Predicting Ophthalmic Clinic Non-attendance using Machine Learning

Finley Breeze  
https://orcid.org/0000-0003-2963-1925

Ruhella Hossain  
University of Auckland  https://orcid.org/0000-0002-1613-9598

Michael Mayo  
James McKelvie (james@mckelvie.co.nz)  
University of Auckland  https://orcid.org/0000-0002-6460-4175

Article

Keywords:

Posted Date: July 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1725499/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

Ophthalmic clinic non-attendance in New Zealand is associated with poorer health outcomes, marked inequities and costs NZD$30m per annum. Initiatives to improve attendance typically involve expensive and ineffective brute-force strategies.

Purpose

To develop machine learning models to accurately predict ophthalmic clinic non-attendance.

Methods

Nationwide ophthalmology clinic attendance data was aggregated for analysis. In addition to a single global model, data were stratified by DHB to train regional models. Model weighting was adjusted to account for the highly imbalanced dataset. Repeated ten-fold cross validation was used to ensure that model selection and evaluation metrics were robust. Logistic regression and XGBoost model performance were evaluated on all NZ publicly funded ophthalmology clinics.

Results

Data included 3.1 million clinic appointments with 5.8% non-attendance rate from 2009 to 2018. Appointments from 2018 onwards were used as a test set outside of the cross-validation experiments to simulate prospective evaluation. Across multiple centres, an average model AUROC of 0.76 and AUPRC of 0.16 was achieved on the test data. DHBs with higher non-attendance rates had improved AUPRC. For example, Tairawhiti DHB had 7.8% non-attendance and achieved AUPRC of 0.29.

Conclusion

It is possible to use machine learning algorithms to make clinically useful predictions of clinic non-attendance. The model in the current study is competitive with previously published models of clinic attendance in the literature using fewer input features. This level of discrimination is high enough to be used in advanced scheduling methods and targeted public health interventions.

Introduction

Non-attended scheduled outpatient appointments are a significant issue in New Zealand (NZ) ophthalmology departments. Non-attendances, also known as “did-not-attends” (DNAs), are defined as an
outpatient event in which the patient fails to attend a scheduled appointment without cancelling or rescheduling in advance.

Non-attendance results in an under-utilisation of limited health resources. Across all outpatient departments in NZ, the total cost of missed specialist appointments is estimated to be in excess of $29 million per year.\textsuperscript{1} In ophthalmology, non-attendance is linked to poorer visual acuity outcomes with statistically significant associations between clinic attendance and ocular conditions such as amblyopia,\textsuperscript{2} age-related macular degeneration,\textsuperscript{3–5} glaucoma\textsuperscript{6,7} and diabetic retinopathy.\textsuperscript{8–10} Māori and Pasifika communities in New Zealand suffer disproportionately, with socioeconomic and ethnic inequities associated with increased non-attendance rates and poor health outcomes.\textsuperscript{11}

Non-attendance is a complicated issue correlated with many different factors relating to ophthalmic services, patient demographics and wider environmental influences.\textsuperscript{12,13} Machine learning methods excel in distilling high-dimensional attendance data into meaningful predictions or probabilities that can be used to address non-attendance. Successfully predicting a patient’s risk of non-attendance will enable better use of resources and improve health outcomes. Non-attendance models can be used to optimise clinic scheduling practices to maximise health outcomes and reduce the risk of non-attendance. They can also be easily scaled to provide significant health benefits and cost savings on a national level.

Although the potential of machine learning is widely recognised, public healthcare has been cautious to adopt machine learning solutions. In order to gain the trust of healthcare professionals, how and why machine learning generated predictions are made must be transparent and understandable. Often the models that achieve the highest performance are also the most complex and are therefore difficult to interpret. This study seeks to compare machine learning algorithm performance whilst accounting for the potential trade off in interpretability with increased model complexity.

The aim of this study was to develop machine learning models to accurately predict non-attendance in all of New Zealand’s publicly-funded ophthalmology clinics. This study uses nationwide multicentre data to develop predictive models and evaluate predictive performance for single District Health Boards (DHBs). In addition, the current study aims to quantify the relative contribution of ethnicity for model predictions and how machine learning models can be optimised to enable equitable health outcomes.

