# Developing a multivariate predictive model of hospital outpatient attendance for patients of all ages

Jonathan Holdship*†[1], Harpreet Dhanoa†[1], Adrian Hopper[1], Claire J. Steves[1,2], Mark Butler[1], Ingrid Wolfe[3], Katie Tucker[1], Carolyn Cooper[1], Jeremy Yates[4]

[1]Guy's and St Thomas' NHS Foundation Trust, London, UK

[2] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK

[3]Department of Women and Children's Health, School of Life Course Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom

[4] Computer Science Department, University College London, London, United Kingdom

* Corresponding author. Email: jonathan.holdship@gstt.nhs.uk

† These authors contributed equally to this work

**Objectives** Patient non-attendance at outpatient appointments (DNA) is a major concern for healthcare providers. Non-attendances increase waiting lists, reduce access to care and may be detrimental not for the patient who did not attend. However, non-targeted interventions to

reduce the DNA rate may not be effective and thus we aim to produce a model which can accurately predict which appointments will be attended.

**Methods** In this work, a random forest classification algorithm was trained to predict whether an appointment will be missed using 7 million past outpatient appointments. The model was applied to patients of all ages and appointments with all specialties at a major London teaching hospital including a validation set covering the COVID-19 pandemic.

**Results** The model achieves an AUROC score of 0.76 and accuracy of 73% on test data. We find that the waiting period between booking an the appointment, the patient's past DNA behaviour, and the levels of deprivation in their local area are important factors in predicting future DNAs.

**Discussion** Our model is strongly predictive of whether a hospital outpatient appointment will be attended. Its performance on both patients who did not appear in the training data and appointments from a different time period which covers the Covid-19 pandemic indicate it generalized well across both face to face and virtual appointments and could be used to target resources and intervention towards those patients who are likely to miss an appointment. Moreover, it highlights the impact of deprivation on patient access to healthcare

**Conclusion** Our model successfully predicts patient attendance at outpatient appointments.

**Funding** This study was partially funded by STFC DiRAC innovation fellowships which funded the work of Jonathan Holdship and Harpreet Dhanoa. An Award/Grant number is not applicable to this funding.

## STRENTGHS AND LIMITATIONS OF THIS STUDY

- We produce the first model which can accurately predict the attendance of all outpatient appoints for all ages and specialities at a large, multisite teaching hospital.
- We show a strong correlation between patient non-attendance and deprivation demonstrating that random forest algorithms can provide insight as well as useful predictions
- We find that the broad target population results in a slightly less accurate model than those from the literature that are built for specific outpatient clinics

# 1 INTRODUCTION

Outpatient services have an increasingly important role in clinical care by secondary and tertiary healthcare providers. Non-attendance at booked clinic appointments is common - averaging 8% in the English NHS [1] - and may disproportionately impact disadvantaged groups or patients living with multimorbidity or frailty. This Did Not Attend (DNA) rate remains high although there is variation in England between NHS hospital trusts and by specialty within NHS hospital trusts [2].

The use of artificial intelligence is growing rapidly in healthcare. Some applications are specific to a small group of patients or a specific technique but may not answer large scale problems where patient centered factors including those linked to health inequalities may be important. In this study we took routine data available at a hospital and national deprivation data to develop a predictive tool to identify those patients at risk of non-attendance at a booked outpatient clinic appointment.

Many factors associated with increased risk of DNA are also associated with increased risk of poor health outcomes so that those patients most likely to miss appointments are likely to be those most at risk of harm from disruption of their care. Clinic DNA also causes costs associated with wasted clinic time and

underutilised staff, with an average cost per missed appointment of £150 in the UK or almost £1 billion annually [1]. Beyond the financial costs, there are opportunity costs for other patients waiting for a clinic appointment which is particularly important when there is high demand.

Untargeted interventions to prevent DNAs are common [3]; NHS patients routinely receive email and SMS reminders about upcoming appointments. However, given the relatively small proportion of patients who DNA, any intensive but untargeted efforts such as phone call reminders may well have high administrative costs relative to the efficiency savings they produce.

Targeted interventions are therefore required and this necessitates identifying patients who have a high risk of DNA. There have been attempts to address this problem using machine learning models. One group at University College Hospital [4] analysed 22,000 MRI appointments at two hospitals, training gradient boosting machines to predict DNAs. They obtained an AUROC score of 0.852. Another, used a neural network to predict appointment attendance for a wider range of appointment types and patients at the Royal Berkshire NHS Foundation Trust. They obtain an AUROC score of 0.71 [5]. Both of these studies indicate there is value in machine learning approaches to this problem as their models perform significantly better than a random approach.

In this work, we apply a similar approach to train machine learning models to predict whether a patient will attend their outpatient appointment. However, we consider a much larger patient group by using every outpatient appointment scheduled at the Guy's and St Thomas NHS Foundation Trust (GSTT) between April 2015 and September 2019 for all ages and specialities. Guy's and St Thomas NHS Foundation Trust (GSTT) is a large multisite teaching hospital in central London with a comprehensive range of secondary and tertiary care specialties. The local population for the hospital includes areas with high levels of deprivation and the mean outpatient DNA rate is higher than NHS average (9.8% in 2017 [6]).

