# Children's Urgent Care in the Right Place Every Time

Sam Relins[1], Ben Cooper[1], Ruaridh Mon-Williams[2], Mark Mon-Williams[1,2], Mathew Mathai[3]

[1]Leeds Institute for Data Analytics, University of Leeds, [2]Bradford Institute for Health Research, [3]Bradford Teaching Hospitals NHS Foundation Trust

July 28, 2021

# Contents

# List of Figures

# List of Tables

# Executive Summary

Bradford Teaching Hospitals Foundation Trust is home to a hospital-at-home service for children and young people (CYP) called the Ambulatory Care Experience (ACE). ACE offers an alternative to hospital referral or admission for CYP that require urgent care. CYP under the ACE team are treated and monitored in their own homes in a "virtual ward", under the care of a consultant paediatrician.

ACE clinicians admit patients where they are confident care can be provided safely in the community with specialist input - this decision is based on their clinical assessment and a standard condition-specific pathway. Although the majority of ACE patients are treated and discharged without requiring a hospital review, some patients are later referred to hospital for ongoing care. This project aims to model these outcomes, with the aim of improving the ACE referral pathway.

## Data

The data consist of referral records of 502 patients treated for asthma/wheeze by the ACE team, labeled with the treatment outcome - discharged without referral to hospital (421 examples or 83.9% of the data) or later referred to hospital (81 examples or 16.1% of the data). The records contain structured numerical and categorical observations, including clinical measurements, patient details and general referral metadata. The data also includes unstructured free-text fields: medical histories, examination details and recommendations; these text features are used to record relevant referral details that aren't captured in the structured data,In later analyses, data were linked from patients' primary and secondary care records (held in the Connected Bradford cloud platform), as well as geospatial data for the Bradford region.

## Main objectives

The primary objective of this project is to model the outcomes of treatment in ACE using the referral data. This objective can be divided into four key goals:

1. Train and evaluate classification models to predict the probability of hospitalisation using the ACE dataset
2. Identify important features from the ACE data that are indicative of increased or decreased risk of hospitalisation
3. Quantify the degree of certainty or uncertainty associated with both the hospitalisation predictions and important features, as estimated from the ACE dataset
4. Determine any increase in model performance with the addition of geographical data, or data from patients' primary and secondary care records

## Approach

The exploration is characterised by four main approaches:

1. **Classification Modelling**: We trained a range of common classification models (e.g. logistic regression, gradient boosted decision trees, naive Bayes) to predict the probability a patient is hospitalised, evaluating for both accuracy and precision/recall. The aim was to test the baseline accuracy, or the predictive potential, of classification models trained using the ACE referral data. A particular focus was addressing the imbalance of positive/negative labels in the dataset.

2. **Text Analysis**: We analysed the free-text from the ACE referral data using natural language processing techniques. The text features were preprocessed and then vectorised using Term Frequency - Inverse Document Frequencies (TF-IDF). These TF-IDF representations were numerically analysed, and then used to train a logistic regression model to predict hospitalisation risk. A lasso constraint was used to force a sparse selection of words that are most predictive of hospitalisation or successful treatment - these words then guided the creation of new text-based features.

3. **Bayesian Analysis**: Important features identified during the prior analyses were used, individually and in combination, to iteratively model hospitalisation risk using samples generated from a Bayesian logistic regression model. Outputs from these analyses indicate the level of confidence we may assign to the association between these features and hospitalisation risk, and the confidence with which we can predict the probability of hospitalisation.

4. **Additional variables**: We sought to find additional data, beyond the ACE referral form, that may aid model predictive performance. These data came from geospatial analyses of air pollution, distances to healthcare facilities, and local measures of socio-economic deprivation, alongside data on patients' medical history and healthcare usage patterns. We trained predictive models using just the data from the ACE referral form, and compared their performance to models trained using the above data alongside the ACE referral data.

## Main Conclusions

Our analysis identifies features in the ACE data that are predictive of increased/decreased risk of hospitalisation. The Bayesian analysis identifies relationships between several of the dataset features and treatment outcomes, and assigns a high level confidence to relationships - it is extremely unlikely these associations have arisen by chance. Risk factors include clinical indicators that would be obvious to the ACE team, such as low oxygen saturations and high respiratory rates, and less obvious features such as the referral originating from a GP surgery. Promising features meriting further investigation included history of eczema or pneumonia, mention of asthma or certain medications in the referral notes, prescription of long-acting bronchodilators, and the patient's home being in an area of high air pollution.

Despite the relationship between key variables and hospitalisation, the current referral data are not sufficient to generate accurate predictions of hospitalisation among the ACE patients. These results confirm that ACE clinicians are making good referral decisions given the referral observations. The ACE team use the referral data to determine which patients are suitable for home treatment, and only those patients feature in the dataset - that machine learning models are unable to use this data to accurately predict the risk of hospitalisation among these patients indicates ACE clinicians haven't missed any obvious predictors of hospitalisation. This does not rule out the possibility that other predictors

may exist that aren't currently considered in the ACE referral criteria; indeed, our examination of additional medical and geospatial variables yielded promising relationships between aspects of a patient's medical history and hospitalisation that aren't considered in the ACE referral criteria.

## Limitations

The data only includes patients accepted by the ACE service, and the patients included are referred via many different treatment pathways; therefore, it is likely the data are subject to a number of hidden biases. The specific methods and results must be considered only within the scope of the ACE service and cannot be generalised.

The dataset is small for this kind of analysis, and data that represent hospitalisations are a small minority (approximately 15%). Results are therefore subject to considerable instability/variability. Though the general findings (particularly those supported by Bayesian analysis) are likely to be robust, many of the outputs generated during the experimentation would differ significantly if the data were re-sampled.

## Recommendations and further work

The current ACE referral criteria are already well utilised to determine suitability for treatment at home. If machine learning is to improve upon referral decisions, additional features should be considered beyond those in the extant referral criteria - text analysis, geospatial analysis and primary care records all yielded variables with promising associations with hospitalisation. However, whilst models trained using these additional variables showed improved predictive performance (relative to models trained just using data from the ACE referral form), absolute predictive performance remained middling.

Discussions with the ACE team have identified a number of potential sources of bias in the ACE dataset - chief among which is the selection bias that follows from using only patients that were accepted for ACE treatment. It is likely that these biases have unforeseen effects on the models and the inferences drawn from the dataset. Therefore, an exploration of the causal relationships in the ACE data, and careful exploration of the different clinical pathways that end in ACE referral are strongly recommended.

Exploring the biases present in the ACE dataset will not make the results that derive from it any less biased. Ultimately a new dataset is required to achieve predictions that can be generalised, a dataset that isn't influenced by the ACE treatment pathway. In collecting such a dataset, a sampling methodology should be designed with the ACE team as subject matter experts - for example we might collect records of hospital admissions for paediatric urgent care from a separate hospital trust, and label them based on an agreed set of clinical characteristics that demonstrate suitability for home care. Such data could then be analysed without the concern of any bias introduced by the ACE service, either via referral or treatment.

# 1 Introduction

## 1.1 The ACE Service

ACE or the Ambulatory Care Experience is a service that delivers paediatric urgent care in the Bradford community. As part of Bradford Teaching Hospitals Foundation Trust's "virtual wards" strategy, ACE is intended to improve patient outcomes and deliver value for money [1]. Indeed, the Care Quality Commission reports have singled out the ACE service as "outstanding practice" and "make[ing] best use of clinical resources" [2]. The service has also been awarded a Health Services Journal "Improvement in Emergency and Urgent Care" award [3].

The primary aim of the ACE service is to avoid hospitalisation. Children who require ongoing specialist assessment, treatment and monitoring can be looked after by ACE from the comfort of their own homes, under the care of a nurse and consultant paediatrician. The team support and monitor the effectiveness of treatments, taking care to note any deterioration in the patient that may indicate the need for hospital treatment. Figure 1.1 details the ACE treatment pathway for CYP with "wheeze/asthma". This approach avoids the traditional pathway of hospitalisation, assessment, observation and discharge without need for further treatment, that is often seen in paediatric urgent care.

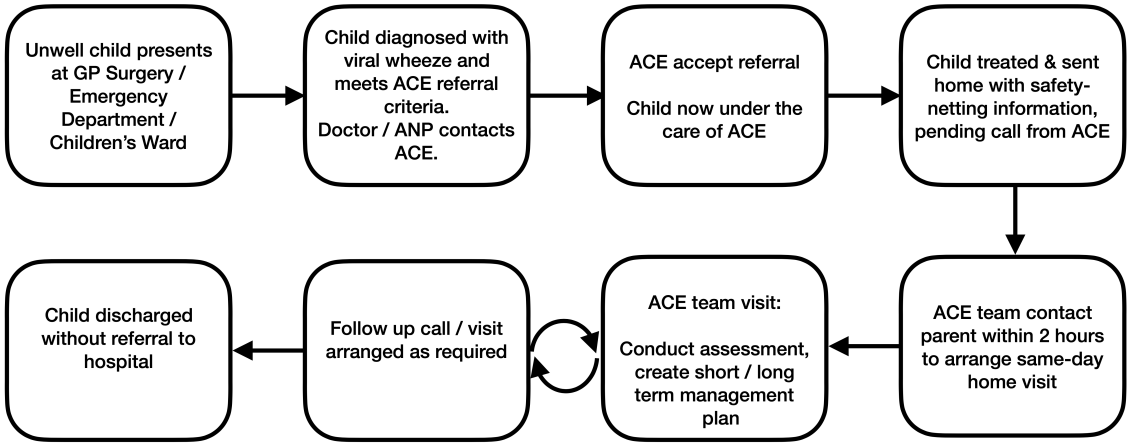



Figure 1.1: The referral pathway for an ACE patient referred with asthma/wheeze and successfully discharged without hospital referral

Key to the success of ACE is deciding which of the referred patients are of "mild to moderate concern". The service can only reduce hospitalisations if their patients are being discharged without later being hospitalised. ACE define a set of referral criteria that indicate suitability for home treatment. Examples of these criteria for the viral wheeze / asthma pathway can be seen in Figure 1.2. Patients are accepted for community treatment if the referrer and ACE team judge that this is appropriate (even if the clinical parameters fall outside the set referral criteria). These referral decisions contribute to the successful discharge of 84% of patients, based on current data.

The principle need for urgent care in this setting is observation and monitoring for deterioration. Thus, the expectation is that some patients will inevitably require later

hospital treatment. Nonetheless, the aim of ACE is to ensure the number of patients hospitalised under their care is as low as possible.

| | | |
|---|---|---|
| **Oxygen saturations in air** | <94% | |
| **Heart Rate (beats per minute)** | **Age:** | **Value:** |
| | 2–5 years | 95–140 |
| | 5–12 years | 80–120 |
| | >12 years | 60–100 |
| **Respiratory Rate (breaths per minute)** | **Age:** | **Value:** |
| | 2–5 months | 25–30 |
| | 5–12 years | 20–25 |
| | >12 years | 15–20 |
| **Auscultation** | Good air entry with some wheeze | |
| **Speech** | Able to complete sentences | |
| **Work of breathing** | Minimal/no recessions | |
| **Conscious level** | normal | |

Figure 1.2: The ACE referral criteria. A set of clinical characteristics that indicate a patient is of mild to moderate concern and is suitable for treatment within ACE. Note: Patients can be accepted for treatment even if their observations fall outside these criteria.

## 1.2 Predicting ACE Treatment Outcomes

ACE clinicians have long been preoccupied with the characteristics that make a paediatric urgent care patient suitable for treatment at home. Having collected 3 years of referral data, they hypothesised that modern machine learning techniques could use this data to predict treatment outcomes within the service. This project aimed to test that hypothesis.

We aimed to answer the following three questions:

1. Can machine learning classification models use ACE referral data to accurately predict which patients required hospital treatment?

   Machine learning has already shown great potential in a range of clinical settings [4]. Repositories of high quality labelled data found in the healthcare setting offer an abundance prospective applications of machine learning. Among these applications are many successful examples of risk modelling [5] [6] [7]. With their own labelled referral records, the ACE team hypothesise that a similar approach might be applied to predict the risk of hospitalisation for patients referred to the ACE service.

2. Are there predictors of hospital referral that can be identified from the ACE referral data?

   An important aspect of predictive modelling, especially so in the healthcare setting, is *why* a prediction was made. The ACE team are particularly interested the *predictors* of hospitalisation risk, alongside the *predictions* themselves - which features increase a patient's risk of a hospital referral, or indicate that a patient is more

suitable for home treatment? Many predictive modelling approaches are capable of offering such insights, be they "explainable" predictive modelling techniques, or post-hoc explainability methods - we aim to use such methods to identify predictors of hospitalisation.

3. Can we quantify our uncertainty / certainty in the findings from questions 1 and 2?

   The size and scope of the ACE service is such that training data is scarce. Any inferences drawn from this data will be supported by very few real examples. As such, it will be as important to quantify what *can't* be said based on the referral data as what *can*. Under what circumstances can we be confident of an outcome or of the effect of a given feature, and when are we less certain? This work aims to test methods that can predict not only the chance of a given outcome but also the error, or anticipated range of resonable values, for this prediction.

# 2 | The Data

The data consist of the referral records of 502 patients treated by ACE for viral wheeze/asthma between December 2017 (when the service began), to March 2020 (when the service was suspended as a result of the Covid-19 pandemic). Examples are labeled with one of two possible treatment outcomes - discharged without referral to hospital (421 examples or 83.9% of the data) or later referred to hospital (81 examples or 16.1% of the data). Each individual record details referral information, data from the proceeding telephone/in-person consultations, and general performance metrics collected by the service. For the purposes of this study, only the data collected at the referral stage is used, given the aim to model the risk of hospitalisation upon referral; these variables are described in Table 2.1 below:

| Feature | Description | Possible Values |
|---|---|---|
| **Referral From** | Place from where patient is referred | GP / A&E / ED (emergency department) / CCDA (children's clinical decision area) Includes optional ANP / paed ANP (Advanced Nurse Practitioner) |
| **Referee's Profession** | Profession of clinician making the referral | Consultant / Doctor / ANP (Advanced Nurse Practitioner) / Registrar |
| **Age** | Age of patient in whole years | 1-16 |
| **Address** | Postcode are from patient's address | BD01 / BD02 etc. |
| **Ethnicity** | Patient's ethnicity | free-text, not limited to pre-determined options |
| **Gender** | Patient biological sex | Male / Female |
| **Allergy** | Allergy information for patient | One or many of NKA (no known allergy) / NKDA (no known drug allergy - distinction between NKA not clear) / Food / Drug / Other |
| **Date of Referral** | Time of year of referral | Spring / Summer / Autumn / Winter |
| **Time of Referral** | Time of day of referral | Morning / Afternoon / Evening |

| Feature | Description | Possible Values |
|---|---|---|
| **Severity of Illness** | Referring clinicians opinion on the severity of child's illness | Mild / Moderate |
| **Activity Level of Child** | How active / energetic the child is (opinion of parent) | Usual / Lower |
| **"Gut Feeling" of Referrer** | The referrer's "gut feeling" on the condition of the patient | Well / Low Concern / Unwell |
| **Oxygen Saturations** | The oxygen saturations in air of patient | Percentage Value |
| **Respiratory Rate** | Number of breaths taken per minute | Integer |
| **Heart Rate** | Heart rate in beats per minute | Integer |
| **Temperature** | Body temperature in degrees centigrade | Float / Non-negative Real Number |
| **Sepsis Red Flags** | Any indications of sepsis that are of concern | None Noted / Low Level |
| **Safeguarding Issues** | Any safety concerns about the patient's home environment | Yes / No |
| **Medical History** | Short description of the patient's related medical history | Free text 50-100 words |
| **Medical History** | Short description of the patient's related medical history | Free text 50-100 words |
| **Examination Summary** | Short summary of the examination conducted by the referring clinician | Free text 50-100 words |
| **recommendation** | Treatment recommendations for patient / patient / guardian to follow prior to contact from ACE | Free text 50-100 words |

Table 2.1: Names and descriptions of the observations or features of the ACE dataset

## 2.1 Data Cleaning / Preprocessing

The dataset as provided by the ACE team is remarkably clean and required little preprocessing. A number of features were omitted from the referral data as they are largely

incomplete (e.g. blood pressure, body weight) leaving those described in Table 2.1. A minority of the remaining features have missing values that were filled or removed using rules detailed in Table 2.2.

| **Missing Observations Per Example** | **Rule:** |
|---|---|
| **2 or more** | Examples removed from the dataset entirely - only 14 such examples, all of which are from overrepresented group that were successfully treated by ace, so removing has little impact |
| **1** | Missing observations inferred from rest of examples based on outcome i.e. observations from example that required hospital treatment taken from the group that required hospital treatment. Median values used for numerical features, majority (mode) category used for categorical features |

Table 2.2: Strategies for filling missing observations from dataset.

