RESEARCH

Predicting decayed, missing or filled teeth in young children: A comparative use of conventional statistical methods and machine learning

Sarah A. Sonal[1*], Martin Lee[2], Jennifer A. Brown[1] and Philip J. Schluter[3,4]

Abstract

Background: Early childhood caries is a preventable chronic disease with a strong socio-economic gradient. The overall arching goal of this research is to establish if routinely collected data can be used to predict dental disease. The primary aim of this study was to compare a conventional statistical technique with a supervised machine learning technique, to establish the most appropriate method for answering this research goal.

Methods: This study utilised routinely collected dental records, hospital admissions and New Zealand Index of Multiple Deprivation data from 21,236 children aged 5-years, in the Canterbury region. Selection was limited to children who turned five years old between 2014 and 2017. The data were split into 3 datasets, a training dataset to build models to predict a count of decayed, missing or filled teeth, a tuning dataset to tune the best of these models, and a testing dataset to compare the models on their predictive abilities. Models were compared on goodness-of-fit, root mean square error (RMSE) and sensitivity and specificity.

Results: The zero-inflated negative binomial and the random forest models performed better at fitting and predicting than the other methods considered. The random forest model performed better at prediction with a RMSE of 2.678 compared to the zero-inflated negative binomial RMSE of 2.727. The sensitivity for the random forest model was 0.203 which was higher than the zero-inflated negative binomial sensitivity of 0.071. Specificity was 0.926 for the random forest model and 0.972 for the zero-inflated negative binomial model. The model building, tuning and testing process for the random forest model was more computationally efficient than for the zero-inflated negative binomial model.

Conclusion: Machine learning, specifically random forests, are a faster approach to modelling routinely collected dental data, with greater precision and accuracy to fit the data and predict dental disease.

Keywords: machine learning; paediatric oral health; early childhood caries

[*]Correspondence:
sarah.sonal@pg.canterbury.ac.nz
[1]Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, 8140, Christchuch, New Zealand
Full list of author information is available at the end of the article

Background

Caries in early childhood is a serious health concern worldwide. It is a preventable phenomenon that can have detrimental effects on physical and mental health. Early childhood caries (ECC) has been associated with increased weight and malnutrition [1]. Fear, depression and anxiety are also associated with poor oral health in children [2, 3, 4] and caries in childhood has been linked to higher rates of caries in adult teeth [5]. According to the California Dental Association, dental caries is the number one chronic health issue in early childhood [6].

Oral health is a key area of development for the World Health Organization (WHO). According to the WHO Oral Health Fact Sheet, oral disease affects approximately half the world's population and is predominantly preventable or treatable in the early disease stages [7]. School age children and youth are among the key target groups for WHO oral health intervention [8]. Although decreasing in prevalence internationally [9, 10, 11], oral diseases are still considered the most common chronic illness for children, especially for those living in deprived circumstances. These children have more missing teeth and are more likely to have unmet treatment needs than those in less deprived circumstances [11, 12]. In developed nations, children who come from higher income families have lower rates of dental disease than those from low income families. Similarly, children whose parents have higher education levels have lower rates of dental diseases than those with less educated parents [11, 12].

Investigating ECC is important to New Zealand public health. In the 2016/17 year, the Ministry of Health identified that 12.3% of 5-9 year olds and 18.0% of 10-14 year olds had a tooth or multiple teeth removed in their lifetime due to dental caries [13]. Utilising data from the Dunedin Multidisciplinary Health and Development study [14], Thompson (2002) showed that children who came from lower socio-economic backgrounds were more likely to have had a tooth or multiple teeth removed in adulthood by age 26 years than children from higher socio-economic homes [15]. This study also showed that children who had dental issues at age five years were at a higher risk of dental extraction by age 26 years than those that did not [15].

Prevention and effective early treatment for ECC is paramount for a child's future oral health. When not prevented or treated in a timely manner, ECC can lead to complications including chronic infections or abscesses, pain and issues with sleeping

[6]. When an ECC is severe, tooth-saving treatments are no longer viable and tooth extraction is required. Premature extraction of primary teeth can result in crowding, teeth migration, early eruption of permanent teeth, remaining teeth destabilisation and potential speech issues [16, 17]. Oral disease has been associated with other complications such as systemic diseases, poor nutrition and a decreased immune system function [11]. On a population level, one potential aid in avoiding ECC is the development of prediction models that can identify those most at risk, for health promotion or early intervention targeting. However, such models only have utility if they have effective predictive properties.

Health data, including oral health datasets, are rapidly increasing in size and scope with electronic record keeping, medical imaging and personal wearable technologies. This growth is unlikely to abate; indeed, further growth and expansion is almost inevitable. As a result, machine learning techniques have developed to better deal with larger datasets. Arguably, we are on the cusp whereby traditional statistical techniques, developed in the epoch of smaller datasets and various sampling strategies, may be superseded by these new techniques. They may provide a better platform for understanding complex variable relationships, including the derivation of prediction models.

Here, a purposeful selection of conventionally employed generalised linear modelling techniques will be employed on a large oral health dataset, and the best ascertained using a priori defined criteria. Concurrently, a selection of modern datamining techniques will be employed, and using the same criteria, the best will be selected. The best performing will then be compared on how well the models fit the data, and on their ability to predict using a data subset not used for building the model. Finally, the adequacy of each of these models for the predictive screening of ECCs in this population will be assessed.

Methods

Study Design

A retrospective analysis of routine oral health data collected between 1st January 2014 to 31st December 2017, inclusive.

Participants and Setting

Children aged 5-years attending the Community Dental Services in the Canterbury and South Canterbury regions of New Zealand for their routine oral health check. Those children living outside the Canterbury region or without a residential address were excluded.

Variables

The primary variable of interest was the count of decayed, missing or filled teeth (dmft) in children's deciduous teeth [18]. Potential exploratory variables of interest, available from routine collected data sources, included: appointment attendance; household deprivation (based on last known residential address); ethnicity; and, co-morbidities/hospital admissions from Canterbury District Health Board (CDHB) data. Table1 contains the full variable list, and associated definitions and classifications.

Procedure

In New Zealand, all children under the age of 18 years are eligible for free dental check-ups and treatments [19]. Community Dental Services is the organisation that provides this service for children in the Canterbury and South Canterbury regions from birth to age 13-years [20]. The 5-years age threshold was selected because children begin primary school at this age, and they also begin receiving dental treatments during school hours. All routine patient and appointment information is collected and entered on the services electronic database, Titanium [21]; a software application designed specifically to assist in the management and delivery of oral health services. Research data were extracted from this Titanium database on two separate occasions. The first extraction was to gather address data for geocoding. This dataset had all identifiable information removed except for address, a new unique identifier was created, and addresses were then geocoded to latitude and longitude using ArcGIS software [22]. Both automatic and manual geocoding, using the multiple options identified by ArcGIS, were undertaken.

The geocoded address information was then used to extract the deprivation data for each child. The chosen dataset for this study was the New Zealand Index of Multiple Deprivation (IMD) developed at the University of Auckland [23]. This was chosen over the NZDep2013 Index of Deprivation [24] produced by Statistics New

Zealand. One reason for this is the Canterbury earthquake sequence of 2010 and 2011 which resulted in significant displacement of the population in the following years, especially in eastern suburbs of Christchurch. This population displacement may affect the reliability of the NZDep2013 Index of Deprivation [25]. The NZDep2013 Index of Deprivation is made from census data, whereas IMD is an index made up from routinely collected data from a combination of sources including government organisations and census data [23]. There are 28 individual measures that are grouped into the following categories: employment, income, crime, housing, health, education and access [23]. Access is a combination category, and it is the distance to the 3 nearest General Practitioner (GP) doctor's office or accident and emergency clinic, petrol stations, supermarkets or schools (excluding high schools) [23].

The extraction of this deprivation information for each patient was completed within ArcGIS. The open source shape file was downloaded from the IMD website [23] and joined to the patient geocoded address data through a spacial based join [22], using the latitude and longitude of the address data to select the datazone for the deprivation dataset. The joined deprivation data was returned to Community Dental Services to remove identifying address information and to change the unique identifier. This dataset is referred to as the IMD dataset and/or the deprivation dataset.

The new unique identifier was replicated to the second extract of data from Community Dental Services, and to hospital admissions for joining of datasets. The second extract of data from community dental services was the clinical data with the new unique identifier, which excluded the address and any other identifying information. This dataset is referred to as the oral health dataset. Changing the unique identifier was to avoid identifiable information being present in the dataset at the same time as the clinical data, and prevented the re-joining of identifying address information to clinical data.

