck cancer from OCPC.
Ou et al, 2017	Not distinguish head and neck carcinoma from OCPC.
Morelli et al, 2017	Generated a new classification and definition of periodontal disease.
Chatzopoulos et al, 2018	Validated the predictive value of 'predictors', no multivariate model was constructed.
Fischer et al, 2018	Validated a clinical tool to differentiate between thin, moderate, and thick gingival biotypes.

Chapter 5 Supplement 10. Study characteristics description

Study Characteristics Most studies were conducted in high-income countries (n=27), were patient oriented (n=26) concerned diagnosis of oral conditions (n=24) and used administrative data sources (n=28). Ten studies focused on prognosis, including progression (n=2), survival (n=7) and treatment outcome (n=1). Data were from cross-sectional studies (n=12), retrospective cohorts (n=10), prospective cohorts (n=10), case-control (n=1) and RCT (n=1).
Oral health outcomes The most investigated outcomes were periodontal disease (n=9 plus 5 for tooth loss) and oral cancer (n=9). Other outcomes included implantitis (n=3), dental caries (n=3), dental pain (n=2), mucositis (n=1), periapical cyst (n=1), and oromandibular dystonia (n=1).
Predictors Used Final prediction models included between 3 and 23 predictors. As shown in Appendix 11, popular predictors for periodontal outcomes were smoking (n=13), age (n=11), sex (n=11), general health conditions (n=9), and socioeconomic indicators (n=7). Regarding oral cancer survival models, popular prognostic factors included age (n=12), sex (n=7), T (tumor size) (n=7) and N (nodule involvement) category (n=9). In addition, studies included biomarkers such as metabolites (n=2), DNA (n=5), and other protein molecules (n=3) as predictors of oral cancer.
Model Derivation Among the 24 model development studies, logistic regression (n=13), Cox proportional hazards regression (n=4), least absolute shrinkage and selection operator (LASSO) regression (n=1), decision tree (n=4), neural network (n=2), support vector machine (n=2), recursive partitioning analysis (n=1) were adopted.
Model Presentation Complete regression formula, including regression coefficients and intercepts were presented in 29 models (out of the 58 models identified corresponding to 51%). Another 18 models (31%) were presented as web-based calculator or mobile application (n=4), questionnaires (n=6), scoring system or chart (n=3), nomogram (n=4) or cariogram (n=1). Machine learning algorithms were presented in 9 models (16%), including decision tree structure (n=5), support vector machine (n=2), and neural network (n=2). Information of model presentation was missing in one model.

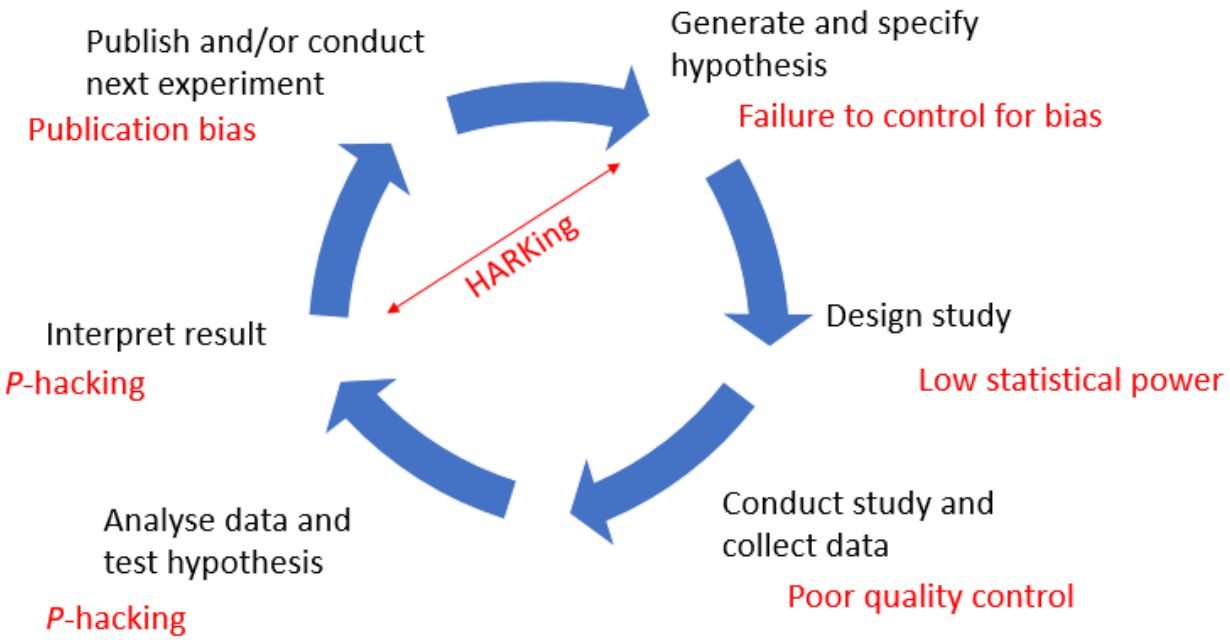
Chapter 5 Supplement 11. Variables included in periodontal- and oral cancer-related prediction models

	Study (Author /Year)	Outcomes	Predictors (n)
Periodontal outcome	Chatzopoulos 2016	Periodontal disease	Age, sex, four questionnaire items
	Eke 2016	Periodontal disease	Age, sex, race, smoking, poverty
	Lee 2018	Periodontal disease	Age(3), sex(3), smoking(3), residence area(3), education(3), stress(3), alcohol(3), hyperlipidaemia(3)
	Leite 2017	Periodontal disease	Sex (4), family income at birth (4), education, number of people living in the house, smoking, BMI (2), diastolic blood pressure (2), frequency of tooth brushing, proportion of teeth with pocket (3), bleeding or calculus (4); number of teeth lost (4); number of DMFT (2)
	Carra 2018	Periodontal disease	12 questionnaire items
	Heaton 2017	Periodontal disease	8 questionnaire items
	Kuboniwa 2016	Periodontal disease	8 metabolites (Ornithine, 5-Oxoproline, Valine, Proline, Spermidine, Hydrocinnamate, Histidine, Cadaverine)
	Su 2017	Periodontal disease	CPI score
	Martinez-Canut 2018A	Tooth loss	Age(2), smoking(2), bruxism(2), severe periodontitis(2); number of teeth loss(2); type of tooth(2); furcation involvement; PPD(2); bone loss(2); mobility(2); crown-to-root ratio(2)
	Martinez-Canut 2018B	Tooth loss	Age, smoking, bruxism, severe periodontitis; number of teeth loss; type of tooth; furcation involvement; PPD; bone loss; mobility; crown-to-root ratio
	Meisel 2018	Tooth loss ǂ	Age, sex, education, financial condition, smoking, antidiabetic drug, mobile dentures, number of dental visits, oral health rating
	Morelli 2018	Tooth loss and periodontitis progression	Age, sex, race, smoking, diabetes, University of North Carolina-Periodontal profile class (UNC-PPC): IAL, direct attachment level, interproximal PPD, direct PPD, interproximal gingival recession, direct GR, BOP, GI, PI, decayed coronal surface, filled coronal surface, decayed root surface, filled root surface, presence/absence of full prosthetic crowns.
	Schwendicke 2018	Tooth loss ǂ	Age, sex, smoking (3), systemic diseases (2), financial condition, PPD (3), bone loss (5), FI (2), calculus (2), crown-root ratio (2), mobility (2), aesthetic zone involvement, adherent to recall interval, root-canal filling
	Gul 2016	Treatment of periodontal disease	6 biomarkers (MMP-8, Elastase, Sialidase, P. gingivalis %, T. forsythia %, F. nucleatum %)
Oral cancer	Bersani 2017	OC survival (among HPV-positive TSCC and BOTSCC patients) Þ	Age, diagnosis, sex, T category and N category, M category, overall stage, CD8+ TILs, HPV16 E2 and E5 mRNA expression, treatment
	Bobdey 2018	OC survival (among surgically treated T4 buccal mucosa cancer patients) õ	Sex, invasion, lymph node involvement, bone infilteration
	OuYang 2017	OC survival (among NPC) Þ	Age (2), sex (2), BMI, T category (2), N category (2), EBV DNA (2), CRP (2), LDH (2), Hb
	Prince 2016	OC survival õ	Age (5), sex(3), race(4), comorbidity(3), tumour site(4), T category (2), N category (3), M category, invasion, Grade (2), smoking, histology, prior tumour, ECS

	Zhang 2017	OC survival Þ	Age, P53, CA9, degree of dysplasia
	Xu 2017	OC survival (among NPC) Þ õ	Age (2), histology (2), T category (2), N category (2), EBV DNA(2), neutrophil-lymphocyte ratio
	Peng 2018	OC survival (among NPC) Þ	Nodal category, overall stage, EVB DNA
	Rao 2016	Oral cancer	Smoking, chewing tobacco, quid with tobacco, quid without tobacco, alcohol consumption, family history of UADT cancer, spiciness of food, fruit consumption, rinsing mouth with water after eating/chewing
	Jover-Esplá 2018	Recurrence of laryngeal glottic cancer	Age, lymph node involvement, alcohol consumption, overall stage
	Orlandi 2018	Mucositis (among NPC)	Oral cavity EUD, Combined parotid glands EUD, BMI

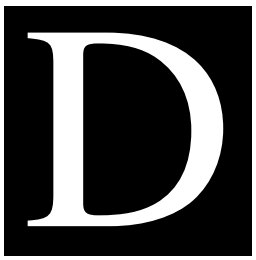
Note: CPI: Community periodontal index; PPD: Probing pocket depth; OC: oral cancer; Þ: Progression-free survival, õ: Overall survival; NPC: Nasopharyngeal carcinoma; HPV: Human papilloma virus; EVB: Epstein-Barr virus; ǂ: Original articles reported outcome as tooth loss, however, the authors assumed tooth loss due to periodontal disease, as the predictors are periodontal-related, and papers are published on periodontal journals.

Chapter 5 Supplement 12. Threats to reproducible science



Appendix Figure 1. Threats to reproducible science. Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication, hypothesizing after the results are known (HARKing), poor study design, low statistical power, analytical flexibility, *P*-hacking, publication bias and lack of data sharing. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Source: Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers E-J, Ware JJ, Ioannidis JPA. 2017. A manifesto for reproducible science. Nature Human Behaviour. 1(1)



Appendices to Chapter 6

Chapter 6 Supplement 1. Demographic characteristics of patients with oral and pharyngeal cancers in SEER cohorts (for imputed data)

	3-year cohort	5-year cohort
	Imputed data (n=21,154)	Imputed data (n=21,000)
Death status		
Alive	13494 (63.8%)	12347 (58.8%)
Dead	7660 (36.2%)	8653 (41.2%)
Survival months		
Mean (SD)	65.9 (44.8)	66.0 (44.9)
Median [min, max]	76.0 [2.00, 143]	77.0 [2.00, 143]
Age(years)		
Mean (SD)	59.1 (12.3)	59.1 (12.3)
Median [min, max]	58.0 [18.0, 105]	58.0 [18.0, 105]
Sex		
Female	5451 (25.8%)	5398 (25.7%)
Male	15703 (74.2%)	15602 (74.3%)
Race		
American Indian/Alaska native	112 (0.5%)	110 (0.5%)
Asian or Pacific Islander	1349 (6.4%)	1335 (6.4%)
Black	2057 (9.7%)	2044 (9.7%)
White	17636 (83.4%)	17511 (83.4%)
Marital status		
Divorced	2746 (13.0%)	2701 (12.9%)
Married (including common law)	12188 (57.6%)	12108 (57.7%)
Separated	254 (1.2%)	254 (1.2%)
Single (never married)	4160 (19.7%)	4123 (19.6%)
Widowed	1806 (8.5%)	1814 (8.6%)

Chapter 6 Supplement 2. Tumour-related characteristics of patients with oral and pharyngeal cancers in SEER cohorts (for imputed data)

	3-year cohort	5-year cohort
	Imputed data (n=21,154)	Imputed data (n=21,000)
Differentiation grade		
Well differentiated; grade I	2911 (13.8%)	2890 (13.8%)
Moderately differentiated; grade II	9801 (46.3%)	9768 (46.5%)
Poorly differentiated; grade III	7893 (37.3%)	7805 (37.2%)
Undifferentiated; anaplastic; grade IV	549 (2.6%)	537 (2.6%)
T category		
T1	5425 (25.6%)	5389 (25.7%)
T2	5544 (26.2%)	5500 (26.2%)
T3	2209 (10.4%)	2202 (10.5%)
T4	3659 (17.3%)	3637 (17.3%)
Tx	4317 (20.4%)	4272 (20.3%)
N category		
N0	7652 (36.2%)	7583 (36.1%)
N1	4285 (20.3%)	4255 (20.3%)
N2	7227 (34.2%)	7191 (34.2%)
N3	836 (4.0%)	831 (4.0%)
Nx	1154 (5.5%)	1140 (5.4%)
M category		
M0	19075 (90.2%)	18944 (90.2%)
M1	809 (3.8%)	806 (3.8%)
Mx	1270 (6.0%)	1250 (6.0%)
Stage		
I	3489 (16.5%)	3436 (16.4%)
Ii	2405 (11.4%)	2337 (11.1%)
Iii	3708 (17.5%)	3689 (17.6%)
Iv	11552 (54.6%)	11538 (54.9%)
Lymph nodes removed		
None	13160 (62.2%)	13063 (62.2%)
Yes	7994 (37.8%)	7937 (37.8%)
Tumour size		
0~1cm	2746 (13.0%)	2684 (12.8%)
1~2cm	4857 (23.0%)	4878 (23.2%)
2~3cm	5433 (25.7%)	5405 (25.7%)
3~4cm	3704 (17.5%)	3680 (17.5%)
4~5cm	2455 (11.6%)	2410 (11.5%)
5~6cm	1009 (4.8%)	1019 (4.9%)
6~7cm	463 (2.2%)	461 (2.2%)
7~8cm	241 (1.1%)	217 (1.0%)
8~9cm	100 (0.5%)	99 (0.5%)
9~10cm	69 (0.3%)	71 (0.3%)
>10cm	77 (0.4%)	76 (0.4%)
Surgical therapy		
Surgery not performed	10488 (49.6%)	10421 (49.6%)
Surgery performed	10666 (50.4%)	10579 (50.4%)
Tumour sites (icd code)		
Lip (C00)	1240 (5.9%)	1227 (5.8%)
Base of tongue (C01)	3989 (18.9%)	3968 (18.9%)
Other parts of tongue (C02)	3500 (16.5%)	3468 (16.5%)
Gum (C03)	612 (2.9%)	609 (2.9%)
Floor of mouth (C04)	1118 (5.3%)	1108 (5.3%)
Palate(C05)	607 (2.9%)	601 (2.9%)
Other oral cavity(C06)	1040 (4.9%)	1029 (4.9%)
Parotid gland (C07)	414 (2.0%)	411 (2.0%)
Other salivary glands (C08)	72 (0.3%)	70 (0.3%)
Tonsil (C09)	4921 (23.3%)	4885 (23.3%)
Oropharynx (C10)	801 (3.8%)	800 (3.8%)
Nasopharynx (C11)	1187 (5.6%)	1176 (5.6%)
Pyriform sinus (C12)	754 (3.6%)	752 (3.6%)
Hypopharynx (C13)	616 (2.9%)	614 (2.9%)
Others (C14)	283 (1.3%)	282 (1.3%)