In Section 2, we describe the data processing and model optimization that we performed. In Section 3, we discuss the results of that optimization and the final performance metrics of the model. Finally, in Section 4 we summarize the work.

# 2 METHODS

In this work, a series of classification models were trained on patient data from GSTT to learn to classify appointments as a DNA or attendance event using data available up to the day before the appointment. The best performing algorithm was then taken forward to develop a final predictive model.

## 2.1 DATA PRE-PROCESSING

The main dataset was a pseudonymised record of 9.4 million individual outpatient appointments for all specialities and for patients of all ages between April 2015 and September 2019 who were referred to GSTT. The starting date were chosen to coincide with a change in data collection at GSTT so that all data was collected in the same way and thus the sample size is all possible records conforming to this constraint. We exclude appointments that were cancelled with notice by either the patient or GSTT. For each appointment, the dataset contained patient demographic details as well as details of the appointment.

This dataset was split by patient in the ratio 80:20 to provide a training and test set respectively. The test set was used only to produce the performance statistics presented in Section 3. Approximately 11% of all appointments were DNA events so to prevent this class imbalance from affecting the model performance, the training data was resampled to contain an equal number of DNA and attendance events. The test data were kept unbalanced to better evaluate the performance of the model in practice.

A secondary data set covering all outpatient appointments from the year 2020 at GSTT was also obtained to act as a temporally distinct test set. This period covers a large part of the recent Covid-19 epidemic which has affected outpatient attendance rates, as well as the modality of outpatient appointments with increased virtual clinics. Thus, if the model still performs well on this data, we can be more certain of its reliability.

## 2.2 Feature Engineering

Many variables such as patient age and the time between booking and the appointment were immediately available. Other variables required minimal processing such as the binary encoding of categorical variables including clinical specialty and patient sex. In this section, we briefly describe how more complex information was extracted from our records.

We leverage the patient's attendance history using the method of Goffman et al. [8]. We encoded a patient's 8 most recent appointments as a sequence of 0s and 1s where a 0 is a DNA and a 1 is an attended appointment. For any sequence, we can calculate the fraction of appointments in our data set where a patient had the same sequence and then did not attend. We treat that fraction of DNA events as a probability of the patient choosing to DNA given their history. Goffman et al. found that in their data set, a sequence of 0000011111 gives a 50% DNA rate but only 20% of people with that sequence will DNA their next appointment if the 1s represent the most recent appointments. Thus, the sequence provides much more insight than the overall DNA rate.

Patient postcodes were mapped to a Lower Layer Super Output Area (LSOA) which is an area of on average 650 households in order to get local deprivation information. The English Indices of deprivation rank every LSOA in seven domains: income, employment, healthcare, education, crime, and housing as well as a combined rank. The 2015 English indices of deprivation were used in this work [9] by using the

patient's LSOA to obtain their local rankings. Where a patient did not have a postcode, we used the medians from our population for their deprivation rankings.

Finally, hospital spell data for patients who had had inpatient or day case treatment over the same time period was used to augment what was known about each patient. For each appointment, the ICD-10 diagnosis codes applied to the patient over the year prior to the appointment were listed. From this, the Global Frailty Score [10] and a modified Elixhauser score [11] were calculated. Scores were assigned to all patients despite the fact that the Global Frailty Score has not been validated on under 75s and paediatric patients typically had a score of zero. This was the simplest way to include the information for adults in a way that works for models which cannot handle missing data.

## 2.3 Model Optimization and Training

In the first step, models from the Scikit-learn python library [7] as well as the XGBoost classifier [12] were trained to classify appointments using the processed data sets. They were each evaluated by their average AUROC score [13] from a 3-fold cross validation using the training data. The models and their AUROC scores are given in Table 1. Without optimization, the random forest and XGBoost classifiers outperformed other models. We therefore chose the random forest [14] for further optimization, preferring its simple explainability for a clinical context.

**Table 1** AUROC score obtained on validation data for four models which were trialled

| Model | AUROC |
|---|---|
| **K-Nearest Neighbours** | 0.62 |
| **Logistic Regression** | 0.71 |
| **Random Forest** | 0.77 |
| **XGBoost Classifier** | 0.77 |

The first optimizing step was feature selection, the model was first trained with all features and then the features were sorted by order of importance for this model. We then iteratively trained the model, including more features from this sorted list each time until the AUROC score becomes constant. After this, the features were reduced further by removing features from each pair that had a Pearson correlation coefficient greater than 0.7. This process removed features that were important but had predictive power which came from information that was already included in other features.

Scikit-learn's randomsearchCV function was then used to select the optimal values for the model hyperparameters including the number of trees, number of features and the granularity of the trees. Those optimal values were selected by maximizing the AUROC score. To conduct the search, a 3-fold cross validation was performed using the entire training set.

### 2.4 PATIENT AND PUBLIC INVOLVEMENT

No patient involved

# 3 RESULTS

---

### 3.1 THE OPTIMIZED RANDOM FOREST

The random search of hyperparameter space yielded the values given in Table 2. Whilst these are the values that give the best result, examination of the grid showed that the model was insensitive to many hyperparameters. Overall, we found a large number of trees improved the model but otherwise, the default values used by the Scikit-learn implementation of the random forest were adequate and many appear in the best scoring model.