The hospital admission data was requested by Community Dental Services. This process was done by utilising the National Health Index number (NHI) of each patient in the cohort and extracting their Christchurch Hospital admission information, both elective and acute admission types. The NHI number is a unique identifier that is assigned to every person who uses health and disability support services in New Zealand. The NHI number and all other identifiable information

was removed by Community Dental Services, and the new unique identifier was attached.

Study size

An approach to test model prediction ability is to have data used for building a model separate from data used for testing prediction ability. This can be taken a step further to have three separate datasets, one for building models, one for tuning model parameters and selecting the best model, and a final set for testing the final model. The three set approach is used in this study because an aim of the study is to compare two types of modelling techniques against each other. This allows for the models to be tested and improved separately before being evaluated against each other. The cohort of four years (2014 to 2017) of five-year-olds was chosen to ensure there were enough data for the three datasets. The research team required approximately 10,000 patients for the training dataset, 5,000 for the tuning/validation set and 5,000 for the testing dataset. A four year cohort allowed for an estimated 5% data to be lost through geocoding and data joining, while minimising potential bias.

Data Cleansing and Joining Methods

Data Cleansing

All data cleansing was conducted within the statistical software R [26].

Ethnicity was found to have high proportions of patients with missing data. Ethnicity was considered a highly important variable, therefore removal of the variable was not acceptable. This variable already had a factor level of "Unknown Ethnicity", therefore the missing data could be converted to "Unknown Ethnicity" without losing any information or patients.

Combining variables into new variables was done during data cleansing. Hospital admission data was transformed from raw International Classification of Diseases (ICD) code [27] data into ambulatory-sensitive hospitalisation (ASH) codes [28]. ASH admissions are hospital admissions that are identified to be reducible through preventative interventions [28]. This was of interest as ECC is a preventable chronic illness, and it was of interest to see if other preventable diseases were related to ECC in this cohort. To do this data conversion, after the hospital admission data were rotated to one line per patient, the ICD codes which make up each ASH code were

added together into a new column, and the ICD codes were removed from the dataset for analysis. They were not removed from the raw data.

Data cleansing also included the transformation of variables. The days admitted in hospital total (days_admit_tot) variable was generated by calculating the length of stay for each hospital admission, and adding these together on rotation of the data. The length of time between referral into dental services and treatment date (exam.delay) was generated by subtracting the appointment date from the referral date for each patient. The remaining NA values for the hospital admission data were recoded as "0" to indicate no hospital admissions.

The full clean dataset was simplified twice from a full dataset to two reduced datasets. The datasets were reduced by removing variables with correlations above $r \geq 0.9$ and then $r \geq 0.7$. This was to identify and reduce potential multicollinearity in the dataset. This process used the findCorrelation function from the Caret package [29] in R. This function identifies variables with pairwise correlations above a prescribed correlation threshold. It then considers which of the two variables in each pairwise correlation has high correlation with other variables, and identifies that as the variable to be removed.

Data Joining

Using dataset joining methods within ArcGIS, each patient had their last known address attached to a deprivation data zone. All other data joining was conducted within R. To join the oral health dataset to the deprivation dataset, a right join was used, with all data from the IMD dataset kept, and the oral health data rows removed if there was no corresponding patient in the IMD dataset. To join this combined dataset to the hospital admission dataset, a left join was used, with all patients data maintained and NA's generated for the patients who did not have hospital admissions. These were cleansed as outlined above.

Statistical Methods

Exploratory Data Analysis

After data cleansing, exploratory data analysis was conducted. Initially, a visual method called a scatter-plot matrix was built in groups of 10 variables, to identify possible multicollinearity. Figure 1 is an example scatter-plot matrix from this visualisation process. Multicollinearity can be an issue as model coefficient estimates

can have high variation between samples of data, with small differences in input data leading to large differences in the model [30]. Overfitting of the model is highly likely with multicollinearity [30].

Conventional Biostatistical Techniques

To begin the traditional modelling, simpler models were attempted before trialling of the more complex models. The first model type attempted was a Poisson model. This model was chosen as the outcome variable, decayed, missing or filled teeth, was discrete and could take the values zero or above [18, 31]. Next a negative binomial model was chosen to model the data. A negative binomial is better able to deal with dispersed data than a Poisson model [31] as it has separate parameters for variance and mean, whereas Poisson has one parameter for both. The data was then modelled with a zero-inflated negative binomial model. Zero-inflated data is count data that has an excess of zeros that a standard distribution model will underestimate [32]. A zero-inflated negative binomial model is a mixture model containing the over-dispersed Poisson (the negative binomial) and a logit component to model the excess zeros [33]. Each zero-inflated negative binomial model was compared to the equivalent negative binomial using the Voung test [34]. The Voung test is a comparative test for establishing if the zero-inflated model fits the data significantly better than the standard model [34].

During the conventional modelling, variables were grouped to build separate models, reduced to parsimonious models using the Akaike information criterion (AIC) [35], then joined into a full model. This is a technique often used in biostatistics called an ensemble model. The variables were grouped based on the three original datasets, group one was the variables originally from the oral health data, group two was the data originally from the IMD and group three was from the hospital admissions. The specific groupings are broken down in Table1.

When building and tuning the models, the data predicted was the validation set, and when comparing the final conventional biostatitics model and the final machine learning model, the predicted data was the testing set.

Machine Learning Techniques

Two styles of machine learning were selected to compare with the traditional statistical models; support vector machines (SVM) [36] and random forests [37]. A SVM

is a supervised learning technique that categorises a binary outcome using a hyperplane [38]. A tree-based approach to supervised learning breaks the predictor space up into segments, which can be represented as a hierarchical tree, and the mean of each segment is used for the prediction [38]. A random forest is an extension of this approach, building multiple trees using bootstrapping methods, which increases the prediction accuracy [38]. Each time a tree is split, a random sample of m predictor variables is selected as possible candidates at that split [38].

The random forest technique was selected to continue the modelling process. The random forest parameters were then tuned. Tuning a model can improve both model fit and prediction accuracy [39]. The first parameter to tune was the number of variables to be randomly selected to try at each split of the tree [40]. This was tuned by using the inbuilt tuning function within the RandomForest package [41], the tuneRF function. The second parameter to tune was the number of random forests to build. This was done by setting up a loop to show between 500, 1000, 1500, 2000 and 2500 trees built, and comparing them on predictive performance.

Prediction

At each stage of the modelling process, for both the zero-inflated negative binomial models and for the random forest models, models were tested for prediction capabilities. This was done by testing prediction ability using the validation dataset. For each model, a RMSE was calculated when predicting this validation set, and the model with the lowest RMSE was selected as the best model.

The best zero-inflated negative binomial model and the best random forest model were compared to each other based on their predictive abilities to the testing dataset, data that were not used in any stage of the model building process.

The models were also compared on their sensitivity and specificity. Sensitivity is the certainty of a correct positive diagnosis for a disease [42]. It is a ratio of the correctly positively diagnosed patients to all patients with the disease [43]. Specificity is the certainty of a correct negative diagnosis for a disease [42], therefore as sensitivity increases the specificity decreases, and vice versa. Generally in health it is important to diagnose as many people with a disease correctly as possible. A patient with a treatable serious disease such as dental decay, miss-diagnosed to disease free would be undesirable, therefore sensitivity is usually regarded as being

more important than specificity. Sensitivity and specificity are always between zero and one, with values closer to one indicating higher sensitivity or specificity. To do this testing, the outcome variable was converted to a binomial outcome of no disease or disease present. The zero-inflated negative binomial model and the random forests were then compared on their sensitivity and specificity when predicting the testing set.

All code for the Data Analyses is included in the supplementary materials.

Ethical approval

The study complied with the ethical standards for human experimentation as established by the Helsinki Declaration 1995 (as revised in Edinburgh 2000) and New Zealand's Health and Disability Ethics Committee (HDEC). HDEC defined this study as minimal risk observational research and it did not require ethics committee review. University of Canterbury Ethics Committee approved the study internally. For data access purposes, a locality agreement was formulated, and evidence of the internal ethics approval was provided to them along with a confirmation letter from HDEC that external ethics approval was not required. These documents are included in the supplementary materials.