Chapter 6 Supplement 3. Hazard ratios of each predictors returned by Cox regression

	exp(coef)	exp(-coef)	lower .95	upper .95
Age	1.0319	0.9691	1.0287	1.0351
SexMale	1.0209	0.9795	0.9430	1.1052
RaceAsian or Pacific Islander	0.6215	1.6091	0.4064	0.9503
RaceBlack	0.8914	1.1219	0.5905	1.3455
RaceWhite	0.6653	1.5030	0.4443	0.9962
Marital_sMarried (including common law)	0.7011	1.4263	0.6370	0.7717
Marital_sSeparated	0.9629	1.0386	0.7255	1.2779
Marital_sSingle (never married)	1.1360	0.8803	1.0198	1.2655
Marital_sWidowed	1.0733	0.9317	0.9409	1.2243
GradePoorly differentiated; Grade III	0.9313	1.0738	0.8663	1.0010
GradeUndifferentiated; anaplastic; Grade IV	0.9324	1.0725	0.7261	1.1972
GradeWell differentiated; Grade I	0.8087	1.2365	0.7221	0.9057
T_nT2	1.1621	0.8605	0.9507	1.4205
T_nT3	1.3497	0.7409	1.0879	1.6746
T_nT4	1.8214	0.5490	1.4871	2.2310
T_nTX	1.5970	0.6262	0.9733	2.6206
N_nN1	1.4340	0.6973	1.2673	1.6227
N_nN2	1.4997	0.6668	1.3168	1.7079
N_nN3	1.8992	0.5265	1.5593	2.3131
N_nNX	1.6248	0.6155	1.0521	2.5093
M_nM1	2.5649	0.3899	2.2253	2.9565
M_nMX	1.6576	0.6033	1.2633	2.1748
StageII	1.3670	0.7315	1.1251	1.6610
StageIII	1.2592	0.7941	1.0296	1.5400
StageIV	1.3917	0.7186	1.1263	1.7195
LN_rYes	0.8905	1.1229	0.8145	0.9737
TS_n0~1cm	0.3714	2.6924	0.2213	0.6235
TS_n1~2cm	0.5840	1.7124	0.3568	0.9559
TS_n2~3cm	0.5996	1.6679	0.3726	0.9649
TS_n3~4cm	0.7123	1.4039	0.4431	1.1451
TS_n4~5cm	0.8386	1.1924	0.5234	1.3437
TS_n5~6cm	0.8903	1.1232	0.5523	1.4352
TS_n6~7cm	0.9833	1.0170	0.6011	1.6086
TS_n7~8cm	1.2497	0.8002	0.7483	2.0870
TS_n8~9cm	1.2781	0.7824	0.7298	2.2384
TS_n9~10cm	0.8710	1.1481	0.4684	1.6198
SurgerySurgery performed	0.6187	1.6163	0.5611	0.6822
ICD_nBase of tounge (C01)	1.1606	0.8616	0.8537	1.5778
ICD_nOther parts of tounge (C02)	2.9980	0.3336	2.2311	4.0285
ICD_nGum (C03)	2.2024	0.4541	1.5841	3.0620
ICD_nFloor of mouth (C04)	3.1474	0.3177	2.3132	4.2824
ICD_nPalate(C05)	2.1562	0.4638	1.5388	3.0214
ICD_nOther oral cavity(C06)	2.9675	0.3370	2.1729	4.0526
ICD_nParotid gland (C07)	1.8092	0.5527	1.2708	2.5757
ICD_nOther salivary glands (C08)	3.1383	0.3186	1.8906	5.2093
ICD_nTonsil (C09)	0.8786	1.1381	0.6471	1.1931
ICD_nOropharynx (C10)	1.6830	0.5942	1.2041	2.3525
ICD_nNasopharynx (C11)	1.3728	0.7284	0.9671	1.9488
ICD_nPyriform sinus (C12)	2.0788	0.4810	1.5005	2.8799
ICD_nHypopharynx (C13)	2.2264	0.4492	1.5960	3.1057

Note: Results are from the complete-case analysis of training datasets (80% of the original data) of 3-year cohort.

Chapter 6 Supplement 4. C-indexes for models predicting the 3- and 5-year disease-specific survival of OPCs in the model development and test datasets (with partition ratio of 8:2, 7:3, 5:5 and 3:7)

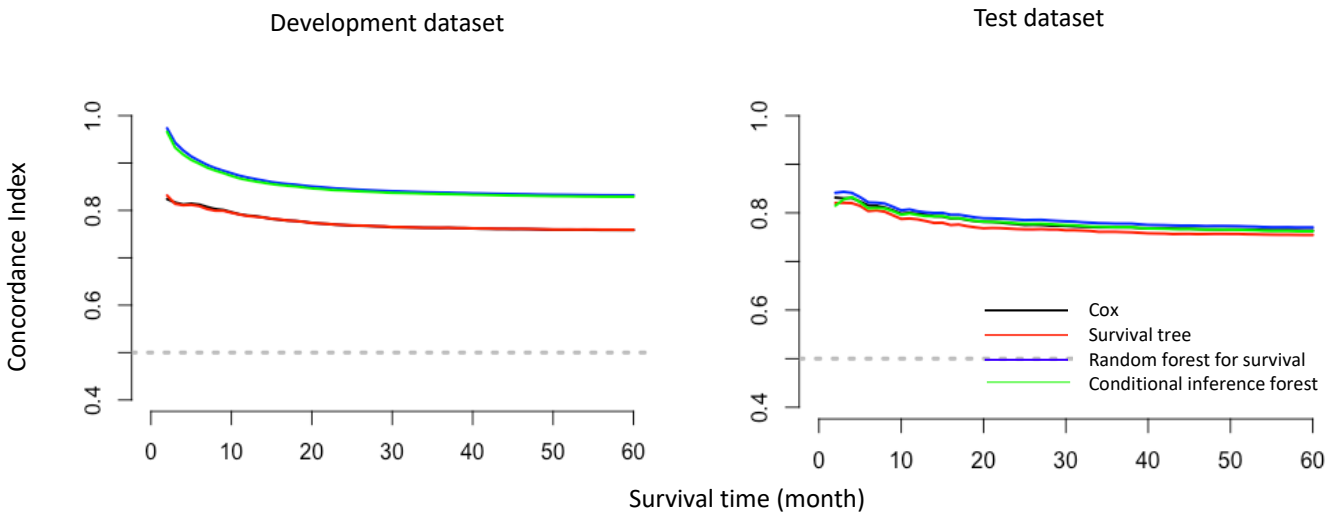
Three-year survival cohort		
	Development dataset (80%) (Median (IQR))	Testing dataset (20%) (Median (IQR))
Data set with complete cases (N=11,888)		
Cox	0.768 (0.767, 0.770)	0.764 (0.758, 0.768)
Survival tree	0.702 (0.701, 0.703)	0.703 (0.693, 0.705)
Random forest for survival	0.834 (0.834, 0.838)	0.766 (0.760, 0.773)
Conditional inference forest	0.833 (0.827, 0.856)	0.759 (0.755, 0.763)
Data set with imputation (N=21,154)		
Cox	0.768 (0.768, 0.769)	0.768 (0.766, 0.768)
Survival tree	0.696 (0.688, 0.706)	0.688 (0.684, 0.700)
Random forest for survival	0.831 (0.791, 0.837)	0.775 (0.771, 0.776)
Conditional inference forest	0.850 (0.838, 0.873)	0.768 (0.767, 0.770)
	Development dataset (70%) (Median (IQR))	Testing dataset (30%) (Median (IQR))
Data set with complete cases (N=11,888)		
Cox	0.769 (0.767, 0.770)	0.763 (0.760, 0.767)
Survival tree	0.702 (0.699, 0.705)	0.699 (0.693, 0.705)
Random forest for survival	0.841 (0.833, 0.841)	0.769 (0.764, 0.774)
Conditional inference forest	0.848 (0.828, 0.852)	0.775 (0.771, 0.777)
Data set with imputation (N=21,154)		
Cox	0.770 (0.769, 0.772)	0.764 (0.760, 0.765)
Survival tree	0.697 (0.690, 0.707)	0.696 (0.688, 0.701)
Random forest for survival	0.840 (0.831, 0.845)	0.776 (0.775, 0.777)
Conditional inference forest	0.843 (0.835, 0.846)	0.767 (0.765, 0.767)
	Development dataset (50%) (Median (IQR))	Testing dataset (50%) (Median (IQR))
Data set with complete cases (N=11,888)		
Cox	0.771 (0.767, 0.772)	0.764 (0.763, 0.767)
Survival tree	0.703 (0.698, 0.707)	0.696 (0.692, 0.700)
Random forest for survival	0.839 (0.828, 0.843)	0.768 (0.766, 0.769)
Conditional inference forest	0.840 (0.839, 0.842)	0.758 (0.754, 0.760)
Data set with imputation (N=21,154)		
Cox	0.771 (0.769, 0.772)	0.764 (0.763, 0.765)
Survival tree	0.698 (0.691, 0.705)	0.694 (0.686, 0.698)
Random forest for survival	0.836 (0.835, 0.842)	0.775 (0.774, 0.775)
Conditional inference forest	0.852 (0.850, 0.856)	0.771 (0.771, 0.772)
	Development dataset (30%) (Median (IQR))	Testing dataset (70%) (Median (IQR))
Data set with complete cases (N=11,888)		
Cox	0.771 (0.766, 0.774)	0.762 (0.760, 0.764)
Survival tree	0.698 (0.697, 0.710)	0.695 (0.688, 0.699)
Random forest for survival	0.842 (0.836, 0.845)	0.757 (0.762, 0.766)
Conditional inference forest	0.853 (0.851, 0.854)	0.752 (0.750, 0.756)
Data set with imputation (N=21,154)		
Cox	0.772 (0.770, 0.773)	0.765 (0.764, 0.766)
Survival tree	0.703 (0.693, 0.708)	0.689 (0.685, 0.698)
Random forest for survival	0.839 (0.830, 0.847)	0.770 (0.770, 0.772)
Conditional inference forest	0.842 (0.839, 0.850)	0.771 (0.768, 0.772)

Table continued

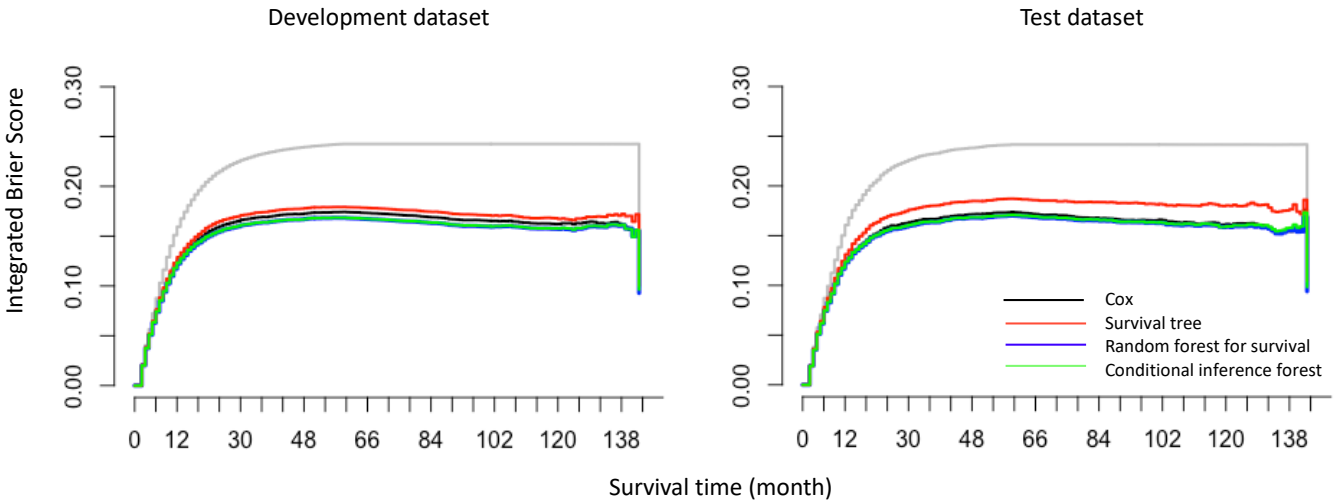
Table continued

Five-year survival cohort		
	Development dataset (80%) (Median (IQR))	Testing dataset (20%) (Median (IQR))
Data set with complete cases (N=11,807)		
Cox	0.762 (0.761, 0.763)	0.756 (0.761, 0.764)
Survival tree	0.694 (0.692, 0.698)	0.688 (0.680, 0.696)
Random forest for survival	0.826 (0.826, 0.833)	0.762 (0.761, 0.764)
Conditional inference forest	0.850 (0.837, 0.856)	0.752 (0.751, 0.764)
Data set with imputation (N=21,000)		
Cox	0.764 (0.761, 0.764)	0.762 (0.762, 0.767)
Survival tree	0.692 (0.690, 0.695)	0.689 (0.683, 0.695)
Random forest for survival	0.829 (0.828, 0.830)	0.773 (0.769, 0.776)
Conditional inference forest	0.849 (0.843, 0.854)	0.767 (0.766, 0.767)
Modelling approaches	Development dataset (70%) (Median (IQR))	Testing dataset (30%) (Median (IQR))
Data set with complete cases (N=11,807)		
Cox	0.762 (0.761, 0.763)	0.758 (0.756, 0.761)
Survival tree	0.693 (0.692, 0.698)	0.688 (0.685, 0.694)
Random forest for survival	0.828 (0.827, 0.832)	0.760 (0.757, 0.763)
Conditional inference forest	0.843 (0.837, 0.845)	0.760 (0.759, 0.762)
Data set with imputation (N=21,000)		
Cox	0.764 (0.761, 0.764)	0.762 (0.762, 0.767)
Survival tree	0.692 (0.690, 0.696)	0.685 (0.690, 0.695)
Random forest for survival	0.828 (0.819, 0.836)	0.772 (0.771, 0.773)
Conditional inference forest	0.850 (0.842, 0.856)	0.769 (0.766, 0.769)
	Development dataset (50%) (Median (IQR))	Testing dataset (50%) (Median (IQR))
Data set with complete cases (N=11,807)		
Cox	0.763 (0.761, 0.765)	0.758 (0.756, 0.760)
Survival tree	0.696 (0.691, 0.701)	0.690 (0.684, 0.695)
Random forest for survival	0.825 (0.817, 0.826)	0.761 (0.759, 0.764)
Conditional inference forest	0.829 (0.821, 0.835)	0.759 (0.754, 0.761)
Data set with imputation (N=21,000)		
Cox	0.764 (0.761, 0.764)	0.762 (0.761, 0.764)
Survival tree	0.694 (0.689, 0.700)	0.689 (0.684, 0.696)
Random forest for survival	0.824 (0.820, 0.829)	0.766 (0.766, 0.767)
Conditional inference forest	0.830 (0.830, 0.832)	0.764 (0.761, 0.765)
	Development dataset (30%) (Median (IQR))	Testing dataset (70%) (Median (IQR))
Data set with complete cases (N=11,807)		
Cox	0.764 (0.761, 0.768)	0.756 (0.754, 0.758)
Survival tree	0.696 (0.690, 0.704)	0.684 (0.681, 0.692)
Random forest for survival	0.830 (0.812, 0.835)	0.755 (0.754, 0.759)
Conditional inference forest	0.854 (0.842, 0.859)	0.756 (0.751, 0.758)
Data set with imputation (N=21,154)		
Cox	0.762 (0.762, 0.763)	0.762 (0.761, 0.762)
Survival tree	0.693 (0.690, 0.702)	0.688 (0.683, 0.692)
Random forest for survival	0.832 (0.826, 0.838)	0.764 (0.764, 0.766)
Conditional inference forest	0.850 (0.842, 0.856)	0.765 (0.759, 0.769)