A small number of features were also subject to simple preprocessing to make their interpretation easier, detailed in Table 2.3.

| **Feautre:** | **Preprocessing details:** |
|---|---|
| **Referral From** | Observations include "A&E" and "ED" (emergency department) which are synonyms - both merged to "ED". Entries mention "ANP" (advanced nurse practitioner) or "paed ANP" - this information is duplicated in "referral profession" feature so is removed. Resulting categories are simply "ED" / "GP" / "CCDA" (children's clinical decision area) |
| **Allergies** | Allergies can be one, or a combination of, "food" / "drug" / "other" - better represented as three separate categorical features "food allergy" / "drug allergy" / "other allergy" with "Y" / "N" categories - "NKA" (no known allergy) and "NKDA" (no known drug allergy) observations are implied given "N" in each of the new allergy features |
| **Ethnicity** | Vast and diverse array of ethnicity descriptions - too diverse for meaningful analysis. Ethnicities are grouped into 3 categories "European" / "Asian" / "Other" - unclear or mixed ethnicities default to "Other" |

Table 2.3: Pre-processing steps performed on selected datset features

Prior to any analysis or modelling, the data were divided into a training dataset and holdout test-set. This is a standard approach to ensure the results of this study are not biased too heavily towards observations from the dataset, and that the findings generalise

well to unseen data. Given the size of the dataset and the scarcity of positively labeled examples, a stratified split of 2/3 training data 1/3 holdout test data was used. Any observations or results discussed from this point are taken from the training data only, unless otherwise specified.

## 2.2 Data summary

We focussed our initial analysis on a comparison between the patients that were successfully discharged from ACE and those that were admitted to hospital. Results of these analyses can be seen in Section A.1. The results show that there isn't a clear distinction between patients that required hospital treatment and those that didn't. Very few of the features, in isolation, show an obvious difference in their distribution between the two groups - statistical significance tests suggest that only "Referral Time" (p-value 0.032), "Gut Feeling of Referrer" (p-value 0.052) and "Oxygen Saturation" (p-value 0.002) have a statistically significant relationship to hospitalisation rates.

It should be noted that a 5% significance test is a high burden to place on data of this type. The ACE dataset is comprised of all available referral data, rather than experimental output or a carefully designed cohort study, which are more typical use-cases for these significance tests. As such, "Referral From" (p-value 0.17) and "Age" (p-value 0.14) also show signs of correlating with referrals to hospital, though the relationship is much weaker than the other features highlighted.

Among the features that do correlate strongly with hospitalisation, are further considerations that diminish their potential as predictors. All of the features that show particularly high/low proportions of hospitalisations among their categories are supported by only a small fraction of the observations. This effect is visualised in Figure 6. Problems may arise from models that rely heavily on these features in making predictions:

1. The tiny fraction observations that underly these particularly high/low hospitalisation proportions suggest these figures will vary significantly if the data is resampled or new observations are added. The predictions and composition of models that rely on these features are likely to be similarly variant/unstable as a result.

2. The small fraction of the examples that fall into these categories suggest they are relatively uncommon and so will have minimal impact, considering the majority of predictions will relate to other categories.

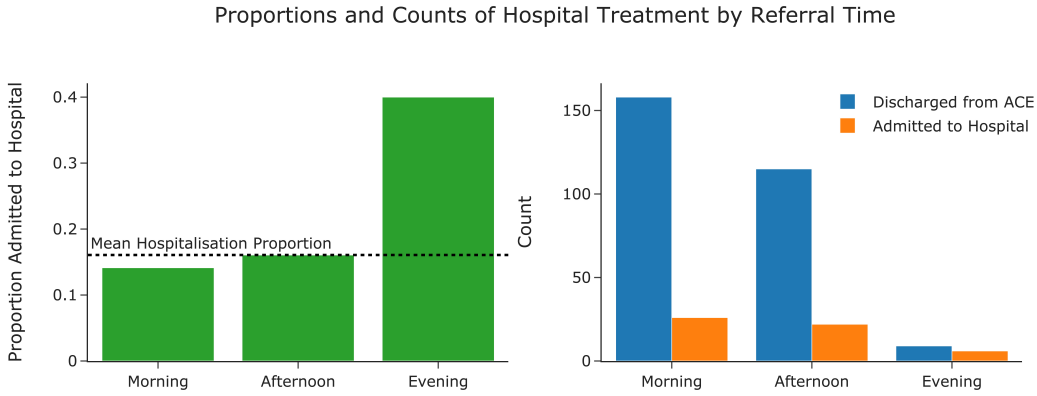



Figure 2.1: Visualisation of the "referral time" feature. The proportion of outcomes for each time is shown on the left, and the individual counts for each group are shown on the right. Particular attention should be paid to the proportions of patients hospitalised that are referred in the evening and the relative numbers of observations that make up this category - the bars on the right-hand side of each plot.

## 2.3 Interaction Effects and Feature Engineering

The analyses thus far consider each of the features in isolation. We expect that the features also interact with one another to influence outcomes. We know, for example, that age and heart rate are features that interact; resting heart rate in children decreases with age, and so an unusually high heart rate for a 10 year-old may be perfectly ordinary for a baby. The permutations of potential interaction effects extend into the hundreds, when considering only pairs of the features in the ACE dataset. Given this, domain expertise is particularly useful in highlighting related features that may interact to affect the outcome. Such domain-specific knowledge was leveraged to design new features from the ACE dataset:

- The ACE referral criteria for the wheeze/asthma pathway (Figure 1.1) define reasonable ranges for heart and respiratory rates broken into age groupings. New features were engineered that represent the ACE referral criteria, for example "low"/ "normal"/"high" categories for heart and respiratory rates.
- The Royal College of Nursing publish Advanced Paediatric Life Support (APLS) guidelines [8], which include definitions of normal heart and respiratory rate ranges divided into more granular age groupings than the ACE guidelines - these guidelines were also used to generate new features.

The engineered features were subjected to the same analyses as the simple categorical features. The results of these analyses reveal only weak relationships with hospitalisation, with low statistical significance (Tables A.4 to A.5).

## 2.4 Discussion of Data Analysis

One might be tempted to say that the lack of strong predictors of hospitalisation seen throughout this analysis, particularly among the specific criteria ACE use to guide refer-

rals, suggests flaws in the criteria ACE use to make referral decisions. It bears repeating, therefore, that **these data are taken from patients accepted for ACE treatment**. When considering that patients in the dataset were already deemed to be of sufficiently low hospitalisation risk to be admitted to ACE, based specifically on the features recorded therein, it is unsurprising that there are few strong predictors of hospitalisation to be found among those same features. Indeed, it is the relative success of ACE decision making that results in the lack of clear predictors of hospitalisation found in the dataset.

These analyses do highlight key challenges when considering using the ACE data to model outcomes. The small fractions of the dataset that support certain feature categories (Figure 2.1) suggest that re-sampling the data will result in significant variance of the distribution of these categories between samples. This variance is likely to affect the robustness and generalisability of any models trained using the data. This should be considered carefully when interpreting the results of all experiments, and particular attention should be given to the variance of any figures reported (where available).

### 2.4.1 Reproducing Results

Code to reproduce the results of these analyses can be found in ***exploration/-initial_exploration.ipynb***

## 2.5 Rejected Referrals

The ace team do collect data on the referrals they reject (88 examples). A significant proportion (approximately 50%) of these referrals are rejected because they do not meet the following minimum requirements to be considered for treatment in the service:

- Age <2 years old
- Referral time after 16:00 - the service remains open until 20:00 and requires patients to have a minimum 4 hour observation period

Such referrals are not suitable for this study as the patients would never have been accepted for ACE treatment under any circumstances.

Other referrals are rejected because the ACE team or referring clinician aren't satisfied they can be safely treated at home. Whilst such examples, appropriately labelled, would be an important addition to the dataset, unfortunately the vast majority (>90%) are missing a significant number of observations. Aside from the fact that these examples aren't labelled, the missing fields are such that these observations would not be useful even if a labelling scheme were devised.

# 3 | Classification Modelling

## 3.1 Task Description

The ACE team hypothesise that the referral data they collect can be used to predict treatment outcomes. Defining this as a machine learning problem, they suspect a relationship exists between the input variables - the referral data, $X$ - and the outcome variable - discharge with / without the need for hospital treatment, $y$. This relationship can be formalised as:

$$y = f(X) + \epsilon \tag{3.1}$$

where $y$ is an unknown function of the input variables and $\epsilon$ is a random error term (independent of with mean zero). Defining the problem in this way, the aim of this experiment is to approximate $f$, and subsequently make predictions of $y$ of the form:

$$\hat{y} = \hat{f}(x) \tag{3.2}$$

where $\hat{y}$ and $\hat{f}$ are approximations of the true underlying $y$ and $f$. This process can be thought of more generally as training a predictive model [9].

The accuracy of such a predictive model depends on two terms, the reducible and irreducible error. The reducible error is the degree to which $\hat{f}(X)$ accurately approximates $f(X)$ - the more accurate the representation, the lower the reducible error. The irreducible error is the $\epsilon$ term - this is independent of $X$ and can be thought of as the unavoidable error - the factors that affect outcomes and that aren't captured in the data. In testing our hypothesis we hope to establish:

1. the degree to which the referral data is able to explain the outcome - the relative sizes of $f(X)$ and $\epsilon$ as proportions of $y$
2. how accurately we might approximate $f$

No "one-size-fits-all" predictive model exists. The modern machine learning toolkit includes vast number of approaches to classification modelling - each has it's own prior assumptions of the form that $\hat{f}$ takes, and thus each has it's own associated benefits and drawbacks. This experiment will test a range of these approaches, with the expectation that one amongst these techniques will establish a reasonable baseline for $\hat{f}$ that minimises the reducible error as much as possible. Unfortunately, the complexity of these modelling techniques and their variety renders their discussion in this report impractical, though the general intuition established above and the following discussion is sufficient to understand the results.

### 3.1.1 Encoding Non Numeric Data

Machine learning models require data to be represented numerically. This is an issue when considering categorical data, such as the "referral from" or "allergy" features in the ACE dataset. There are a number of approaches that represent categorical data numerically [10] [11], some of which cannot be used in this setting given the small size of the dataset. The following approaches will be used in this experiment:

- **One-hot encoding**: Each categorical feature is split into its respective categories, each with a simple 1/0 or "on"/"off" value. For example, the following data:

| Patient | Referral Time |
|---|---|
| **1** | Morning |
| **2** | Afternoon |
| **3** | Morning |
| **4** | Evening |

would be one-hot encoded as:

| Patient | Referral Time Morning | Referral Time Morning | Referral Time Morning |
|---|---|---|---|
| **1** | 1 | 0 | 0 |
| **2** | 0 | 1 | 0 |
| **3** | 1 | 0 | 0 |
| **4** | 0 | 0 | 1 |

- **Target encoding**: Each category is given a numerical value based on the proportion of target variable it represents, in this case the proportion of patients that require hospital treatment. For example, if 15% of patients referred in the evening require hospital treatment, the "evening" category is replaced by the figure 0.15. Care must be taken to avoid "leakage" when using this approach - that is, encodings should not be calculated using the label for the example in question, or from the labels of examples that will be used to evaluate model performance.

The free-text fields, such as "medical history" and "examination summary" present a greater challenge, and will therefore be excluded from this experiment. Further analysis and modelling of these text features can be found in Chapter 4.

### 3.1.2 Balancing Target Labels

As discussed in Chapter 2, examples of children that required hospital treatment are far fewer in number than those successfully treated by ACE. This presents a significant challenge when attempting to train a classification model to accurately predict the probability of a hospital referral [12]. For example, models trained on imbalanced data can achieve relatively high prediction accuracy by predicting the majority label only - so, any model that predicts every patient will be treated successfully by ACE will be approximately 86% accurate. Optimising for prediction accuracy alone is likely to result in many such models.

To mitigate these issues, the following data preparation techniques will be tested, each of which attempts to address the imbalance of labels in the dataset:

- **Weighted labels**: Models are "punished" during training for making incorrect predictions. This penalty can be weighted depending on the label, so a model can be more heavily "punished" for making incorrect predictions of the minority label. The size of weighting is usually determined by the proportion of majority/minority labels

in the dataset - so a label that is five times less common than another is weighted five times more heavily. Note: Label weighting is only available in modelling techniques that use certain optimisation approaches, and thus is not available for some of the modelling techniques tested in thie experiment.

- **Synthetic Minority Oversampling Technique (SMOTE)**: SMOTE [13] generates new synthetic examples of the minority label to balance the proportion of labels in the dataset. New samples are generated by selecting a random minority example, and a small number of "neighbours" for that example - other examples that are the most similar to the selected example. One of the neighbours is then randomly chosen, and a synthetic data point is sampled by interpolating between the random example and the selected neighbour. This process is repeated until the number of examples of each label match.

- **Undersampling**: Similar in spirit to oversampling, undersampling is the removal of examples from the majority label until the number of examples with each label match. There are many approaches to systematically select examples to remove - in this experiment random undersampling will be used.

### 3.1.3 Evaluating Models

An imbalance of labels also makes model evaluation more challenging. Simple accuracy is not an effective measure of performance if the proportion of labels is skewed heavily in one direction. Given this, it is important to use metrics that measure the proportion of the imbalanced labels that are correctly classified:

- **Precision**: This is the proportion of examples that are classified correctly, among those that are predicted to have a positive outcome - in terms of the ACE task, this is the proportion of patients that actually require hospital treatment, out of those predicted to need hospital treatment

- **Recall**: This is the proportion of positive examples that are classified correctly (ignoring every negative example) - in terms of the ACE task, this is the proportion of the children that need hospital treatment that are correctly identified.

- **F1 Score**: F1 is a combination of precision and recall. F1 calculates the harmonic mean between the precision and recall, offering a balance between these metrics:

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{3.3}$$

  In isolation, one can achieve a perfect precision score by predicting one positive example correctly and all the others negative, or a perfect recall score by simply predicting every example as positive - F1 avoids this by evaluating the two metrics together. As F1 is a harmonic mean of two proportions, it also takes on values between 0 and 1 and can be easily interpreted like precision and recall.

- **AUC/ROC**: This is a this represents the degree of "seperability" of the model predictions based on the true labels - it measures the degree to which a model is capable of separating between the two labels correctly. The theory is too complex to discuss here, but an interpretation of the metric can be easily explained. Perfect seperability will achieve an AUC of 1. 0.5 indicates no separation or that a model is choosing randomly. Values below 0.5 indicate that the model is skewed toward

making incorrect decisions. Values closer to zero are rarely seen, given that one could simply flip the predictions to achieve an accurate model, but occasionally models will stray slightly below 0.5 - these results should be interpreted in much the same way as those at 0.5 or slightly above.

## 3.2 Experimental set-up

### 3.2.1 General set-up

Predictive models were trained (with the help of the popular Scikit-Learn[14], and Imblearn [15] Python packages) using a combination of each of the following modelling techniques, and approaches to categorical encoding and label balancing:

| | |
|---|---|
| **Predictive Modelling Technique*** | Logistic Regression<br>Support Vector Machines<br>K-Nearest Neighbours<br>Random Forest Classifier<br>Gradient Boosted Decision Trees<br>Ada-Boost Classifier<br>Gaussian Naive Bayes Classifier<br>Quadratic Discriminant Analysis |
| **Categorical Encoding Technique** | One-Hot Encoding<br>Mean Target Encoding |
| **Label Balancing Technique** | Balanced (Weighted) Labels<br>SMOTE<br>Random Undersampling |

**Amongst these models, a wide variety of hyperparameters specific to each technique were tested. These are too numerous to detail here, but details can be seen in* ***models/sklearn_models.py*** *from the project repo*

A grid search method was used to test each combination of model, hyperparameters, categorical encoding approach, and label balancing approach. Models were scored using a 3-fold cross validation method, given that dividing the training data any further would result in too few positive examples in each validation fold. Variability of outcomes was a significan issue in early experimentation - to establish a reliable estimate of the variance of cross validation results, the training data was shuffled and the 3-fold cross validation scoring was repeated 10 times.

The models, hyperparameters and data preparation methods that performed best in cross validation were then tested against the holdout test set, to estimate how well the cross-validation scores represent the prediction scores for data that wasn't used during training, and how well the models generalise. Only the best performing models in cross validation were scored against the holdout test set, to minimise the risk of biasing the results to those that perform best against the test set.

### 3.2.2 Precautions

Particular care was taken to write a custom cross validation loop that accounted for the following complexities of this experiment:

- Synthetic samples were added to the training folds only - care was taken to ensure models were validated against genuine training examples only, without any added synthetic data
- Target encoding was calculated using a "leave-one-out" method from the training folds only to avoid "data leakage" - no holdout validation examples were used to calculate target encodings
- Target/one-hot encoding was completed after generating synthetic examples - otherwise the SMOTE algorithm would treat the encoded categories as numeric, and interpolates between them creating erroneous "sub-categories"
- The random fold samples were kept identical when training each individual model and configuration to ensure an unbiased comparison of each model

## 3.3 Results

Results cross validation results from each of the classification models can be seen in Table A.6 and the test set results of the best performing models can be seen in Figure 3.1. None of the classification models are able to make useful predictions of hospital outcomes from the ACE data during cross validation or against the holdout test set. Those models that achieve good overall accuracy (>70%) do so at the expense of identifying patients that required hospital treatment - recall (<30%) and precision (<25%). Conversely, models that are able to identify greater numbers of patients that require hospital treatment, do so at the expense of overall accuracy. None of the models achieve an F1 score above 0.3 or an AUC above 0.55, indicating the low degree to which the models are able to separate patients that were successfully treated by ACE from those that required hospital referral.

| | | F1 | AUC | Acc | Rec | Prec | True +ve | True -ve | False +ve | False -ve |
|---|---|---|---|---|---|---|---|---|---|---|
| **SMOTE** | **Logistic Regression** | 0.158 | 0.452 | 0.605 | 0.222 | 0.122 | 6 | 92 | 43 | 21 |
| | **Random Forest** | 0.067 | 0.422 | 0.654 | 0.074 | 0.061 | 2 | 104 | 31 | 25 |
| **Balanced** | **Logistic Regression** | 0.244 | 0.519 | 0.617 | 0.37 | 0.182 | 10 | 90 | 45 | 17 |
| | **Random Forest** | 0.208 | 0.533 | 0.765 | 0.185 | 0.238 | 5 | 119 | 16 | 22 |

Figure 3.1: Test set scores for the best performing classification models as determined by the cross validation scores

The observed standard deviations between the cross validation folds also indicate that model's predictions vary significantly depending on the data they see during training. This indicates that the decisions, or heuristics, of the classification models are not robust to small changes in the training dataset. These results support the issues discussed in Section 2.2 - there are very few examples that exhibit the features that are most indicative of hospitalisation risk, and model results vary dramatically depending on the inclusion/exclusion of these examples during training.