Results

Participants

Overall, there were records for 21,236 children in the full oral health dataset. Although address data for 78 (0.4%) children was unable to be geocoded, no other data were lost during data joining. The final dataset included data from 21,158 children, 5,197 from 2014, 5,241 from 2015, 5,380 from 2016 and 5,340 from 2017.

Descriptive Statistics of participants

Of the 22,158 children, 10,666 (50.4%) were boys, 10,483 (49.6%) were girls and 9 (0.04%) had sex missing. The majority, 14,256 (67.4%), identified with European ethnicity, 2,421 (11.4%) with Māori, 949 (4.5%) with Pasifika, 1,996 (9.4%) with Asian and 388 (1.8%) with an other ethnicity. Ethnicity was unknown for 1,148 (5.4%) of the children. Ethnicity was prioritised using New Zealand Ministry of Health methods. 4,554 (21.5%) of the children came from the least deprived IMD decile rank of 1, 3,466 (16.4%) from decile 2, 2,109 (10.0%) from decile 3, 2,421

(11.4%) from decile 4, 1,896 (19.0%) from decile 5, 1,581 (7.5%) from decile 6, 2,018 (9.5%) from decile 7, 1,308 (6.2%) from decile 8, 1,311 (6.2%) from decile 9 and 423 (2.0%) from the most deprived ranked neighbournood of IMD decile 10. Table2 shows a breakdown of key statistics for sex, ethnicity and deprivation for the dataset.

Statistics New Zealand publishes demographic data for each region of New Zealand from a 4 yearly census. According the the 2013 data [44], the sex split in Canterbury is 49.4% male and 50.6% female, which is similar to the study data. Ethnic breakdown in the data is slightly different to the census levels, with 86.9% European, 8.1% Māori, 2.5% Pasifika, 6.9% Asian and 2.7% other ethnicity. This was accepted as there may have been changes in ethnic diversity since the 2013 census.

Outcome Data

The distribution of dmft was heavily right skewed, with median 0 (Q1=0, Q3=3) and maximum 20. Table2 presents the dmft distribution by sex, ethnicity and deprivation. There was no difference in dmft between boys and girls ($p=0.11$). Compared to the dominant European population, Māori ($p<0.01$), Pasifika ($p<0.01$), Asian ($p<0.01$), other ($p<0.01$) and unknown ethnicity ($p<0.01$) all had higher mean dmft. Compared to the least deprived IMD decile of 1, deciles 2 ($p<0.01$), 3 ($p<0.01$), 4 ($p<0.01$), 5 ($p<0.01$), 6 ($p<0.01$), 7 ($p<0.01$), 8 ($p<0.01$), 9 ($p<0.01$), 10 ($p<0.01$) all had higher mean dmft.

Figure 2 shows the density histograms for the outcome variable in the training set, the validation set and the testing set. From these plots, the histogram shows a high density of patients with zero decayed missing or filled deciduous teeth.

Main Results

Data Joining and Exploratory Data Analysis

Most of the addresses were successfully geocoded automatically. A small proportion of addresses were manually geocoded from multiple options identified by ArcGIS. Patient records that failed to geocode had to be dropped from the study as one of the criteria for inclusion was the patient living in the Canterbury region. Without an address, there was insufficient evidence that the patient was a resident.

Scatterplot matrices showed, in many case, that there was strong correlations and multicollinearity between predictor variables. Two reduced datasets were used in model building as multicollinearity was present in the dataset.

Conventional Biostatistical Techniques

Simpler models were attempted before more complex models. First, a Poisson model was attempted but this model did not converge due to skewness of the data and because the variance of the data was significantly larger than the mean. Then a negative binomial was used, which was able to fit the data, but it was not a particularly good fit due to zero-inflation of the dataset. Lastly, a zero-inflated negative binomial model was applied.

There were several issues found at this stage of the analysis. The zero-inflated model could not handle the many variables at the same time. First, the data had to be divided into groups to be modelled, then the best variables combined into one model. When using group two data (see Table1) to build the model with the full dataset, the IMD decile rank excluding housing variable had to be removed manually due to non-convergence issues. The zero-inflated negative binomial model built using group three data (see Table1) had two variables manually excluded due to fitting issues. These two variables were Kidney/urinary infections and Vaccine-preventable disease: Measles, Mumps and Rubella (MMR). This is likely due to very low patient numbers in these admission types, only 8 children had admissions for Kidney/urinary infections and 3 children with MMR related admissions.

The ensemble model performed better than individual group models when compared on improvement of Root Mean Square Error(RMSE). The RMSE was calculated using each model to predict unknown data and the rmse function from the ModelMetrics package [45]. Using the Voung test each zero-inflated model was confirmed to be superior to the standard negative binomial.

Table 3 shows the output information for the final zero-inflated model. This output information shows the variables that make up the final parsimonious model. All coefficients are statistically significant at an alpha value of 0.05 except housing decile and Asthma hospital admissions. When the non-statistically significant variables were removed, the AIC increased - and so they were retained.

Machine Learning Techniques

On initial investigation of these methods with the dataset, random forests outperformed SVM on RMSE and were faster to build. Therefore random forests were chosen to continue with for the analysis.

Three random forest models were built using the full dataset, the first reduced dataset of size 28 variables (correlated variables with $r \geq 0.9$ were removed) and the second reduced dataset of 26 variables (correlated variables with $r \geq 0.7$ were removed). Table4 shows the RMSE for each model. The second reduced dataset had the lowest RMSE of 2.556, therefore this model was used going forward with further model tuning.

The model was then tuned for the number of variables randomly selected to try at each tree split (*mtry*) and the number of trees to build. Using a *mtry* parameter of 4 was found to reduce the Out-Of-Bag (OOB) error the most (OOB=0.311). OOB error is used to measure the prediction error for a random forest model [40]. OOB error is the mean difference between the predicted outcomes and the real outcomes for the data not used for building the model [40]. The second reduced dataset performed the best, with a RMSE of 2.535. The ideal *mtry* of 4 was used when tuning the number of trees. The second reduced dataset was again the best performing model when using 1,500 trees, with a RMSE of 2.539 when predicting the validation data. The random forest model explained 16.98% of the variation in dmft.

Figure 3 shows the variable importance plot for the random forest. This plot shows that the 10 most important variables used to predict the number of primary teeth with dental decay were delay from referral to community dental and appointment (exam.delay), ethnicity (ethnicity), number of dental exams (nexam), number of did not attend appointments (ndna), decile rank for education (edurankd), decile rank for housing (hourankd), decile rank for crime (crirankd), decile rank for health (hlthrankd), decile rank for access (accrankd) and number of cancelled appointments (ncanc). Of note, sex was ranked 11th. This is unexpected as the descriptive statistics comparing sex and dmft directly did not show difference between females and males. This may be due to a disproportionate number of hospital admissions between the sexes.

Other Analyses

Each final model was tested and compared with residual diagnostic plots, prediction effectiveness and sensitivity and specificity analysis. Residual diagnostic plots and full explanations can be found in the supplementary materials. Overall, the residual diagnostic plots are slightly better for the random forest model than the zero-inflated negative binomial model, providing evidence of a better model fit.

Sensitivity and Specificity Analysis with Receiver Operating Characteristic Curves

All sensitivity and specificity testing was conducted using the testing dataset. A test dataset is one not used to build the models and is considered new data.

The random forest model has higher sensitivity of 0.203 compared to the sensitivity for the zero-inflated negative binomial model of 0.071. For the random forest model, 20.3% of five year old patients with dental decay would be diagnosed correctly as having disease, compared to 7.1% from the negative binomial model.

Both models had high specificity. For the zero-inflated negative binomial model, 2.8% of five year old patients who are dental disease free would be miss-classified as having dental disease compared to 7.4% from the random forest model.

Figure 4 shows the receiver operator curve (ROC) for the two final models. The line indicating the random forest model is on the outside of the zero-inflated negative binomial model for the length of the curve, which indicates that it is performing better for sensitivity and specificity overall. There is a point in the middle of the curve where the lines touch, indicating no difference in sensitivity and specificity. Overall the plot indicates that the random forest performs better for sensitivity and specificity.

Model Prediction Testing

Table 4 shows the RMSE for both the zero-inflated negative binomial model building process and the random forest model building process. For the zero-inflated model building, the second reduced dataset had a lower RMSE than the other datasets. This is also the case for the random forest model building process. When predicting the unknown testing dataset with the final models, the RMSE is significantly lower for the random forest model at 2.678 than the zero-inflated negative binomial model at 2.727.