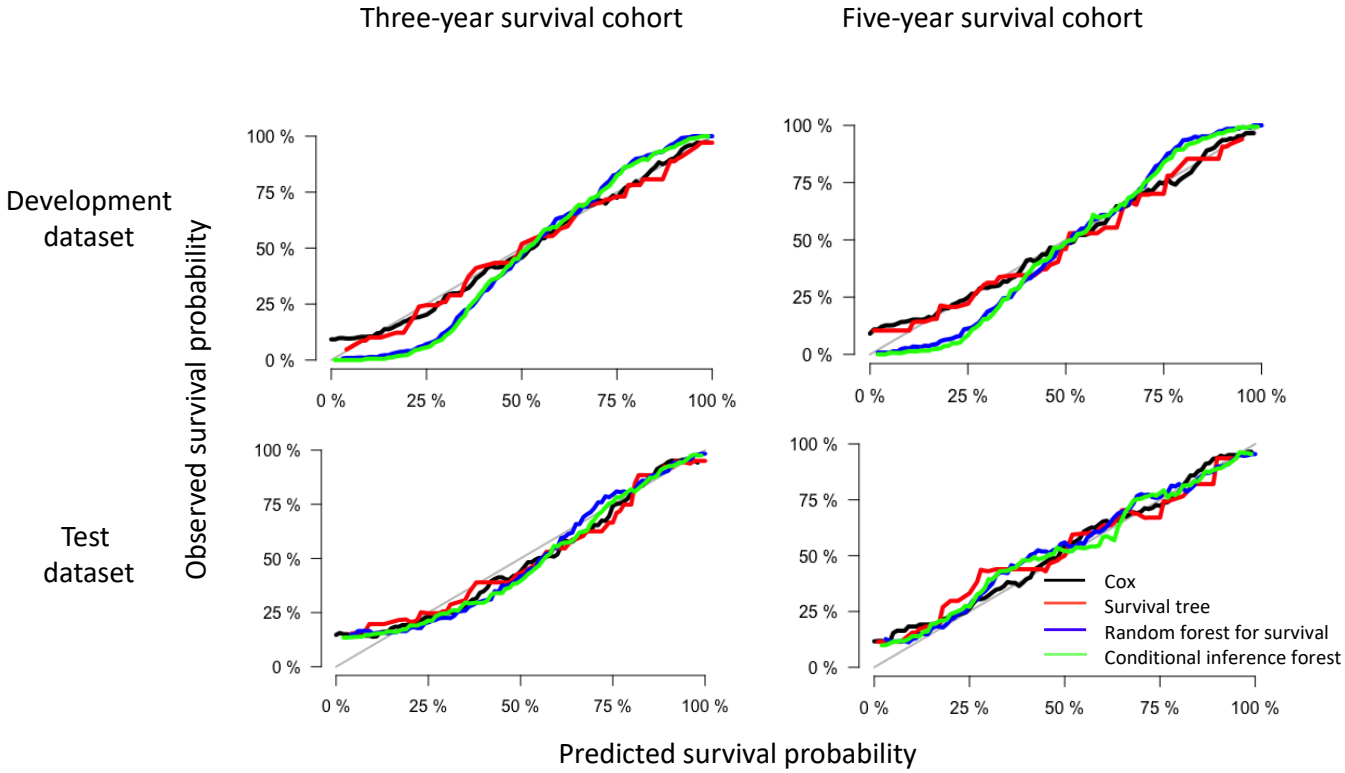
Chapter 6 Supplement 5. Overtime C-index for predicting 3- and 5-year disease-specific survival of oral and pharyngeal cancers with various models in the imputed datasets



Chapter 6 Supplement 6. The prediction error curves for various models in predicting disease-specific survival of oral and pharyngeal cancers based on Integrated Brier Score in the imputed datasets



Chapter 6 Supplement 7. Calibration plots for predicting 3- and 5-year disease-specific survival of oral and pharyngeal cancers with various models in the imputed datasets



Chapter 6 Supplement 8. Time-dependent receiver operator curves for predicting 1- to 5-year disease-specific survival of oral and pharyngeal cancers with Cox models

Description of the methods for plotting the time-dependent ROC

The time-dependent receiver operator curves (ROC) are extensions of the standard ROC curves (developed for binary data) and are developed for situations where the event status (e.g. death) occurs at various time point during the study period, and it is suitable to time-to-event analysis. The time-dependent ROC can be constructed based on the cumulative sensitivity (Se^C) and dynamic specificity (Sp^D), which have been well defined in the literature*:

Let T_i denote the predicted time of event onset and η_i is the predicted 'risk' (represented by hazard ratios of predictor values at baseline) for individual i, ($i = 1, \ldots, n$). At each observed time point t, each individual is classified as a case or control (e.g. has event/no event at that time point in between time periods $T_i = 0$ and t). A case is defined as any individual experiencing the event between baseline $t = 0$ and time t and a control as an individual remaining event-free at time t. The cases and controls change over time and each individual may play the role of control at the earlier time (when the event time is greater than the target time, i.e. $T_i > t$) but then contributes as a case for later times (when the event time is less than or equal to the target time, i.e. $T_i \leq t$). For an observed threshold c, the cumulative sensitivity of our model is defined as the probability that the individual has an predicted 'risk' greater than c among the individuals who experienced the event before time t, and the dynamic specificity is the probability that an individual has a predictor value less than or equal to c among those event-free individuals beyond time t. Thus:

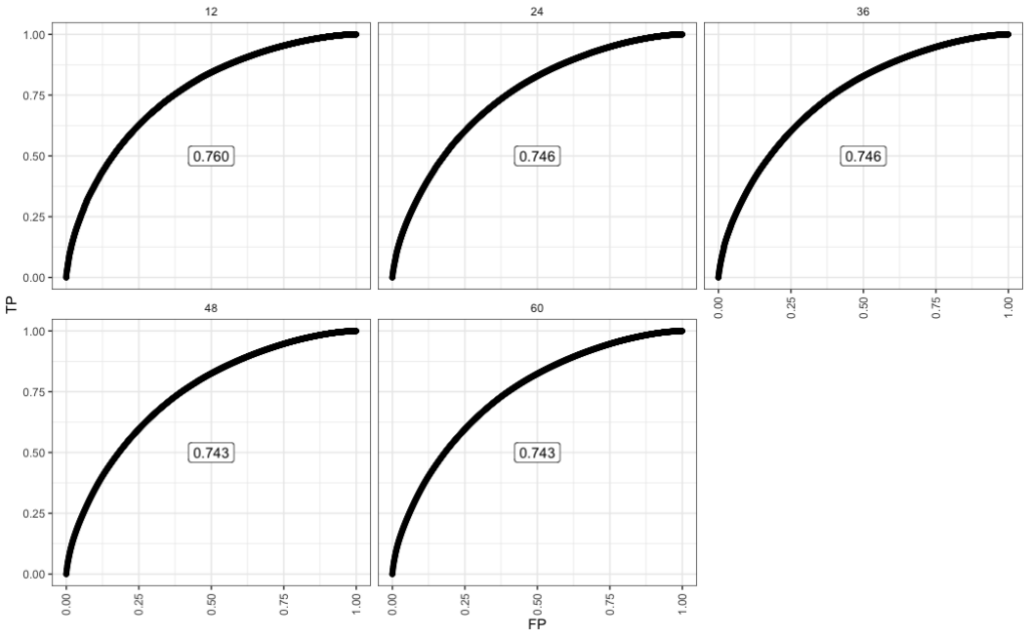
$$Se^C(c,t) = P(\eta_i > c | T_i \leq t)$$

$$Sp^D(c,t) = P\ (\eta_i \leq c | T_i \geq t)$$

$$AUC^{C,D}\ (t) = P\big(\eta_i > \eta_j \big| T_i \leq t, T_j > t\big), i \neq j$$

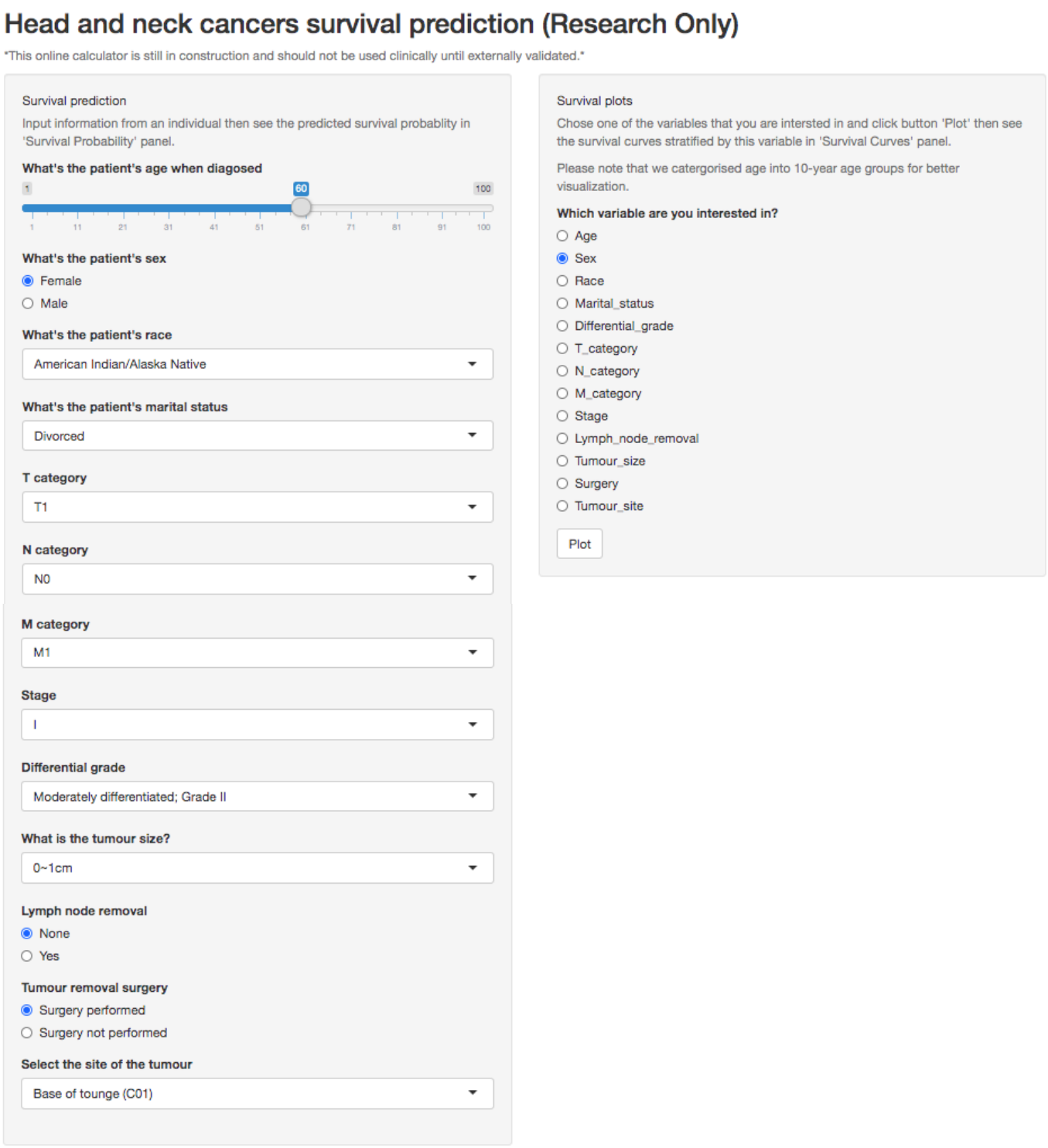
Results

Once the time-dependent setting is applied, the death status is observed and predicted at each time point which yields different values of sensitivity and specificity throughout the investigated time period. So that we could choose to plot the time-dependent ROC curves at a specific time of interest. For example, the following plots give the cumulative prediction performance at 12, 24, 36, 48 and 60 months.



Source: Kamarudin, Adina Najwa, Trevor Cox, and Ruwanthi Kolamunnage-Dona. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC medical research methodology 17.1 (2017): 53.

Chapter 6 Supplement 9. Snapshot of a web-based calculator for OPCs survival probability



Chapter 6 Supplement 10. Sensitivity analysis to investigate the effect of unmeasured factor on the estimation of hazard ratios

Methods

Here's an example call to the sensitivity analysis (R Code):

```
obsSensSCC(cox1, which=1, g0=c(0.1,0.5,2), p0=seq(0,1,0.2), p1=seq(0,1,0.2), logHaz=F)
```

where:

- obsSensSCC = a sensitivity analysis for three variables: outcome Y is a survival outcome, exposure X is a categorical variable such as sex, and latent variables U are categorical variables)
- model = the Cox regression model
- which = the parameter in the regression model that specifies the predictor, e.g. 2 refers to the second predictor, which was sex-male in our analysis
- g0 = strength of the relationship between U and the outcome (specified here as a hazard ratio); also called gamma
- p0 = prevalence of U in unexposed group (or when exposure = 0)
- p1 = prevalence of U in the exposed group (or when exposure = 1)
- logHaz = whether log of the hazard or the hazard ratio should be returned

In the sensitivity analysis, a range of **g** is chosen to include the unadjusted $\boldsymbol{\beta}$ before adjusting for unmeasured predictors. Together with a range of ***p***, $\boldsymbol{\beta}$ is then estimated for different values of **g** and ***p***. For example, a range of **g** of 0.1 to 2 was chosen as the hypothetical effect of unmeasured predictors that could explain away the $\boldsymbol{\beta}$ or reduce it to a specific level. If the confidence intervals of the adjusted $\boldsymbol{\beta}$ do not include 1, this suggests a direct beneficial relationship between survival outcome and predictors. On the other hand, if the confidence interval included 1, then the unmeasured predictor could explain the relationship between survival outcome and predictors.

Results

Let's look at two examples:

The following table presents the impact of unmeasured predictor on the hazard ratios of the association between two predictors and 3-year survival outcome. For the predictor 'whether the surgery was performed or not', we found that the range of $\boldsymbol{\beta}$ did not include 1 across all scenarios, which means the added unmeasured predictor did not impact on the effect of 'Surgery' on the outcome. However, for the predictor 'T category, T3', when larger proportion of T3 patients have the unmeasured predictor than the non-T3 patients (e.g. this meets the chemotherapy scenario), the effect might be changed only when the unmeasured predictors has effect 2-fold larger than the existing predictor 'T category, T3'.

Sensitivity analysis on variable SurgerySurgery performed on a Hazard Ratio scale

, , Gamma = 0.1

P1 \ P0	0.2	0.4	0.6	0.8
0.2	0.643 (0.579,0.714)	0.502 (0.452,0.558)	0.361 (0.325,0.401)	0.220 (0.198,0.244)
0.4	0.824 (0.742,0.915)	0.643 (0.579,0.714)	0.462 (0.416,0.513)	0.281 (0.253,0.313)
0.6	1.146 (1.032,1.273)	0.895 (0.805,0.994)	0.643 (0.579,0.714)	0.391 (0.352,0.435)

, , Gamma = 0.5

P1 \ P0	0.2	0.4	0.6	0.8
0.2	0.643 (0.579,0.714)	0.572 (0.515,0.635)	0.500 (0.450,0.556)	0.429 (0.386,0.476)
0.4	0.723 (0.651,0.804)	0.643 (0.579,0.714)	0.563 (0.507,0.625)	0.482 (0.434,0.536)
0.6	0.827 (0.744,0.918)	0.735 (0.662,0.816)	0.643 (0.579,0.714)	0.551 (0.496,0.612)

, , Gamma = 2

P1 \ P0	0.2	0.4	0.6	0.8
0.2	0.643 (0.579,0.714)	0.750 (0.675,0.833)	0.857 (0.772,0.953)	0.965 (0.868,1.072)
0.4	0.551 (0.496,0.612)	0.643 (0.579,0.714)	0.735 (0.662,0.816)	0.827 (0.744,0.918)
0.6	0.482 (0.434,0.536)	0.563 (0.507,0.625)	0.643 (0.579,0.714)	0.723 (0.651,0.804)

Sensitivity analysis on variable T_nT3 on a Hazard Ratio scale

, , Gamma = 0.1

P1 \ P0	0.2	0.4	0.6	0.8
0.2	1.293 (1.023,1.635)	1.009 (0.798,1.276)	0.726 (0.574,0.917)	0.442 (0.349,0.558)
0.4	1.657 (1.311,2.095)	1.293 (1.023,1.635)	0.930 (0.735,1.175)	0.566 (0.447,0.715)
0.6	2.305 (1.823,2.915)	1.799 (1.423,2.275)	1.293 (1.023,1.635)	0.787 (0.623,0.995)

, , Gamma = 0.5

P1 \ P0	0.2	0.4	0.6	0.8
0.2	1.293 (1.023,1.635)	1.150 (0.909,1.454)	1.006 (0.796,1.272)	0.862 (0.682,1.090)
0.4	1.455 (1.151,1.840)	1.293 (1.023,1.635)	1.132 (0.895,1.431)	0.970 (0.767,1.226)
0.6	1.663 (1.315,2.103)	1.478 (1.169,1.869)	1.293 (1.023,1.635)	1.109 (0.877,1.402)

, , Gamma = 2

P1 \ P0	0.2	0.4	0.6	0.8
0.2	1.293 (1.023,1.635)	1.509 (1.193,1.908)	1.724 (1.364,2.180)	1.940 (1.534,2.453)
0.4	1.109 (0.877,1.402)	1.293 (1.023,1.635)	1.478 (1.169,1.869)	1.663 (1.315,2.103)
0.6	0.970 (0.767,1.226)	1.132 (0.895,1.431)	1.293 (1.023,1.635)	1.455 (1.151,1.840)

**** g(Gamma)** refers to the effect estimate of the association between predictors and an unmeasured covariate; **p** refers to the correlation between survival outcome and unmeasured covariate. Effect estimates not including 1 represent conditions where survival outcome is associated with the presence of predictors, whereas those including 1 represent conditions where survival outcome is not associated with predictors.*

Chapter 6 Supplement 11. Pruning parameters for survival tree and random forests for survival

Model	Pruning Parameters
Survival tree	(‘minsplit’, lower=1,upper=20), corresponds to the minimum number of observations that must exist in a node in order for a split to be attempted. (‘maxdepth’,lower=1,upper=30), corresponds to the maximum depth of a tree. Depth is the length of the longest path from a Root node to a Leaf node.
Random forests for survival	(‘ntree', lower=1000, upper=2000), corresponds to the total number of trees in the forest. ('mtry',lower = 1,upper = 12), corresponds to the number of variables tested in any split. ('nsplit',lower = 0, upper=20), corresponds to the size of random split points for each ‘mtry’ candidate. ('splitrule',values = 'logrank', special.vals = list('logrank','logrankscore','random')), corresponds to the split rule and formula. ('nodedepth',lower = …, upper =…), corresponds to the length of the longest path from a root to a leaf of any tree in the forest. The default behaviour is that this parameter is ignored. ('nodesize',lower =…, upper=…), corresponds to the minimum number of unique cases (data points) in a terminal node of any tree in the forest. The default behaviour is that this parameter is ignored.
Conditional Inference Forest	(‘ntree', lower=1000, upper=2000), corresponds to the total number of trees in the forest. ('mtry',lower = 1,upper = 12), corresponds to the number of variables tested in any split. ('minsplit',lower = 0, upper=20), corresponds to the minimum size of random split points for each ‘mtry’ candidate. (‘teststat’, values= ‘quad’, special.vals = list(‘quad', ‘max')), corresponds to a character specifying the type of the test statistic to be applied. (‘mincriterion’), corresponds to the depth of the trees. Usually unstopped and unpruned trees are used in random forests. To grow large trees, set it to a small value. The default behaviour is that this parameter is ignored.