**Materials And Methods**

The study was conducted according to the tenets of the Declaration of Helsinki and the National Ethics Advisory Committee guidelines\textsuperscript{14} and met the criteria for exemption from formal review by the New Zealand Health and Disability Ethics Committee.\textsuperscript{15}

**Dataset**
De-identified appointment data from the New Zealand Ministry of Health (MoH) database were used for modelling. The dataset included all scheduled public-funded ophthalmology appointments in New Zealand for 10 years (1st July 2009 to 30th June 2019). The outcome variable was patient non-attendance, coded by the MoH as a Did Not Attend (DNA). These are defined as an outpatient event in which the patient fails to attend a scheduled appointment without cancelling or rescheduling in advance.

The dataset included eight other variables that can be broadly categorised into three groups; patient demographic, clinic-related and time-based factors. Demographic variables included the age, gender, ethnicity and NZ Deprivation Index (NZDep). NZDep is an area-based measure of socioeconomic deprivation in New Zealand based on nine Census variables.\(^{16}\)

Clinic-related variables were clinic type (e.g. diabetic clinic, minor operation), whether the appointment was for a First Specialist Assessment (FSA) or a follow up visit, and DHB where the clinic appointment took place. There are 20 DHBs in New Zealand, which are responsible for providing and funding the provision of health services for their district. The only time-based variable in the dataset was the year the appointment took place. Appointment year was only used to partition the dataset into the 2009–2017 training set and 2018 testing set, then removed from further analysis.

To test the hypothesis that regional models would perform better than a nationwide model, the dataset was stratified into DHB subsets. This allowed for models to be trained on the entire nationwide dataset and compared with models restricted to a regional subset.

**Feature engineering**

All features apart from patient age were categorical. The categorical variables were one-hot-encoded into binary features (0 or 1) for each category. Chi-squared was used for significance testing. As the only numerical feature, patient age was standardised.

**Development, Validation and Scoring of Predictive Models**

Model training and evaluation was performed using machine learning packages *scikit-learn*\(^ {17}\) and *XGBoost*\(^ {18}\) in Python (version 3.7.13).

Two different model types were investigated; logistic regression and eXtreme Gradient Boosting machine (XGBoost). Appointments from the 1st July 2009 until the 30th June 2018 were used as the training set. To compare models and tune hyperparameters, a repeated 10-fold cross-validation of the training set was used. Folds were stratified to preserve the percentage of samples for each class. Best performing hyperparameters for each algorithm in cross-validation were then used to predict the testing set examples from the financial year 2018 (1st July 2018–30th June 2019).

To account for the imbalanced dataset, the weights of the training samples were set to be inversely proportional to the training set’s class distribution. This gave the rare class of non-attendance greater weight in predictions.
Relevant hyperparameters settings used in logistic regression were \( C = 1.0 \) (inverse of regularisation strength) with max iterations set to 10000. For XGBoost, maximum tree depth was six with 100 estimators.

The models were predominantly evaluated with the area under the receiver operating characteristics curve (AUROC). Other scoring metrics were also recorded; area under the precision-recall curve (AUPRC) - calculated as average precision, accuracy, sensitivity and specificity.

**Feature importance**

Feature importance for both logistic regression and XGBoost was calculated using SHapley’s Additive exPlanations (SHAP) values.\(^{19}\) The XGBoost algorithm is a tree-based ensemble method, so coefficients for individual features cannot be generated. SHAP values are a model agnostic post-hoc explanation technique that can determine feature importances and their impact on predictions.

**Results**

The final testing set consisted of 407,574 NZ ophthalmology appointments with a non-attendance rate of 5.7% \((n = 23,309\text{ missed appointments})\).

All categorical variables in the dataset had a statistically significant association with attendance using chi-squared testing. Bonferroni correction was applied to compensate for multiple comparisons, where \( p < 0.002 \) was considered significant. The difference between the mean age of non-attended appointments (57.7 years, std = 25.2) and attendances (43.7 years, std = 26.7), was also statistically significant \((p < 0.001)\) using Two Sample T-testing. After one-hot encoding there were 22 input features in total.

**Nationwide XGBoost Model**

Figure 1 summarises the performance of this nationwide model compared with regionally restricted models on the testing set from each DHB (i.e. appointments from 2018). On average, models that were restricted to regional data had statistically better performance.

An AUROC of above 0.7 was achieved on the testing set for 19 out of 20 DHBs. Models achieved a mean AUROC of 0.756 (std = 0.060) for logistic regression and 0.764 (std = 0.058) for XGBoost. The top performing models were Auckland, Bay of Plenty, Lakes and Waitemata - achieving AUROC of 0.90, 0.88, 0.83, 0.81 respectively. The ophthalmology non-attendance rates recorded for these DHBs were all extremely low \(< 0.8\%\). DHBs with non-attendance rates higher than 3% did not achieve AUROC scores above 0.8.