Discussion

The final random forest model had a higher sensitivity (0.203) than the final zero-inflated negative binomial (0.071), and had slightly lower specificity (0.926) than the final zero-inflated negative binomial model (0.972). This means that if these models were used to predict if a child had dental disease, 79.7% of children would be misdiagnosed as disease free when they actually do have dental disease, when the random forest model is used. 92.9% of children would be misdiagnosed as disease free if the zero-inflated negative binomial model is used. Neither of these figures are particularly good, although the random forest is the better of the two. When the random forest model is used, 7.4% of patients would be diagnosed as having dental disease when they do not, whereas the zero-inflated negative binomial model would misdiagnose 2.8% of patients who do not have dental disease as having disease. There is a small difference in specificity favouring the zero-inflated negative binomial model, and a larger difference in sensitivity favouring the random forest model. The final random forest model has a lower RMSE of 2.678 than the final zero-inflated negative binomial model RMSE of 2.727. This shows that the random forest model is outperforming the zero-inflated negative binomial model for predicting the exact number teeth with dental issues. This shows that a random forest model could be used to predict the number of teeth expected to have disease, which could then be used to estimate treatment appointment length based on number of diseased teeth. This could help prevent appointments running over time or prevent patients having to return for second or third treatment appointments.

Conventional biostatistical techniques have been used to identify factors associated with poorer oral health. For example, Cruvinel (2010) [46] used logistic regression to predict dmft and permanent tooth decayed missing or filled teeth (DMFT) for children born prematurely or at term. Unlike the current study where routinely collected data was used to predict dmft, Cruvinel (2010) [46] used a sample of 80 children, 40 premature and 40 born full term, therefore traditional modelling techniques are more appropriate. Javali (2007) [47] investigated the use of generalised linear modelling for predicting DMFT over multiple linear regression, given the skew that the DMFT outcome variable was found to have. This was a bigger dataset compared to the previous study, with 7,188 patients interviewed and examined. This is would have taken many hours of examinations and following up

participants, although they were able to include variables of interest such as oral hygiene habits that were not possible in the current study. Given Javali (2007) [47] was over 10 years before the current study, and supervised learning has only just begun to infiltrate the health industry, it is not surprising that these larger data techniques were not considered. Negative binomial regression has also been used to predict DMFT in Hong Kong [48]. A sample of 324 young people was used to predict DMFT using examination results and clinical examination. This sample size does not lend itself to supervised learning, unlike the current study.

Machine learning has been used in the oral health area for diagnostic purposes. Angelino (2017) [49] used a hand-held light emitting diode device with a machine learning visual processing algorithm to diagnose plaque. Deep learning has been used to predict oral malodour using saliva samples, with a 97% success rate [50]. Image processing has also been used to diagnose gum disease [51]. These studies have required examination and investigations to be used in conjunction with machine learning, whereas the current study utilised routinely collected data to gain insights into oral health.

Looking at the model performance purely from a statistical performance point of view, neither modelling technique has resulted in a model that is an effective predictor of dmft for Canterbury children. A recent systematic review of dental caries risk assessment models discussed variables that made up predictive models. These models were made up of variables including socio-demographic factors (age, ethnicity and socio-economic status), behavioural factors (diet, fluoride, dental appointment attendance), blinical factors (oral hygiene habits, caries history, cleaning method, systemic health, medication), and microbiological and salivary factors (certain bacteria levels, salivation rate) [52]. Models including the full range of these variables had varying sensitivity scores from 0.41 to 0.75. The current study had all socio-demographic factors listed, appointment attendance and systemic disease requiring hospitalisation. This does not cover a large proportion of these predictor variables, therefore the current study with sensitivity scores of 0.203 for the random forest model and 0.071 for the zero-inflated negative binomial model fit with the expected level of sensitivity. Specificity on the other hand is high compared to other models that have the full range of variables, which ranged from 0.71 and 0.88, which the current study outperforms with the sensitivity scores for the random forest and the

zero-inflated negative binomial at 0.926 and 0.972 respectively. Given that the data used for the current study does not include any information about patient daily oral health practises or dietary intake, it would be unrealistic to expect the models to be out performing models with more data types.

Limitations

A limitation of the zero-inflated negative binomial model is that the functions to build this model could not handle wide data, that is, data with many explanatory variables. The data had to be divided into three subsets, and each subset used to build separate models, then joined into an ensemble model. This has become a relatively common approach in epidemiological modelling [53]. Many variables had to be removed manually along the way, as specific variables would not allow the model to converge. All of the variables that were removed manually are a potential source of bias and uncertainty. This is partially shown in the way in which the random forest has outperformed the zero-inflated negative binomial model in RMSE and sensitivity. There is also likely bias in the zero-inflated negative binomial that is not visible or measured. The residual diagnostic plots were all far from the ideal diagnostic plots for a good model. A good model would have non-linear residuals that are normally distributed about the model, can take values over the full range of the predictor values and without any particular values that are influencing the model more than the others [54]. This means that there is an aspect of the data that is not being modelled by the zero-inflated negative binomial model or the random forest model.

Other limitations with this model were the length of time taken to build and simplify each model, the number of steps required to simplify each model and the process of building a model from 3 separate simplified models. The zero-inflated negative binomial model took two to three hours to build each model as variables had to be manually removed from the model. When the variables were removed manually, this took a lot of time a trial and error to find which variables were causing the issues. Once the variables were identified that could be kept in the model, it took under 1 minute to build the model.

The number of steps required in building the ensemble final model to ensure no mistakes were made took three days. The automated simplification of the models

took more than four hours, therefore over-night processing was required for this step. This is not the best way to build a predictive model. It is best to have a streamlined process to build, tune and test the model, which was not achievable for this model type. This is a benefit of artificial intelligence methodologies, as the processes are automated without a loss of precision.

A limitation of the random forest model is that it cannot be written out as a formula and used directly to predict the outcome variable like a conventional model can. A single tree from the 1,500 trees built in the random forest could be extracted, drawn in a hierarchical diagram and used to predict the count of dmft, but this would not be as accurate as the full random forest.

During the model building process the random forest model was slower to build than the zero-inflated negative binomial model, at approximately five minutes per model. This is a disadvantage over the zero-inflated negative binomial model, and could be considered slow when compared to other machine learning techniques. Tuning the random forest took longer than building the forest. To tune the mtry parameter took 20 to 30 minutes. The tuning process to decide the number of trees took longer again as the loop had to build a random forest with an increasing number of trees each time. This took approximately 2 hours. Although these times are slow, they did not require user input while the processes were automating, whereas the zero-inflated negative binomial model building process had to be constantly monitored. Predicting using the random forest was fast, within seconds. Sensitivity and specificity was also fast, within seconds.

An expected sources of bias was from patients removed from the final dataset while joining the three initial datasets together (oral health data; hospital admission data; and deprivation data), and through data cleansing for missing data.

During automated geocoding most of the addresses were geocoded to a Canterbury address automatically or manually, leaving a small proportion unmatchable. The unmatched addresses were examined manually and appeared to have no recorded address. The loss of this data was considered acceptable, and although it may have introduced a small amount of bias, given that the original premise was Canterbury based children, without an address there is insufficient evidence to establish if the children are residents of the region.

During data cleansing, ethnicity was found to have missing data. If these patients were removed from the dataset, this could have potentially introduced bias into the dataset, as there may be a reason why all the patients with missing data are similar. This is called a "missing not a random" situation [55]. As it was not possible to discover why the data were missing and if the patients were similar in some way, the decision was made to keep all patients in the dataset and deal with the missing values in a different way to avoid and/or minimise bias.

Bias was also reduced by having only a four year cohort of data. This was to help ensure that patients in each year were not growing up in differing social environments, especially given social changes due to natural disasters experienced in 2010-2011. If the cohort had been extended too close to these years, the dataset would be expected to have bias due to address displacement. This is acknowledged to potentially be a source of bias that is not controllable, but was minimised.

One limitation of using routinely collected data is that the data is gathered without the study questions in mind. This means that the data may not contain all the variables of interest, and is likely to have confounding factors, both measured and unmeasured. For example, it may be of interest to know employment status of the child's parents, whether the child has other siblings with or without dental issues, the oral health of the parents, the diet of the child and their parents, genetic predisposition to oral health issues, oral health habits of the child, and many other non-measured variables. All known confounding factors and any unknown confounding factors could be affecting both models, introducing bias. The IMD data also does not measure a specific individuals deprivation, it is an average score based on the patients last known address. The assumption that the deprivation is true for each individual based on their last known address is based on the assumption that those in each IMD location are homogeneous. This may not hold true, and there is no way to test this within the limitations of this study.