Chapter 6 Supplement 12. The step-by-step practical procedure of developing a ST, RF and CF algorithm

The development of a ST algorithm can be summarized as follows:

Start function *F* build survival tree
  Create an initial survival tree with root node t_0
  Create an empty stack *S* of open nodes
  while *S* is not empty **do**
    $t = t_0 + t_1$,
    if stopping criterion is met for *t* **end**
      else Find the split on *F* that maximizes the survival difference between children nodes
    Partition data in to two child nodes of *t*
    end
  end
End

For the tree growing and pruning purposes, one needs a splitting statistic (log-rank) that handles the dependence of failure times and a measure (C-index) to evaluate the performance of the tree.

The development of a RF algorithm can be summarized as follows:

Start
Select the number of trees to build, *ntree*
  for *i* = 1 to *ntree* **do**
    Generate a *B* bootstrap sample (usually two thirds) of the original data, one third is left as out-of-bag (OOB) data
    Train a tree model on this sample
    for *each split* **do**
      Randomly select *k* (< *K*) of the original predictors
      Select the best predictor among the *k* predictors
        for *each splitting point of the best k* **do**
          Compare the survival curves of the two groups using one splitting rule *r* among a) log-rank splitting rule, b) log-rank score splitting rule, or c) random log-rank splitting rule.
          Select the best splitting rule and partition the data
        end for determining splitting rule
    end for determining *one split* in *one tree*
    Use tree stopping criteria to determine when a tree is complete.
    Using OOB data, the prediction ability is calculated and represented by C-index. The cumulative hazard function (CHF) is also calculated for each tree.
  end for *one tree*
The individual C-index are then averaged to obtain the ensemble C-index. The individual CHFs are then averaged to obtain the ensemble CHF.
End

The development of a conditional inference tree can be summarized as follows:

1: For case weights *w*, test the global null hypothesis of independence between any of the p covariates and the response variable. Stop if this hypothesis cannot be rejected otherwise the the j^{th} covariate X with strongest associate to the outcome.
2. Select a set $A \in X$ in order to split X into two disjoint sets. The weights w_L and w_R determine the two subgroups with $w_{L,i} = w_i I(X_{j,i} \in A)$ and $w_{R,}I = w_{iI}(X_{j,i} \neq A)$ for all i=1,2,…n.
3. Recursively repeat steps 1 and 2 with modeified case weights w_L and w_R, respectively.

Source:

1. *Ishwaran H, Kogalur UB, Blackstone EH, and Lauer MS. Random survival forests. Ann. Appl. Statist., 2:841–860, 2008.*
2. *Kuhn, M. and Johnson, K., Applied predictive modeling (Vol. 26). New York: Springer. 2013.*
3. *Torsten Hothorn, Kurt Hornik & Achim Zeileis. Unbiased Recursive Partitioning: A Conditional Inference Framework, Journal of Computational and Graphical Statistics, 15:3, 651-674, 2006.*

Chapter 6 Supplement 13. Type of regression methods used to impute each variable when using Multiple Imputation of Covariates by Substantive Model (smcfcs) package

Variables	**Methods**
Race, Marital status, Lymph node removal, Surgery, tumour site	'mlogit' multinomial logistic regression for unordered categorical variables
Grade, TNM category, Stage, Tumour size	'podds' proportional odds regression for ordered categorical variables
Age, Sex, Survival time, Death status	'', fully observed, does not need to be imputed

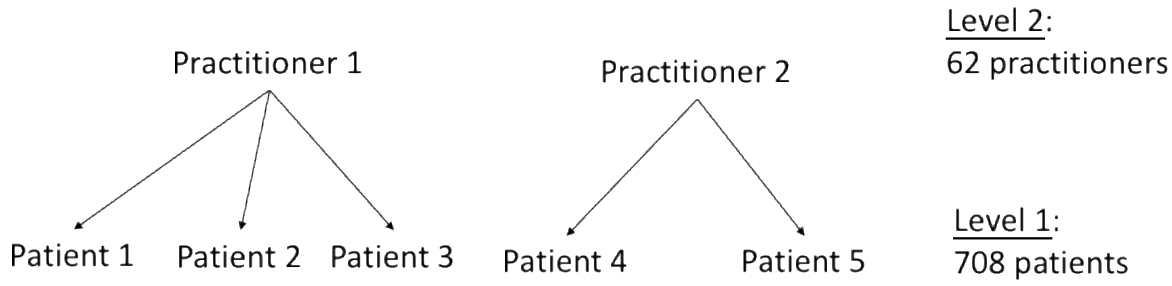
Chapter 6 Supplement 14. TRIPOD checklist for study reporting

Section/Topic	Item	Development or Validation?	Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	1
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	1-2
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model, or both.	1-2
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation datasets, if applicable.	10
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	10
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	10
	5b	D;V	Describe eligibility criteria for participants.	10-11
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	11
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured.	11
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	NA
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	13
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	11
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	11-12
	10c	V	For validation, describe how the predictions were calculated.	NA
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	13
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	NA
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	2-3
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	3-5
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors, and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	3-4, Table 1
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Table S3
	15b	D	Explain how to use the prediction model.	8
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	6-8
Model updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	9
Interpretations	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	10
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	8, 10
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and datasets.	14
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	14

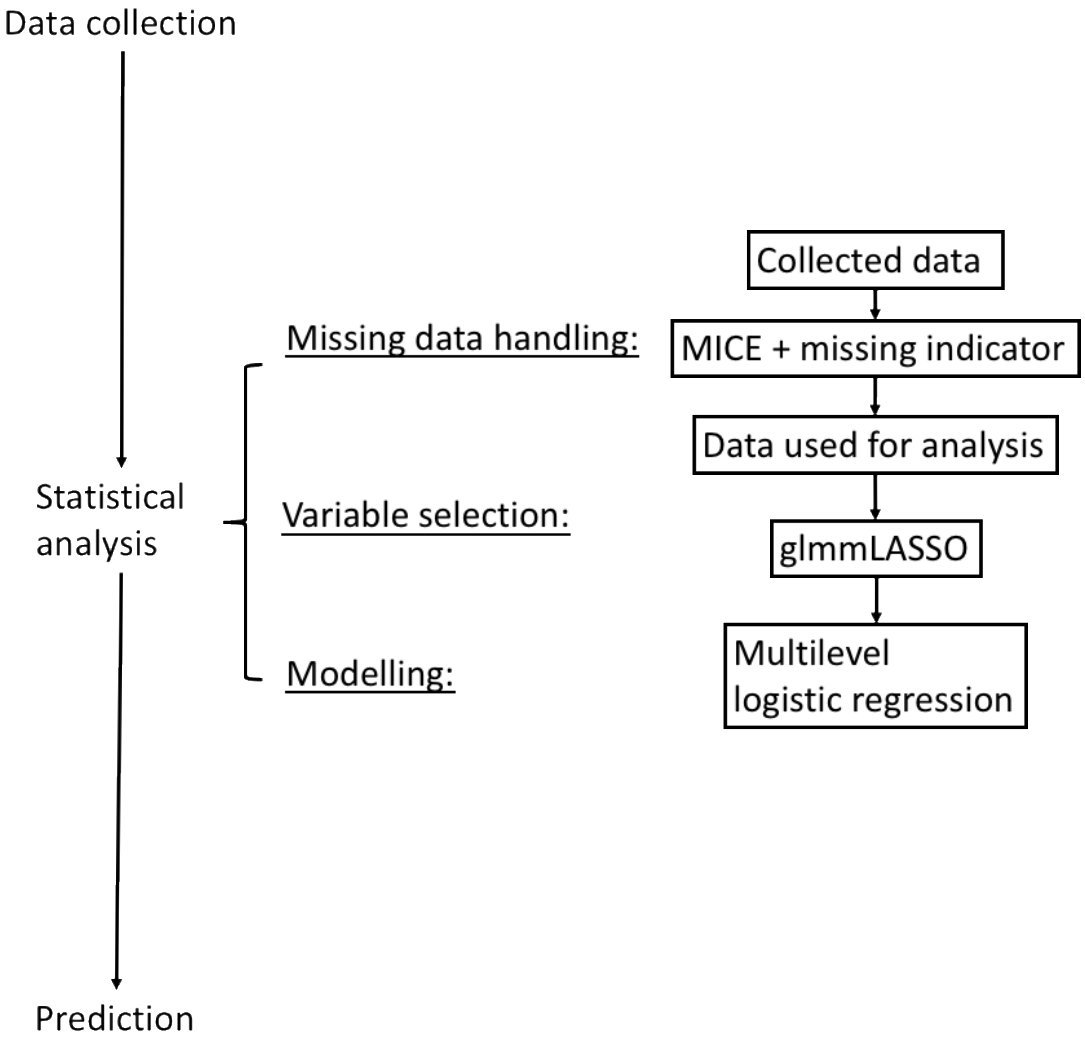
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.

Appendices to Chapter 7

Chapter 7 Supplement 1. Schematic structure of the dataset



Chapter 7 Supplement 2. Visualisation of the adopted statistical analysis



Chapter 7 Supplement 3. R codes used for analysis

Here we use an example of one-week pain prediction:
https://github.com/dumizai/Predicting_pain_following_RCT

Chapter 7 Supplement 4. Models' performance

Measure	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
H	0.43	0.38	0.15	0.41	0.49	0.12
Gini	0.70	0.64	0.36	0.69	0.78	0.32
AUC	0.85	0.82	0.68	0.85	0.89	0.66
AUCH	0.86	0.83	0.70	0.86	0.90	0.68
KS	0.54	0.51	0.31	0.54	0.64	0.24
MER	0.15	0.17	0.24	0.09	0.09	0.10
MWL	0.17	0.18	0.25	0.09	0.07	0.15
Spec.Sens95	0.46	0.33	0.11	0.45	0.57	0.14
Sens.Spec95	0.50	0.42	0.10	0.43	0.51	0.10
ER	0.16	0.17	0.24	0.10	0.10	0.11
Sens	0.48	0.45	0.03	0.21	0.23	0
Spec	0.96	0.95	0.99	0.99	0.98	1
Precision	0.8	0.74	0.5	0.65	0.64	NA
Recall	0.48	0.45	0.03	0.21	0.23	0
TPR	0.48	0.45	0.03	0.21	0.23	0
FPR	0.04	0.05	0.01	0.01	0.02	0
F	0.60	0.56	0.06	0.32	0.34	NA
Youden	0.44	0.40	0.02	0.20	0.21	0
TP	48	45	3	15	16	0
FP	12	16	3	8	9	0
TN	299	295	308	573	572	581
FN	52	55	97	55	54	70

Model 1 & 4: combined set of predictors; Model 2 & 5: clinical set of predictors; Model 3 & 6: sociodemographic set of predictors. *H*: the *H*-measure. Gini: the Gini coefficient. AUC: the Area Under the ROC Curve. AUCH: the Area Under the convex Hull of the ROC Curve . KS: the Kolmogorov-Smirnoff statistic. MER: the Minimum Error Rate. MWL: the Minimum cost-Weighted Error Rate. Spec.Sens95: Specificity when Sensitivity is held fixed at 95%. Spec.Sens95: Sensitivity when Specificity is held fixed at 95%. TPR: Ture Positive Rate. FPR: False Positive Rate.

Chapter 7 Supplement 5. Models' variance and R^2

Model 1 with combined set of variables		Variance-fixed	Variance-random
		63.39535	0.3290074
	R^2 -total	R^2- fixed	R^2- random
	0.95	0.95	

Model 2 with only clinical variables		Variance-fixed	Variance-random
		20.77257	0.139763
	R^2-total	R^2- fixed	R^2- random
	0.86	0.86	

Model 3 with only social variables		Variance-fixed	Variance-random
		0.2676462	0.3572107
	R^2-total	R^2- fixed	R^2- random
	0.16	0.07	

Model 4 with combined set of variables		Variance-fixed	Variance-random
		14.46139	N/A
	R^2-total	R^2- fixed	R^2- random
	0.81	0.81	

Model 5 with clinical variables		Variance-fixed	Variance-random
		30.97322	0. 0812948
	R^2-total	R^2- fixed	R^2- random
	0.90	0.90	

Model 6 with only social variables		Variance-fixed	Variance-random
		0.2055059	0.04684805
	R^2-total	R^2- fixed	R^2- random
	0.07	0.06	

Chapter 7 Supplement 6. Models specification

NOTE: Data dictionary is available at:
https://github.com/dumizai/Predicting_pain_following_RCT

Model1: One-week pain prediction, combined set of predictors

FIXED EFFECTS:

	Est.	S.E.	z val.	p
(Intercept)	-0.58	1.45	-0.40	0.69
Q04P0STRES1	1.94	0.58	3.36	0.00
Q04P0NWPAIN1	-1.76	0.72	-2.44	0.01
Q04P0NWPAIN2	-1.16	0.63	-1.84	0.07
Q04P0NWPAIN3	-0.49	0.54	-0.91	0.36
Q04P0NWPAIN4	0.49	0.50	0.97	0.33
Q04P0NWPAIN5	1.58	0.69	2.30	0.02
Q04P0NWPAIN6	0.01	0.60	0.01	0.99
Q04P0NWPAIN7	-0.49	0.72	-0.68	0.49
Q04P0NWPAIN8	1.15	0.78	1.48	0.14
Q04P0NWPAIN9	0.49	0.87	0.57	0.57
Q04P0NWPAIN10	1.45	1.17	1.24	0.21
Q04P0NOACDAY1	-0.36	0.69	-0.52	0.60
Q04P0NOACDAY2	1.87	0.74	2.53	0.01
Q04P0NOACDAY3	-0.91	0.97	-0.94	0.35
Q04P0NOACDAY4	-0.79	1.31	-0.60	0.55
Q04P0NOACDAY5	0.54	1.14	0.47	0.64
Q04P0EDU2	-0.37	1.17	-0.32	0.75
Q04P0EDU3	-0.46	1.15	-0.40	0.69
Q04P0EDU4	0.13	1.14	0.12	0.91
Q04P0EDU5	-0.64	1.17	-0.55	0.58
Q04P0NOREC1	0.12	0.69	0.18	0.86
Q04P0NOREC2	-1.51	1.01	-1.49	0.14
Q04P0NOREC3	-0.17	0.82	-0.20	0.84
Q04P0NOREC4	1.43	1.45	0.99	0.32
Q04P0NOREC5	1.81	1.10	1.64	0.10
Q04P0NOREC6	-0.39	1.28	-0.31	0.76
Q04P0NOREC7	-1.12	1.48	-0.76	0.45
Q04P0NOREC8	1.18	1.68	0.70	0.48
Q04P0NOREC9	19.84	6647.94	0.00	1.00
Q04P0NOWRK1	-0.65	0.68	-0.95	0.34
Q04P0NOWRK2	0.54	0.83	0.65	0.51
Q04P0NOWRK3	0.50	0.87	0.57	0.57
Q04P0NOWRK4	-1.75	1.50	-1.17	0.24
Q04P0NOWRK5	-1.68	1.27	-1.33	0.18
Q04P0NOWRK6	0.48	1.32	0.37	0.71
Q04P0NOWRK7	0.61	1.40	0.44	0.66
Q04P0NOWRK8	-1.04	1.90	-0.54	0.59
Q04P0NOWRK9	-55.72	11369.08	-0.00	1.00