**Table 2**. Optimized hyperparameter values for the random forest model

| Hyperparameter | Value |
|---|---|
| **Number of Trees** | 120 |
| **Minimum samples per split** | 2 |
| **Minimum samples per leaf** | 1 |
| **Max Features per split** | auto |
| **Max depth** | 81 |
| **Criterion** | Gini |
| **Bootstrap** | True |

## 3.2 FEATURES

The fewer features included in the model, the more likely it is to be replicable in other settings. Of the 260 features in our data set, 100 were required to obtain the same AUROC score on the training data as the full set. Once correlated pairs were removed, this left 63 features which are listed in Table 3 as well as Appendix Table 1.

In Figure 1, we show how the AUROC score obtained on the validation data improves as a function of the number of features included. Most interestingly, we find just six features can produce a model almost as accurate as the final model, with an AUROC score on the validation data of 0.75 compared to 0.76. These top six features include the patient's age, total past DNA rate and the probability obtained from their DNA history sequence as well as measures of their hospital utilization over the past 3 months and past year and the overall IMD ranking of their postcode.

## 3.3 Model Performance

The final model was first tested on the 20% of data that was withheld from the training process. Our optimized random forest with all features included has an accuracy of 73.1% and an AUROC score of 0.76. We estimate the confidence interval on the AUROC score by bootstrapping, calculating the AUROC on 100 test sets sampled from the original set with replacement, we find a variance in the AUROC that is smaller than the reported precision and thus do not report an interval. The AUROC compares well with the predictive model of Nelson et al. [4] which has an AUROC score of 0.85 when we consider that their model was limited to MRI appointments only and thus would have less variance in their data set compared to our analysis. In fact, another recent work by Dashtban & Li [5] uses a deep learning approach to this problem for a set of outpatient appointments more comparable to this work and obtain an AUROC score of 0.71.

The model was further tested on a set of patient appointments that were completely separated in time from the training data. Whilst the test and training data were taken from the year 2015-2019, this second test set included all outpatient appointments at GSTT in 2020. On this test set, the model achieved an AUROC score of 0.75 and so it can be concluded the model is transferable to other time periods.

Figure 2 is a normalized confusion matrix showing that 7% of appointments are DNA events correctly predicted by our model and just 4% of appointments are missed DNAs. However, 27% of appointments will be incorrectly flagged as DNA. This means the vast majority of DNA events would be caught but an expensive intervention may waste of resources.

**Table 3** The twenty most important features for the random forest model.

| Feature | Variable Type | Importance |
|---|---|---|

| | | |
|---|---|---|
| **Time between booking and appointment** | Continuous | 0.139 |
| **Total past DNA rate of patient** | Continuous | 0.060 |
| **Mean DNA rate of specialty** | Continuous | 0.062 |
| **Patient Age** | Continuous | 0.050 |
| **IMD Living Environment Rank** | Continuous | 0.045 |
| **IMD Employment Rank** | Continuous | 0.045 |
| **IMD Barriers to Housing and Services Rank** | Continuous | 0.044 |
| **IMD Crime Rank** | Continuous | 0.044 |
| **Month of appointment** | Integer 1-12 | 0.043 |
| **IMD Education, Skills and Training Rank** | Continuous | 0.043 |
| **Hour of appointment in day** | Integer 1-24 | 0.037 |
| **3 Month running cost of patient contacts with hospital** | Continuous | 0.035 |
| **Day of appointment in year** | Integer 1-365 | 0.029 |
| **Number of contacts between trust and patient in 3 months** | Continuous | 0.028 |
| **Total number of past missed appointments of patient** | Continuous | 0.026 |
| **Number of unique ICD10 codes from the patient's spells in 3 months** | Continuous | 0.020 |
| **Elixhauser Score** | Continuous | 0.016 |
| **Number of outpatient appointment cancellations in past 6 months** | Continuous | 0.013 |
| **Length of patient's last spell** | Continuous | 0.012 |
| **Number of outpatient appointments the patient has on same day** | Continuous | 0.010 |

# 4 DISCUSSION

## 4.1 RISK STRATIFICATION

The model predicts the probability of an appointment being a DNA which is converted to a binary prediction using a threshold. Figure 3 shows the ROC curve for this model, which demonstrates how changing that threshold alters the model's performance, allowing us to limit false positives by demanding higher probabilities. For example, as illustrated by the blue lines in Figure 3, choosing a threshold of 0.8 reduces the false positive rate to ~1% at the cost of reducing the true positive rate to 10%. By choosing to accept more DNA events going unflagged, one can ensure that most of the predicted DNA events are DNA events.

Perhaps more usefully, patients can be ranked by their probability of DNA. This means that in practice, resources can be first directed to those most likely to DNA by using administrative resources to contact those with a very large probability before moving on to contact those who are less likely to DNA. The fact the model's AUROC score is much greater than 0.5 indicates that following a strategy of first intervening in high probability cases will represent a large improvement in efficiency over randomly contacting patients.

## 4.2 IMPROVING THE MODEL

It is likely that some of the ~30% of appointments that are misclassified could be correctly predicted with additional information. For example, in one early iteration of the model, the CityMapper API was queried to obtain the travel time from each postcode in London to St Thomas' hospital for an arbitrary off-peak time to give a measure of the difficulty a patient would have in reaching their appointment. Whilst this improved the AUROC score of the model by 5%, it was dropped to limit the model features to data easily

obtained by GSTT. Nevertheless, it does indicate that a more specific measure of the travel time to the location of an outpatient appointment could improve the model.