There is also potential bias from siblings in the dataset. It is not known if there are siblings (or children with the same address regardless of relationship) in the dataset as all identifying information was removed from the dataset prior to receipt. Siblings or children living at the same address grow up in the same environment which may introduce bias into the data as they are not independent observations.

The hospital admission data is from hospital admissions from the Canterbury region, it does not include acute or elective admissions outside this region. This means that there may be patients who were admitted to hospital nationally or internationally outside Canterbury, and there is no information on this in this dataset. There is also no general practitioner data used in this study, which means that co-morbidities controlled or treated by a general practitioner are not accounted for or identified. These are potential sources of missing data. The CDHB has a world renowned integrated health system in which many acute and chronic conditions are managed outside the hospital system, until a patient has no other option other than hospitalisation [56]. This means there could be children who have conditions that fall under the disease categories used in this dataset, but they are not identified as having the condition as they are managed in the community.

The variables present in the data can also have misclassification issues which may introduce bias into the dataset which may result in a biased and inaccurate model. For example, the ethnicity variable does not allow for multiple ethnicities. This is a flaw in the current oral health database, which should be addressed.

Conclusion

Conventional epidemiology techniques and modern machine learning techniques were both employed and compared on their relative advantages and disadvantages to model a large routinely collected health dataset. They were compared on how well the models fit the data, and on their ability to predict dmft of a dataset not used for model building.

The Random Forest performed slightly better overall than the zero-inflated negative binomial in the residual diagnostic plots which indicates higher model reliability. Sensitivity and Specificity were tested and displayed using a ROC curve. The random forest model displayed higher sensitivity (0.203) than the zero-inflated negative binomial model (0.071). The zero-inflated negative binomial model performed slightly higher in specificity (0.972) than the random forest (0.926). Predictive ability was accessed using an unknown testing set, and assessed by comparing RMSE. The random forest model outperformed the zero-inflated negative binomial model, with a RMSE of 2.678 compared to 2.727.

The process of modelling using these two techniques were quite different from each other. The zero-inflated negative binomial model required an ensemble model technique, whereas the random forest model could handle the data as it was (i.e. all in one go). Where the random forest had parameters that could be tuned to better the model, the zero-inflated negative binomial could only be tuned by simplification. The zero-inflated negative binomial model took a long time to build and tune with a high level of user interaction, the random forest was slightly slower to build and tune but was predominantly automated.

Routinely collected data can be used to model and predict children with dental caries. However, there are likely other contributing factors not being measured by the routinely collected dental appointment data. The random forest model was able to use the data to explain 16.98% of the variation in the level of dmft, therefore there are other contributing factors not collected as part of the clinical data. The random forest modelling process was easier and more efficient than the zero-inflated negative binomial modelling process.

A random forest model could be used in the future to model routinely collected data as in this study, along with additional data on diet, oral health habits, water fluoridation and other factors known to contribute to early childhood caries and dmft. Variables identified by this study are of interest for this process, and future research could lead to better performing predictive models. Although the models generated in this study are not at a level to accurately predict dmft, it has shown that supervised learning techniques tend to perform better than conventional bio-statistical techniques when modelling large health data. This modelling technique, and identified risk factors, also have the potential to predict the severity of dental issues before the child attends an appointment, which could be extended in length if severe ECC are expected. This future model may be effective to predict dental caries in other New Zealand centres with similar population demographics. A future national random forest model could be built using data from all regions, not just data from the Canterbury region. The future model may also be effective to predict dental caries in other countries.

Competing interests

The authors declare that they have no competing interests, Dr Martin Lee is an employee of the Canterbury District Health Board from which the data was sourced.

Author's contributions

S.A.S is the guarantor. The research question and study design was initiated and developed by S.A.S. in conjunction with M.L and P.J.S., and with assistance and refinement from J.A.B. Data were extracted by M.L. S.A.S. led all statistical analyses with assistance from M.L., and all authors contributed to interpreting and reviewing the results. S.A.S. led the drafting of each manuscript iteration, which was then reviewed and commented upon by all authors. All authors agreed with its final contents.

Acknowledgements

We would like to thank Nicholas Kay for his support, and Bronwyn Kay and Michael Hobbs for proofing the document. We also thank the Canterbury District Health Board for access to these data.

Author details

[1]Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, 8140, Christchuch, New Zealand. [2]Canterbury District Health Board, Community Dental Service, Sylvan St., P.O.Box 731, 8024, Christchurch, New Zealand. [3]University of Canterbury, School of Health Sciences, Private Bag 4800, 8140, Christchurch, New Zealand. [4]The University of Queensland, School of Clinical Medicine, St Lucia, QLD, 4072, Brisbane, Australia.

References

1. Hayden C, Bowler JO, Chambers S, Freeman R, Humphris G, Richards D, et al. Obesity and dental caries in children: a systematic review and meta-analysis. Community Dentistry and Oral Epidemiology. 2013;41:289–308.
2. Torriani DD, Ferro RL, Bonow MLM, Santos IS, Matijasevich A, Barros AJ, et al. Dental Caries Is Associated with Dental Fear in Childhood: Findings from a Birth Cohort Study. Caries Research. 2014 07;48(4):263–70.
3. de Souza Barbosa T, Gavião MBD, Castelo PM, Leme MS. Factors Associated with Oral Health-related Quality of Life in Children and Preadolescents: A Cross-sectional Study. Oral Health and Preventive Dentistry. 2016;14(2):137–48.
4. Goettems ML, Shqair AQ, Bergmann VF, Cadermatori MG, Correa MB, Demarco FF. Oral health self-perception, dental caries, and pain: the role of dental fear underlying this association. International Journal of Paediatric Dentistry. 2018;28(3):319–325.
5. Jordan AR, Becker N, Jöhren HP, Zimmer S. Early childhood caries and caries experience in permanent dentition: A 15-year cohort study. Swiss Dental Journal. 2016;126(2):114–119.
6. California Society of Pediatric Dentistry. The consequences of untreated dental disease in children;. Available from: https://www.cda.org/Portals/0/pdfs/untreated_disease.pdf/.
7. World Health Organization. Oral Health; 2018. Accessed 28 Oct 2018. Available from: http://www.who.int/news-room/fact-sheets/detail/oral-health.
8. World Health Organization. Oral health important target groups;. Accessed 28 Oct 2018. Available from: http://www.who.int/oral_health/action/groups/en/.
9. World Health Organization. World oral health report 2003: continuous improvement of oral health in the 21st century – the approach of the WHO Global Oral Health Programme. 2003;.
10. Petersen PE, Bourgeois D, Ogawa H, Estupinan-Day S, Ndiaye C. The global burden of oral diseases and risks to oral health. Bulletin of the World Health Organization. 2005 09;83(9):661–9. Available from: http://search.proquest.com.ezproxy.canterbury.ac.nz/docview/229548088?accountid=14499.
11. World Health Organization. Equity, social determinants and public health programmes. Geneva: World Health Organization; 2010.
12. Tickle M. The 80:20 phenomenon: help or hindrance to planning caries prevention programmes? Community Dental Health. 2002 March;19(1):39–42.
13. Ministry of Health. Annual data explorer 2016/17, New Zealand health survey [Data File]; 2017. Available from: https://minhealthnz.shinyapps.io/nz-health-survey-2016-17-annual-update/.
14. Silva PA, Stanton WR. From child to adult: The Dunedin multidisciplinary health and development study. Auckland: Oxford University Press.; 1996.