Q_09P1PAIN1	0.24	0.39	0.62	0.54
Q_09P1PAIN2	0.66	0.60	1.10	0.27
Q_09P1PAIN3	-0.11	0.76	-0.15	0.88
Q_09P1PAIN4	1.60	1.17	1.37	0.17
Q_09P1PAIN5	0.69	0.78	0.88	0.38
Q_09P1PAIN6	-0.53	0.96	-0.55	0.58
Q04P0RACEv2Black	0.11	0.90	0.12	0.90
Q04P0RACEv2Asian	-0.05	1.45	-0.03	0.97
Q04P0RACEv2Other	21.48	7888.22	0.00	1.00
STATEFL	0.19	0.59	0.32	0.75
STATEMN	-0.18	0.50	-0.36	0.72
STATENA	-0.71	0.88	-0.80	0.42
STATEOR	0.94	0.77	1.22	0.22
STATEWI	1.54	1.00	1.54	0.12
Q_07D1SWELL1	-1.70	0.81	-2.10	0.04
Q_07D1BLEED1	1.14	0.52	2.19	0.03
PMNO1	-1.10	0.77	-1.43	0.15
Q_09P1NUMB2	-1.26	0.69	-1.82	0.07
Q_09P1NUMB3	-19.36	8861.34	-0.00	1.00

RANDOM EFFECTS:

Group	Parameter	Std. Dev.
data_1wk.PRACID	(Intercept)	0.18
DDShispanic	(Intercept)	0.55
DDSgender	(Intercept)	0.01

Grouping variables:

Group	# groups	ICC
data_1wk.PRACID	57	0.01
DDShispanic	2	0.08
DDSgender	2	0.00

Model2: One-week pain prediction, clinical set

FIXED EFFECTS:

	Est.	S.E.	z val.	p
(Intercept)	-2.77	0.60	-4.65	0.00
Q04P0STRES1	1.52	0.51	3.00	0.00
Q04P0NWPAIN1	-1.68	0.70	-2.40	0.02
Q04P0NWPAIN2	-0.97	0.58	-1.67	0.10
Q04P0NWPAIN3	-0.32	0.52	-0.62	0.53

Q04P0NWPAIN4	0.45	0.46	0.98	0.33
Q04P0NWPAIN5	1.43	0.61	2.33	0.02
Q04P0NWPAIN6	0.21	0.56	0.38	0.70
Q04P0NWPAIN7	-0.65	0.69	-0.95	0.34
Q04P0NWPAIN8	1.40	0.74	1.91	0.06
Q04P0NWPAIN9	0.45	0.81	0.55	0.58
Q04P0NWPAIN10	1.27	1.09	1.17	0.24
Q04P0NOACDAY1	-0.13	0.65	-0.20	0.84
Q04P0NOACDAY2	1.58	0.62	2.56	0.01
Q04P0NOACDAY3	-1.08	0.89	-1.22	0.22
Q04P0NOACDAY4	0.04	1.07	0.03	0.97
Q04P0NOACDAY5	0.57	1.05	0.54	0.59
Q04P0NOREC1	0.29	0.67	0.43	0.66
Q04P0NOREC2	-1.66	0.97	-1.71	0.09
Q04P0NOREC3	-0.09	0.79	-0.12	0.91
Q04P0NOREC4	1.39	1.38	1.00	0.32
Q04P0NOREC5	1.85	1.03	1.80	0.07
Q04P0NOREC6	-0.03	1.18	-0.02	0.98
Q04P0NOREC7	-0.26	1.37	-0.19	0.85
Q04P0NOREC8	0.39	1.40	0.28	0.78
Q04P0NOREC9	17.91	3997.95	0.00	1.00
Q04P0NOWRK1	-0.57	0.65	-0.88	0.38
Q04P0NOWRK2	0.59	0.77	0.76	0.44
Q04P0NOWRK3	0.44	0.81	0.55	0.58
Q04P0NOWRK4	-1.68	1.40	-1.20	0.23
Q04P0NOWRK5	-1.59	1.18	-1.35	0.18
Q04P0NOWRK6	0.29	1.15	0.25	0.80
Q04P0NOWRK7	0.65	1.28	0.51	0.61
Q04P0NOWRK8	0.15	1.52	0.10	0.92
Q04P0NOWRK9	-36.26	6486.32	-0.01	1.00
Q_07D1BLEED1	1.14	0.48	2.36	0.02
STATEFL	0.46	0.56	0.82	0.41
STATEMN	0.27	0.46	0.58	0.56
STATENA	-0.11	0.86	-0.13	0.90
STATEOR	1.22	0.72	1.69	0.09
STATEWI	1.59	0.92	1.72	0.09

RANDOM EFFECTS:

Group	Parameter	Std. Dev.
data_1wk.PRACID	(Intercept)	0.37

Grouping variables:

Group	# groups	ICC
data_1wk.PRACID	57	0.04

Model3: One-week pain prediction, sociodemographic set

FIXED EFFECTS:

	Est.	S.E.	z val.	p
(Intercept)	-0.78	1.43	-0.54	0.59
Q04P0GENDERFemale	-0.06	0.26	-0.24	0.81
Q04P0HISPNon-hispanic	-0.57	0.68	-0.85	0.40
Q04P0DENINS1	0.22	0.34	0.63	0.53
Q04P0INCOME2	0.74	0.87	0.85	0.39
Q04P0INCOME3	0.64	0.83	0.77	0.44
Q04P0INCOME4	0.96	0.80	1.20	0.23
Q04P0EDU2	-0.46	0.84	-0.54	0.59
Q04P0EDU3	-0.79	0.83	-0.95	0.34
Q04P0EDU4	-0.16	0.82	-0.20	0.84
Q04P0EDU5	-0.66	0.87	-0.76	0.45
Q04P0RACEv2Black	0.06	0.56	0.11	0.91
Q04P0RACEv2Asian	1.02	0.98	1.05	0.30
Q04P0RACEv2Other	2.56	1.18	2.16	0.03
DecGrad3	-0.63	0.32	-1.97	0.05
DecGrad4	-0.42	0.45	-0.94	0.35
DecGrad5	-0.41	0.52	-0.78	0.43

RANDOM EFFECTS:

Group	Parameter	Std. Dev.
DDShispanic	(Intercept)	0.41
DDSgender	(Intercept)	0.43

Grouping variables:

Group	# groups	ICC
DDShispanic	2	0.05
DDSgender	2	0.05

Model4: Six-month pain prediction, combined set

FIXED EFFECTS:

	Est.	S.E.	z val.	p
(Intercept)	-5.28	1.14	-4.63	0.00
Q04P0OUTC2	1.08	0.32	3.36	0.00

Q04P0OUTC3	0.19	0.76	0.25	0.80
Q04P0TEMP1	0.79	0.33	2.35	0.02
Q04P0NWPAIN1	1.37	0.50	2.76	0.01
Q04P0NWPAIN2	-0.34	0.64	-0.53	0.60
Q04P0NWPAIN3	-0.43	0.66	-0.65	0.52
Q04P0NWPAIN4	-0.14	0.73	-0.19	0.85
Q04P0NWPAIN5	-0.75	0.72	-1.04	0.30
Q04P0NWPAIN6	0.03	0.66	0.04	0.97
Q04P0NWPAIN7	-1.50	0.96	-1.56	0.12
Q04P0NWPAIN8	-0.29	0.88	-0.33	0.74
Q04P0NWPAIN9	0.54	0.89	0.61	0.54
Q04P0NWPAIN10	0.46	0.55	0.83	0.41
Q04P0NOACDAY1	-0.62	0.76	-0.83	0.41
Q04P0NOACDAY2	1.66	0.53	3.12	0.00
Q04P0NOACDAY3	-0.78	0.91	-0.85	0.39
Q04P0NOACDAY4	-0.48	1.21	-0.40	0.69
Q04P0NOACDAY5	-0.04	1.00	-0.04	0.97
Q04P0INCOME2	1.05	0.87	1.21	0.23
Q04P0INCOME3	0.17	0.85	0.20	0.84
Q04P0INCOME4	-0.27	0.81	-0.34	0.73
Q05D0DEEPLOCdistal	-0.09	0.42	-0.21	0.84
Q05D0DEEPLOCmesial	0.63	0.39	1.60	0.11
STATEFL	0.38	0.74	0.51	0.61
STATEMN	0.77	0.64	1.21	0.23
STATENA	1.73	0.75	2.30	0.02
STATEOR	1.20	0.81	1.49	0.14
STATEWI	-0.26	1.46	-0.18	0.86
Q04P0RACEv22	0.10	0.86	0.11	0.91
Q04P0RACEv24	-1.10	1.24	-0.89	0.37
Q04P0RACEv26	1.49	0.76	1.96	0.05
Q11_P1NWPAIN1	0.49	0.55	0.89	0.37
Q11_P1NWPAIN2	0.93	0.64	1.45	0.15
Q11_P1NWPAIN3	1.77	0.63	2.79	0.01
Q11_P1NWPAIN4	0.53	1.00	0.53	0.59
Q11_P1NWPAIN5	3.73	0.75	4.95	0.00
Q11_P1NWPAIN6	-17.01	4272.31	-0.00	1.00
Q11_P1NWPAIN7	-0.94	0.86	-1.09	0.28
Q11_P1NWPAIN8	-0.88	0.79	-1.12	0.26
Q11_P1NODAC1	1.38	0.49	2.79	0.01
Q11_P1NODAC2	-16.45	2638.92	-0.01	1.00
Q11_P1NODAC3	1.32	0.69	1.93	0.05
Q11_P1NODAC4	1.81	1.00	1.82	0.07
Q11_P1NODAC5	1.06	0.96	1.11	0.27
Q11_P1NODAC6	-0.25	1.24	-0.20	0.84
Q11_P1NODAC7	2.01	0.77	2.60	0.01

RANDOM EFFECTS:

Group	Parameter	Std. Dev.

data_6m_00.PRACID	(Intercept)	0.13
DecGrad	(Intercept)	0.00

Grouping variables:

Group	# groups	ICC
data_6m_00.PRACID	59	0.00
DecGrad	4	0.00

Model5: Six-month pain, clinical set

FIXED EFFECTS:

	Est.	S.E.	z val.	p
(Intercept)	-6.68	1.17	-5.73	0.00
Q04P0OUTC2	0.95	0.35	2.70	0.01
Q04P0OUTC3	0.18	0.87	0.21	0.84
Q04P0TEMP1	0.75	0.37	2.05	0.04
Q04P0NWPAIN1	1.27	0.54	2.35	0.02
Q04P0NWPAIN2	-0.73	0.70	-1.03	0.30
Q04P0NWPAIN3	-1.14	0.75	-1.54	0.12
Q04P0NWPAIN4	-0.39	0.84	-0.46	0.65
Q04P0NWPAIN5	-1.10	0.80	-1.38	0.17
Q04P0NWPAIN6	-0.30	0.72	-0.41	0.68
Q04P0NWPAIN7	-2.31	1.14	-2.02	0.04
Q04P0NWPAIN8	-1.05	0.98	-1.07	0.28
Q04P0NWPAIN9	0.29	0.96	0.31	0.76
Q04P0NWPAIN10	0.76	0.62	1.22	0.22
Q04P0WSTPAIN1	0.18	1.03	0.17	0.87
Q04P0WSTPAIN10	0.15	0.79	0.19	0.85
Q04P0WSTPAIN2	1.20	0.85	1.41	0.16
Q04P0WSTPAIN3	0.83	0.94	0.88	0.38
Q04P0WSTPAIN4	1.58	1.00	1.57	0.12
Q04P0WSTPAIN5	0.89	0.92	0.96	0.33
Q04P0WSTPAIN6	0.68	0.87	0.78	0.44
Q04P0WSTPAIN7	0.50	0.79	0.63	0.53
Q04P0WSTPAIN8	1.04	0.78	1.33	0.18
Q04P0WSTPAIN9	1.82	0.80	2.28	0.02
Q04P0NOACDAY1	-0.35	0.74	-0.47	0.64
Q04P0NOACDAY2	1.60	0.63	2.52	0.01
Q04P0NOACDAY3	-0.65	0.91	-0.71	0.48
Q04P0NOACDAY4	0.02	1.54	0.01	0.99
Q04P0NOACDAY5	0.67	1.08	0.62	0.54
Q05D0DEEPLOCdistal	-0.36	0.46	-0.80	0.43
Q05D0DEEPLOCmesial	0.46	0.43	1.07	0.28

STATEFL	0.44	0.83	0.53	0.60
STATEMN	0.76	0.74	1.03	0.30
STATENA	2.11	0.84	2.50	0.01
STATEOR	1.15	0.89	1.30	0.19
STATEWI	-1.92	1.84	-1.05	0.29
Q11_P1NWPAIN1	-0.25	0.67	-0.37	0.71
Q11_P1NWPAIN2	1.10	0.71	1.54	0.12
Q11_P1NWPAIN3	1.75	0.76	2.31	0.02
Q11_P1NWPAIN4	1.30	1.09	1.19	0.23
Q11_P1NWPAIN5	3.62	0.91	4.00	0.00
Q11_P1NWPAIN6	-20.64	9012.90	-0.00	1.00
Q11_P1NWPAIN7	1.17	1.09	1.07	0.28
Q11_P1NWPAIN8	1.16	1.07	1.08	0.28
Q11_P1WSTPAIN1	1.71	0.86	1.98	0.05
Q11_P1WSTPAIN10	0.23	1.22	0.19	0.85
Q11_P1WSTPAIN2	1.25	0.89	1.40	0.16
Q11_P1WSTPAIN3	0.42	0.89	0.47	0.64
Q11_P1WSTPAIN4	0.15	0.91	0.16	0.87
Q11_P1WSTPAIN5	1.85	0.83	2.22	0.03
Q11_P1WSTPAIN6	2.19	0.99	2.20	0.03
Q11_P1WSTPAIN7	2.42	0.86	2.80	0.01
Q11_P1WSTPAIN8	1.51	1.01	1.49	0.14
Q11_P1WSTPAIN9	-0.27	1.06	-0.25	0.80
Q11_P1NODAC1	1.05	0.63	1.66	0.10
Q11_P1NODAC2	-20.34	9425.71	-0.00	1.00
Q11_P1NODAC3	0.71	0.97	0.73	0.47
Q11_P1NODAC4	1.57	1.34	1.17	0.24
Q11_P1NODAC5	0.27	1.43	0.19	0.85
Q11_P1NODAC6	-2.24	1.55	-1.45	0.15
Q11_P1NODAC7	3.89	1.11	3.51	0.00
Q11_P1NOWRK1	0.44	0.70	0.63	0.53
Q11_P1NOWRK2	0.72	1.12	0.64	0.52
Q11_P1NOWRK3	0.45	0.87	0.52	0.60
Q11_P1NOWRK4	-17.57	4297.49	-0.00	1.00
Q11_P1NOWRK5	0.88	1.04	0.85	0.40
Q11_P1NOWRK6	-1.04	2.16	-0.48	0.63
Q11_P1NOWRK7	-2.64	1.35	-1.95	0.05