Visualising the model predictions further emphasises the poor performance of the classification models. Figures 3.2 to 3.3 show plots of the test set predictions from the best performing models. The distributions of predictions barely differ between patients that were successfully discharged from ACE and patients that were referred to hospital. The logistic regression model also lacks confidence in its predictions - the vast majority of the predictions made fall within the mid range of probabilities, indicating that the model will rarely deviate from an approximate 40-60% chance of hospitalisation.

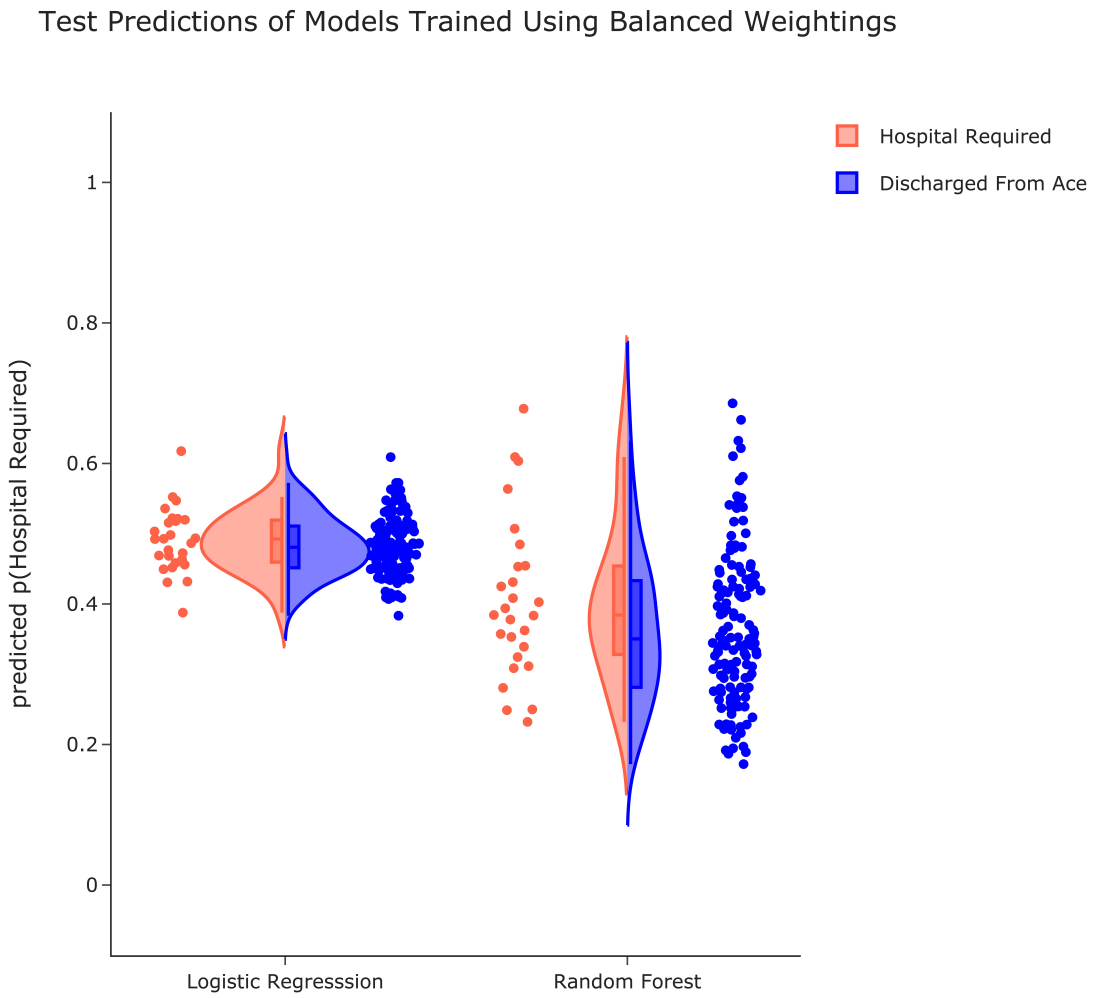



Figure 3.2: Combination scatter/violin/box plots of the test set predictions from the best performing models trained using weighted labels. Note: the scattered points have been "jittered" - randomly displaced along the x axis to make visualisation easier - thus the x axis position has no meaning.

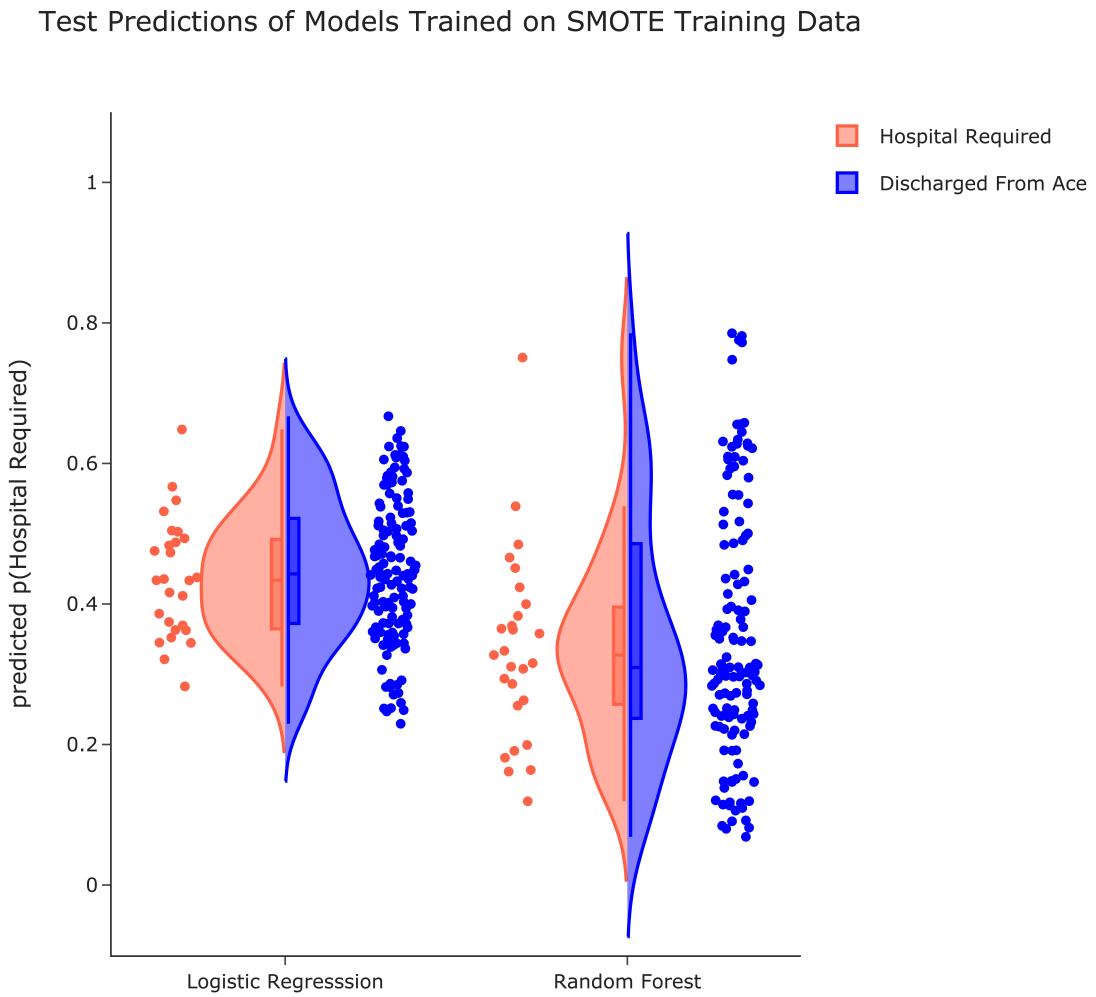



Figure 3.3: Combination scatter/violin/box plots of the test set predictions from the best performing models trained using SMOTE training data. Scatter points are "jittered" as in Figure 3.2

### 3.3.1 Reproducing results

Code to reproduce these results can be found in ***models/sklearn_models.py*** and ***models/plot_predictions.ipynb*** from the project repository

## 3.4 Conclusions

A broad range of data preparation methods and modelling techniques were used in this experiment. It is reasonable, therefore, to assume that we have established a good baseline for $\hat{f}$, our approximation of the true relationship between the referral data and hospitalisation outcomes. As we were unable to generate anything approximating accurate predictions of hospitalisation, we can reasonably reject the initial hypothesis of the ACE team - it is not possible to predict treatment outcomes using the referral data they have provided. The

irreducible error presented by this problem appears to be such that little can be determined about the outcomes from the input data. (It should be noted that the experiments thus far have excluded the free-text features, which are analysed in Chapter 4).

These results are unsurprising when considered in the context of the findings of the initial data analysis. The lack of obvious predictors of hospitalisation found in the data analysis is reflected by the poor prediction performance of the models trained using the dataset. It bears repeating that the absence of predictors in the dataset serves to affirm the referral decisions made by the ACE team, given that the dataset comprises only patients that were accepted for treatment within ace. If instead we had trained extremely accurate classification models, this would indicate the presence of obvious indicators of hospitalisation in the referral data that ACE clinicians were oblivious to.

# 4 | Free-Text Analysis

## 4.1 Task Description and Methods

The ACE dataset contains free text notes that detail the medical history, initial examination information, and the treatment recommendations made for each patient on referral. As with all free-text, these notes are unstructured - there is no direct approach to represent them numerically. Because of this, they do not appear in the initial data analysis (Chapter 2) or initial modelling (Chapter 3). That isn't to say that these notes don't contain useful information, or that it isn't possible to leverage the information therein, only that doing so requires a different approach. Natural Language Processing (NLP) includes many such approaches to analysing and modelling free text data [16]. In this experiment we aimed to analyse the free text data to identify words or concepts that are predictors of hospitalisation.

The following are brief introductions to the concepts that are used in this experiment:

### 4.1.1 Bag of Words

Free text analysis will often begin by representing the text as a "bag of words". This is a list (or vector) representation, in which each position (or index location) represents a word, and the integer value at each position is a count of the respective word as it appears in the text. For example, the sentences:

**sentence 1**: *"the cat sat on the mat"*
**sentence 2**: *"the dog sat with the ball"*

can be represented by counts of the words that appear in them, using a common vocabulary as follows:

| | **the** | **cat** | **dog** | **sat** | **on** | **with** | **mat** | **ball** |
|---|---|---|---|---|---|---|---|---|
| **Sentence 1** | 2 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| **Sentence 2** | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |

This approach decomposes text into a numeric or vector representation, thus offering a means to analyse text statistically.

### 4.1.2 Pre-processing

Applying the bag of words approach unconstrained will usually lead to huge vector representations of text (consider the variety of words in common English). These huge vectors are difficult to interpret statistically. As such, the process of "tokenising" - representing text numerically - will often involve a number of pre-processing techniques that reduce the size of the common vocabulary, thereby reducing the length and complexity of the resulting vector representations. The following pre-processing techniques are used in this experiment:

- **Stopwords**: The most common words in natural language often do not have any meaning when taken in isolation. Examples include "and", "the", "it", "is", "of"

etc. It is common when performing simple NLP analyses to compose a list of such words and remove them from the vocabulary.

- **Stemming**: The English language includes a number of conjugations or inflections of words that have a common stem. For example the stem of the words "given"/"giving"/"gave" is "give". Counting these words individually is to treat them as separate and independent, which may not be desirable given their common meaning. Stemming reduces examples of such inflections to a common root, reducing the size of the vocabulary whilst retaining most of the information or meaning from the words that are redacted.

Other pre-processing decisions *add* additional words to the vocabulary. Considering only individual words eliminates the contextual relationships of pairs or groupings of words. When building a vocabulary "n-grams" can also be considered - these are groupings of n words that are considered as an individual element of the vocabulary in their own right. Bigrams, for example, are the pairs of words that appear in the text - bigrams from the first example sentence we saw above would be "The cat"/ "cat sat" /"sat on"/"on the"/"the mat". By considering such "n-grams", one can include some of the contextual relationships between the words in a text, that considering individual words would ignore. We might, for example, encounter the bigram "severe asthma" in the ACE notes, which has an important meaning that may not be captured by only noting the presence of the two words independently.

**Note**: Accounting for n-grams, a vocabulary can now contain word groupings as well as individual words. Therefore, to avoid confusion the individual elements of a vocabulary are often referred to as "tokens" rather than words.

### 4.1.3 TF-IDF: Term Frequency / Inverse Document Frequency

Given a numeric "bag of words" representations of text, we inevitably wish to analyse the word composition. Our intuition may be to assume that words that appear more frequently are more important. This assumption has two key flaws; flaws that can be addressed by using Term Frequency / Inverse Document Frequency (TF-IDF):

1. Using raw word counts will inevitably lead to higher counts of words in longer texts, leading to a bias that assumes words that appear in longer texts are more important. Calculating the term frequency, the relative frequency of a word as it appears in the text, overcomes this issue.

2. The wider context of the other texts that are being used for comparison is important. The importance of the word "patient" will depend on how common it is in general - its meaning will likely be very important if it only appears in a few specific texts, and will be of less importance if it appears in every other text in our corpus. The inverse document frequency calculates the inverse of the number of documents (notes in this context) that a given word appears in. It provides a measure of how common or rare a word is within the context of a wider corpus of texts.

By combining term frequency and inverse document frequency, we can establish an estimate of word importance that accounts for the length of the text, and the vocabulary of the wider corpus of texts being analysed.

### 4.1.4 Purpose

A note of caution: The motivation behind this experiment is **not to establish the free-text notes as predictors themselves**. Free text features are problematic as predictors for a number of reasons (the following points are not exhaustive):

- They vary considerably depending on the way a note is phrased, regardless of any difference in the underlying meaning
- They lack any pre-determined format, so information important to a referral may be present in some notes, but may be omitted in others
- Author's style of writing has a considerable effect on the composition of notes, particularly with a small group of authors like the ACE team, and can introduce unintended biases as a result

Because of issues like those above (and many others), robust classification models that use free text require training corpora that are many times larger than the examples available in the ACE dataset. As such, it is not feasible to use the free text notes as predictors in their entirety. Instead, the purpose of this experiment is to analyse the free-text notes and to identify any individual words/phrases that show a promising correlation with hospitalisation. This analysis can be used as a guide to collect more robust data from patient records, that have a greater provenance or accuracy than the content of the notes.

## 4.2 Experimental set-up

### 4.2.1 Text Pre-processing

We took the three free-text features from the ACE dataset - "medical history", "examination summary" and "recommendation" - and used the following pipeline to represent them numerically:

- **Stopwords**: A pre-determined list of stopwords were removed from the notes. The NLTK (natural language toolkit) python package [16] list of English language stopwords were used appended with an additional list of words specific to this task (details can be found in ***/exploration/text_analysis.ipynb*** from the project repo).
- **Stemming**: We stemmed the words to reduce the variety of inflections/conjugations - the NLTK "SnowballStemmer" was used [reference]
- **Vectorising**: We built a vocabulary of individual words and bigrams (pairs of words) and reduced this vocabulary to the 500 most common words/bigrams. Notes were "vectorised" using this vocabulary - represented as raw counts of these 500 tokens.
- **TF-IDF**: TF-IDF values for each of the vectorised notes were then calculated.

Note: the above pipeline was applied to each text feature individually i.e. vocabularies and inverse document frequencies were established for each of the text notes ("medical history", "examination summary", "recommendation") separately.

### 4.2.2 Analysis

We analysed the TF-IDF values visually to identify differences in the distribution of notes relating to patients that were hospitalised and those that were discharged successfully from ace.

We then modelled these positive/negative outcomes using a "lasso" logistic regression model, fitting one model for each text feature individually. The "lasso" form of the model enforces constraints that severely limit the number of features the model can use to make predictions. The aim was to produce a sparse group (small number) from the 500 most common words that are most predictive of hospitalisation/successful treatment, along with coefficient values that indicate the effect of this relationship (e.g. to increase/decrease the risks of hospitalisation).

The performance of the regression models was evaluated using a cross validation strategy that split the data into 3 folds. A range of heavy regularisation parameters was explored to identify better performing models from those that use a sparse range of features ($< 15$), optimising for F1 score. The outcome labels were weighted (as in the experiments in Chapter 3) to address the imbalance of positive / negative examples.

## 4.3 Results

### 4.3.1 Visual Analysis

Visual analysis of the TF-IDF values can be seen in Figure A.2. The distributions of TF-IDF values indicate clear differences in the importance of different tokens between patients that were referred to hospital and those that weren't. These results are, however, noisy:

- the standard errors for each value are large, particularly those taken from patients hospitalised, suggesting significant differences in the composition of notes
- Many words / tokens appear prominently in both the notes of hospitalised and discharged patients, some of which don't follow intuition i.e. mentions of "good" in the examination notes of patients that were hospitalised as well as those that weren't

As such it is hard to establish the presence of tokens that strongly differentiate between the notes of patients that did or didn't need hospital treatment from these analyses.