For DHBs exceeding 3% recorded non-attendance, the top performing models were XGBoost classifiers trained on regional data for Tairāwhiti, Taranaki and Whanganui with AUROCs of 0.790, 0.785 and 0.781 respectively. The nationally trained XGBoost also achieved an AUROC of 0.790 on the Wairarapa testing set which was better than the Wairarapa regionally trained model (0.773).
Regional Models: Logistic Regression and XGBoost

Comparing the upper and lower figure, AUROC scores on the testing set for all 20 DHBs lie outside the confidence intervals of repeated 10-fold cross-validation results on the training set (2009–2017).

Global interpretability - Overall feature importance

Local interpretability - Individual appointment

Discussion

This study used national and regional datasets to develop machine learning models which predict patient non-attendance in New Zealand’s public-funded ophthalmology clinics. To our knowledge, this is the first major study predicting non-attendances for ophthalmology across multiple centres and clinic types. The current study compares model performance across multiple centres and gives comprehensive explanations of predictions compared with previous studies.

The current study has four main findings; first, clinically useful predictions of non-attendance can be made with basic demographic and clinic information in New Zealand; secondly, XGBoost had better predictive performance than logistic regression on this dataset (p < 0.001); thirdly, restricting models to regional training data produces better predictive performance in some DHBs but not in others; and lastly, SHAP values offer a model-agnostic method of interpreting how features impact predictions of non-attendance.

Basic demographic and clinic data held in Electronic Health Records (EHR) can be used for predictive modelling of non-attendance. Models in the current study achieved AUROC scores between 0.7 and 0.9. This performance is competitive with other published predictive models of non-attendance, particularly for predicting FSA non-attendance.

Other published models of patient clinic non-attendance have reported AUROC scores over 0.820–23, with one study recently reporting 0.970 using a deep neural network.24 In these studies, several features were derived from previous patient appointment information. In particular, patient’s prior non-attendance history was found to be the most important feature. This indicates a high level of predictive performance can only be reliably achieved for patients with a history of previous clinic visit attendance. For example, a 2021 study developed a model specifically for follow-up patients which achieved an AUROC of 0.90 using XGBoost. However, for new patients, the best AUROC achieved was 0.74 using LASSO regression (XGBoost for new patients had an AUROC of 0.64).20

In comparison, the current study only used features derived from demographic and clinic-type information which does not require data from previous appointments. As a result, the models in the current study were effective in predicting attendance for both new and follow-up patient visits. In terms of health outcomes, accurately predicting FSA non-attendance is particularly important. FSA appointments
are typically when a clinical diagnosis is made and treatment is initiated. In some cases, a delay in diagnosis and/or initiating timely treatment may lead to irreversible vision loss.

The experimental results reported by other studies bring into question the robustness of their validation frameworks as they may not reflect performance on future, real-world data. In the studies discussed above, the metrics reported are from k-fold cross-validation on the training dataset. Although cross-validation ensures models are never tested on data they were trained on, it does not simulate forward testing of the model. It is possible the models may have overfit the training samples, making the reported AUROC scores overly optimistic for predicting attendance to future appointments. Studies using a large number of features also further increase the risk of overfitting due to models becoming unnecessarily complex and potentially fitting noise in the training data.

The current study used a robust validation strategy to avoid overfitting. A 10-fold cross-validation was used to determine the best hyperparameters and demonstrate performance on appointments from 2009–2017. The highest performing models were then used to predict the 2018 testing set appointments - with these being the main results reported. A notable finding in the current study is that all AUROC scores on the testing set (2018 examples) were outside of the confidence intervals for 10-fold cross-validation results on the training set (2009–2017). This indicates that cross-validation does not necessarily reflect the model’s performance on future appointments in the current study. Therefore, reserving the 2018 test set simulates forward testing of models and may better reflect their effectiveness in real-world applications.

The second main finding is that the improved performance achieved by XGBoost over logistic regression is statistically significant. In previous studies predicting non-attendance, regression-based models have been used extensively. XGBoost is a decision-tree based ensemble algorithm that uses a gradient boosting framework with regularisation. XGBoost algorithm is considered best-in-class for a wide range of structured data problems\textsuperscript{18} and has been the highest performing algorithm in other studies predicting patient non-attendance\textsuperscript{20}.