RANDOM EFFECTS:

Group	Parameter	Std. Dev.
data_6m_00.PRACID	(Intercept)	0.29
DecGrad	(Intercept)	0.00

Grouping variables:

Group	# groups	ICC
data_6m_00.PRACID	59	0.02

DecGrad	4	0.00

Model6: Six-month pain, sociodemographic set

FIXED EFFECTS:

	Est.	S.E.	z val.	p
(Intercept)	-3.44	1.37	-2.52	0.01
Q04P0GENDER2	0.24	0.28	0.87	0.39
Q04P0HISP2	0.23	0.61	0.38	0.70
Q04P0DENINS1	0.20	0.35	0.56	0.58
Q04P0INCOME2	0.63	0.71	0.89	0.37
Q04P0INCOME3	-0.13	0.71	-0.18	0.86
Q04P0INCOME4	-0.28	0.68	-0.41	0.68
Q04P0EDU2	0.83	1.10	0.75	0.45
Q04P0EDU3	0.65	1.09	0.60	0.55
Q04P0EDU4	1.14	1.08	1.05	0.29
Q04P0EDU5	0.83	1.13	0.74	0.46
Q04P0RACEv22	-0.64	0.76	-0.84	0.40
Q04P0RACEv24	-0.63	1.08	-0.58	0.56
Q04P0RACEv26	1.40	0.61	2.30	0.02

RANDOM EFFECTS:

Group	Parameter	Std. Dev.
DecGrad	(Intercept)	0.22
DDShispanic	(Intercept)	0.00
DDSgender	(Intercept)	0.00

Grouping variables:

Group	# groups	ICC
DecGrad	4	0.01
DDShispanic	2	0.00
DDSgender	2	0.00

Chapter 7 Supplement 7. TRIPOD checklist for study reporting

Section/Topic	Item	Development or Validation?	Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	1
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	1-2
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model, or both.	1-2
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation datasets, if applicable.	9
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	9
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	9
	5b	D;V	Describe eligibility criteria for participants.	9-10
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	10
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured.	10
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	NA
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	11
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	10
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	10-11
	10c	V	For validation, describe how the predictions were calculated.	NA
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	11-12
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	NA
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	2-3
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	3-5
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors, and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
	15b	D	Explain how to use the prediction model.	7-8
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	6-7
Model updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	9
Interpretations	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	13-14
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and datasets.	12
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	12

TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.

References

Abernathy, J. R., Graves, R. C., Bohannan, H. M., Stamm, J. W., Greenberg, B. G., and Disney, J. A. (1987). Development and application of a prediction model for dental caries. *Community Dentistry and Oral Epidemiology*, 15(1):24–28.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., et al. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, 34(3):341–382.

Ahmad, M. A., Eckert, C., and Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560.

Ahmed, I., Debray, T. P., Moons, K. G., and Riley, R. D. (2014). Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Medical Research Methodology*, 14:3.

Ahrens, A., Hansen, C. B., and Schaffer, M. E. (2020). lassopack: Model selection and prediction with regularized regression in stata. *The Stata Journal*, 20(1):176–235.

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P., McGinn, T., and Guyatt, G. (2017). Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *Journal of American Medical Association*, 318(14):1377–1384.

Albandar, J. M. (2002). Global risk factors and risk indicators for periodontal diseases. *Periodontology 2000*, 29:177–206.

Allison, P. (2012). Handling missing data by maximum likelihood. sas global forum statistics and data analysis.

Almeida, A. M. d., Castel-Branco, M. M., and Falcao, A. (2002). Linear regression for calibration lines revisited: weighting schemes for bioanalytical methods. *Journal of Chromatography B*, 774(2):215–222.

Alonso, B. F., Nixdorf, D. R., Shueb, S. S., John, M. T., Law, A. S., and Durham, J. (2017). Examining the sensitivity and specificity of 2 screening instruments: Odontogenic or temporomandibular disorder pain? *Journal of Endodontics*, 43(1):36–45.

Altman, D. G. (2009). Prognostic models: a methodological framework and review of models for breast cancer. *Cancer Investment*, 27(3):235–43.

Altman, D. G. and Lyman, G. H. (1998). Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research and Treatment*, 52(1-3):289–303.

Altman, D. G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, 19(4):453–473.

Altman, D. G., Vergouwe, Y., Royston, P., and Moons, K. G. (2009). Prognosis and prognostic research: validating a prognostic model. *British Medical Journal*, 338:b605.

American Academy of Periodontology, . (2008). American academy of periodontology statement on risk assessment. *Journal of Periodontology*, 79(2):202.

American Cancer Society, . (2021). Key statistics for oral cavity and oropharyngeal cancers @https://www.cancer.org/cancer/oral-cavity-and-oropharyngeal-cancer/about/key-statistics.html. Accessed on: 24-4-2021.

Anderson, K. M., Odell, P. M., Wilson, P. W., and Kannel, W. B. (1991). Cardiovascular disease risk profiles. *American Heart Journal*, 121(1):293–298.

Arias, A., de la Macorra, J. C., Hidalgo, J. J., and Azabal, M. (2013). Predictive models of pain following root canal treatment: a prospective clinical study. *International Endodontic Journal*, 46(8):784–793.

Aromataris, E. and Munn, Z. (2017). Joanna briggs institute reviewers' manual.

Austin, P. C., Pencinca, M. J., and Steyerberg, E. W. (2017). Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the cox proportional hazards regression model. *Statistical Methods in Medical Research*, 26(3):1053–1077.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452.

Banerjee, P., Dehnbostel, F. O., and Preissner, R. (2018). Prediction is a balancing act: importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets. *Frontiers in Chemistry*, 6:362.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Columbia Law Review*, 104:671.

Bartlett, J. W., Seaman, S. R., White, I. R., and Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4):462–487.

Baselli, G., Codari, M., and Sardanelli, F. (2020). Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? *European Radiology Experimental*, 4:1–7.

Beck, J. D. (1994). Methods of assessing risk for periodontitis and developing multifactorial models. *Journal of Periodontology*, 65(5 Suppl):468–478.

Bedogni, G. (2009). Clinical prediction models-a practical approach to development, validation and updating. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 172:944–944.

Beek, M. v. d., Hoeksma, J., and Prahl-Andersen, B. (1991). Vertical facial growth: a longitudinal study from 7 to 14 years of age. *The European Journal of Orthodontics*, 13(3):202–208.

Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science. *Circulation Research*, 116(1):116–126.

Benjamin, R. (2019). Assessing risk, automating racism. *Science*, 366(6464):421–422.

Beran, D., Chappuis, F., Damasceno, A., Jha, N., Pesantes, M. A., Singh, S. B., Somerville, C., Suggs, L. S., and Miranda, J. J. (2019). High-quality health systems: time for a revolution in research and research funding. *The Lancet Global Health*, 7(3):e303–e304.

Bersani, C., Mints, M., Tertipis, N., Haeggblom, L., Sivars, L., Ahrlund-Richter, A., Vlastos, A., Smedberg, C., Grun, N., Munck-Wikland, E., Nasman, A., Ramqvist, T., and Dalianis, T. (2017). A model using concomitant markers for predicting outcome in human papillomavirus positive oropharyngeal cancer. *Oral Oncology*, 68:53–59.

Bouwmeester, W., Zuithoff, N. P., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E. W., Altman, D. G., and Moons, K. G. (2012). Reporting and methods in clinical prediction research: a systematic review. *PLoS Medicine*, 9(5):1–12.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57.

Brier, G. (01 Jan. 1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The binormal assumption on precision-recall curves. In *2010 20th International Conference on Pattern Recognition*, pages 4263–4266. IEEE.

Burns, P. B., Rohrich, R. J., and Chung, K. C. (2011). The levels of evidence and their role in evidence-based medicine. *Plastic and Reconstructive Surgery*, 128(1):305.

Bøvelstad, H. M. and Borgan, (2011). Assessment of evaluation criteria for survival prediction from genomic data. *Biometrical Journal*, 53(2):202–216.

Canullo, L., Radovanovic, S., Delibasic, B., Blaya, J. A., Penarrocha, D., and Rakic, M. (2017). The predictive value of microbiological findings on teeth, internal and external implant portions in clinical decision making. *Clinical Oral Implants Research*, 28(5):512–519.

Canullo, L., Tallarico, M., Radovanovic, S., Delibasic, B., Covani, U., and Rakic, M. (2016). Distinguishing predictive profiles for patient-based risk assessment and diagnostics of plaque induced, surgically and prosthetically triggered peri-implantitis. *Clinical Oral Implants Research*, 27(10):1243–1250.

Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gulmezoglu, A. M., Howells, D. W., Ioannidis, J. P. A., and Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet*, 383(9912):156–165.

Chang, S.-W., Abdul-Kareem, S., Merican, A. F., and Zain, R. B. (2013). Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC Bioinformatics*, 14(1):1–15.

Chen, H.-C., Kodell, R. L., Cheng, K. F., and Chen, J. J. (2012). Assessment of performance of survival prediction models for cancer prognosis. *BMC Medical Research Methodology*, 12(1):102.

Chen, J. H. and Asch, S. M. (2017). Machine learning and prediction in medicine - beyond the peak of inflated expectations. *New England Journal of Medicine*, 376(26):2507–2509.

Chen, Q., Wu, J., Li, S., Lyu, P., Wang, Y., and Li, M. (2016). An ontology-driven, case-based clinical decision support model for removable partial denture design. *Scientific Reports*, 6(1):1–8.

Chen, R. and Snyder, M. (2013). Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(1):73–82.

Choi, J., Dekkers, O. M., and le Cessie, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*, 34(1):23–36.

Choodari-Oskooei, B., Royston, P., and Parmar, M. K. (2012). A simulation study of predictive ability measures in a survival model i: explained variation measures. *Statistics in Medicine*, 31(23):2627–2643.

Chouldechova, A. and Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L. M., Moons, K. G., and Altman, D. G. (2014). External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, 14.

Collins, G. S., Mallett, S., Omar, O., and Yu, L. M. (2011). Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Medicine*, 9:103.

Collins, G. S., Omar, O., Shanyinde, M., and Yu, L. M. (2013). A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *Journal of Clinical Epidemiology*, 66(3):268–277.

Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Annals of Internal Medicine*, 162(1):55–63.

Counsell, C. and Dennis, M. (2001). Systematic review of prognostic models in patients with acute stroke. *Cerebrovascular Diseases*, 12(3):159–70.

Damato, B. and Taktak, A. (2007). Chapter 2 - survival after treatment of intraocular melanoma. In Taktak, A. F. and Fisher, A. C., editors, *Outcome Prediction in Cancer*, pages 27–41. Elsevier, Amsterdam.

Damen, J. A., Hooft, L., Schuit, E., Debray, T. P., Collins, G. S., Tzoulaki, I., Lassale, C. M., Siontis, G. C., Chiocchia, V., Roberts, C., Schlussel, M. M., Gerry, S., Black, J. A.,

Heus, P., van der Schouw, Y. T., Peelen, L. M., and Moons, K. G. (2016). Prediction models for cardiovascular disease risk in the general population: systematic review. *British Medical Journal*, 353:i2416.

Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1):1–25.

David, C. R. et al. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34(2):187–220.

Dickerman, B. A. and Hernán, M. A. (2020). Counterfactual prediction is not only for causal inference. *European Journal of Epidemiology*, 35(7):615–617.

Dima, S., Wang, K.-J., Chen, K.-H., Huang, Y.-K., Chang, W.-J., Lee, S.-Y., and Teng, N.-C. (2018). Decision tree approach to the impact of parents' oral health on dental caries experience in children: a cross-sectional study. *International Journal of Environmental Research and Public Health*, 15(4):692.

Du, M., Bo, T., Kapellas, K., and Peres, M. A. (2018). Prediction models for the incidence and progression of periodontitis: A systematic review. *Journal of Clinical Periodontology*, 45(12):1408–1420.

Du, M., Haag, D., Song, Y., Lynch, J., and Mittinty, M. (2020). Examining bias and reporting in oral health prediction modeling studies. *Journal of Dental Research*, 99(4):374–387.

Du, M., Nair, R., Jamieson, L., Liu, Z., and Bi, P. (2019). Incidence trends of lip, oral cavity, and pharyngeal cancers: Global burden of disease 1990–2017. *Journal of Dental Research*, 99(2):143–151.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Edwards, R. R., Doleys, D. M., Fillingim, R. B., and Lowery, D. (2001). Ethnic differences in pain tolerance: clinical implications in a chronic pain population. *Psychosomatic Medicine*, 63(2):316–23.

Efron, B. (1988). Logistic regression, survival analysis, and the kaplan-meier curve. *Journal of the American statistical Association*, 83(402):414–425.

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115(530):636–655.

Eke, P. I., Dye, B. A., Wei, L., Slade, G. D., Thornton-Evans, G. O., Borgnakke, W. S., Taylor, G. W., Page, R. C., Beck, J. D., and Genco, R. J. (2015). Update on prevalence of periodontitis in adults in the united states: Nhanes 2009 to 2012. *Journal of Periodontology*, 86(5):611–22.

Eke, P. I., Page, R. C., Wei, L., Thornton-Evans, G., and Genco, R. J. (2012). Update of the case definitions for population-based surveillance of periodontitis. *Journal of Periodontology*, 83(12):1449–1454.

Eke, P. I., Zhang, X., Lu, H., Wei, L., Thornton-Evans, G., Greenlund, K. J., Holt, J. B., and Croft, J. B. (2016). Predicting periodontitis at state and local levels in the united states. *Journal of Dental Research*, 95(5):515–22.

Escarce, J. J., Morales, L. S., and Rumbaut, R. G. (2006). The health status and health behaviors of hispanics.

Feres, M., Louzoun, Y., Haber, S., Faveri, M., Figueiredo, L. C., and Levin, L. (2018). Support vector machine-based differentiation between aggressive and chronic periodontitis using microbial profiles. *International Dental Journal*, 68(1):39–46.

Fletcher Mercaldo, S. and Blume, J. D. (2020). Missing data and prediction: the pattern submodel. *Biostatistics (Oxford, England)*, 21(2):236–252.

Fouodo, C. J., König, I. R., Weihs, C., Ziegler, A., and Wright, M. N. (2018). Support vector machines for survival analysis with r. *R Journal*, 10(1).

Frérot, M., Lefebvre, A., Aho, S., Callier, P., Astruc, K., and Aho Glélé, L. S. (2018). What is epidemiology? changing definitions of epidemiology 1978-2017. *PLoS One*, 13(12):1–27.

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., and Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492.

Geersing, G.-J., Bouwmeester, W., Zuithoff, P., Spijker, R., Leeflang, M., and Moons, K. (2012). Search filters for finding prognostic and diagnostic prediction studies in medline to enhance systematic reviews. *PloS One*, 7(2):e32844.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435.

Gerds, T. A., Cai, T., and Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(4):457–479.

Giannobile, W. V., Beikler, T., Kinney, J. S., Ramseier, C. A., Morelli, T., and Wong, D. T. (2009). Saliva as a diagnostic tool for periodontal disease: current state and future directions. *Periodontology 2000*, 50:52–64.

Goldstein, B. A., Navar, A. M., and Carter, R. E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European Heart Journal*, 38(23):1805–1814.

Gradella, C. M., Bernabe, E., Bonecker, M., and Oliveira, L. B. (2011). Caries prevalence and severity, and quality of life in brazilian 2- to 4-year-old children. *Community Dentistry and Oral Epidemiology*, 39(6):498–504.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Greenhalgh, T. (1997). How to read a paper: papers that report diagnostic or screening tests. *British Medical Journal*, 315(7107):540–543.

Greenland, S., Daniel, R., and Pearce, N. (2016). Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *International Journal of Epidemiology*, 45(2):565–575.

Grol, R., Baker, R., and Moss, F. (2002). Quality improvement research: understanding the science of change in health care. *BMJ Quality & Safety*, 11(2):110–111.

Gursoy, U. K., Kononen, E., Pussinen, P. J., Tervahartiala, T., Hyvarinen, K., Suominen, A. L., Uitto, V. J., Paju, S., and Sorsa, T. (2011). Use of host- and bacteria-derived salivary markers in detection of periodontitis: A cumulative approach. *Disease Markers*, 30(6):299–305.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123.