### 4.3.2 Regression Analysis

The results of the regression analysis draws a cleaner distinction between the most prominent tokens in the notes of patients that required hospital treatment. Figure 4.1 displays the tokens and coefficient values used by each of the individual logistic regression models. Remarkably, the models for medical history and examination summary use only one word token out of 500 - "asthma" and "salbutamol" respectively - to predict hospitalisation outcomes.

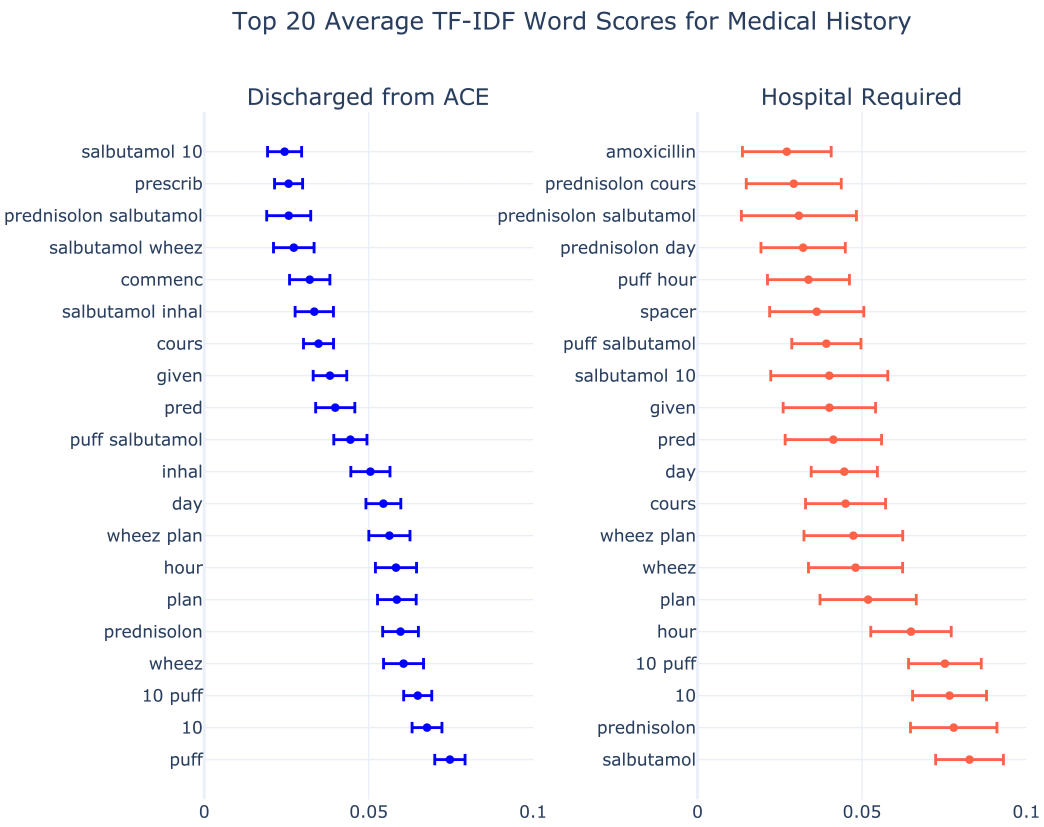



Figure 4.1: Coefficient values from the lasso logistic regression models trained on the TF-IDF scores from the individual text features. Positive values indicate that the associated word is predictive of a greater risk of hospitalisation. Conversely, negative values indicate that a word is predictive of lower risk of hospitalisation

The cross-validation scores of the regression models can be seen in Figure 4.2. Performance varies significantly between the different models and the different folds of the cross validation. The examination summary model achieves a better F1 score (F1 = 0.339, ROC=0.604) than any of the other models seen thus far, and remains accurate (accuracy = 0.751). The medical history model is accurate (accuracy = 0.736) but achieves a disappointing cross validation F1 score (F1 = 0.185, ROC=0.539), and the F1 score shows very high variance (F1 variance = 0.131) suggesting these results are very unstable, and that this model performs much better on certain samples of the data and much worse on others. The recommendation model shows poor results that barely deviate from chance allocation of labels, despite using the most features (accuracy = 0.659, F1 = 0.118, ROC = 0.454).

| | $\mathbf{F_1}$ | **AUC** | **Accuracy** | **Recall** | **Precision** |
|---|---|---|---|---|---|
| **Medical History** | 0.18 (0.13) | 0.54 (0.03) | 0.75 (0.06) | 0.22 (0.16) | 0.16 (0.11) |
| **Examination Summary** | 0.34 (0.01) | 0.6 (0.01) | 0.74 (0.03) | 0.41 (0.03) | 0.29 (0.03) |
| **Recommendation** | 0.12 (0.04) | 0.45 (0.02) | 0.66 (0.07) | 0.15 (0.07) | 0.1 (0.02) |

Figure 4.2: Cross validation scores for the TF-IDF logistic regression classifiers for each of the text features. Figures in brackets are the standard deviations of the respective figures

### 4.3.3 Reproducing results

Supporting scripts to reproduce these results can be found in ***models/text_analysis.ipynb*** in the the project repo.

## 4.4 Conclusions

This analysis indicates that mentions of "asthma" and "salbutamol" in a patient's medical history and examination summary respectively are potential indicators of hospitalisation risk. Models trained using only the TF-IDF scores relating to these words were able to achieve a prediction accuracy comparable to, and sometimes better than, models that use the rest of the ACE dataset. It is possible that a history of asthma, or recent treatment with salbutamol may be indicative of a greater risk of hospitalisation during ACE treatment. This is certainly a hypothesis worth exploring further.

However, results should be considered with caution for the following reasons:

- **Instability**: The high standard errors in the TF-IDF word scores and similarly high variance in the performance of the regression models indicate that these results are highly unstable. Indeed, different samples of the training data, and different parameterisations of the text processing and modelling stages result in significant shifts in the observed results. The primary findings - the strength of "asthma" and "salbutamol" as predictors - generally remain robust to such changes, but the data analysis and model classification scores and other coefficient values shift significantly. Many other promising predictors were identified in earlier experiments, to be later eliminated from consideration as a result of this instability.

- **Context**: These analyses considered words and word pairings in isolation. A significant amount of context is lost when isolating individual elements of text in this way. Take for example the simple case of negation: the TF-IDF score of the word "asthma" in a patients medical history will be identical in the sentences "patient has a history of asthma" and "patient has no history of asthma" despite their opposite meaning. Many other variations in the contexts that words appear aren't captured by these analyses. This lack of context as a factor in this analysis may have a significant effect on the observed results.

# 5 | Bayesian Analysis

## 5.1 Task descriptions

Our initial experiments defined the predictive task as taking the form:

$$y = f(x) + \epsilon \tag{5.1}$$

where we attempt to estimate $f$ as accurately as possible. This is often referred to as a "frequentist" approach, after the frequentist branch of statistics from which it derives. By defining the problem this way, we are only able to explain the outcomes $y$ in terms of $f(X)$. The error term $\epsilon$ represents the aspects of the outcome $y$ that remain unexplained after considering the input data. Thus, by using this approach we are unable to say anything about $\epsilon$, the irreducible error.

Assuming one can achieve accurate predictions by estimating $f$, the irreducible error is a minor issue: $\epsilon$ will be relatively small and $\hat{f}(X)$ "explains" $y$ relatively well. Unfortunately, this has not been the case with our attempts to model ACE treatment outcomes, and so key parts of the hypothesis remain unanswered:

1. We are unable to quantify the degree to which treatment outcomes can be modelled using the ACE referral data. Thus far we have seen a "best guess" based on the outputs from a range of popular machine learning methods.

2. We have not been able to establish which of the referral observations reliably indicate an increased or decreased risk of hospitalisation. Though we can certainly establish predictive heuristics from the models trained in our initial experiments, the lack of predictive accuracy among these models suggests these heuristics would not be useful.

We aim to address these issues in this experiment.

### 5.1.1 The Bayesian Approach and its Benefits

Ideally, we require a methodology that not only models the outcome $y$ but can also account for the uncertainty or error of this model. Fortunately, we can achieve this using a Bayesian approach [17]. Instead of formalising the problem as having fixed or deterministic outcomes $y$, we can represent the outcomes as a random variable $Y$. By doing this, the error or uncertainty becomes part of the model: the variance or uncertainty in the outcome of the random variable. The power of this approach is best represented with a simple example:

Let there be two ACE patients, **A** and **B**, and we wish to determine their suitability for ACE treatment. We might use a simple heuristic that patients with a **<0.3** chance of hospitalisation are suitable to be treated at home. If we defined the problem using a frequentist approach, we might establish a model that predicts the probability of hospitalisation for patients **A** and **B** as **0.1** and **0.2** respectively - thus, under this model both patients **A** and **B** would be accepted for ACE treatment. A Bayesian approach, on the other hand, would return a predictive distribution of probabilities for both patients, such as:

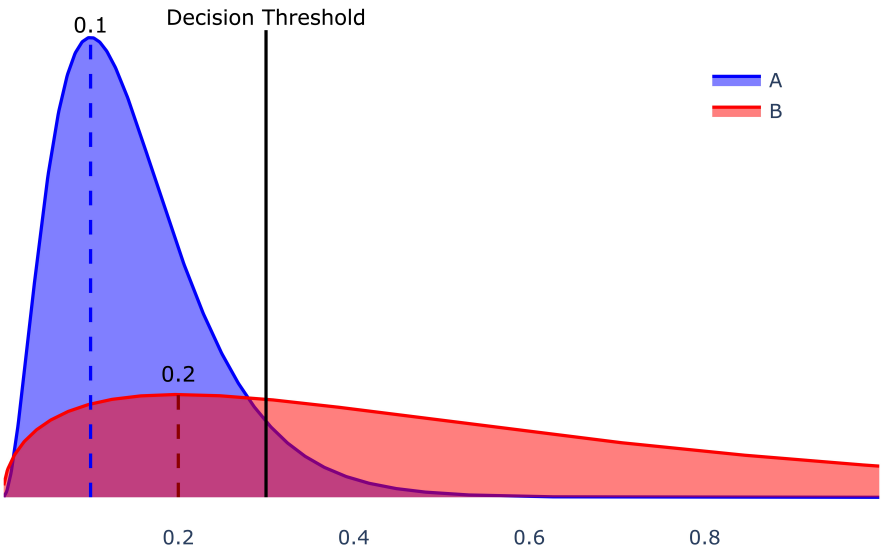



Figure 5.1: An example of the distribution of predictions that a Bayesian model might produce

The example result in Figure 5.1 have the same maximum likelihood estimates (MLEs) as the frequentist predictions: **MLE(A) = 0.1** and **MLE(B) = 0.2**, but their interpretation is very different. The chance that patient **A** will be hospitalised is likely to be **<0.3**, but the same cannot be said of patient **B**. As a result, we may well make a different decision based on the Bayesian output.

### 5.1.2 Bayesian Logistic Regression

A Bayesian approach can be applied to many of the popular frequentist machine learning techniques. In this experiment, we will model the data using logistic regression, one of the techniques that proved most successful in our initial experiments. Logistic regression assumes that the observations have a linear relationship with the outcomes, which takes the form:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{5.2}$$

Where the $x$ terms are individual features from the observations and the $\beta$ terms are model coefficients (sometimes referred to as parameters) that correspond to each feature (and $\beta_0$ the intercept term). $z$, a linear combination of the input features and the coefficient terms, is then used to calculate the probability of the outcome $Y$ using a sigmoid function:

$$p(Y = 1) = \frac{1}{1 + e^{-z}} \tag{5.3}$$

The coefficients interact with the observations to indicate what effect they have on the outcome. For example, if we observe the coefficient for "referred from GP surgery" is

**+1.6**, this indicates that this feature serves to increase a patient's risk of hospital referral. Conversely, should we observe the coefficient for "oxygen saturation" is **-0.75**, this would indicate that higher oxygen saturations reduce risk of hospitalisation (and conversely lower saturations increase risk). The sigmoid function takes the sum of these linear interactions $z$, which can take any real values ($z \in [-\infty, \infty]$), and constrains them to probabilities between zero and one ($p(Y = 1) \in [0, 1]$).

In Bayesian logistic regression, the coefficients are represented as random variables. The model is then defined by a set of three key functions that build from one to the next:

1. **Prior**: A "prior" distribution is chosen for each of the coefficients that is intended to define a plausible range of possible values that the coefficient could take, based on the prior knowledge of some subject matter expert. For example, we might say that the "heart rate" coefficient is more likely to be positive, and less likely to be negative, given our understanding that higher heart rates are indicative of greater clinical risks - we would assign that coefficient a prior probability distribution that reflected these assumptions.

2. **Likelihood**: The prior distributions, the actual observations in the data, and our chosen distribution of the outcome variable , are all combined to form a likelihood function. The likelihood can be thought of as the probability that we might observe a particular set of referral features and a subsequent outcome, given the assumptions we have made about the problem.

3. **Posterior**: The likelihood function is used to establish a "posterior" estimate for each of the coefficients - these are distributions of likely values for the coefficients, based on the prior distributions, the observed values in the data and the outcomes. This can be thought of intuitively as taking our initial assumptions of how the data interact to affect the outcomes (the priors), and then gradually adjusting these assumptions as we observe real data (the posteriors).

The power of Bayesian methodology is that it defines model parameters probabilistically. Not only can we represent how the features of our data interact to affect the outcomes, but also how certain or uncertain we are of these interactions. This can be particularly important in a situation where there is little training data. Bayesian predictions and coefficient estimates can take account of the number of observations on which they are based - so, a parameter estimate based on many real-world observations will be less variable, or more confident, than another parameter based on very scarce observations.

The Bayesian approach also eliminates the need to balance the labels in the dataset. **The fact that there are fewer examples that require hospital treatment is actually useful information to a Bayesian model**. The number of examples that exhibit a particular observation informs the level of confidence the model can assign to the associated parameter estimates - parameter estimates based on scarce examples will therefore be less confident or have higher variance. Therefore, a lack of examples in the training data helps the model to "say what it can't be confident about".

### 5.1.3 Highest Posterior Density Intervals

Given that Bayesian models output a distribution of parameter estimates, we require a statistic that can represent these distributions succinctly. One such statistic that is frequently used in this setting is the HPD or Highest Posterior Density interval. The HPD is the narrowest interval in which a given proportion of a distribution lies - so a 95% HPD

for a parameter estimate would be the shortest interval in which there is a 95% chance of the true parameter value falling in that interval. This concept can be easily represented visually:

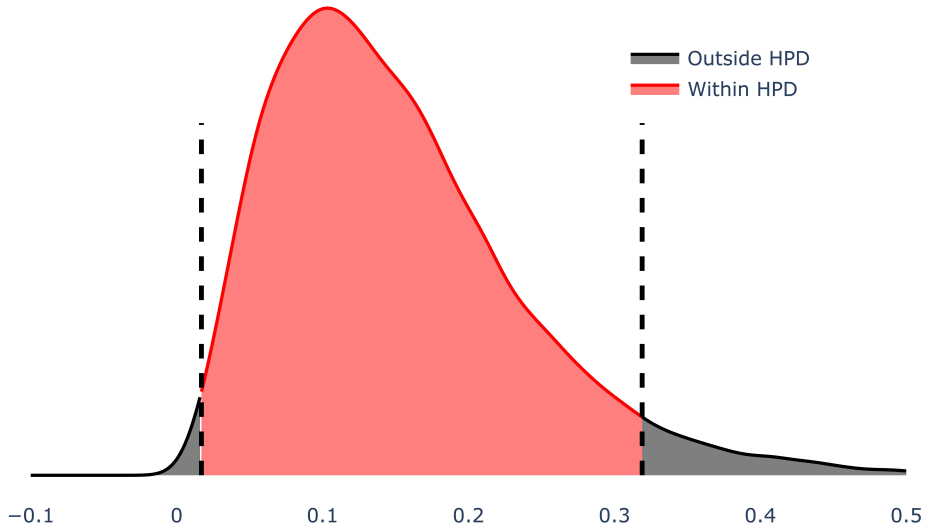



Figure 5.2: An example of a probability density plot for a hypothetical posterior distribution highlighted with a 95% highest posterior density interval

The HPD is a simple way to represent plausible upper and lower bounds for our parameter estimates. This is particularly useful when testing the null hypothesis that a parameter has no effect on outcomes. Parameter estimates whose HPDs have a greater overlap with zero indicate a greater chance that they are having no effect on outcomes - conversely, a HPDs that are entirely positive or negative indicate reasonable certainty that the feature is having the associated effect.

## 5.2 Experimental set-up

### 5.2.1 MCMC Sampling

The complexity of Bayesian modelling in such that only a brief introduction to the methods can be given here. We used a method called Markov Chain Monte Carlo (MCMC) to establish estimates of our parameter distributions. Many different MCMC algorithms have been established in recent years that perform the same basic functionality: they iteratively generate samples for each of the parameters from the likelihood function. These samples are used to obtain estimates of the posterior distribution for each of the parameters, and subsequently calculate individual predictive distributions for the outcomes of different observations.