Another potential area of improvement is the encoding of the patient's diagnostic history. Measures such as the Elixhauser score use the patient's ICD-10 coding history to determine their general health. ICD-10 codes are subject to large error rates [15] and coding is typically done with a view to generating financial data rather than predictive variables. An alternative system such as SNOMED [16] in which clinical staff directly produce diagnostic data with an aim towards predictive analytics may produce more accurate models.

## 4.3 Broader Model Applicability

This model shows a reasonable accuracy of DNA prediction across different time periods and across both virtual and face to face appointments. This allows outpatient clinics to target inventions in additional to routine care based on the patient's predicted DNA risk. Each of the features used in the model are routinely collected in NHS hospital trusts and thus the model can be applied in other NHS hospitals in England. There is no patient information included in the random forests so they can be shared without concerns for patient privacy.

However, when moving to a different population, a more successful approach is likely to result from retraining the same algorithm using historical data from the hospital that intends to use it rather than replicating this model which results from training on GSTT data. This is due to the weighting given to each feature. Features like DNA rate, ICD-10 coding rates and patient demographics vary greatly between trusts. Therefore, if the informative part of a feature is whether a patient deviates from the average, a decision that has learned thresholds from GSTT will fail at a trust with a much different distribution.

Further, it is not possible to assume the accuracy or completeness with which each trust records any given feature is similar. For example, the accuracy of ICD-10 coding varies greatly between trusts [17]. It’s

therefore possible a model will learn to give less weight to a less accurate feature if trained on the data from a another provider.

### 4.4 THE EFFECTS OF DEPRIVATION

One of the most striking aspects of the model is the importance given to measures of deprivation. Five of the ten most informative variables shown in Table 3 are measures of patient deprivation. This indicates that it is the most deprived patients whose health is affected by DNA and that, by targeting patients for intervention, some of the effect of broader inequality on patient health could be mitigated.

Our population is an interesting one for studying the effects of deprivation; 42% of GSTT patients are from areas in the lowest 3 IMD deciles but many patients are also from the highest deciles. Figure 4 shows the London borough of Southwark which one of the two boroughs from which the majority of GSTT patients are drawn. Each LSOA is plotted on a colour scale showing the IMD decile of that area and is compared to the the DNA rate of that LSAO. The IMD ranking is ordered such that the lower deciles are the more deprived areas. These heavily overlap with the areas with the highest DNA rate. In fact, the DNA rate and IMD ranking of an LSAO have a Spearman correlation coefficient of -0.73 highlighting the need for more investigation into the link between deprivation and outpatient attendance.

## 5 CONCLUSION

In order to predict whether patients would attend future outpatient appointments, a random forest classifier model was trained on historical data from Guy's and St Thomas NHS Foundation Trust. The input data to the model included basic appointment details, summaries of the patient medical history, their past attendance record and information about the area in which they live. The model achieved a 73% accuracy and AUROC of 0.76 in adults and children.

The model has considerable predictive power, producing predictions that are much better than random. Moreover, it performed equally well on two different test sets. The first being a set of appointments from the same time period as the training data but using a set of patients who did not appear in that training data. The second test set included all appointments from the 2019/2020 financial year which is a year after the most recent appointments in the training data and had the considerable difference of including appointments scheduled during the Covid-19 pandemic, which included a much higher proportion of `virtual' outpatient appointments.

Despite the model accuracy, the false positive rate is a large issue due to the small fraction of patients who actually DNA. To counter this, the model's output probabilities can be used rather than the binary prediction. The false positive rate drops dramatically with increasing probability of DNA such that predictions with a probability > 80% have a false positive rate of 3%. It is therefore expected that this model could be used to prioritize patients for intervention based on their probability of DNA.

The random forest model is also informative in that it allows us to assess which factors have the greatest effect on whether a patient will DNA with clinical scheduling factors the most important. This model highlights the overlap between high LSOA deprivation and DNA rate. Any intervention should factor in the relationship between deprivation and outpatient attendance as this effect could reinforce inequalities of health outcomes which could be reinforced further by digital inequality.

Future work will focus on implementing the algorithm in clinical settings to determine the best use of the information the model provides. It is also likely the model could be improved by including auxiliary data that is not part of the routinely collected data at GSTT. The ultimate goal is to use this predictive insight into likelihood of non-attendance to improve patient management and increase healthcare efficiency. Given the importance of socio-geographic factors, at least for this London hospital, this could help reduce inequalities in healthcare. Implementation trials, for example using targeted engagement approaches for

highest risk groups identified, are needed to test the best methods to integrate such prediction algorithms into local routine healthcare.

**Acknowledgments**

JH and HD were supported by STFC DiRAC innovation fellowships. The authors also thank CityMapper for the use of their API to test whether travel times affected the model.

**Author Contributions**

JH and HD contributed to the conceptualization of the study, data curation and analysis and the writing of the manuscript. AH contributed to the conceptualization of the study, interpretation of the data and critical review of the manuscript. CJS contributed to the conceptualization of the study and interpretation of the data. MB contributed interpretation of the data and critical review of the manuscript. IW contributed interpretation of the data. KT contributed through project administration. CC contributed through project administration. JY contributed to the study methodology.