XGBoost scored higher than logistic regression with statistically significant improvements (95% CI) for all DHBs except West Coast (WC). On the testing set, XGBoost scored higher than logistic regression for all DHBs except West Coast and Wairarapa. Across all DHBs, the difference between the mean AUROC for XGBoost and logistic regression was statistically significant. These results support the use of machine learning approaches for epidemiological predictive modelling which is becoming increasingly more prevalent in the literature\textsuperscript{25}.

Other studies have predominantly used data from a single institution, so it is unclear whether nationwide-multicentre data would improve predictive performance. The current study aggregated nationwide attendance records from all publicly funded ophthalmology clinics in NZ and demonstrated that the highest predictive accuracy was achieved by restricting the model to individual regions. An XGBoost model was trained on nationwide data, with 2,915,667 training examples, and compared to the
performance of XGBoost models restricted to datasets from each DHB region. The average DHB had 145,783 training examples with a maximum of 483,102 (Counties Manukau) and minimum of 20,524 (West Coast). The mean performance of regionally-restricted models for each DHB was higher than the model trained on the complete nationwide dataset ($p = 0.04$). For example, Whanganui DHB appears to noticeably benefit from a regionally trained model. The Whanganui regional model may have learnt unique regional patterns in the 68,145 Whanganui examples from 2009–2017 that better predicts the 2018 non-attendances. Overall, these findings suggest the nationwide model does not generalise well to all regions, even with a limited feature set. Regional variations in clinic booking protocols, data quality, and appointment coding may have contributed to the reduced performance of the nationwide model. Different booking protocols and appointment coding also likely explain the extremely low non-attendance rates recorded by Auckland, Waitemata, Bay of Plenty and Lakes DHBs.

Previous studies using complex machine learning models to predict non-attendance are limited by the lack of transparency in explaining predictions.$^{20–24}$ The current study uses SHapley Additive exPlanations (SHAP values) to give global feature importance and local interpretations to explain why individual predictions were made. Feature importance differs slightly between logistic regression and XGBoost models. These differences are likely due to how the algorithms incorporate features into the final model. For both models, age is the most important feature with younger age being predictive of non-attendance. This is followed by ethnicity with European and Asian ethnicity being predictive of attendance, and Māori and Pasifika predictive of non-attendance. The highest deprivation quintile also ranks amongst the most important features for both algorithms. These findings reflect significant ethnic and socioeconomic inequities in access to ophthalmic care. It is likely the NZ health system does not effectively engage these communities with the current healthcare delivery model.$^{26}$

Clinic type ranks at several positions in features predictive of non-attendance in the current study. ‘Minor operation’ and ‘injection’ clinics have a strong negative impact on predictions (favour attendance), whereas ‘diabetic’ clinics have a strong positive impact (favour non-attendance). Interestingly, clinic types involving procedures favoured predictions of attendance - potentially reflecting the patient's perceived need for the appointment. Gender was shown to have lower absolute SHAP values, indicating it is less important in predicting non-attendance and attendance.

Using SHAP local interpretations enables hospitals to identify the factors which have the greatest impact on predictions for specific individual patient appointments. Models integrated into current booking systems would make this information readily available to help clinic staff identify and reduce factors that contribute to inequities in care, better engage with patients and develop targeted interventions. For example, clinic staff may choose to contact patients in advance of all appointments predicted to be high risk, or offer other convenient appointment slots at the time of booking. This personalised patient-focused approach is particularly important when non-attending patients can quickly become lost in the attendance records of the national healthcare system.
Factors that consistently carry high global importance in the model should prompt wider action. For example, minority ethnicity is predictive of non-attendances in the current study. Other studies predicting non-attendance have excluded race and race proxy (e.g. language) features to avoid ethnic bias in models.\textsuperscript{24} However, this strategy may not eliminate bias or inequity.\textsuperscript{27,28} In explainable models such as those used in the current study, making ethnicity information readily available promotes transparency surrounding inequities in care and is critical in developing culturally safe interventions.

The current study has limitations that could be addressed in future work. The main limitation was the nationwide dataset lacked information on previous appointments (e.g. previous non-attendances). Time-based variables, e.g. lead-time for appointment booking, time of day, day of week, and month for the scheduled appointment, have also been predictive features in prior studies. Data for these variables were not available for analysis in the current study. Using these features, the current models may have achieved better performance for follow-up patient visits. As a large number of models were trained in order to compare DHBs, analysis was limited to two algorithms. Further studies utilising other machine learning techniques may be useful. With non-attendance being a relatively infrequent event, class imbalance was a challenge in this dataset.