We used the Python probabilistic programming package PyMC3 to define our logistic regression models and to simulate our MCMC samples. PyMC3 defaults to a sampling algorithm called "JAGS" or "Just Another Gibs Sampler". The same hyperparameters and sampler specification were used to sample each model, details of which are in ***models/-pyMC3/bayes_models.ipynb***. We chose to use the default diffuse priors (sometimes referred to as "non informative"), to ensure that our coefficient estimates are informed primarily by the information (or lack thereof) in the dataset. The intention was to have the models produce confident parameter estimates only when the information in the training data supported these estimates.

To begin, a logistic regression model was defined and sampled for each of the individual features in the dataset. The samples from these models were used to estimate 95% HPD intervals for each of the coefficients values, which indicate the single effects of these features on the outcomes in isolation.

A subsequent model was then trained iteratively using a mix of a "greedy" forward selection approach, and a subsequent simple backward selection. Starting from the best single-feature model, additional features were added iteratively selecting for those models that resulted in the highest performance. At each point a feature was added, subsequent models were tested by removing one feature from the model at that stage. Performance was evaluated using estimated log point wise predictive density (ELPD), calculated using Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO-CV). The ELPD takes into account the "effective number of parameters" in each model, effectively punishing greater model complexity. The training algorithm was allowed to run until the size of the model was such that samples became unstable - exhibiting divergences and low effective sample sizes - as signified by the warning features of the PyMC3 package. Coefficient HPD intervals and test set predictions were calculated using the samples from this best performing model.

### 5.2.2 Dataset Features

Given the results of the text analysis, we felt it important to test these findings alongside the other structured observations to facilitate a direct comparison. As such, we engineered structured features representing the predictors identified in the text analysis. Simple regular expression text searches were used to create simple yes / no features identifying the mention of "asthma" in the "medical history" feature and "salbutamol" in the "examination summary" feature. The regular expressions were defined to exclude any mentions that included negation before or after the keyword e.g. "no history of asthma", "salbutamol not given".

A number of the structured features were also excluded from the experiment, as it was not possible to simulate reliable samples from models that used these features. All the "address" features, the "APLS respiratory rate low" and the "APLS heart rate low" features were each excluded. The difficulties sampling models with these features relate to their relative rarity. Very few of the examples fall into these categories and so the samples from models that included these features were very unstable - they resulted in many "divergences" (the estimates become too extreme and run toward infinity) and low effective sample sizes (high autocorrelation between subsequent samples which reduces the accuracy of the parameter estimates).

## 5.3 Results

### 5.3.1 Coefficient Estimates/Feature Importances

Results for the single feature models can be found in Table A.7 and the coefficients of the best multi-feature model can be seen in Figure 5.3. Features with positive coefficient estimates indicate an increase the risk of hospitalisation and the negative estimates indicate a greater likelihood of successful discharge. These coefficient estimates indicate that the features taken from the best Bayesian logistic regression model are reliable predictors of hospitalisation - each of the HPDs are skewed heavily in either the positive or negative direction. The negative skew and relative magnitude of the intercept estimate should also be noted - the model begins from a baseline of significantly low risk of hospitalisation, and this risk only increases when an example shares several of the features that indicate an increased risk.

| | **Lower HPD** | **Upper HPD** | **Examples count** |
|---|---|---|---|
| **Intercept** | -6.59 | -2.01 | n/a |
| **Mentions Salbutamol** | 0.64 | 2.05 | 72.0 |
| **Mentions Asthma** | 0.13 | 1.69 | 73.0 |
| **Respiratory Rate** | 1.57 | 5.81 | n/a |
| **Referral From GP** | 0.39 | 3.35 | 193.0 |
| **Age Range Secondary** | 0.28 | 2.9 | 16.0 |
| **Gut Feeling Unwell** | 1.0 | 7.5 | 3.0 |
| **Referral Date Winter** | 0.03 | 1.4 | 111.0 |
| **Oxygen Saturation** | -4.27 | -0.18 | n/a |
| **Referral From ED** | -0.14 | 3.01 | 90.0 |

Figure 5.3: 95% highest posterior density intervals for the coefficients of the best performing logistic regression model. Features are listed in the order they were added to the model which, given the way in which the feature search algorithm worked, is an indication of the order of feature importance - most important top, least important bottom. Also included are counts of the examples that exhibit these features where relevant.

Interesting results can be seen by comparing the effects of the features as individual predictors and in the best performing model. Several of the features that appear in the best model are comparatively strong individual predictors ("mentions salbutamol", "mentions asthma", "referral from GP", "Oxygen Saturations"). Others that appear in the best model are less effective predictors individually ("respiratory rate", "age range - secondary", "gut feeling - unwell", "referral date - winter", "referral from ED") - these features appear to be more effective predictors when identifying "residual" risks - that is, the risk of hospitalisation "left over" after the "explanation" of more effective predictors.

The estimates for "referral from ED" are of particular interest, and serve to explore this "residual" effect further. The coefficient values from the individual model suggests this feature is more likely to have a negative effect on hospitalisation risk than positive when taken in isolation - HPD = [-1.10 - 0.34], though this estimate lacks confidence. In the presence of the other features in the best model, however, the relationship reverses, and the coefficient estimates are mostly positive - HPD = [-0.15, 3.04]. This suggests that the other features in the best model offer a better "explanation" of the reduction in risk that an emergency department referral signifies on its own, and that in the presence of these features an emergency department referral actually points to an increase in the risk of hospitalisation.

There are also features that are relatively effective features individually, that aren't used by the best model. These features are likely to be highly covariant with one, or a combination, of the features in the best model. An intuitive example is the "referral from GP" feature which is used by the best model, and the "referral profession GP" feature which, whilst it is a good individual predictor, is not used by the final model - both clearly identify very similar information and so the information from the better of the two features renders the other obsolete.

A comparison of the coefficient values for the categorical features in the best model can be seen in Figure 5.4. This comparison shows that the features are estimated to have very different effects on the risk of hospitalisation, and that the confidence of these estimates varies significantly. Note that **the relative size of the coefficient should not be confused with its effectiveness at predicting outcomes** - the text features "mentions salbutamol" and "mentions asthma" are among those that result in the most accurate predictions despite having the smallest comparative coefficient values. Broader or more diffuse estimates are either the result of a relatively small number of supporting examples (such as the "gut feeling unwell" - only 3 exaples exhibit this observation) or from features that don't correlate as clearly with one or other outcome, such as "referral from GP".

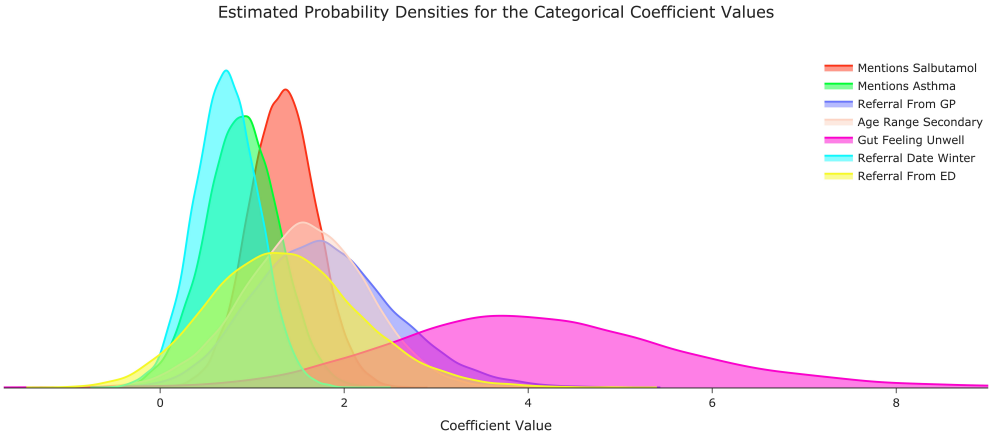



Figure 5.4: Estimated probability densities for the coefficient values of each categorical feature as taken from the best Bayesian logistic regression model. Results correspond with the HPD values in Figure 5.3

It should be highlighted that a comparison of the coefficients can be problematic. Each individual model has a unique intercept term and so the coefficients from different models cannot be compared, given the relative scales of the values vary. A comparison of the coefficient values within one model can only be made if the features have outcome values that can be sensibly compared - yes/no categorical features are good candidates as their value indicates the simple presence or abscence of a particular observation. Coefficient estimates of categorical features should not be compared with those of numerical features, nor should numerical observations of different phenomena be compared - it makes no sense to be comparing heart rates with oxygen saturations. As such, individual visulations of the coefficient estimates for the remaining features can be seen in Figure A.3.

### 5.3.2 Model Predictions

The predictions of the Bayesian logistic regression model are far better than those of the classification models from our initial experiments (Chapter 3). The test set predictions of the Bayesian model achieve an **AUC of 0.672** which is significantly higher than the best performing frequentist classification model (Balanced Random Forest Classifier - **AUC = 0.563**). It should be noted that the Bayesian model was trained using additional text features that were not part of the training data for the classification models.

Despite improved predictions, the Bayesian model remains unable to confidently identify patients that go on to be referred to hospital. A density plot of the individual sample test set predictions from the Bayesian model can be seen in Figure 5.5 and the median test set predictions can be seen in Figure 5.6. These outputs demonstrate that the model is able to confidently identify patients that are likely to be successfully discharged from ACE, signified by the very high peak at a probability **<0.05** and the narrow range of predictions at low probabilities for successfully treated patients. The predicted probabilities for hospitalised patients, though higher that those discharged, still remain low - the peak of predictions remains low (approximately **0.12**), an ideal peak would be much higher (further to the right). The model predicts a much broader range of probabilities as the risk of hospitalisation increases, indicating that it's confidence decreases with hospitalisation risk.

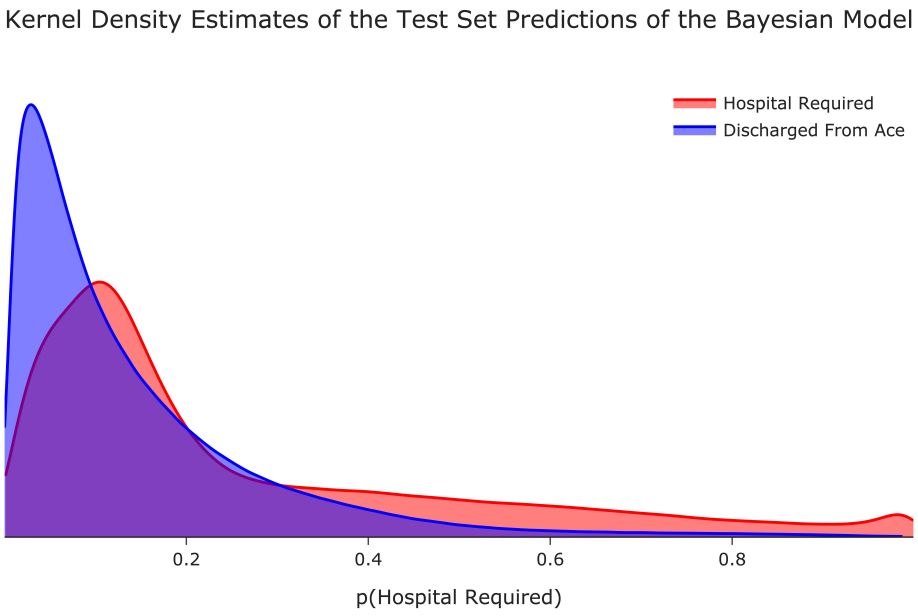



Figure 5.5: Kernel density estimates for all the sample predictions from the Bayesian logistic regression model grouped by the true label - successfully treated by ACE or referred to hospital

### 5.3.3 Reproducing results

Code to reprodce the results of this experiment can be found in ***models/pyMC3/bayes_models.ipynb***

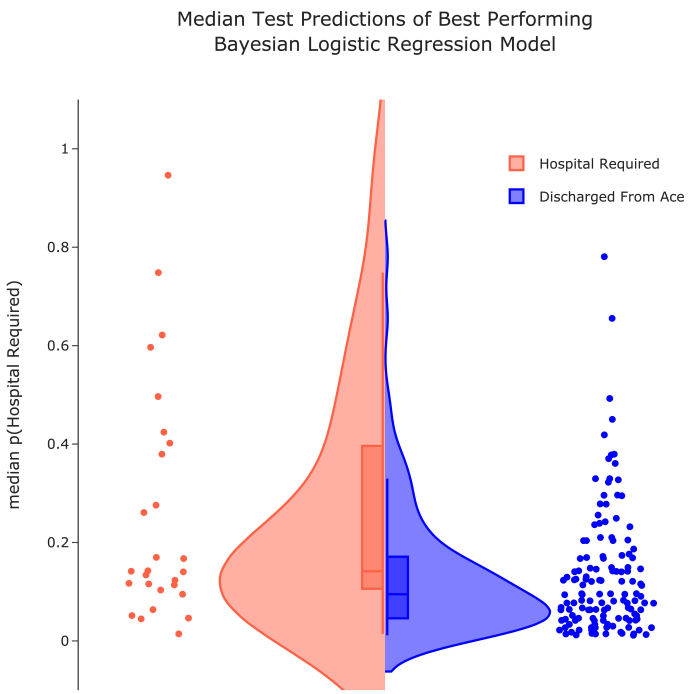



Figure 5.6: Combination Violin/Box/Scatter plot of the median predictions from the Bayesian logistic regression model, grouped by the true label. A trend of higher probabilities of hospitalisation can be seen among those examples that were referred to hospital, demonstrating the better performance of the Bayesian model over the models in Chapter 3.

## 5.4 Conclusions

The predictions of the Bayesian model clearly demonstrate that it is not possible to make confident predictions of hospitalisation from the ACE referral data. Though the predictions of the Bayesian model do show some separation between hospitalised patients and those discharged from ACE, it remains the case that the model assigned many of the hospitalised patients a low predicted risk and several patients that were successfully discharged a higher predicted risk. Generally, our model can either confidently predict successful treatment, or give a diffuse or uncertain range of hospitalisation probabilities. These findings support those in the initial data analysis (Chapter 2) and the frequentist classification modelling (Chapter 3), and serve to reiterate that the ACE team are making good use of the referral data to identify patients at higher risk of hospitalisation.

Parameter estimates from the Bayesian logistic regression model clearly show a relationship between certain key features in the ACE referral data and treatment outcomes. Particularly effective were the features taken from the text notes that were added for this experiment. It is likely that the text features are more effective as they are not part of the ACE referral criteria, and thus they have not been specifically used when making the referral decisions that underpin each of the examples in the dataset. The parameter estimates of the final model indicate that the new text features interact with the other referral features to indicate hospitalisation risk. This highlights the promise of exploring other features from patient's medical histories that might prove to be powerful indicators

of outcomes within the ACE service.

# 6 Additional features

## 6.1 Task Description

The analyses presented so far have used variables in, or derived from, the ACE referral form. Consequently, these variables represent data that the ACE admissions team have access to at the point of deciding whether to accept a referral. Considering this, it may therefore be unsurprising that predictive models trained on that data are not able to outperform experienced healthcare professionals.

The aim of this chapter is to investigate additional sources of data beyond the ACE referral form. By linking this data to that derived from the ACE referral form, we intend to build a model that may have increased predictive power over those built just using ACE data. We also aim to flag any variables of high predictive power that may be useful for the ACE team to consider as part of their acceptance criteria.

## 6.2 Experimental set-up

### 6.2.1 Data sources

The additional features used throughout this chapter fall into two main groups: Those derived from primary and secondary care (PSC) records, and those derived from geographical and geospatial analyses. PSC data were obtained from Connected Bradford (cBradford), an anonymised database linking routine electronic health data across primary, secondary, social and community care for over 700,000 individuals across Bradford Teaching Hospitals NHS Foundation Trust.

The geospatial analyses fell into three categories: Air quality measures, indices of multiple deprivation (IMD) and distance measures. Estimated background air pollution maps based on 2018 were downloaded from the DEFRA website (found here). IMD data from 2019 were downloaded from the Ministry of Housing, Communities and Local Government's GIS team (found here). Co-ordinates of GP surgeries and hospitals in the Bradford area were obtained using Google Maps.

### 6.2.2 Features from ACE referral form

The following features derived from the ACE referral form were included in the modelling process: Age, gender, illness severity, activity level, oxygen saturation, respiratory rate, heart rate, temperature, safeguarding concerns, food allergies, drug allergies, other allergies, whether the patient met the ACE referral criteria, mention of 'asthma' in the medical history, mention of 'salbutamol' in the examination notes, whether the referral was made by the GP, respiratory rate outside of the 'normal' APLS category, heart rate outside of the 'normal' APLS category, and whether the examiner's 'gut feeling' was anything other than 'normal'. Specific definitions of these variables can be found in the previous sections of this report.

### 6.2.3 Variables of interest

After data extraction and variable definition (detailed below in Section 6.2.4 and Section 6.2.5), variables were tested in univariate binary logistic regression models with hospitalisation as the response variable. Those showing a both significant ($p<0.05$) association with hospitalisation, and occurrence in at least 5% of the total patient population were termed 'variables of interest'. In total, 37 'variables of interest' beyond the variables extracted from the ACE referral form were identified and carried forward to further analysis.

### 6.2.4 Primary/secondary care features

Raw data were extracted from cBradford using SQL scripts built on a subset of the database (containing 5% of the total patients), then run on the main database. Four main categories of data were extracted, and are detailed below.

#### 6.2.4.1 Demographics

Gender and ethnicity were extracted from cBradford, although these variables also already existed in the ACE referral form. Ethnicity was grouped into 'Pakistani', 'White' and 'Other'. Neither of these variables influenced hospitalisation risk (see Table A.8), and thus none were carried forward to modelling.