15. Thomson WM, Poulton R, Milne BJ, Caspi A, Broughton JR, Ayers KMS. Socioeconomic inequalities in oral health in childhood and adulthood in a birth cohort. Community Dentistry and Oral Epidemiology. 2004 10;32(5):345–353. Available from: http:https://doi.org/10.1111/j.1600-0528.2004.00173.x.
16. Petcu A, Balan A, Haba D, Stefanache AMM, Savin C. Implications of Premature Loss of Primary Molars. Pediatric Dentistry. 2016;6(2):130–134.
17. Gray DJ. Preservation of the teeth indispensable to comfort and appearance, health and longevity, being a new edition of dental practice. Richard and John E. Taylor; 1842.
18. Lo E. Epidemiology: The DMF Index;. Accessed 27 Dec 2018. Available from: https://www.dentalcare.com/en-us/professional-education/ce-courses/ce368/epidemiology-the-dmf-index.
19. Ministry of Health. Publicly funded dental care; 2017. Accessed 28 Jan 2018. Available from: https://www.health.govt.nz/your-health/services-and-support/health-care-services/visiting-dentist/publicly-funded-dental-care.
20. Canterbury District Health Board. Community Dental Service; 2017. Accessed 28 Jan 2018. Available from: http://www.cdhb.health.nz/Hospitals-Services/Community-Rural-Health-Services/Community-Dental/Pages/default.aspx.
21. Titanium Solutions software. Auckland, New Zealand; 2005. Available from: https://www.titanium.solutions/.
22. Environmental Systems Research Institute (ESRI); 2012. ArcGIS Release 10.1.
23. Exeter DJ, Zhao J, Crengle S, Lee A, Browne M. The New Zealand indices of multiple deprivation (IMD): a new suite of indicators for social and health research in Aotearoa, New Zealand. PLoS One; 2017. Accessed 20 Feb 2018. Available from: https://doi.org/10.1371/journal.pone.0181260.
24. Atkinson J, Salmond C, Crampton P. NZDep2013 Index of Deprivation. 2014 May;Accessed 20 Feb 2018.
25. Smith BJ. Understanding pre-school children's Community Dental Service appointment failure: a mixed-methods study; 2016. Accessed 28 Jan 2018. Available from: https://ir.canterbury.ac.nz/bitstream/handle/10092/13165/Smith%2C%20Belinda%20MHLSc%20Thesis.pdf?sequence=1.
26. R Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2015. Available from: http://www.R-project.org/.
27. World Health Organization. Classification of Diseases; 2018. Accessed 3 Jan 2019. Available from: https://www.who.int/classifications/icd/en/.
28. Health Quality and Safety Commission New Zealand. Childhood ambulatory sensitive hospitalisations; 2016. Accessed 3 Jan 2019. Available from: https://www.hqsc.govt.nz/our-programmes/health-quality-evaluation/projects/atlas-of-healthcare-variation/childhood-ambulatory-sensitive-hospitalisations/.
29. Kuhn M. caret: Classification and regression training; 2017. R package version 6.0-78. Available from: https://CRAN.R-project.org/package=caret.
30. Alin A. Multicollinearity. WIREs Computational Statistics. 2010;2:370–374.
31. Hilbe JM. Modeling count data;.
32. Chin HC, Quddus MA. Modeling Count Data with Excess Zeroes: An Empirical Application to Traffic Accidents. Sociological Methods & Research. 2003;32(1):90–116.
33. NCSS. Zero-inflated negative binomial regression;. Available from: https://ncss-wpengine.netdna-ssl.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/ZeroInflated$_{N}egative_{B}inomial_{R}egression.pdf$.
34. Vuong QH. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica. 1989;57(2):307–333. Available from: http://www.jstor.org/stable/1912557.
35. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer; 2002.
36. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Miscellaneous functions of the department of statistics (e1071), TU Wien; 2014. R package version 1.6-4. Available from: http://CRAN.R-project.org/package=e1071.
37. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22. Accessed 20 Dec 2018. Available from: http://CRAN.R-project.org/doc/Rnews/.
38. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. 1st ed. Springer; 2013.
39. Probst P, Wright M, Boulesteixr AL. Hyperparameters and tuning strategies for random forest; 2018. Accessed 20 Dec 2018. Available from: https://arxiv.org/pdf/1804.03515.pdf.

40. Breiman L, Cutler A. Breiman and Cutler's random forests for classification and regression; 2018. Accessed 20 Dec 2018. Available from: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.
41. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22. Accessed 20 Dec 2017. Available from: https://CRAN.R-project.org/doc/Rnews/.
42. Parikh R, Mathai A, Parikh S, Sekhar GC, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian Journal of Ophthalmology. 2008;56(1):45–50. Accessed 20 Dec 2018. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636062/.
43. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care and Pain. 2008;8(6):221–223. Accessed 20 Dec 2018. Available from: http://dx.doi.org/10.1093/bjaceaccp/mkn041.
44. NZ S. 2013 Census QuickStats about a place: Canterbury Region; 2013. Accessed 20 Jan 2019. Available from: http://archive.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-about-a-place.aspx?request$_value = 14703tabname = Culturaldiversity$.
45. Hamner B, Frasco M. Metrics: Evaluation metrics for machine learning; 2017. R package version 0.1.3. Available from: https://CRAN.R-project.org/package=Metrics.
46. Cruvinel VRN, Gravina DBL, Azevedo TDPL, Bezerra ACB, Toledo OAd. Prevalence of dental caries and caries-related risk factors in premature and term children. Brazilian Oral Research. 2010 09;24:329 – 335. Available from: http://www.scielo.br/scielo.php?script=sci$_artextpid = S1806 - 83242010000300012nrm = iso$.
47. Javali S, Pandit P. Use of the generalized linear models in data related to dental caries index; 2007. Accessed 5 Jan 2019. Available from: http://www.ijdr.in/article.asp?issn=0970-9290;year=2007;volume=18;issue=4;spage=163;epage=167;aulast=Javali;t=6.
48. Lu HX, Wong MCM, Lo ECM, McGrath C. Risk indicators of oral health status among young adults aged 18Â years analyzed by negative binomial regression. BMC Oral Health. 2013 August;13(1):40. Available from: https://doi.org/10.1186/1472-6831-13-40.
49. Angelino K, Shah P, Edlund DA, Mohit M, Yauney G. Clinical validation and assessment of a modular fluorescent imaging system and algorithm for rapid detection and quantification of dental plaque. BMC Oral Health. 2017 Dec;17(1):162.
50. Nakano Y, Suzuki N, Kuwata F. Predicting oral malodour based on the microbiota in saliva samples using a deep learning approach. BMC Oral Health. 2018 Jul;18(1):128.
51. Rana A, Yauney G, Wong LC, Gupta O, Muftu A, Shah P. Automated segmentation of gingival diseases from oral images. In: 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT); 2017. p. 144–147.
52. Cagetti MG, Bontà G, Cocco F, Lingstrom P, Strohmenger L, Campus G. Are standardized caries risk assessment models effective in assessing actual caries status and future caries increment? A systematic review. BMC Oral Health. 2018 Jul;18(1):123. Available from: https://doi.org/10.1186/s12903-018-0585-4.
53. Stafford RJ, Schluter P, Kirk M, Wilson A, Unicomb L, Ashbolt R, et al. A multi-centre prospective case-control study of campylobacter infection in persons aged 5 years and older in Australia. 2006;135(6):978–88.
54. Kim B. Understanding Diagnostic Plots for Linear Regression Analysis. University of Victoria Library; 2015. Accessed 5 Feb 2019. Available from: https://data.library.virginia.edu/diagnostic-plots/.
55. Rothman K, Greenland S, Lash T. Modern Epidemiology, 3rd Edition. Lippincott Williams Wilkins; 2008.
56. Charles A. Developing accountable care systems, Lessons from Canterbury, New Zealand. Kings Fund; 2017. Accessed 21 Jan 2018. Available from: https://www.kingsfund.org.uk/publications/developing-accountable-care-systems.