Hankey, B. F., Ries, L. A., and Edwards, B. K. (1999). The surveillance, epidemiology, and end results program: a national resource. *Cancer Epidemiology and Prevention Biomarkers*, 8(12):1117–1121.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Harrell, F. (2020). Road map for choosing between statistical modeling and machine learning. https://www.fharrell.com/post/stat-ml/ Accessed on 05-02-2021.

Harrell, F. E., J., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–87.

Harrell, Frank E., J., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of American Medical Association*, 247(18):2543–2546.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hayden, J. A., Côté, P., and Bombardier, C. (2006). Evaluation of the quality of prognosis studies in systematic reviews. *Annals of Internal Medicine*, 144(6):427–437.

Hayden, J. A., Cote, P., Steenstra, I. A., Bombardier, C., and Group, Q.-L. W. (2008). Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. *Journal of Clinical Epidemiology*, 61(6):552–60.

Hayden, J. A., van der Windt, D. A., Cartwright, J. L., Côté, P., and Bombardier, C. (2013). Assessing bias in studies of prognostic factors. *Annals of Internal Medicine*, 158(4):280–286.

Heitz-Mayfield, L. J. (2005). Disease progression: identification of high-risk groups and individuals for periodontitis. *Journal of Clinical Periodontology*, 32 Suppl 6:196–209.

Hemingway, H., Croft, P., Perel, P., Hayden, J. A., Abrams, K., Timmis, A., Briggs, A., Udumyan, R., Moons, K. G., Steyerberg, E. W., et al. (2013). Prognosis research strategy (progress) 1: a framework for researching clinical outcomes. *British Medical Journal*, 346.

Hernán, M. A., Hsu, J., and Healy, B. (2019). A second chance to get causal inference right: a classification of data science tasks. *Chance*, 32(1):42–49.

Hernán, M. A. and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764.

Higgins, J. and Green, S. (2011). Cochrane handbook of systematic reviews of intervention.

Hill, N., Frappier-Davignon, L., and Morrison, B. (1979). The periodic health examination. *Canadian Medical Association Journal*, 121:1193–254.

Hingorani, A. D., van der Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G., Steyerberg, E. W., Schroter, S., Sauerbrei, W., Altman, D. G., and Hemingway, H. (2013). Prognosis research strategy (progress) 4: stratified medicine research. *British Medical Journal*, 346:e5793.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., and Brindle, P. (2008). Predicting cardiovascular risk in england and wales: prospective derivation and validation of qrisk2. *British Medical Journal*, 336(7659):1475–1482.

Holtfreter, B., Albandar, J. M., Dietrich, T., Dye, B. A., Eaton, K. A., Eke, P. I., Papapanou, P. N., Kocher, T., and Joint, E. U. U. S. A. P. E. W. G. (2015). Standards for reporting chronic periodontitis prevalence and severity in epidemiologic studies: Proposed standards from the joint eu/usa periodontal epidemiology working group. *Journal of Clinical Periodontology*, 42(5):407–12.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Howe, C. J., Cole, S. R., Lau, B., Napravnik, S., and Eron Jr, J. J. (2016). Selection bias due to loss to follow up in cohort studies. *Epidemiology (Cambridge, Mass.)*, 27(1):91.

Hughes, R. A., Heron, J., Sterne, J. A., and Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology*, 48(4):1294–1304.

Ingui, B. J. and Rogers, M. A. (2001). Searching for clinical prediction rules in medline. *Journal of the American Medical Informatics Association*, 8(4):391–397.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

Ishwaran, H., Kogalur, U. B., and Kogalur, M. U. B. (2020). Package 'randomforestsrc'.

Jaja, B. N. R., Cusimano, M. D., Etminan, N., Hanggi, D., Hasan, D., Ilodigwe, D., Lantigua, H., Le Roux, P., Lo, B., Louffat-Olivares, A., Mayer, S., Molyneux, A., Quinn, A., Schweizer, T. A., Schenk, T., Spears, J., Todd, M., Torner, J., Vergouwen, M. D. I., Wong, G. K. C., Singh, J., and Macdonald, R. L. (2013). Clinical prediction models for aneurysmal subarachnoid hemorrhage: A systematic review. *Neurocritical Care*, 18(1):143–153.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., 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. *The Lancet*, 392(10159):1789–1858.

Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E., and Vergouwe, Y. (2008). Updating methods improved the performance of a clinical prediction model in new patients. *Journal of Clinical Epidemiology*, 61(1):76–86.

Jover-Espla, A. G., Palazon-Bru, A., Folgado-de la Rosa, D. M., Severa-Ferrandiz, G., Sancho-Mestre, M., de Juan-Herrero, J., and Gil-Guillen, V. F. (2018). A predictive model for recurrence in patients with glottic cancer implemented in a mobile application for android. *Oral Oncology*, 80:82–88.

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent roc curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, 17(1):53.

Kassebaum, N., Smith, A., Bernabé, E., Fleming, T., Reynolds, A., Vos, T., Murray, C., Marcenes, W., and Collaborators, G. . O. H. (2017). Global, regional, and national prevalence, incidence, and disability-adjusted life years for oral conditions for 195 countries, 1990–2015: a systematic analysis for the global burden of diseases, injuries, and risk factors. *Journal of Dental Research*, 96(4):380–387.

Kattan, M. W., Yu, C. H., Stephenson, A. J., Sartor, O., and Tombal, B. (2013). Clinicians versus nomogram: Predicting future technetium-99m bone scan positivity in patients with rising prostate-specific antigen after radical prostatectomy for prostate cancer. *Urology*, 81(5):956–961.

Kayaoglu, G., Gürel, M., Saricam, E., Ilhan, M. N., and Ilk, O. (2016). Predictive model of intraoperative pain during endodontic treatment: prospective observational clinical study. *Journal of Endodontics*, 42(1):36–41.

Kehlet, H., Jensen, T. S., and Woolf, C. J. (2006). Persistent postsurgical pain: risk factors and prevention. *The Lancet*, 367(9522):1618–25.

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., and Altman, D. G. (2012). Improving bioscience research reporting: the arrive guidelines for reporting animal research. *Osteoarthritis Cartilage*, 20(4):256–60.

Kim, D. W., Lee, S., Kwon, S., Nam, W., Cha, I.-H., and Kim, H. J. (2019). Deep learning-based survival prediction of oral cancer patients. *Scientific Reports*, 9(1):1–10.

Kim, M. (2007). An effective under-sampling method for class imbalance data problem. In *Proceedings of the 8th Symposium on Advanced Intelligent Systems*, pages 825–829.

Kioi, M. (2017). Recent advances in molecular-targeted therapy for oral cancer. *International Journal of Oral and Maxillofacial Surgery*, 46:27.

Krois, J., Graetz, C., Holtfreter, B., Brinkmann, P., Kocher, T., and Schwendicke, F. (2019). Evaluating modeling and validation strategies for tooth loss. *Journal of Dental Research*, 98(10):1088–1095.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Kundu, S., Mazumdar, M., and Ferket, B. (2017). Impact of correlation of predictors on discrimination of risk models in development and external populations. *BMC Medical Research Methodology*, 17(1):63.

Kvamme, H., Borgan, Ø., and Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30.

Kye, W., Davidson, R., Martin, J., and Engebretson, S. (2012). Current status of periodontal risk assessment. *ournal of Evidence-Based Dental Practice*, 12(3 Suppl):2–11.

Lai, H., Su, C.-W., Yen, A. M.-F., Chiu, S. Y.-H., Fann, J. C.-Y., Wu, W. Y.-Y., Chuang, S.-L., Liu, H.-C., Chen, H.-H., and Chen, L.-S. (2015). A prediction model for periodontal disease: Modelling and validation from a national survey of 4061 taiwanese adults. *Journal of Clinical Periodontology*, 42(5):413–421.

Lang, N. P., Suvan, J. E., and Tonetti, M. S. (2015). Risk factor assessment tools for the prevention of periodontitis progression a systematic review. *Journal of Clinical Periodontology*, 42(Suppl 16):S59–70.

Lash, T. L., Fox, M. P., MacLehose, R. F., Maldonado, G., McCandless, L. C., and Greenland, S. (2014). Good practices for quantitative bias analysis. *International Journal of Epidemiology*, 43(6):1969–85.

Law, A. S., Nixdorf, D. R., Aguirre, A. M., Reams, G. J., Tortomasi, A. J., Manne, B. D., Harris, D. R., and National Dental, P. C. G. (2015). Predicting severe pain after root canal therapy in the national dental pbrn. *Journal of Dental Research*, 94(3 Suppl):37S–43S.

Law, A. S., Nixdorf, D. R., Rabinowitz, I., Reams, G. J., Smith, J. A., J., Torres, A. V., and Harris, D. R. (2014). Root canal therapy reduces multiple dimensions of pain: a national dental practice-based research network study. *Journal of Endodontics*, 40(11):1738–45.

Lee, J. H., Jeong, S. N., and Choi, S. H. (2018a). Predictive data mining for diagnosing periodontal disease: the korea national health and nutrition examination surveys (knhanes v and vi) from 2010 to 2015. *Journal of Public Health Dentistry*.

Lee, J.-H., Kim, D.-h., Jeong, S.-N., and Choi, S.-H. (2018b). Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. *Journal of Periodontal & Implant Science*, 48(2):114.

Lee, K. J., Tilling, K. M., Cornish, R. P., Little, R. J., Bell, M. L., Goetghebeur, E., Hogan, J. W., Carpenter, J. R., et al. (2021). Framework for the treatment and reporting of missing data in observational studies: The treatment and reporting of missing data in observational studies framework. *Journal of Clinical Epidemiology*, 134:79–88.

Lee, Y.-h., Bang, H., and Kim, D. J. (2016). How to establish clinical prediction models. *Endocrinology and Metabolism*, 31(1):38.

Leite, F. R. M., Peres, K. G., Do, L. G., Demarco, F. F., and Peres, M. A. A. (2017). Prediction of periodontitis occurrence: Influence of classification and sociodemographic and general health information. *Journal of Periodontology*, 88(8):731–743.

Leushuis, E., van der Steeg, J. W., Steures, P., Bossuyt, P. M., Eijkemans, M. J., van der Veen, F., Mol, B. W., and Hompes, P. G. (2009). Prediction models in reproductive medicine: a critical appraisal. *Human Reproduction Update*, 15(5):537–52.

Li, G. and Wang, X. (2019). Prediction accuracy measures for a nonlinear model and for right-censored time-to-event data. *Journal of the American Statistical Association*, 114(528):1815–1825.

Lindskog, S., Blomlof, J., Persson, I., Niklason, A., Hedin, A., Ericsson, L., Ericsson, M., Jarncrantz, B., Palo, U., Tellefsen, G., Zetterstrom, O., and Blomlof, L. (2010). Validation of an algorithm for chronic periodontitis risk assessment and prognostication: risk predictors, explanatory values, measures of quality, and clinical use. *Journal of Periodontology*, 81(4):584–93.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360.

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110:63–73.

Marteau, T. M. and Weinman, J. (2006). Self-regulation and the behavioural response to dna risk information: a theoretical analysis and framework for future research. *Social Science & Medicine*, 62(6):1360–1368.

Martinez-Canut, P., Alcaraz, J., Alcaraz, J., J., Alvarez-Novoa, P., Alvarez-Novoa, C., Marcos, A., Noguerol, B., Noguerol, F., and Zabalegui, I. (2018). Introduction of a prediction model to assigning periodontal prognosis based on survival time. *Journal of Clinical Periodontology*, 45(1):46–55.

Martinez-Canut, P. and Llobell, A. (2018). A comprehensive approach to assigning periodontal prognosis. *Journal of Clinical Periodontology*, 45(4):431–439.

Mason, W. (2001). Statistical analysis: Multilevel methods. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of the Social Behavioral Sciences*, pages 14988–14994. Pergamon, Oxford.

McNutt, M. (2014). Journals unite for reproducibility. *Science*, 346(6210):679–679.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Michos, E. D., Vasamreddy, C. R., Becker, D. M., Yanek, L. R., Moy, T. F., Fishman, E. K., Becker, L. C., and Blumenthal, R. S. (2005). Women with a low framingham risk score and a family history of premature coronary heart disease have a high prevalence of subclinical coronary atherosclerosis. *American Heart Journal*, 150(6):1276–1281.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Moerbeek, M., van Breukelen, G. J., and Berger, M. P. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, 56(4):341–350.

Mogensen, UB; Ishwaran, H. G. T. (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50(11):1–23.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *British Medical Journal*, 339:b2535.

Moons, K. G., de Groot, J. A., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., and Collins, G. S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the charms checklist. *PLoS Medicine*, 11(10).

Moons, K. G., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., and Grobbee, D. E. (2012). Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*, 98(9):683–90.

Moons, K. G. M. and Grobbee, D. E. (2002). When should we remain blind and when should our eyes remain open in diagnostic studies? *Journal of Clinical Epidemiology*, 55(7):633–636.

Moons, K. G. M., Royston, P., Vergouwe, Y., Grobbee, D. E., and Altman, D. G. (2009). Prognosis and prognostic research: what, why, and how? *British Medical Journal*, 338.

Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., and Mallett, S. (2019). Probast: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*, 170(1):W1–W33.

Morelli, T., Moss, K. L., Preisser, J. S., Beck, J. D., Divaris, K., Wu, D., and Offenbacher, S. (2018). Periodontal profile classes predict periodontal disease progression and tooth loss. *Journal of Periodontology*, 89(2):148–156.

Munafò, M. (2019). Raising research quality will require collective action. *Nature*, 576(7786):183–183.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Cambridge, MA.

Nasejje, J. B., Mwambi, H., Dheda, K., and Lesosky, M. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Medical Research Methodology*, 17(1):115.

Nattino, G., Finazzi, S., and Bertolini, G. (2016). A new test and graphical tool to assess the goodness of fit of logistic regression models. *Statistics in Medicine*, 35(5):709–720.

Norregaard, C., Gronhoj, C., Jensen, D., Friborg, J., Andersen, E., and von Buchwald, C. (2018). Cause-specific mortality in hpv+ and hpv- oropharyngeal cancer patients: insights from a population-based cohort. *Cancer Medicine*, 7(1):87–94.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13):1216.

OuYang, P. Y., Zhang, L. N., Xiao, Y., Lan, X. W., Zhang, X. M., Ma, J., and Xie, F. Y. (2017). Validation of published nomograms and accordingly individualized induction chemotherapy in nasopharyngeal carcinoma. *Oral Oncology*, 67:37–45.

Ozden, F., Ozgonenel, O., Ozden, B., and Aydogdu, A. (2015). Diagnosis of periodontal diseases using different classification algorithms: a preliminary study. *Nigerian Journal of Clinical Practice*, 18(3):416–421.

Page, R. Martin, J. (2007). Quantification of periodontal risk and disease severity and extent using the oral health information suite (ohis). *PERIO - Periodontal Practice Today*, 4(3):17.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal*, 372.

Page, R. C. and Beck, J. D. (1997). Risk assessment for periodontal diseases. *International Dental Journal*, 47(2):61–87.

Page, R. C. and Eke, P. I. (2007). Case definitions for use in population-based surveillance of periodontitis. *Journal of Periodontology*, 78:1387–1399.

Page, R. C., Krall, E. A., Martin, J., Mancl, L., and Garcia, R. I. (2002). Validity and accuracy of a risk calculator in predicting periodontal disease. *Journal of American Dental Association*, 133(5):569–76.

Pak, J. G. and White, S. N. (2011). Pain prevalence and severity before, during, and after root canal treatment: a systematic review. *Journal of Endodontics*, 37(4):429–38.

Papantonopoulos, G., Gogos, C., Housos, E., Bountis, T., and Loos, B. G. (2017). Prediction of individual implant bone levels and the existence of implant "phenotypes". *Clinical Oral Implants Research*, 28(7):823–832.

Parker, K., Menasce, J., Morin, R., and Lopez, M. (2015). Chapter 7: The many dimensions of hispanic racial identity. *Pew Research Center*.

Patton, L. L. (2017). At the interface of medicine and dentistry: shared decision-making using decision aids and clinical decision support tools. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, 123(2):147–149.