**Ethics Approval**

This work was approved by Guy's and St Thomas' NHS Foundation Trust's Quality Improvement & Patient Safety department. It is registered in the trust database with Audit/Service evaluation number 116298.

**Conflicts of Interest**

The authors declare that there is no conflict of interest regarding the publication of this article.

**Data Availability**

The data used in this work is not available as, even in its anonymised form, it is the personal data of GSTT patients.

# REFERENCES


[1] NHS Benchmarking Network 2019 Outpatients project, 2020. [Online]. https://www.nhsbenchmarking.nhs.uk/news/2019-outpatients-project-results-published.

[2] NHS Digital Hospital Outpatient Activity 2018-19 - 2019. [Online]. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/2018-19.

[3] P. Kiruparan, N. Kiruparan, and D. Debnath, "Impact of pre-appointment contact and short message service alerts in reducing 'Did Not Attend' (DNA) rate on rapid access new patient breast clinics: A DGH perspective, "BMC Health Services Research, vol. 20, no. 1, p. 757,Aug. 2020,issn: 14726963.doi:10.1186/s12913-020-05627-2. [Online]. https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-020-05627-2.

[4] A. Nelson, D. Herron, G. Rees, and P. Nachev, "Predicting scheduled hospital attendance with artificial intelligence" npj Digital Medicine, vol. 2, no. 1, pp. 1–7, Dec. 2019,issn: 2398-6352.doi:10.1038/s41746-019-0103-3. [Online]. https://doi.org/10.1038/s41746-019-0103-3.

[5] M. Dashtban and W. Li, “Deep Learning for Predicting Non-attendance in Hospital Outpatient Appointments,” inProceedings of the 52nd Hawaii International Conference on System Sciences, Hawaii International Conference on System Sciences, 2019.doi:10.24251/HICSS.2019.451. [Online]. http://hdl.handle.net/10125/59809.

[6] M. Bartlett, S. Blazer, G. Hobson, and I. Abbs, “The power of digital communications: improving outpatient attendances in south London,”Future Healthcare Journal, vol. 5, no. 1,pp. 43–46, Feb. 2018 ISSN: 2514-6645.doi:10.7861/futurehosp.5-1-43. [Online].

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P.Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in{P}ython,”Journal of MachineLearning Research, vol. 12, pp. 2825–2830, 2011.

[8] R. M. Goffman, S. L. Harris, J. H. May, A. S. Milicevic, R. J. Monte, L. Myaskovsky, K. L.Rodriguez, Y. C. Tjader, and D. L. Vargas, “Modeling Patient No-Show History and Predicting Future Outpatient Appointment Behavior in the Veterans Health Administration,”Military Medicine, vol. 182, no. 5, e1708–e1714, May 2017,issn: 0026-4075.doi:10.7205/MILMED-D-16-00345. [Online]. https://academic.oup.com/milmed/article/182/5-6/e1708-e1714/4158896.

[9]English indices of deprivation 2015 - GOV.UK. [Online]. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015.

[10] J. T. Soong, J. Kaubryte, D. Liew, C. J. Peden, A. Bottle, D. Bell, C. Cooper, and A. Hopper “Dr Foster global frailty score: An international retrospective observational study developing and validating a risk prediction model for hospitalised older persons from administrative datasets” BMJ Open, vol. 9, no. 6, p. 26 759, Jun. 2019,issn: 20446055.doi:10.1136/bmjopen-2018-026759. [Online]. http://bmjopen.bmj.com/.

[11] C. Van Walraven, P. C. Austin, A. Jennings, H. Quan, and A. J. Forster, “A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data,”Medical Care, vol. 47, no. 6, pp. 626–633, Jun. 2009,issn: 00257079.doi:10.1097/MLR.0b013e31819432e5. [Online]. https://pubmed.ncbi.nlm.nih.gov/19433995/.

[12] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings ofthe 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,ser. KDD ’16, New York, NY, USA: ACM, 2016, pp. 785–794,isbn: 978-1-4503-4232-2.doi:10.1145/2939672.2939785. [Online]. http://doi.acm.org/10.1145/2939672.2939785.

[13] T. Fawcett, “An introduction to ROC analysis,”Pattern Recognition Letters, vol. 27, no. 8,pp. 861–874, Jun. 2006,issn: 01678655.doi:10.1016/j.patrec.2005.10.010.

[14] L. Breiman, “Random forests,”Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001,issn:08856125.doi:10.1023/A:1010933404324. [Online]. https://link.springer.com/article/10.1023/A:1010933404324.

[15] M. Farzandipour, A. Sheikhtaheri, and F. Sadoughi, “Effective factors on accuracy of principaldiagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10),”International Journal of Information Management, vol. 30, no. 1, pp. 78–84, Feb. 2010,issn:02684012.doi:10.1016/j.ijinfomgt.2009.07.002.

[16]Data Analytics with SNOMED CT - Data Analytics with SNOMED CT - SNOMED Con-fluence. [Online]. https://confluence.ihtsdotools.org/display/DOCANLYT/Data+Analytics+with+SNOMED+CT.