**Conclusion**

Machine learning models can be developed to provide clinically useful predictions of non-attendance in NZ ophthalmology clinics with relatively basic demographic and clinic data. Findings from the current study support the use of XGBoost for predicting non-attendances due to higher AUROC scores for 16 out of 20 DHBs compared with other algorithms. To address the problem of model interpretability, the authors suggest using SHAP values for feature importance and explaining predictions. The findings outlined in the current study have the potential to result in more efficient clinic management, reduced costs, development of targeted interventions for equitable access to care and ultimately improve visual outcomes.

**Declarations**

**ACKNOWLEDGEMENTS:**

The authors would like to thank the New Zealand Ministry of Health for providing the national dataset.

**CONFLICT OF INTEREST:**

None to report.

**FUNDING:**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
AUTHOR CONTRIBUTION STATEMENT:

FB was responsible for analysing data, designing the protocol, machine learning model development, creating tables and figures, interpreting results, and writing the report.

RH was responsible for acquisition and extraction of relevant data, analysing data and providing feedback on the report.

MM was responsible for designing the protocol, machine learning model development, interpreting results and providing feedback on the report.

JM was responsible for supervising and coordinating the study, designing the protocol, interpreting results and writing the report.

References


26. Scott N. A Maori cultural reluctance to present for care, or a systems and quality failure? How we pose the issue, informs our solutions. NZ Med J. 2014 May 2;127(1393):8–11.


Table

Table 1 is available in the Supplementary Files section.

Figures

Figure 1

Comparison of model AUROC scores for each District Health Board (DHB) 2018 testing data. DHBs are sorted in order of ascending recorded non-attendance rate, Min = 0.076% (Waitemata), Max = 9.8%
Green dots are models restricted to training data from that DHB ('DHB_restricted') and blue dots are models trained on nationwide data ('Nation_wide'). The mean AUROCs of the nationwide model and regional models are displayed with the dashed lines. The difference between these means was statistically significant (p=0.04) using the Two-Sample T-test.
AUROC of Logistic Regression and XGBoost for all DHB-restricted models. The upper figure shows District Health Board-restricted model scores in repeated 10-fold cross-validation on the 2009-2017 training data. As this generates 100 sets of results for each model, error bars are given. The lower figure shows the AUROC for the 2018 testing set. The mean AUROC of XGBoost (orange) and logistic regression (pink) type is displayed with the dashed lines. The difference between the means is statistically significant (p=0.002) on the test set using the Two-sample T-test.

Figure 3

Beeswarm plots of global feature importances for nationwide logistic regression (left) and XGBoost (right) using SHAP values. SHAP values reflect the degree to which the feature positively or negatively impacted the model. Other than the numerical feature for age (“Age_scaled”), feature names are formatted by the variable name (e.g. ‘Clinic’) and followed by the category that has been binary encoded (e.g. ‘Clinic_Diabetic’). Features are listed in order of decreasing importance in model predictions. For each feature listed, the adjacent dots correspond to samples from the 2018 testing set. Red represents a high feature value (i.e. ‘1’ indicating presence of the binary feature or older age for ‘Age_scaled’) and blue represents a low feature value (i.e. ‘0’ indicating absence of the binary feature or younger age). For example, testing samples where the clinic type was ‘injection’ are represented as a red dot beside ‘Clinic_Injection’ and a blue dot beside other clinic type binary features (i.e. ‘Clinic_Diabetic’, ‘Clinic_Optometrist’ etc). As XGBoost is a non-linear model, red dots appear spread out in horizontal lines due to the SHAP value (i.e. impact) of that binary feature varying in different predictions. In logistic regression (a linear model), binary features have the same impact in every prediction so dots are grouped at exactly the same SHAP value.
Figure 4

Waterfall plots showing the top 9 features affecting the prediction of 2 specific appointments from the 2018 testing examples. $E[f(X)]$ (expected value) is the average model output over the training examples. $f(x)$ is the model output for the specific testing example. Using a classification threshold of 0.5, the model predicts attendance in the upper figure ($f(x)<0.5$) and non-attendance in the lower figure ($f(x)>0.5$). These were both correct predictions of the testing example. SHAP values are used to calculate the relative
contribution of each feature, explaining the difference between these values. Due to weighting the minority class (non-attendance), these values do not directly represent probabilities of non-attendance.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- ALLRESULTSSupplementaryInfo.docx
- Table1.xlsx