#### 6.2.4.2 Co-morbidities

As this work uses data from the 'asthma/wheezy child' ACE pathway, we extracted information on a number of conditions known to be co-morbid with asthma [18]. Full details of co-morbidity data extracted are shown in Table A.9. From co-morbidity occurrence dates, binary variables were created for each condition determining whether diagnosis occurred within the month, six months, or year before ACE acceptance, as well as whether a diagnosis of the condition had ever been made.

Co-morbidity 'variables of interest' were any diagnoses of eczema, pneumonia, bronchitis, and of any other 'common respiratory conditions' (Influenza, common cold, hay fever, sinusitis, croup and streptococcal pharyngitis), as well as diagnosis of eczema in the last year (before ACE acceptance), pneumonia in the last year, and eczema in the last six months. Further details on these variables are found in Table A.15.

#### 6.2.4.3 Prescriptions

Data were extracted on the prescription (and frequency) of four main categories of medication: Fast-acting bronchodilators (*relievers*), slow-acting bronchodilators (*preventers*), antihistamines, and prednisolone (see Table A.10). Data were also extracted on the number of inhalers prescribed prior to ACE acceptance.

As with co-morbidity data, prescription data were used to construct binary variables measuring if (and if so, how many times) prescription occurred during the year, six months and month prior to acceptance. Variables of interest (Table A.16) were the prescription of slow bronchodilators at all, within the last year, and within the last month, and the prescription of prednisolone within the last year, and within the last six months.

Also of interest were several binarised variables (see Section 6.2.6): >12 total inhalers prescribed before ACE acceptance, >4 total courses of prednisolone, >3 courses of prednisolone in the last year, and >4 courses of prednisolone in the six months.

#### 6.2.4.4 Visits

Data were extracted on the number of visits to different healthcare facilities made by patients prior to ACE admission. These were the emergency room (ER), general practice (GP), a hospital as an outpatient (OP), and a hospital as an inpatient (IP) (Table A.11). The number of visits within 3 years, one year, six months and one month of ACE acceptance were totalled.

Binarised variables of interest (Table A.17) consisted of >4 visits to the ER in the last year, >6 visits to the GP in the last year, >4 visits to the GP in the last six months, >9 IP visits in total, >4 IP visits in the last 3 years, >2 IP visits in the last year, >1 IP visit in the last 6 months, >8 visits of any kind in the last year, >4 visits of any kind in the last six months, and >6 visits of any kind in the last month.

### 6.2.5 Geospatial features

The geospatial measures were calculated at the Lower Layer Super Output Area (LSOA) level, as this is the finest spatial scale to which patient data was available. Each LSOA is of comparable population, with a minimum of 1,000 people and a mean of 1,500. As the full ACE dataset contained patients from 172 unique LSOAs, LSOA could not be included in any model as its own covariate. Moreover, increased hospitalisation associated with certain specific LSOAs would give little information on *why* those patients were at higher risk of hospitalisation. Thus, we calculated three main families of continuous geospatial features which allowed analysis of the factors influencing any geographic variation in hospitalisation risk.

#### 6.2.5.1 Air quality measures

The three main metrics of air quality used were nitrogen dioxide ($NO_2$), particles $< 10\mu$m ($PM_{10}$), and particles $< 2.5\mu$m ($PM_{2.5}$). These metrics were in the form of estimated mean concentrations (in $\mu g{\cdot}m^3$) obtained from DEFRA background data (see Section 6.2.1) on a $1{\times}1$ $km^2$ grid. For each postcode within an LSOA, we measured the nearest air quality estimate, then took the value for the LSOA to be the mean of all postcodes within the LSOA.

Summaries of the three measures are shown in Table A.12. Whilst none of the three measures showed a linear relationship with hospitalisation risk, $NO_2$ and $PM_{10}$ counted as 'variables of interest' after binarisation (Table A.18). As such, $NO_2 > 18.75$ $\mu g{\cdot}m^3$ and $PM_{10} > 13.5$ $\mu g{\cdot}m^3$ were carried forward to modelling.

#### 6.2.5.2 Indices of Multiple Deprivation

The data on indices of multiple deprivation were already provided at the LSOA level. They comprised 16 variables in total: 9 scores and 6 sub-domain scores, which are combined to calculate the total IMD score. The 9 main scores are: 'Barriers to housing and services', 'Crime', 'Education, skills and training', 'Employment', 'Living environment', 'Health deprivation and disability', 'Income deprivation affecting children', 'Income deprivation affecting older people', and 'Income'. The 6 sub-domain scores are: 'Adult skills', 'Children and young people', 'Geographical barriers', 'Indoors', 'Outdoors' and 'Wider barriers' (Table A.13).

Whilst these scores are also available as deciles, we kept them as continuous numerical scores in order to conserve model degrees of freedom. None of the scores showed a linear

relationship with hospitalisation risk, but three showed a relationship following binarisation - 'Children and young people' sub-domain score, 'Education, skills and training' score and 'Income deprivation affecting children' score (Table A.19).

#### 6.2.5.3 Distance measures

From the centroid of each LSOA, we measured the distance to each GP surgery, as well as the distance to the closest hospital (either St Luke's or the Bradford Royal Infirmary). We measured simply Euclidean distance; as we were measuring from LSOA centroids, rather than patient's postcodes or houses directly, there was little advantage to be gained from measuring actual travel distances or times.

The log-transformed distance from the LSOA in which a patient lived to their GP surgery showed a significant relationship with hospitalisation risk, and was thus included as a variable of interest (Table A.14). After binarisation, both standard and log-transformed distances to GP surgeries and to the nearest hospital were all included as variables of interest (Table A.20).

### 6.2.6 Binarisation of continuous variables

To allow for potential non-linearities in the relationship between continuous variables and hospitalisations, we performed an automated binarisation process to yield a split point, above which the variable would be classed as 'high', and below which it would be low. The best binarisation threshold for each variable was found by iterating the following process:

1. Select a candidate binarisation threshold.
2. Binarise the variable and fit a logistic regression model with the binary variable as the explanatory variable, and hospitalisation as the response variable.
3. Calculate a 95% confidence interval on the resulting coefficient
4. Record the difference between the upper and lower confidence limits, and whether the confidence interval overlaps with zero

The above process was followed for all possible binarisation thresholds. The best threshold was taken as the one that produces the narrowest confidence interval (i.e., smallest difference between lower and upper limit) where the limit does not include zero. If all tested confidence intervals overlapped with zero, the variable was not carried forward to modelling. Binarisation thresholds for all variables of interest are shown in the tables in Section A.5.

### 6.2.7 Outcome measure

As with previous chapters, our outcome measure is whether hospitalisation was required following acceptance onto the ACE service. Unlike in previous chapters, however, not all data from the ACE referral form dataset can be used, as only a subset of these patients (n = 446) had data on cBradford available for linkage. Of these, n = 78 (17.4%) were admitted to hospital.

### 6.2.8 Determining model structure

Two datasets, and thus two models, were created and assessed. The first, termed 'Original' comprised only variables extracted or engineered from the ACE referral form. The second, termed 'Additional', also contained the additional features detailed above.

The variables retained in each models (and thus, the structure of each model) was determined by least absolute shrinkage and selection operator (lasso) logistic regression, with hospitalisation as the response variable, and all other variables as explanatory variables. Lasso regression works by a shrinkage parameter $\lambda$ which penalises very small coefficient estimates, setting them to zero; the greater the value of $\lambda$, the fewer non-zero coefficients remain in the final model [19].

We profiled a range of potential $\lambda$ values and selected a final value which yielded a model with the best adherence to the '10 events per variable' rule-of-thumb [20]. This rule suggests that for a binary outcome, a prediction model should have one additional degree of freedom per 10 positive cases. With a total of 78 positive cases, this allows for a model with 8 degrees of freedom (equating to 8 variables, since all variables in both datasets are either binary or continuous, adding only one degree of freedom).

### 6.2.9 Bootstrap aggregation modelling

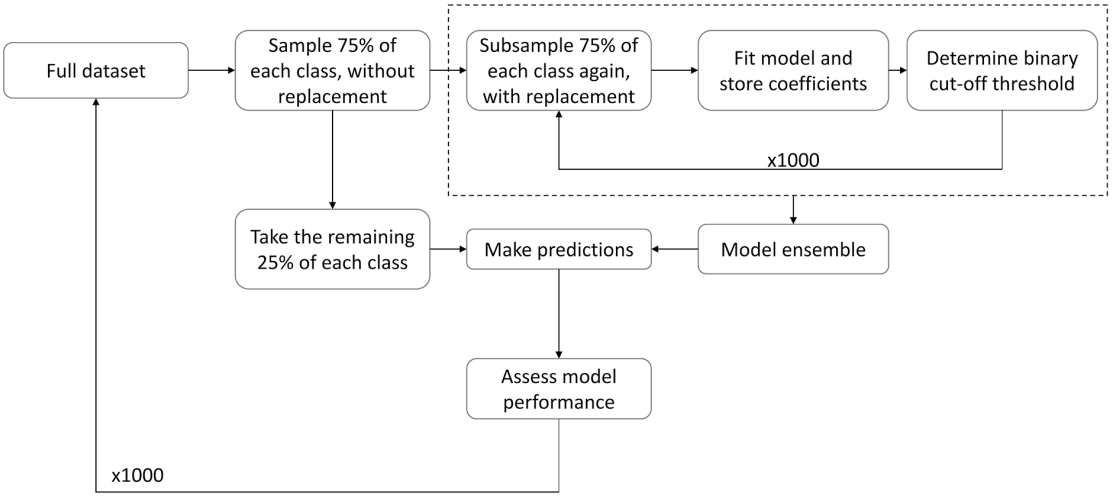



Figure 6.1: Flowchart detailing the iterative bootstrap aggregation modelling process. The dotted box represents the 'inner loop' for determining model coefficients. The remainder of the flowchart is the 'outer loop' for assessing model performance. This process was carried out separately on the 'Original' and 'Additional' datasets.

Modelling was carried out using a bootstrap aggregation process [21], in which an ensemble of models are trained on sub-samples of the full dataset. These are then used to generate predictions via a 'majority votes' system, where each model in the ensemble predicts the outcome for each individual, and the most-predicted class is taken as the final prediction. The full modelling process is detailed below, and schematically in Figure 6.1. This process was carried out twice; once for the 'Original' dataset and lasso-determined model structure, and once for the 'Additional' dataset and lasso-determined model structure.

In summary, the process consists of an 'inner' and 'outer' loop. The inner loop is iterated upon to prevent model overfitting, given the small absolute number of positive

cases (i.e, hospitalisations), and to produce robust estimates for each variable in the model. The outer loop is iterated upon in order to generate distributions of each metric of model predictive performance, to allow us to fully assess the extent to which these metrics differ between the two models.

#### 6.2.9.1 Outer loop: Assessing model performance

The outer loop begins by splitting the full dataset into a training set comprising 75% of positive cases and 75% of negative cases, and a validation set comprising the remaining 25%. Using the training set, an ensemble of logistic regression models are created (see Section 6.2.9.2), each with the structure determined by the lasso process (see Section 6.2.8), but with their own values for each coefficient and prediction threshold.

Hospitalisation is then predicted using each model in the ensemble for each individual patient in the validation set. Each model uses its own prediction threshold and 'votes' for the patient to be predicted as requiring hospitalisation or not. The majority case is then taken as that patient's predicted outcome.

With these predicted outcomes, 12 metrics of predictive performance were then calculated. These metrics are calculated from four rates: False positive ($F^+$), true positive ($T^+$), false negative ($F^-$), and true negative ($T^-$), shown in Figure 6.2.

| | **Observed** | |
|---|---|---|
| **Predicted** | Hospitalised | Discharged |
| Hospitalised | $T^+$ | $F^+$ |
| Discharged | $F^-$ | $T^-$ |

Figure 6.2: Schematic depiction of false and true positives and negatives. $T^+$ = true positive, $F^+$ = false positive, $T^-$ = true negative, $F^-$ = false negative

**Recall** (sometimes termed 'sensitivity') and **precision** are defined in Section 3.1.3, and both measure the ability of the model to predict positive cases correctly. Recall is calculated as

$$Recall = \frac{T^+}{T^+ + F^-} \tag{6.1}$$

and precision is calculated as

$$Precision = \frac{T^+}{T^+ + F^+} \tag{6.2}$$

Similarly, **specificity** is the proportion of negative predictions which are true negatives, calculated as

$$Specificity = \frac{T^-}{T^- + F^+} \tag{6.3}$$

**Accuracy** is calculated as the number of correct predictions ($T^+ + T^-$) divided by the total number of predictions made. In cases such as the ACE data, where positive cases are much rarer than negative cases, accuracy can be as high as the prevalence of the largest category (in this case, 82.6%) by a model that simply constantly guesses the

largest category. This can be remedied by **balanced accuracy**, which gives the average proportion of correct predictions in both classes, calculated as

$$Balanced\ accuracy = \frac{Recall + Specificity}{2} \tag{6.4}$$

Similar to balanced accuracy is **AUC**, which measures between 0.5 and 1 the ability of the model to discriminate between postive and negative cases (see Section 3.1.3). **Prevalence** gives the number of *predicted* positives as a proportion of total predictions, calculated as

$$Prevalence = \frac{T^+ + F^+}{T^+ + F^+ + T^- + F^-} \tag{6.5}$$

Closely related to detection prevalence is **detection rate**, the number of *true* positives as a proportion of total predictions.

Two metrics can quantify the ability of the model to predict negative cases and positive cases respectively. **Negative predictive value** (NPV) is the proportion of predicted negatives which are true negatives, calculated as

$$NPV = \frac{Recall \times (1 - Prevalence)}{((1 - Recall) \times Prevalence) + (Specificity \times (1 - Prevalence)} \tag{6.6}$$

Similarly, **Positive predictive value** (PPV) is the proportion of predicted positives which are true positives, calculated as

$$PPV = \frac{Recall \times Prevalence}{(Recall \times Prevalence) + ((1 - Specificity) \times (1 - Prevalence)} \tag{6.7}$$

Note that PPV should also be equal to recall, though its derivation is different. Finally, we calculated two combined metrics of model performance: **F1** represents the model's trade-off between precision and recall, explained in Section 3.1.3 and defined in Equation (3.3). **Cohen's kappa** ($\kappa$) measures the increase in performance of the model being assessed, compared to a theoretical model that guesses randomly, but following the frequencies of each outcome [22]. For a binary outcome, it is calculated as

$$\kappa = \frac{2 \times ((T^+ \times T^-) - (F^- \times F^+))}{((T^+ + F^+) \times (F^+ + T^-)) + ((T^+ + F^-) \times (F^- + T^-))} \tag{6.8}$$

and yields a value usually between 0 and 1, where greater values indicate increasing performance (though negative values are possible).

The outer loop is repeated 1,000 times to allow a distribution of values to be built for each model performance metric. This allows the assessment of differences in these metrics between the 'Original' and 'Additional' models.

#### 6.2.9.2 Inner loop: Determining model coefficients

The inner loop begins by receiving the training set from the outer loop (see Section 6.2.9.1). The training set is then sub-sampled to 75% of its full size *with replacement*, maintaining proportions of positive to negative cases. Using this sub-sampled dataset, a logistic regression model is fitted with the structure determined in Section 6.2.8, yielding an estimate for each model coefficient.

To determine the optimal threshold of the model for converting the log-odds of hospitalisation generated by the model into a binary predicted outcome, we profile a range

of potential thresholds. At each candidate threshold value, the model is used to predict the log-odds of hospitalisation for the same individuals used to train the model. These predictions are then binarised using the candidate threshold, and balanced accuracy is determined (see Equation (6.4)). The prediction threshold giving the greatest balanced accuracy is then recorded along with the model coefficient estimates.

This loop is carried out 1,000 times per training set, in order to produce a model ensemble which is robust to the specifics of the small number of positive cases upon which each individual model is trained.