[17] CHKS,The quality of clinical coding in the NHS, 2014. [Online]. http://www.chks.co.uk/userfiles/files/The_quality_of_clinical_coding_in_the_NHS.pdf

## FIGURE CAPTIONS

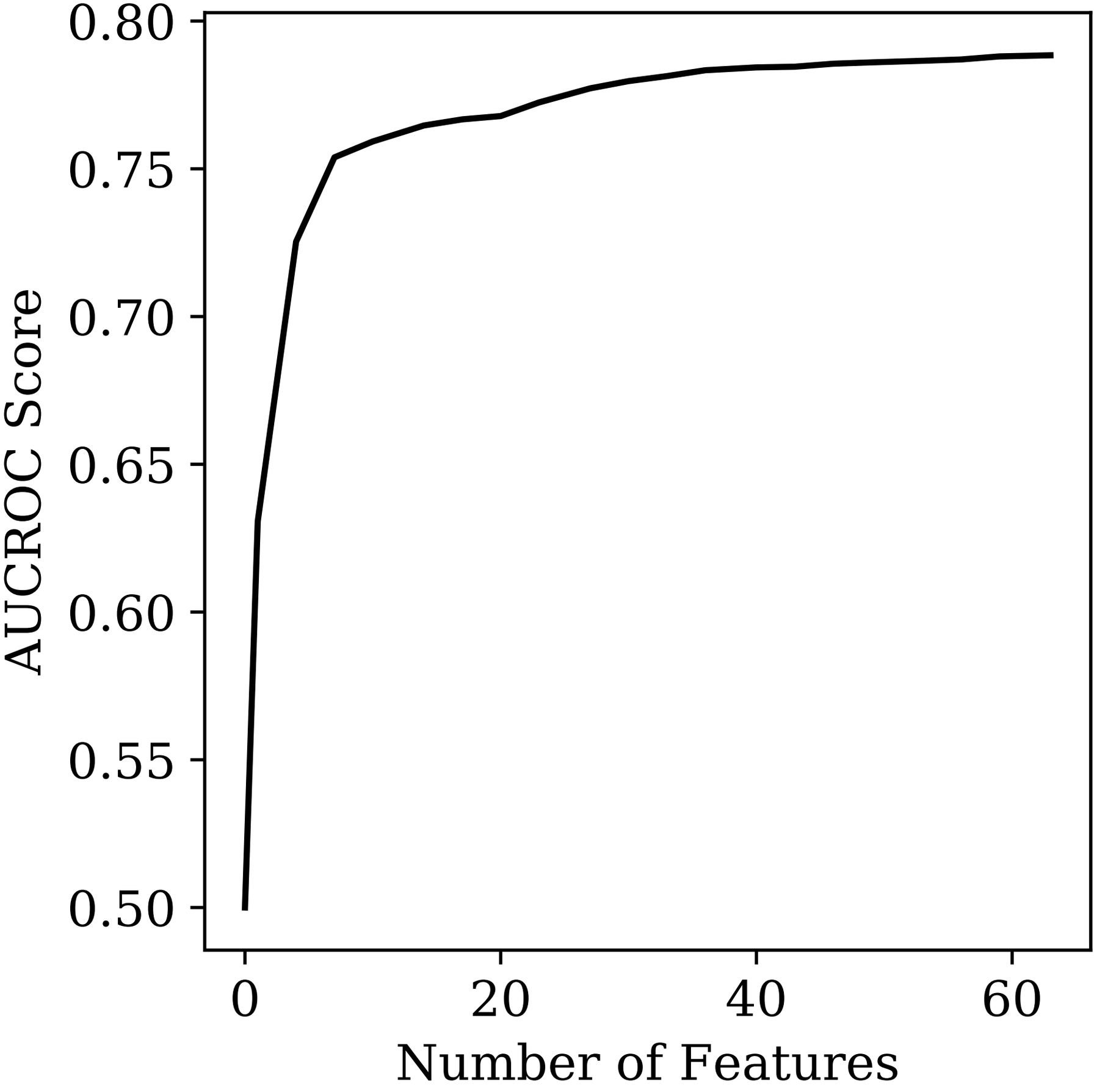
All figures were created by Jonathan Holdship and therefore we have permission to reuse the image as required.