Figures

Figure 1 Scatter plot matrix example used for initial exploratory data analysis

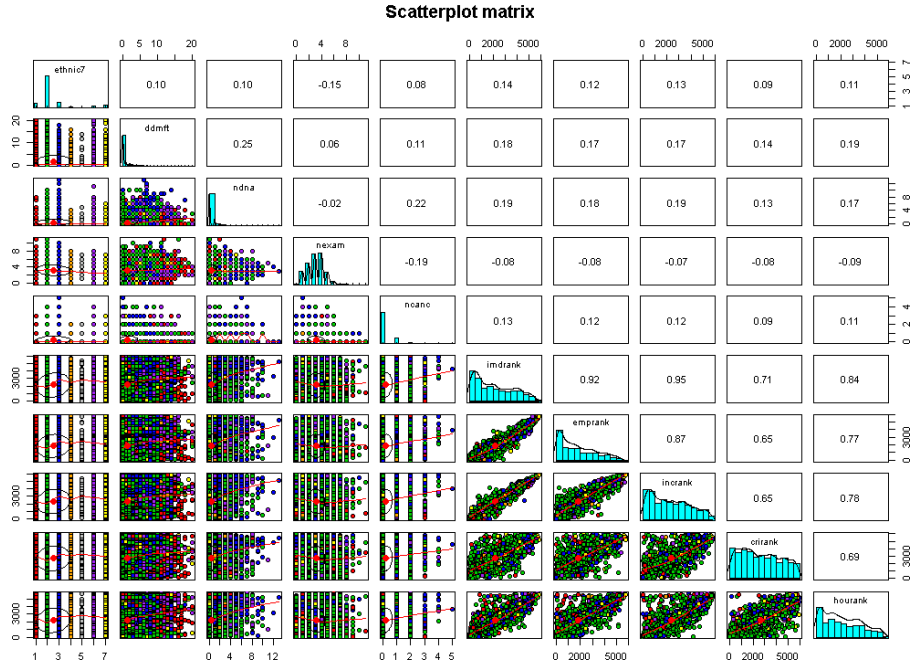


Figure 2 Density histograms for Decayed, Missing or Filled Primary Teeth For each subset of the data, The model training set(train), the model tuning set (valid) and the final model testing set (test).

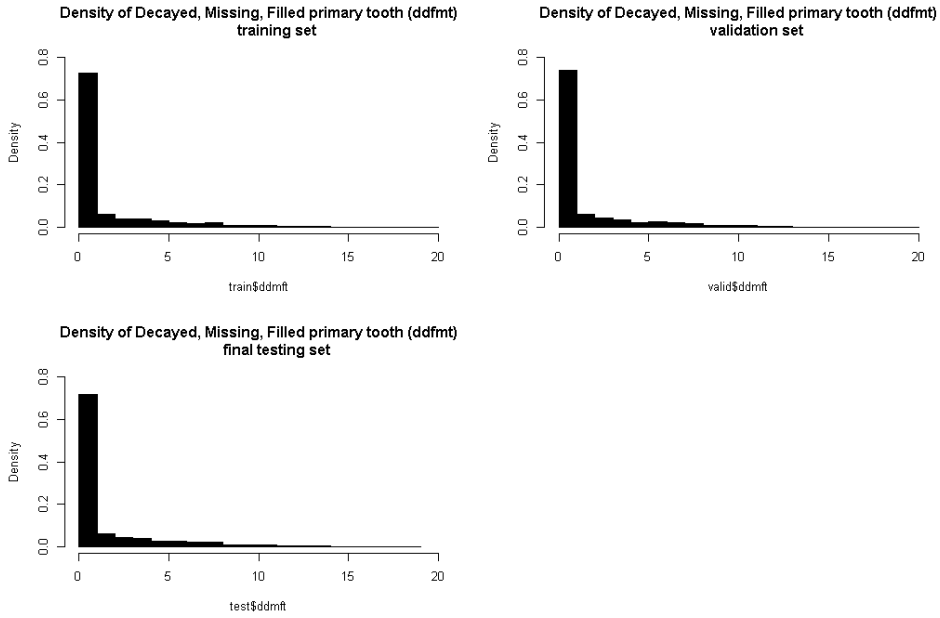


Figure 3 Final random forest variable importance plot A visual display of the important variables listed from highest importance to lowest importance

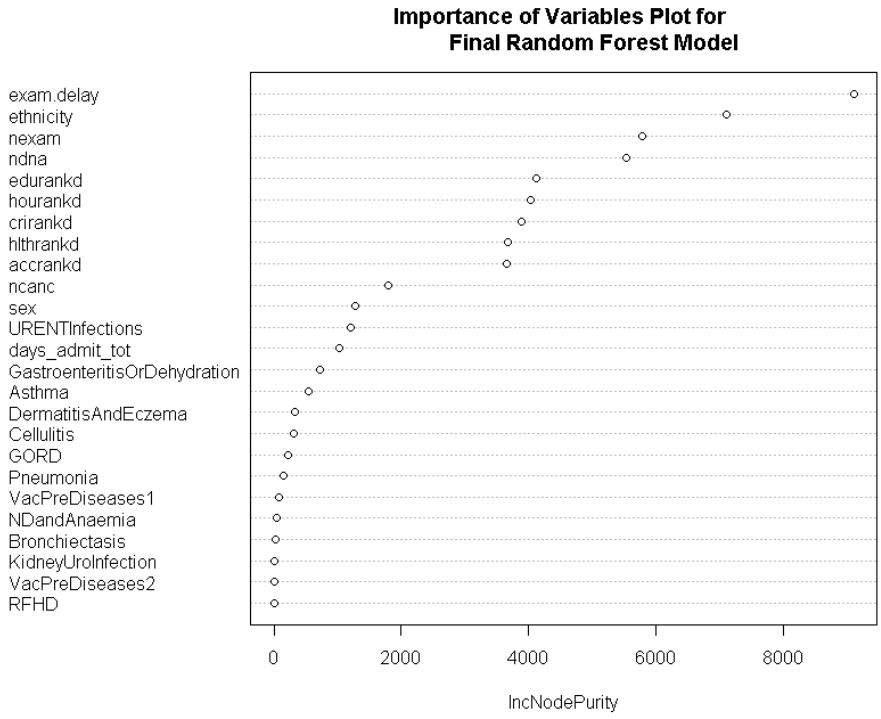
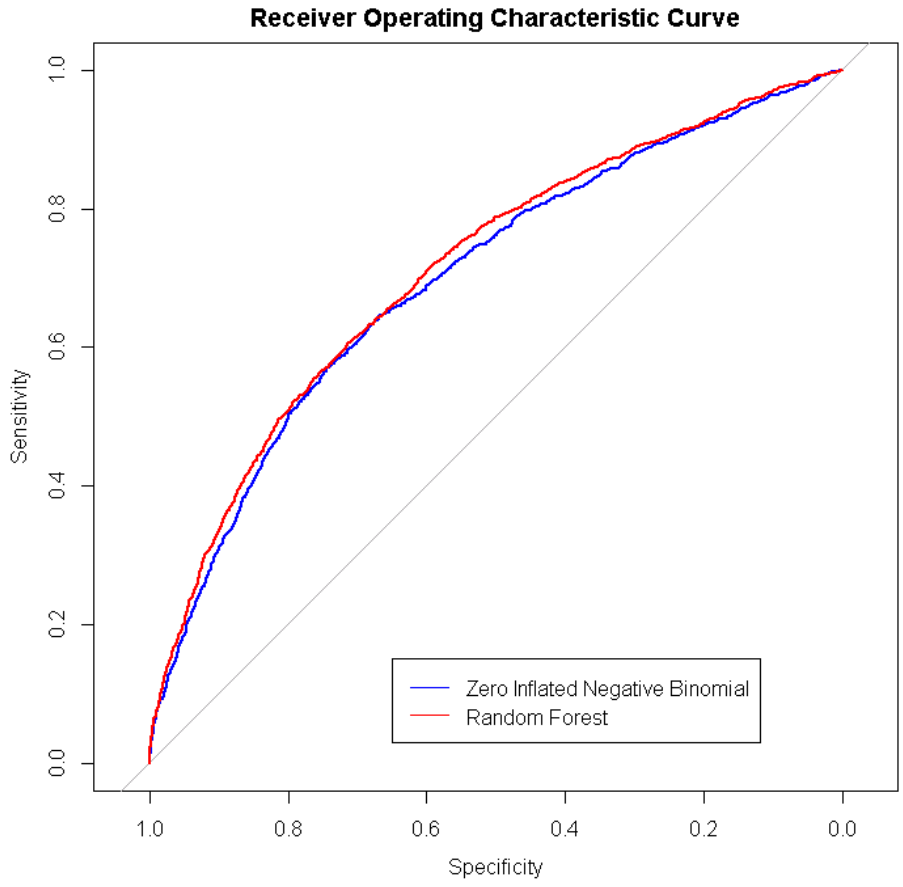


Figure 4 Receiver Operator Curve for the final zero-inflated negative binomial model and the final random forest model



Tables

Table 1 Table of variables in the full clean dataset, with explanations.