Paulus, J. K. and Kent, D. M. (2017). Race and ethnicity: A part of the equation for personalized clinical decision making? *Circulation-Cardiovascular Quality Outcomes*, 10(7).

Peissig, P. L., Santos Costa, V., Caldwell, M. D., Rottscheit, C., Berg, R. L., Mendonca, E. A., and Page, D. (2014). Relational machine learning for electronic health record-driven phenotyping. *Journal of Biomedical Informatics*, 52:260–70.

Pencina, M. J. and D'Agostino, R. B. (2004). Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123.

Pencina, M. J., D'Agostino, R. B., S., D'Agostino, R. B., J., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–72; discussion 207–12.

Pencina, M. J., D'Agostino Sr, R. B., and Song, L. (2012). Quantifying discrimination of framingham risk functions with different survival c statistics. *Statistics in Medicine*, 31(15):1543–1553.

Peng, L., Chen, Y. P., Xu, C., Tang, L. L., Chen, L., Lin, A. H., Liu, X., Sun, Y., and Ma, J. (2018). A novel scoring model to predict benefit of additional induction chemotherapy to concurrent chemoradiotherapy in stage ii-iva nasopharyngeal carcinoma. *Oral Oncology*, 86:258–265.

Peres, M. A., Macpherson, L. M., Weyant, R. J., Daly, B., Venturelli, R., Mathur, M. R., Listl, S., Celeste, R. K., Guarnizo-Herreño, C. C., Kearns, C., et al. (2019). Oral diseases: a global public health challenge. *The Lancet*, 394(10194):249–260.

Petersen, I., Welch, C. A., Nazareth, I., Walters, K., Marston, L., Morris, R. W., Carpenter, J. R., Morris, T. P., and Pham, T. M. (2019). Health indicator recording in uk primary care electronic health records: key implications for handling missing data. *Clinical Epidemiology*, 11:157.

Polycarpou, N., Ng, Y. L., Canavan, D., Moles, D. R., and Gulabivala, K. (2005). Prevalence of persistent pain after endodontic treatment and factors affecting its occurrence in cases with complete radiographic healing. *International Endoddontic Journal*, 38(3):169–78.

Prince, V., Bellile, E. L., Sun, Y., Wolf, G. T., Hoban, C. W., Shuman, A. G., and Taylor, J. M. (2016). Individualized risk prediction of outcomes for oral cavity cancer patients. *Oral Oncology*, 63:66–73.

Rahman, M. S., Ambler, G., Choodari-Oskooei, B., and Omar, R. Z. (2017). Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Medical Research Methodology*, 17(1):60–60.

Ramseier, C. A., Anerud, A., Dulac, M., Lulic, M., Cullinan, M. P., Seymour, G. J., Faddy, M. J., Burgin, W., Schatzle, M., and Lang, N. P. (2017). Natural history of periodontitis: Disease progression and tooth loss over 40years. *Journal of Clinical Periodontology*, 44(12):1182–1191.

Rao, S. K., Mejia, G. C., Logan, R. M., Kulkarni, M., Kamath, V., Fernandes, D. J., Ray, S., and Roberts-Thomson, K. (2016). A screening model for oral cancer using risk scores: development and validation. *Community Dentistry and Oral Epidemiology*, 44(1):76–84.

Reilly, B. M. and Evans, A. T. (2006). Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. *Annals of Internal Medicine*, 144(3):201–209.

Ren, L. and Zhang, Y. (2014). Sliding contact fracture of dental ceramics: principles and validation. *Acta Biomaterialia*, 10(7):3243–3253.

Riley, R. D., Hayden, J. A., Steyerberg, E. W., Moons, K. G., Abrams, K., Kyzas, P. A., Malats, N., Briggs, A., Schroter, S., Altman, D. G., et al. (2013). Prognosis research strategy (progress) 2: prognostic factor research. *PLoS Medicine*, 10(2):e1001380.

Ritter, A. V., Preisser, J. S., Puranik, C. P., Chung, Y., Bader, J. D., Shugars, D. A., Makhija, S., and Vollmer, W. M. (2016). A predictive model for root caries incidence. *Caries Research*, 50(3):271–8.

Robitzsch, A., Grund, S., Henke, T., and Robitzsch, M. A. (2017). Package 'miceadds'. *R Package: Madison, WI, USA*.

Rockhill, B., Kawachi, I., and Colditz, G. A. (2000). Individual risk prediction and population-wide disease prevention. *Epidemiologic Reviews*.

Rosella, L. C., Corey, P., Stukel, T. A., Mustard, C., Hux, J., and Manuel, D. G. (2012). The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. *Population Health Metrics*, 10.

Ross, P. L., Gerigk, C., Gonen, M., Yossepowitch, O., Cagiannos, I., Sogani, P. C., Scardino, P. T., and Kattan, M. W. (2002). Comparisons of nomograms and urologists' predictions in prostate cancer. *Seminars in Urologic Oncology*, 20(2):82–8.

Rothman, K.J.; Greenland, S. L. T. (2008). *Modern epidemiology*. LWW, Philadelphia, PA, 3rd edition.

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141.

Royston, P., Moons, K. G., Altman, D. G., and Vergouwe, Y. (2009). Prognosis and prognostic research: Developing a prognostic model. *British Medical Journal*, 338(b604).

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rudin, C. and Radin, J. (2019). Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2).

Ryo, M. and Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere*, 8(11):e01976.

Sackett, D. L. (1997). Evidence-based medicine. In *Seminars in Perinatology*, volume 21, pages 3–5. Elsevier.

Saghiri, M. A., Garcia-Godoy, F., Gutmann, J. L., Lotfi, M., and Asgar, K. (2012). The reliability of artificial neural network in locating minor apical foramen: a cadaver study. *Journal of Endodontics*, 38(8):1130–1134.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3):e0118432.

Sakamoto, Y., Matsushita, Y., Yamada, S., Yanamoto, S., Shiraishi, T., Asahina, I., and Umeda, M. (2016). Risk factors of distant metastasis in patients with squamous cell carcinoma of the oral cavity. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, 121(5):474–80.

Savage, A., Eaton, K. A., Moles, D. R., and Needleman, I. (2009). A systematic review of definitions of periodontitis and methods that have been used to identify this disease. *Journal of Clinical Periodontology*, 36(6):458–467.

Scannapieco, F. A. and Gershovich, E. (2020). The prevention of periodontal disease—an overview. *Periodontology 2000*, 84(1):9–13.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15.

Schelldorfer, J., Bühlmann, P., and DE GEER, S. V. (2011). Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

Schemper, M. and Stare, J. (1996). Explained variation in survival analysis. *Statistics in Medicine*, 15(19):1999–2012.

Schmid, M. and Potapov, S. (2012). A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in Medicine*, 31(23):2588–2609.

Schulz, K. F., Altman, D. G., Moher, D., and Group, C. (2011). Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery*, 9(8):672–7.

Schwendicke, F., Schmietendorf, E., Plaumann, A., Salzer, S., Dorfer, C. E., and Graetz, C. (2018). Validation of multivariable models for predicting tooth loss in periodontitis patients. *Journal of Clinical Periodontology*, (45):701–710.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, pages 35–47.

Senneby, A., Mejare, I., Sahlin, N. E., Svensater, G., and Rohlin, M. (2015). Diagnostic accuracy of different caries risk assessment methods. a systematic review. *Journal of Dentistry*, 43(12):1385–93.

Shariat, S. F., Karakiewicz, P. I., Roehrborn, C. G., and Kattan, M. W. (2008). An updated catalog of prostate cancer predictive tools. *Cancer*, 113(11):3075–99.

Sharma, N. and Om, H. (2013). Data mining models for predicting oral cancer survivability. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, 2(4):285–295.

Shield, K. D., Ferlay, J., Jemal, A., Sankaranarayanan, R., Chaturvedi, A. K., Bray, F., and Soerjomataram, I. (2017). The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. *CA: A Cancer Journal for Clinicians*, 67(1):51–64.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Singal, A. G., Mukherjee, A., Elmunzer, B. J., Higgins, P. D. R., Lok, A. S., Zhu, J., Marrero, J. A., and Waljee, A. K. (2013). Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *American Journal of Gastroenterology*, 108(11):1723–1730.

Slade, GD; Spencer, A. R. K. (2007). Australia's dental generations. the national survey of adult oral health 2004–06. Report Dental Statistic and Research Series No. 34.

Smith, B. H., Penny, K. I., Purves, A. M., Munro, C., Wilson, B., Grimshaw, J., Chambers, W. A., and Smith, W. C. (1997). The chronic pain grade questionnaire: validation and reliability in postal research. *Pain*, 71(2):141–147.

Snell, K. I., Hua, H., Debray, T. P., Ensor, J., Look, M. P., Moons, K. G., and Riley, R. D. (2016). Multivariate meta-analysis of individual participant data helped externally

validate the performance and implementation of a prediction model. *Journal of Clinical Epidemiology*, 69:40–50.

Snow, G. and Snow, M. G. (2015). Package 'obssens'.

Song, S. (2009). The subject of multiculturalism: Culture, religion, language, ethnicity, nationality, and race? In *New waves in political philosophy*, pages 177–197. Springer.

Speight, P., Elliott, A., Jullien, J., Downer, M., and Zakzrewska, J. (1995). The use of artificial intelligence to identify people at risk of oral cancer and precancer. *British Dental Journal*, 179(10):382–387.

Sperrin, M. and Martin, G. P. (2020). Multiple imputation with missing indicators as proxies for unmeasured variables: simulation study. *BMC Medical Research Methodology*, 20(1):185.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 339(b2393).

Steyerberg, E. W. et al. (2019). *Clinical prediction models*. Springer.

Steyerberg, E. W., Moons, K. G., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G., Group, P., et al. (2013). Prognosis research strategy (progress) 3: prognostic model research. *PLoS Medicine*, 10(2):e1001381.

Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an abcd for validation. *European Heart Journal*, 35(29):1925–1931.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models a framework for traditional and novel measures. *Epidemiology*, 21(1):128–138.

Stiglic, G., Kocbek, P., Fijacko, N., Sheikh, A., and Pajnkihar, M. (2019). Challenges associated with missing data in electronic health records: a case study of a risk prediction model for diabetes using data from slovenian primary care. *Health Informatics Journal*, 25(3):951–959.

Su, C. W., Yen, A. M., Lai, H., Chen, H. H., and Chen, S. L. (2017). Receiver operating characteristic curve-based prediction model for periodontal disease updated with the calibrated community periodontal index. *Journal of Periodontology*, 88(12):1348–1355.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., and Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3(1):1–10.

Tan, H., Peres, K. G., and Peres, M. A. (2016). Retention of teeth and oral health-related quality of life. *Journal of Dental Research*, 95(12):1350–1357.

Taylor, P., Lopez, M. H., Martinez, J. H., and Velasco, G. (2012). When labels don't fit: Hispanics and their views of identity. *Washington, DC: Pew Hispanic Center*.

Thanathornwong, B., Suebnukarn, S., and Ouivirach, K. (2016). Decision support system for predicting color change after tooth whitening. *Computer Methods and Programs in Biomedicine*, 125:88–93.

Therneau, T., Atkinson, B., Ripley, B., and Ripley, M. B. (2015). Package 'rpart'.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Toll, D. B., Janssen, K. J. M., Vergouwe, Y., and Moons, K. G. M. (2008). Validation, updating and impact of clinical prediction rules: A review. *Journal of Clinical Epidemiology*, 61(11):1085–1094.

Tonetti, M. and Claffey, N. (2005). on behalf of the european workshop in periodontology group c. advances in the progression of periodontitis and proposal of definitions of a periodontitis case and disease progression for use in risk factor research. *Journal of Clinical Periodontology*, 32(Suppl 6):205–208.

Tonetti, M. S., Greenwell, H., and Kornman, K. S. (2018). Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *Journal of Clinical Periodontology*, 45 Suppl 20:S149–S161.

Tonetti, M. S., Jepsen, S., Jin, L., and Otomo-Corgel, J. (2017). Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: A call for global action. *Journal of Clinical Periodontology*, 44(5):456–462.

Trombelli, L., Farina, R., Ferrari, S., Pasetti, P., and Calura, G. (2009). Comparison between two methods for periodontal risk assessment. *Minerva Stomatologica*, 58(6):277–287.

Tseng, W. T., Chiang, W. F., Liu, S. Y., Roan, J., and Lin, C. N. (2015). The application of data mining techniques to oral cancer prognosis. *Journal of Medical Systems*, 39(5):59.

Twisk, J. W. (2006). *Applied multilevel analysis: a practical guide for medical researchers*. Cambridge university press.

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., and Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176.

van Geloven, N., Swanson, S. A., Ramspek, C. L., Luijken, K., van Diepen, M., Morris, T. P., Groenwold, R. H., van Houwelingen, H. C., Putter, H., and le Cessie, S. (2020). Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology*, 35:619–630.

Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J., and Habbema, J. D. (2002). Validity of prognostic models: when is a model clinically useful? *Seminars in Urologic Oncology*, 20(2):96–107.

Verstraete, E. H., Blot, K., Mahieu, L., Vogelaers, D., and Blot, S. (2015). Prediction models for neonatal health care-associated sepsis: a meta-analysis. *Pediatrics*, 135(4):e1002–14.

Vittinghoff, E. and McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, 165(6):710–718.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotzsche, P. C., Vandenbroucke, J. P., and Initiative, S. (2014). The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies. *International Journal of Surgery*, 12(12):1495–9.

Von Korff, M., Ormel, J., Keefe, F. J., and Dworkin, S. F. (1992). Grading the severity of chronic pain. *Pain*, 50(2):133–149.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31:841.

Walters, S. J. (2012). Analyzing time to event outcomes with a cox regression model. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):310–315.

Walton, R. and Fouad, A. (1992). Endodontic interappointment flare-ups: a prospective study of incidence and related factors. *Journal of Endodontics*, 18(4):172–7.

Wandner, L. D., Scipio, C. D., Hirsh, A. T., Torres, C. A., and Robinson, M. E. (2012). The perception of pain in others: how gender, race, and age influence pain expectations. *The Journal of Pain*, 13(3):220–227.

Wang, F., Zhang, H., Wen, J., Zhou, J., Liu, Y., Cheng, B., Chen, X., and Wei, J. (2018). Nomograms forecasting long-term overall and cancer-specific survival of patients with oral squamous cell carcinoma. *Cancer Medicine*, 7(4):943–952.

Watt, R. G., Daly, B., Allison, P., Macpherson, L. M., Venturelli, R., Listl, S., Weyant, R. J., Mathur, M. R., Guarnizo-Herreño, C. C., Celeste, R. K., et al. (2019). Ending the neglect of global oral health: time for radical action. *The Lancet*, 394(10194):261–272.

White, I. R. and Royston, P. (2009). Imputing missing covariate values for the cox model. *Statistics in Medicine*, 28(15):1982–1998.

Wiggins, J. S. (1981). Clinical and statistical prediction: Where are we and where do we go from here? *Clinical Psychology Review*, 1(1):3–18.

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., Mallett, S., and Groupdagger, P. (2019). Probast: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1):51–58.

Wong, S. S.-L., Wilczynski, N. L., Haynes, R. B., Ramkissoonsingh, R., Team, H., et al. (2003). Developing optimal search strategies for detecting sound clinical prediction studies in medline. In *AMIA Annual Symposium Proceedings*, volume 2003, page 728. American Medical Informatics Association.

Wright, M. N., Dankowski, T., and Ziegler, A. (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statitics in Medicine*, 36(8):1272–1284.

Zellner, M., Abbas, A. E., Budescu, D. V., and Galstyan, A. (2021). A survey of human judgement and quantitative forecasting methods. *Royal Society Open Science*, 8(2):201187.

Zhang, Y.-P., Zhang, L.-N., and Wang, Y.-C. (2010). Cluster-based majority under-sampling approaches for class imbalance learning. In *2010 2nd IEEE International Conference on Information and Financial Engineering*, pages 400–404. IEEE.

Zhou, Y. and McArdle, J. J. (2015). Rationale and applications of survival tree and survival ensemble methods. *Psychometrika*, 80(3):811–833.

Zihni, E., Madai, V. I., Livne, M., Galinovic, I., Khalil, A. A., Fiebach, J. B., and Frey, D. (2020). Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *Plos One*, 15(4):e0231166.