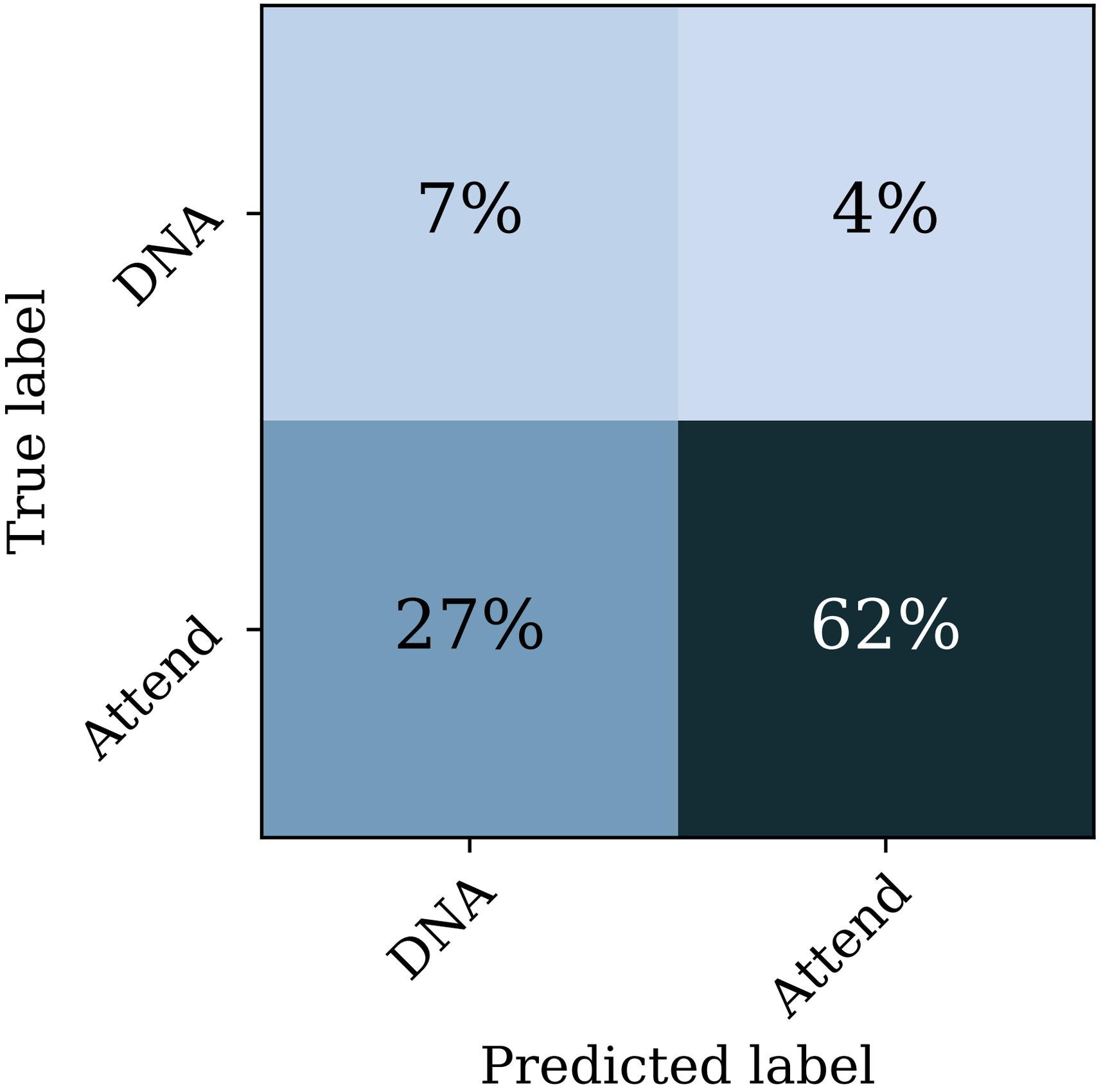
*Figure 1* AUROC score as a function of the number of features included in the model.

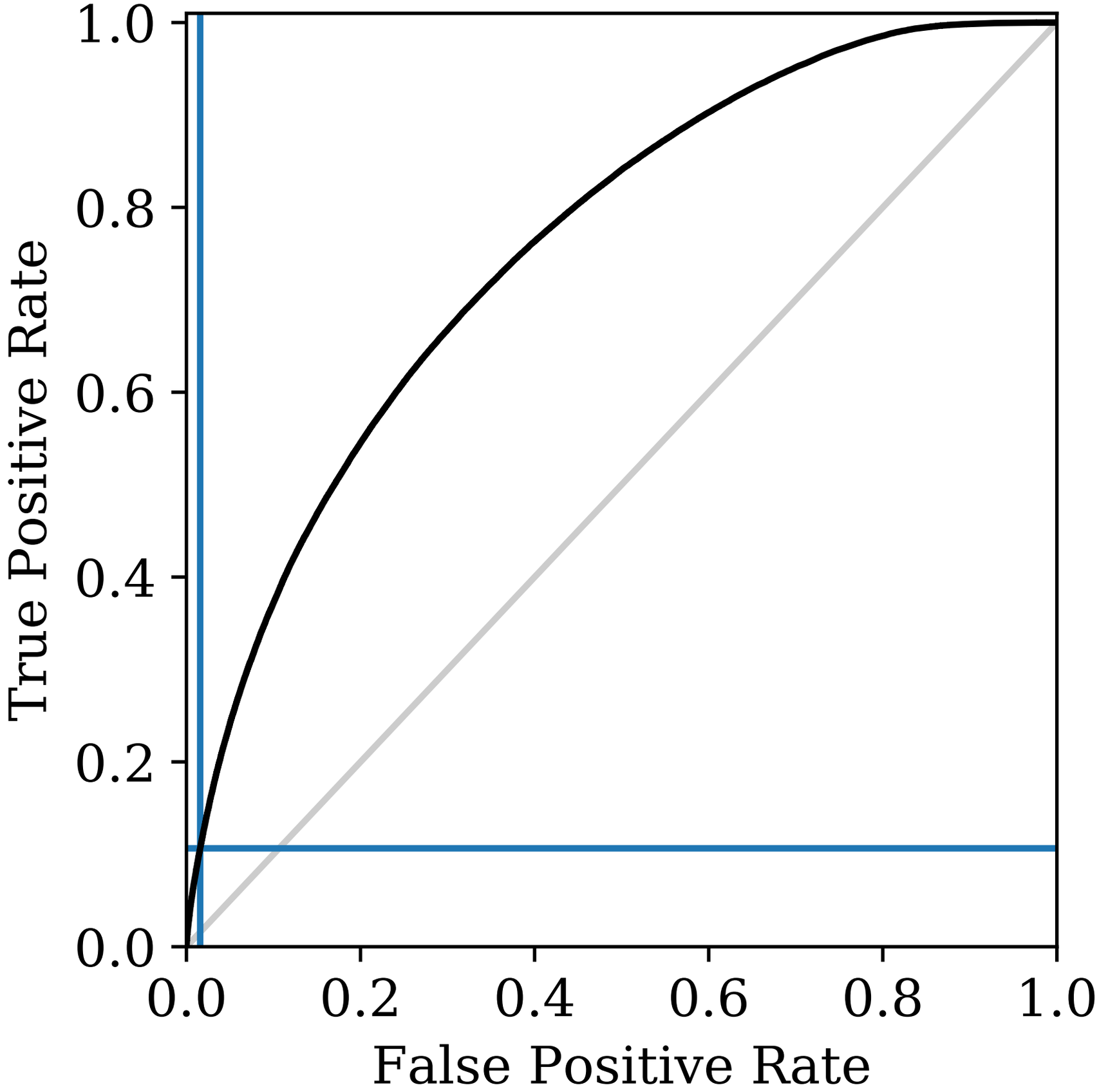
*Figure 2* Normalized confusion matrix which shows the proportion of appointments in the test data which had a given combination of predicted and actual attendance value.