Table 6.1: Prevalence and model coefficients for retained variables for the 'Original' and 'Additional' models. For binary variables, we present the number of patients for which the variable is true, and the percentage of patients in that group (discharged or hospitalised) which that represents. For continuous variables, we present the median value with lower and upper quartiles. Model coefficients are the result of the bootstrap aggregation process, and are presented as medians with lower and upper highest density intervals (HDI)

| Variable ('Original' model) | Discharged from ACE | Admitted to Hospital | Coefficient |
|---|---|---|---|
| Intercept | — | — | -2.61 [-2.96, -2.27] |
| Abnormal respiratory rate | 78 (21.2%) | 23 (29.5%) | 0.55 [0.2, 0.94] |
| Food allergy | 38 (10.3%) | 11 (14.1%) | 0.37 [-0.16, 0.79] |
| Heart rate | 120 [109, 130] | 120 [117, 132] | 0.19 [0.03, 0.41] |
| Moderate illness severity | 44 (12%) | 15 (19.2%) | 0.35 [-0.08, 0.8] |
| Mentions asthma | 53 (14.4%) | 20 (25.6%) | 0.92 [0.56, 1.38] |
| Mentions salbutamol | 59 (16%) | 19 (24.4%) | 0.36 [-0.01, 0.74] |
| Oxygen saturation | 97 [96, 98] | 96 [95, 97] | -0.32 [-0.49, -0.15] |
| Referral from GP | 199 (54.1%) | 55 (70.5%) | 0.72 [0.36, 1.09] |
| Variable ('Additional' model) | Discharged from ACE | Admitted to Hospital | Coefficient |
| Intercept | — | — | -2.99 [-3.36, -2.61] |
| Any eczema diagnosis | 160 (43.7%) | 49 (62.8%) | 0.77 [0.41, 1.11] |
| High local NO2 | 58 (16.1%) | 22 (28.2%) | 0.75 [0.34, 1.1] |
| Mentions asthma | 53 (14.4%) | 20 (25.6%) | 0.52 [0.07, 0.94] |
| Oxygen saturation | 97 [96, 98] | 96 [95, 97] | -0.42 [-0.57, -0.24] |
| Any pneumonia diagnosis | 40 (10.9%) | 19 (24.4%) | 1.05 [0.60, 1.46] |
| Referral from GP | 199 (54.1%) | 55 (70.5%) | 0.60 [0.27, 1.00] |
| Slow bronchodilators in last year | 78 (21.3%) | 29 (37.2%) | 0.37 [0.04, 0.77] |

## 6.3 Results

### 6.3.1 Retained variables

Our lasso logistic regression retained the following variables from the 'Original' dataset: Abnormal respiratory rate (i.e., not in the 'normal' APLS category), presence of food allergies, increasing heart rate, illness severity classified as 'moderate' upon examination, mention of 'asthma' in medical history, mention of 'salbutamol' in the examination notes, decreasing oxygen saturation, and if the referral was made by the patient's GP. Further

details on these variables can be found in the upper portion of Table 6.2, and a graphical representation of the resulting model coefficients can be seen in the upper panel Figure 6.3.

In the 'Additional' dataset, the mention of asthma, oxygen saturation and referral from GP were retained, as in the 'Original' dataset. Further to this, lasso logistic regression also retained the following new variables: Any diagnosis of eczema, any diagnosis of pneumonia, prescription of slow bronchodilators in the year prior to ACE acceptance, and the patient's home being in an area of high $NO_2$ air pollution ($>18.5$ $\mu g \cdot m^3$). Further details on these variables can be found in the lower portion of Table 6.2, and a graphical representation of the resulting model coefficients can be seen in the lower panel Figure 6.3. No variables from the demographic, prescriptions, distance or socio-economic deprivation groups were retained by the model.

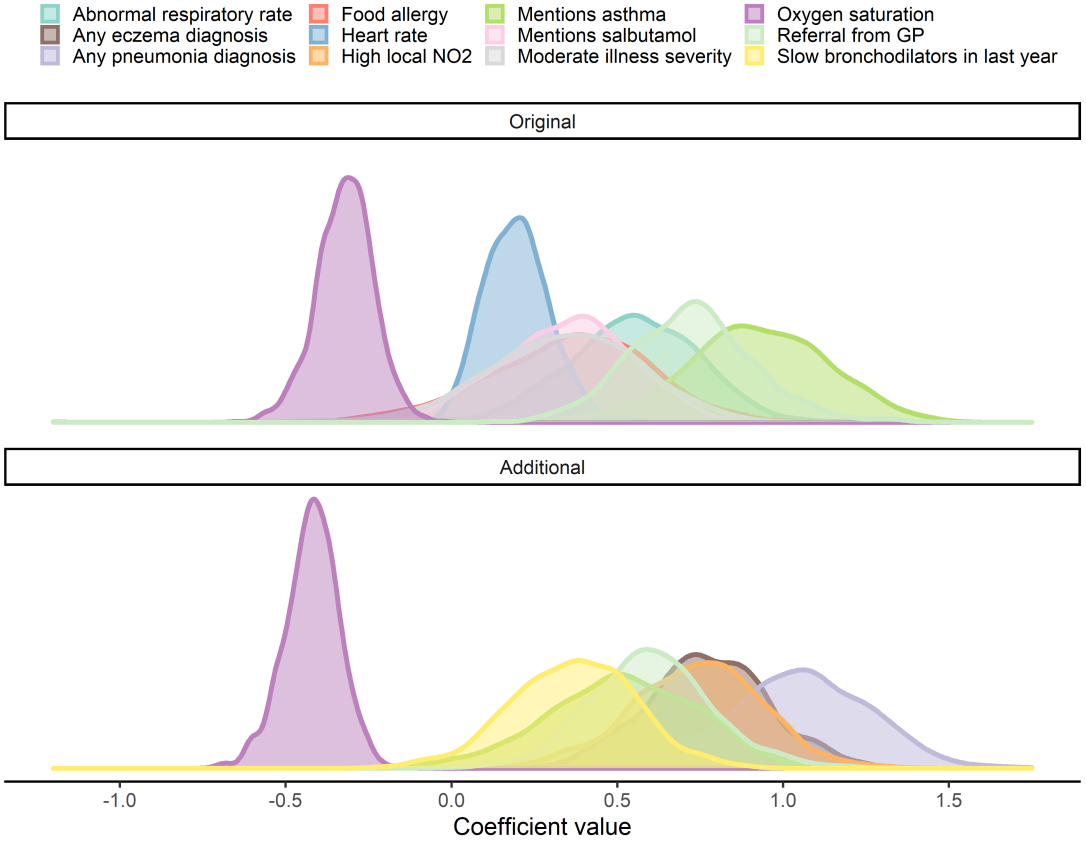



Figure 6.3: Kernal density plots of estimated coefficients for 'original' and 'additional' models. Features shared between the two models are coloured consistently between the two plots.

### 6.3.2 Model performance

In all but one measure of model performance, the 'Additional' model showed superior performance to the original model. Table 6.2 shows these differences in detail, and Figure 6.4 depicts these differences graphically. Prevalence, the number of predicted hospitalisations as a proportion of the total number of predictions made, did not differ between the two models, and nor would we expect it to. As the training data contained fixed and unbalanced proportions of hospitalisations to non-hospitalisations, it would be expected that

Table 6.2: Comparison of model performance metrics (see Section 6.2.9.1) between 'Original' and 'Additional' models. Metrics are presented as median values with upper and lower quartiles. 'Ratio' is the ratio of the median performance metric value of the 'Additional' model to the median performance metric value of the 'Original' model.

| Metric | 'Original' model | 'Additional' model | Ratio |
|---|---|---|---|
| Accuracy | 0.66 [0.63, 0.69] | 0.69 [0.66, 0.72] | 1.05 |
| AUC | 0.60 [0.56, 0.64] | 0.64 [0.60, 0.68] | 1.07 |
| Balanced Accuracy | 0.60 [0.56, 0.64] | 0.64 [0.60, 0.68] | 1.07 |
| Prevalence | 0.34 [0.31, 0.38] | 0.33 [0.30, 0.37] | 0.97 |
| Detection Rate | 0.09 [0.08, 0.11] | 0.10 [0.09, 0.12] | 1.11 |
| F1 | 0.34 [0.30, 0.39] | 0.39 [0.35, 0.44] | 1.15 |
| Kappa | 0.14 [0.09, 0.20] | 0.21 [0.15, 0.26] | 1.50 |
| Negative Predictive Value | 0.87 [0.85, 0.88] | 0.88 [0.86, 0.90] | 1.01 |
| Positive Predictive Value | 0.26 [0.23, 0.29] | 0.30 [0.27, 0.33] | 1.15 |
| Precision | 0.26 [0.23, 0.29] | 0.3 [0.27, 0.33] | 1.15 |
| Recall | 0.50 [0.45, 0.60] | 0.55 [0.50, 0.65] | 1.10 |
| Specificity | 0.70 [0.65, 0.73] | 0.72 [0.68, 0.75] | 1.03 |

prevalence be equal in the two models.

Our results also show that this increase in performance comes from the increased ability of the 'Additional' model to predict hospitalisations, as opposed to increased ability to detect patients who will not be hospitalised. We see, for example, a 1.15-fold increase in the positive predictive value (from 0.26 [0.23, 0.29] in the 'Original' model to 0.30 [0.27, 0.33] in the 'Additional' model), but only a 1.01-fold increase in the negative predictive value (from 0.87 [0.85, 0.88] in the 'Original' model to 0.88 [0.86, 0.90] in the 'Additional' model). This can also be seen in the 1.10-fold increase in recall (from 0.50 [0.45, 0.60] in the 'Original' model to 0.55 [0.50, 0.65] in the 'Additional' model), with again only a 1.03-fold increase in specificity (from 0.70 [0.65, 0.73] in the 'Original' model to 0.72 [0.68, 0.75] in the 'Additional' model).

Whilst the 'Additional' model does out-perform the 'Original' model, we must still consider its objective performance as a classifier. A median Cohen's kappa of 0.21 [0.15, 0.26] denotes only 'fair' performance [22], and similarly, a median AUC of 0.64 [0.60, 0.68] is generally considered to denote 'moderate' predictive performance.

## 6.4 Discussion

As with the results in presented in earlier chapters, our results as presented here do not allow the development of a functional model to predict hospitalisation for patients referred to the ACE service. Whilst the models containing additional variables do perform better than those trained on just the data from the ACE referral form, their objective performance is still, at best, mediocre. The insights from these models, however, do allow us to make some recommendations and considerations, both for improvements to the ACE admissions pathway, and for future study in this area.

### 6.4.1 Considerations for ACE admissions

Of the many additional variables that correlated with increased hospitalisation risk, only a small number were retained in the final model. Whilst this does not mean necessarily that

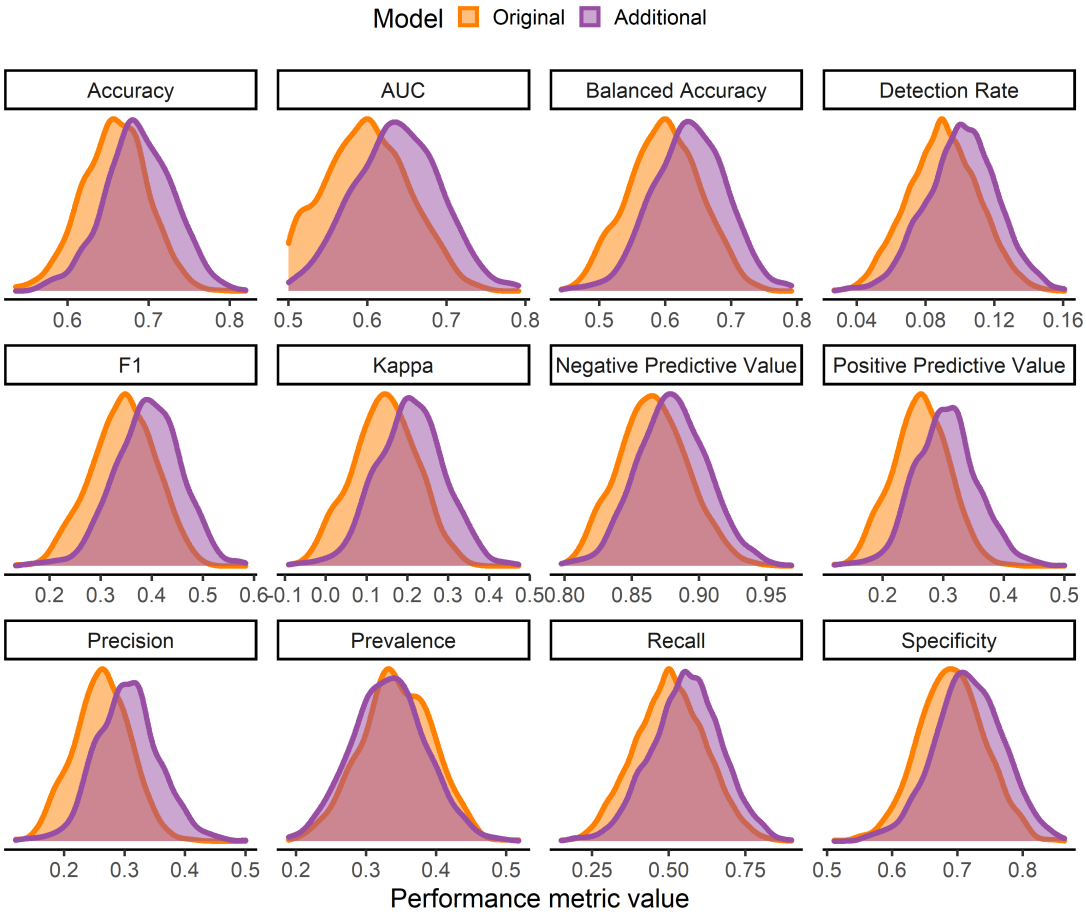



Figure 6.4: Kernal density plots comparing model performance metrics (see Section 6.2.9.1) between 'original' and 'additional' models.

non-retained variables were spurious correlations, it *does* imply the importance of those that were retained. In terms of medical history, past diagnoses of pneumonia, eczema, and food allergies were retained in the models, along with prescription of long-acting bronchodilators. This information would or could likely be known either to ACE admissions staff, or to parents of patients, at the point of referral. Therefore, we recommend that investigation of these co-morbidities and medical histories be investigated further for their association with ACE referral and subsequent hospitalisation, and that the ACE team consider overtly asking about eczema, pneumonia and bronchodilator prescription when assessing a patient for acceptance onto the service.

Similarly, in both models, patients referred to ACE from their GP were at greater risk of hospitalisation. Whilst it is unclear from the data what the causal pathway is here, this finding suggests that the origin of a patient's referral should also be considered when assessing prospective ACE patients.

### 6.4.2 Considerations for future study

This analysis has highlighted a number of fruitful avenues for future research to investigate more thoroughly. First, the mention of 'asthma' in the examination notes was not only retained it both models, but was also retained in the 'Additional' model despite the

presence of variables containing actual information on asthma diagnosis. This suggests that the mention of asthma in the notes is more important than whether or not the patient actually has asthma. This could be, for example, because notes might only tend to mention a patient's asthma if it is relevant, problematic or not under control. Additional analysis of the examination notes such as a sentiment analysis may be able to indicate where mentions of asthma are linked with language indicating that the patient's asthma is potentially problematic, thus improving its predictive power.

Second, the 'Additional' model identified patients from areas of high background air pollution ($NO_2$) as being at higher risk of hospitalisation. Whilst a link between air pollution and hospitalisation for asthma-like respiratory conditions makes logical sense, our results to not permit us to interpret this as a causal relationship, for several reasons. First, our air quality data are estimated, rather than measured. Whilst measured data are available, there are few recording stations in the Bradford area, so estimated local concentrations were determined to be more accurate than concentrations measured further away, as air pollution levels tend to be geographically heterogeneous. Second, our analysis was restricted to the LSOA levels. This comes with a number of assumptions, including the assumption that air pollution is heterogeneous across a given LSOA, and the assumption that the pollution levels in a given LSOA are in fact the levels that children are exposed to. Further study could therefore account for how much children move between neighbouring LSOAs, especially, for example, for school. Finally, the 'High $NO_2$' feature could simply be capturing any number of other geospatial variables, such as socio-economic deprivation, local access metrics, or the quality of local healthcare services. Subsequent work should therefore focus on demonstrating a direct causal link between exposure to increased air pollution, and a greater likelihood of being hospitalised with asthma or related respiratory conditions.

Finally, our work identified a number of variables across prescriptions, healthcare visits, accessibility measures and socio-economic deprivation which correlated with increased hospitalisation risk in a univariate model, but were not retained by final models. One major reason for this is likely due to the number of variables being retained by the model being highly constrained by the low number of positive cases, even with mitigation strategies against overfitting. Future work in this area with a greater number of cases could more thoroughly investigate relationships between the additional 'variables of interest' highlighted here, and the risk of hospitalisation.

# A | Appendices

## A.1 The Data

Table A.1: Hospitalisation frequency and number of examples for each of the original categorical features in the ACE dataset. **The sample mean proportion of hospitalised patients is 0.1614**. Note that very few of the features have a proportion of hospitalised patients significantly above/below the sample mean, and those that do are supported by very few observations.

| Feature | Values | P(Hospital Required) | Total Examples |
|---|---|---|---|
| **Referral From** | **CCDA** | 0.067 | 45 |
| | **ED** | 0.172 | 87 |
| | **GP** | 0.177 | 203 |
| **Referral Profession** | **ANP** | 0.159 | 82 |
| | **Consultant** | 0.086 | 35 |
| | **Doctor** | 0.172 | 157 |
| | **Registar** | 0.177 | 62 |
| **Gender** | **Female** | 0.155 | 129 |
| | **Male** | 0.164 | 207 |
| **Referral Date** | **Autumn** | 0.125 | 112 |
| | **Spring** | 0.154 | 52 |
| | **Summer** | 0.169 | 59 |
| | **Winter** | 0.195 | 113 |
| **Referral Time** | **Afternoon** | 0.161 | 137 |
| | **Evening** | 0.4 | 15 |
| | **Morning** | 0.141 | 184 |
| **Illness Severity** | **Mild** | 0.155 | 290 |
| | **Moderate** | 0.205 | 44 |
| **Activity Level** | **Lower** | 0.188 | 112 |
| | **Usual** | 0.145 | 221 |
| **Gut Feeling** | **Low Concern** | 0.149 | 188 |
| | **Unwell** | 0.667 | 3 |
| | **Well** | 0.162 | 142 |
| **Sepsis** | **Low Level** | 0.158 | 19 |
| | **None Noted** | 0.161 | 317 |
| **Safeguarding** | **No** | 0.167 | 288 |
| | **Yes** | 0.125 | 48 |
| **Food Allergy** | **No** | 0.159 | 290 |
| | **Yes** | 0.174 | 46 |
| **Drug Allergy** | **No** | 0.163 | 301 |
| | **Yes** | 0.143 | 35 |
| **Other Allergy** | **No** | 0.167 | 305 |
| | **Yes** | 0.097 | 31 |

| | | | |
|---|---|---|---|
| **Group Ethnicity** | **Asian** | 0.174 | 184 |
| | **European** | 0.158 | 120 |
| | **Other** | 0.094 | 32 |

Table A.2: Chi$^2$ significance statistics testing the relationship between each of the original categorical features from the ACE referral data and hospitalisation outcomes. Results are ordered by p-value, lowest to highest - this can be interpreted as "most significant" to "least significant".

| | **Chi$^2$** | **p** | **dof** |
|---|---|---|---|
| **Referral Time** | 6.881 | 0.032 | 2 |
| **Gut Feeling** | 5.929 | 0.052 | 2 |
| **Referral From** | 3.446 | 0.179 | 2 |
| **Activity Level** | 0.719 | 0.397 | 1 |
| **Other Allergy** | 0.579 | 0.447 | 1 |
| **Group Ethnicity** | 1.307 | 0.52 | 2 |
| **Illness Severity** | 0.371 | 0.542 | 1 |
| **Referral Date** | 2.078 | 0.556 | 3 |
| **Safeguarding** | 0.266 | 0.606 | 1 |
| **Referral Profession** | 1.738 | 0.628 | 3 |
| **Sepsis** | 0.082 | 0.774 | 1 |
| **Gender** | 0.005 | 0.943 | 1 |
| **Drug Allergy** | 0.004 | 0.952 | 1 |
| **Food Allergy** | 0.002 | 0.963 | 1 |

Figure A.1: Box plots (left) and stacked KDE plots (right) for each of the numerical features, grouped by examples that required hospital treatment and those that were successfully discharged from ACE. The stacked KDE plots maintain the original proportions of patients that were referred to hospital / successfully discharged from ACE and serve as a good indicator of the proportions of examples that represent the numeric features at different values.