Variable Code	Variable Explanation	Levels	Ensemble Model
ethnicity	Ethnicity	1: European, 2: Māori, 3: Pacifika, 4: Asian, 5: Other, 6: Unknown	1
ddfmt	decayed, missing and filled primary teeth	Count of dmft	
ndna	Number of Did Not Attend appointments	Count of dna	1
sex	sex of the patient	1: male, 2: female, 3:unknown	1
nexam	Number of dental examinations	count of examinations	1
ncanc	Number of cancelled dental Appointments	count of cancelled Appointments	1
imdrankd**	Overall deprivation decile	decile rank between 1 and 10*	2
emprankd***	Employment decile	decile rank between 1 and 10*	2
incrankd***	Income decile	decile rank between 1 and 10*	2
crirankd	Crime decile	decile rank between 1 and 10*	2
hourankd	Housing decile	decile rank between 1 and 10*	2
hlthrankd	Health decile	decile rank between 1 and 10*	2
edurankd	Education decile	decile rank between 1 and 10*	2
accrankd	Access to Services decile	decile rank between 1 and 10*	2
imdnoempd**	IMD decile with Employment Domain removed	decile rank between 1 and 10*	2
imdnoincd**	IMD decile with Income Domain removed	decile rank between 1 and 10*	2
imdnocrid**	IMD decile with Crime Domain removed	decile rank between 1 and 10*	2
imdnohoud**	IMD decile with Housing Domain removed	decile rank between 1 and 10*	2
imdnohlthd**	IMD decile with Health Domain removed	decile rank between 1 and 10*	2
imdnoedud**	IMD decile with Education Domain removed	decile rank between 1 and 10*	2
imdnoaccd**	IMD decile with Access Domain removed	decile rank between 1 and 10*	2
days_admit_tot	days admitted in hospital total	count of days	3
exam.delay	length of time between referral into dental services and treatment date	count of days	3
Asthma	Hospital admissions for Asthma	count of admissions	3
Bronchiectasis	Hospital admissions for Bronchiectasis	count of admissions	3
DertmatitisAnd Eczema	Hospital admissions for Dermatitis or Eczema	count of admissions	3
GastroenteritisOr Dehydration	Hospital admissions for gastroenteritis or dehydration	count of admissions	3
GORD	Hospital admissions for Gastro-oesphageal reflux disease	count of admissions	3
KidneyUroInfection	Hospital admissions for Kidney or Urological infection	count of admissions	3
Pneumonia	Hospital admissions for respiratory infections: pneumonia	count of admissions	3
RFHD	Hospital admissions for Rheumatic fever/heart disease	count of admissions	3
URENTInfections	Hospital admissions for upper respiratory and ear, nose or throat Infections	count of admissions	3
VacPreDiseases1	Hospital admissions for Vaccine-preventable diseases: meningitis, whooping cough, hepatitis B, pneumococcal disease , other	count of admissions	3
VacPreDiseases2	Hospital admissions for Vaccine-preventable diseases: MMR	count of admissions	3

* All deciles 1: least deprived, 10: most deprived ** These variables removed when r>0.9 ***These variables removed when r>0.7

Table 2 Distribution of dmft by sex, ethnicity and deprivation

	Summary of dmft			
	n (%)	mean (sd)	median [Q1,Q3]	maximum
Sex				
Male	10,666 (50.41%)	1.538 (2.864)	0 [0,2]	20
Female	10,483 (49.55%)	1.475 (2.818)	0 [0,2]	20
Unknown	9 (0.04%)	1.000 (2.291)	0 [0,1]	7
Ethnicity				
European	14,256 (67.38%)	0.985 (2.185)	0 [0,1]	20
Māori	2,421 (11.44%)	2.507 (3.287)	1 [0,4]	18
Pasifika	949 (4.49%)	3.530 (4.022)	2 [0,6]	20
Asian	1,996 (9.43%)	2.883 (4.072)	1 [0,5]	20
Other	388 (1.83%)	1.807 (3.124)	0 [0,2]	16
Unknown	1,148 (5.43%)	1.705 (3.008)	0 [0,2]	20
Deprivation (IMD decile)				
1	4554 (21.52%)	0.963 (2.255)	0 [0,1]	19
2	3466 (16.38%)	1.151 (2.488)	0 [0,1]	20
3	2109 (9.97%)	1.321 (2.624)	0 [0,1]	16
4	2421 (11.44%)	1.347 (2.661)	0 [0,2]	18
5	1896 (8.96%)	1.589 (2.997)	0 [0,2]	20
6	1581 (7.47%)	1.580 (2.823)	0 [0,2]	20
7	2018 (9.54%)	2.051 (3.237)	0 [0,3]	19
8	1308 (6.18%)	2.152 (3.365)	0 [0,3]	20
9	1311 (6.20%)	2.683 (3.524)	1 [0,5]	19
10	423 (2.00%)	3.288 (3.781)	2 [0,6]	16
Total	21158 (100%)	1.507 (2.841)	0 [0,2]	20

Table 3 Results of the zero-inflated negative binomial model: decayed missing filled teeth predicted by ethnicity + number of did not attend appointments + number of examinations + number of cancelled appointments + education decile + housing decile + born in Christchurch + number of acute hospital admissions + number of diaper dermatitis hospital admissions + number of other and non-specified dermatitis hospital admissions * statistically significant coefficients with p-value < 0.05

Variables	coefficient	standard error	p-value
Intercept	1.924	0.093	0.000 *
Ethnicity-Māori	-0.970	0.085	0.000 *
Ethnicity-Pasifika	-1.249	0.138	0.000 *
Ethnicity-Asian	-1.105	0.084	0.000 *
Ethnicity-Other	-0.451	0.174	0.009 *
Ethnicity-Unknown	-0.648	0.108	0.000 *
Number of did not attend appointments	-0.450	0.037	0.000 *
Number of dental examinations	-0.162	0.019	0.000 *
Number of cancelled appointments	-0.301	0.058	0.000 *
Education decile rank	-0.080	0.011	0.000 *
Housing decile rank	-0.017	0.012	0.166
Asthma	-0.153	0.084	0.067
Upper Respiratory/ Ear, Nose and Throat Infections	-0.136	0.042	0.001 *

Table 4 Table of Root Mean Squared Error for both the conventional and machine learning techniques

Model	Root Mean Squared Error (3 d.p.)
Zero-inflated negative binomial models	
Full Dataset, Group 1, reduced by AIC	2.568*
Reduced Dataset, r>0.9, Group 1, reduced by AIC	2.568*
Reduced Dataset, r>0.7, Group 1, reduced by AIC	2.568*
Full Dataset, Group 2, reduced by AIC	2.730
Reduced Dataset, r>0.9, Group 2, reduced by AIC	2.730
Reduced Dataset, r>0.7, Group 2, reduced by AIC	2.729
Full Dataset, Group 3, reduced by AIC	2.800**
Reduced Dataset, r>0.9, Group 3, reduced by AIC	2.800**
Reduced Dataset, r>0.7, Group 3, reduced by AIC	2.800**
Ensemble model, reduced by AIC	2.544
Ensemble model, reduced by AIC	2.727***
Random Forest Model	
Full Dataset	2.584
Reduced Dataset, p>0.9	2.566
Reduced Dataset, p>0.7	2.566
Full Dataset, tuned mtry	2.576
Reduced Dataset, p>0.9, tuned mtry	2.550
Reduced Dataset, p>0.7, tuned mtry	2.535
Full Dataset, tuned mtry, tuned trees (2,500 trees)	2.574
Reduced Dataset, p>0.9, tuned mtry, tuned trees (2,500 trees)	2.546
Reduced Dataset, p>0.7, tuned mtry, tuned trees (1,500 trees)	2.539
Reduced Dataset, p>0.7, tuned mtry, tuned trees (1,500 trees)	2.678***

*Group 1 models are all the same model as no variables were removed during data simplification

** Group 3 models are the same model when reduced by AIC

***Root Mean Squared Error when predicting testing dataset, all others predicting validation set

Supplementary Materials

Supplementary Material 1: R Code for Data Cleansing, joining and analysis

This file is a .txt file of the r commands for the data cleansing, joining and analysis.

Supplementary Material 2A: Locality Agreement

This file is a .pdf file of the locality agreement for ethnics approval.

Supplementary Material 2B: University of Canterbury internal ethics approval

This file is a .pdf file of the internal University of Canterbury ethics approval.

Supplementary Material 2C: HDEC out of scope letter

This file is a .pdf file of the confirmation letter from HDEC that this study did not require ethics approval.

Supplementary Material 3: Residual diagnostics

This file is a .pdf file of the full explanation and plots for the residual diagnostics.