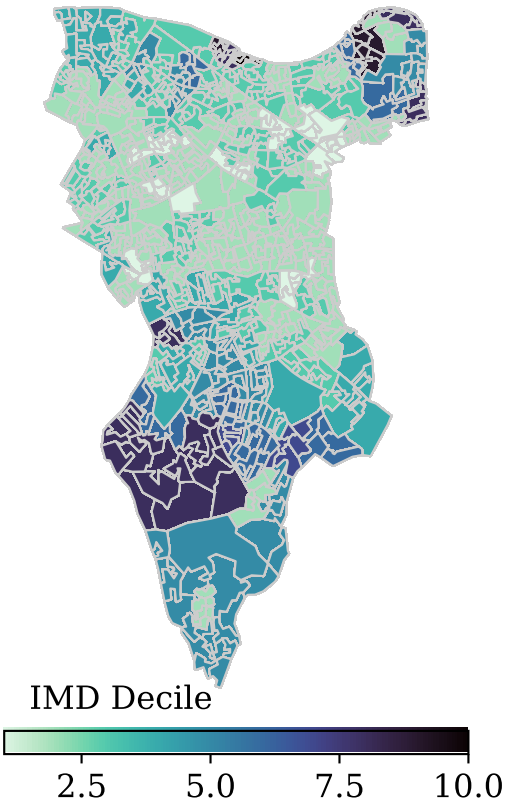
*Figure 3* Receiver-operating characteristic curve, showing the fraction of DNA events captured as a function of the false positive rate. The grey diagonal line represents the performance of a model that makes a random guess and the blue lines indicate the false positive rate (1\%) and true positive rate(10\%) obtained by choosing a threshold probability of 0.8.

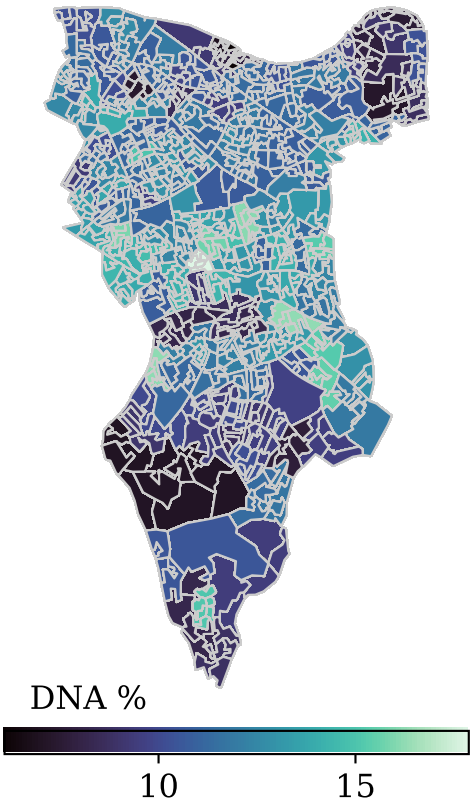
*Figure 4* The London Borough of Southwark, with each LSOA coloured by the IMD decile (left) or DNA rate (right) of that area. Lighter colours show more deprived areas. The color scale is reversed for the DNA rate to better show the negative correlation between the two.


0.80
0.75
0.70
0.65
0.60
0.55
0.50
AUCROC Score
0
20
40
60
Number of Features


7%
4%
27%
62%
DNA
Attend
True label
DNA
Attend
Predicted label


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate


IMD Decile
2.5
5.0
7.5
10.0


DNA %
10
15