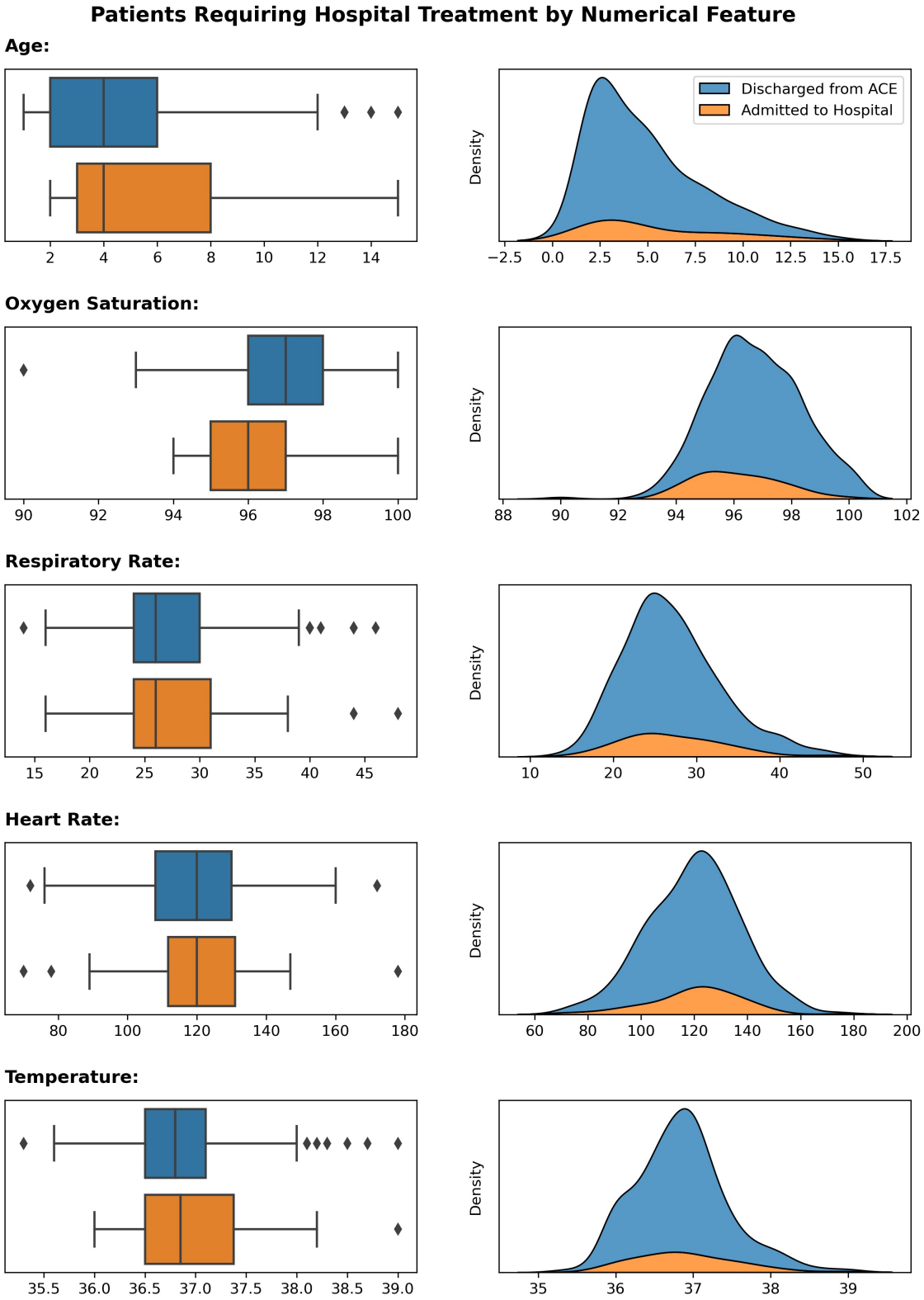


Table A.3: Pearson's R statistics for the numeric/continuous features in the ACE referral data. The high statistical significance of the Oxygen Saturation feature can be seen in Figure A.1. The proportions of patients hospitalised/discharged change significantly at either tail of the distribution - the lower tail has a much greater proportion of patients hospitalised and the opposite for the higher tail. It should be noted that the bulk of the examples lie in the center of the distribution and don't exhibit high or low oxygen saturations.

| | r | p |
|---|---|---|
| **Oxygen Saturation** | -0.164 | 0.003 |
| **Age** | 0.081 | 0.141 |
| **Temperature** | 0.048 | 0.409 |
| **Respiratory Rate** | 0.028 | 0.617 |
| **Heart Rate** | 0.017 | 0.752 |

Table A.4: Hospitalisation Frequency and number of examples for each of the features engineered using the ACE referral criteria and APLS observation guidelines

| Feature | Values | P(Hospital Required) | Total Examples |
|---|---|---|---|
| **APLS Resp Rate** | **High** | 0.207 | 82 |
| | **Low** | 0.0 | 4 |
| | **Normal** | 0.148 | 250 |
| **APLS Heart Rate** | **High** | 0.242 | 33 |
| | **Low** | 0.5 | 2 |
| | **Normal** | 0.15 | 301 |
| **Ox Sat Low** | **No** | 0.162 | 333 |
| | **Yes** | 0.0 | 3 |
| **Age Range** | **Pre School** | 0.161 | 180 |
| | **Primary** | 0.141 | 142 |
| | **Secondary** | 0.357 | 14 |
| **ACE Heart Rate** | **high** | 0.154 | 65 |
| | **Low** | 0.286 | 7 |
| | **Normal** | 0.159 | 264 |
| **ACE Resp Rate** | **high** | 0.17 | 112 |
| | **Low** | 0.175 | 40 |
| | **Normal** | 0.152 | 184 |
| **Meets ACE Criteria** | **No** | 0.17 | 182 |
| | **Yes** | 0.149 | 154 |

Table A.5: $Chi^2$ statistics for the engineered features using the ACE referral criteria and APLS guidelines

| | chi2 | p | dof |
|---|---|---|---|
| **Age Range** | 4.421 | 0.11 | 2 |
| **APLS Heart Rate** | 3.621 | 0.164 | 2 |
| **APLS Resp Rate** | 2.386 | 0.303 | 2 |
| **ACE Heart Rate** | 0.839 | 0.657 | 2 |
| **Meets ACE Criteria** | 0.139 | 0.709 | 1 |
| **ACE Resp Rate** | 0.226 | 0.893 | 2 |
| **Ox Sat Low** | 0.001 | 0.978 | 1 |

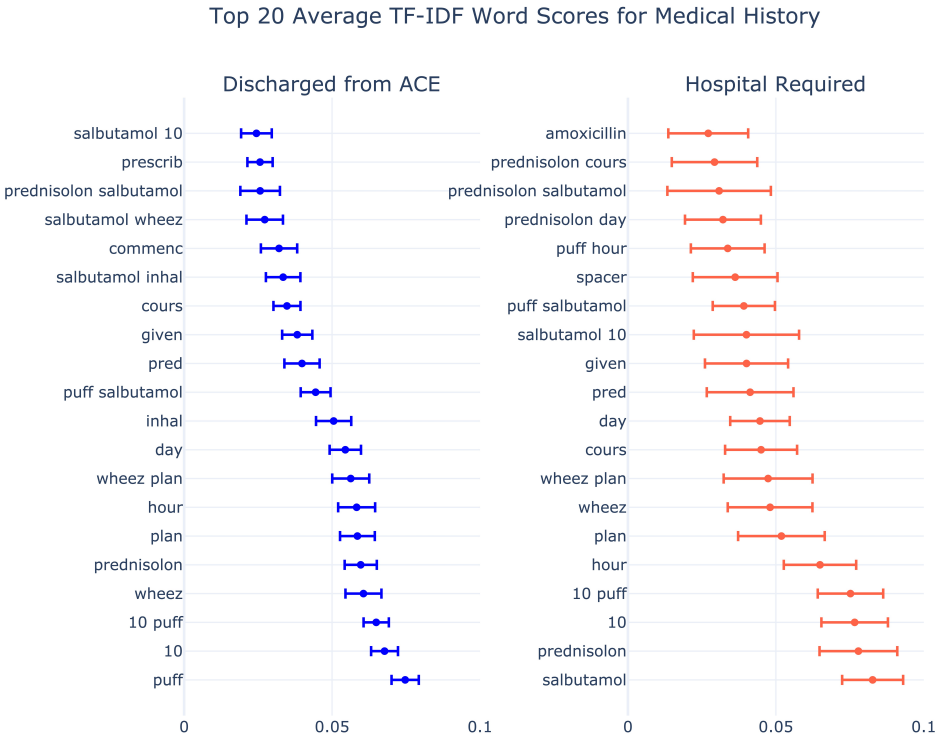
## A.2 Classification Modelling

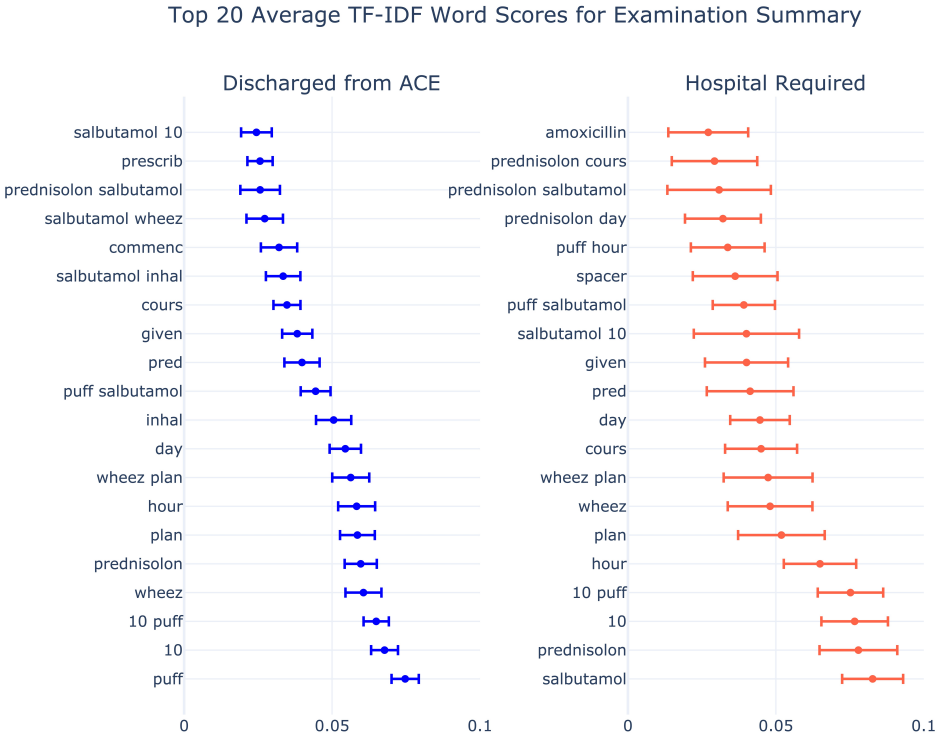
| | $F_1$ | AUC | Accuracy | Recall | Precision |
|---|---|---|---|---|---|
| **One Hot - Weighted Labels** | | | | | |
| **K Nearest Neighbours** | 0.154 (0.071) | 0.502 (0.038) | 0.742 (0.027) | 0.144 (0.073) | 0.169 (0.074) |
| **Support Vector Machines** | 0.269 (0.056) | 0.535 (0.061) | 0.576 (0.082) | 0.474 (0.140) | 0.191 (0.041) |
| **Random Forest Classifier** | 0.209 (0.066) | 0.530 (0.035) | 0.745 (0.033) | 0.209 (0.079) | 0.218 (0.065) |
| **Gradient Boosting Classifier** | 0.153 (0.092) | 0.519 (0.039) | 0.788 (0.029) | 0.119 (0.073) | 0.233 (0.162) |
| **Ada Boost classifier** | 0.173 (0.081) | 0.518 (0.041) | 0.762 (0.034) | 0.154 (0.078) | 0.212 (0.109) |
| **Gaussian Naieve Bayes** | 0.238 (0.055) | 0.521 (0.047) | 0.645 (0.066) | 0.335 (0.096) | 0.189 (0.048) |
| **Logistic Regression** | 0.265 (0.049) | 0.534 (0.050) | 0.596 (0.060) | 0.441 (0.097) | 0.191 (0.037) |
| **Quadratic Discriminant Analysis** | 0.036 (0.048) | 0.497 (0.016) | 0.815 (0.016) | 0.022 (0.031) | 0.106 (0.151) |
| **One Hot - SMOTE** | | | | | |
| **K Nearest Neighbours** | 0.248 (0.052) | 0.512 (0.057) | 0.557 (0.047) | 0.444 (0.111) | 0.173 (0.035) |
| **Support Vector Machines** | 0.292 (0.048) | 0.552 (0.054) | 0.555 (0.071) | 0.546 (0.076) | 0.200 (0.040) |
| **Random Forest Classifier** | 0.267 (0.050) | 0.555 (0.035) | 0.717 (0.039) | 0.313 (0.072) | 0.238 (0.051) |
| **Gradient Boosting Classifier** | 0.255 (0.063) | 0.547 (0.043) | 0.717 (0.041) | 0.293 (0.076) | 0.231 (0.064) |
| **Ada Boost classifier** | 0.278 (0.059) | 0.537 (0.061) | 0.539 (0.093) | 0.535 (0.130) | 0.192 (0.049) |
| **Gaussian Naieve Bayes** | 0.196 (0.060) | 0.465 (0.051) | 0.540 (0.059) | 0.352 (0.141) | 0.138 (0.038) |
| **Logistic Regression** | 0.285 (0.066) | 0.556 (0.062) | 0.634 (0.055) | 0.441 (0.108) | 0.213 (0.051) |
| **Quadratic Discriminant Analysis** | 0.256 (0.058) | 0.519 (0.063) | 0.561 (0.061) | 0.456 (0.112) | 0.179 (0.042) |
| **One Hot - Undersampling** | | | | | |
| **K Nearest Neighbours** | 0.246 (0.063) | 0.505 (0.071) | 0.532 (0.059) | 0.463 (0.126) | 0.169 (0.043) |
| **Support Vector Machines** | 0.276 (0.046) | 0.535 (0.056) | 0.528 (0.084) | 0.546 (0.138) | 0.187 (0.032) |
| **Random Forest Classifier** | 0.275 (0.047) | 0.536 (0.054) | 0.544 (0.055) | 0.524 (0.106) | 0.187 (0.032) |
| **Gradient Boosting Classifier** | 0.262 (0.045) | 0.520 (0.050) | 0.530 (0.066) | 0.506 (0.111) | 0.178 (0.033) |
| **Ada Boost classifier** | 0.267 (0.050) | 0.528 (0.055) | 0.540 (0.054) | 0.511 (0.126) | 0.182 (0.034) |
| **Gaussian Naieve Bayes** | 0.269 (0.048) | 0.534 (0.053) | 0.561 (0.069) | 0.493 (0.121) | 0.187 (0.031) |
| **Logistic Regression** | 0.267 (0.051) | 0.525 (0.062) | 0.531 (0.062) | 0.517 (0.106) | 0.181 (0.035) |
| **Quadratic Discriminant Analysis** | 0.269 (0.067) | 0.529 (0.070) | 0.557 (0.083) | 0.489 (0.125) | 0.189 (0.054) |

| | $F_1$ | AUC | Accuracy | Recall | Precision |
|---|---|---|---|---|---|
| **Mean Target - Weighted Labels** | | | | | |
| **K Nearest Neighbours** | 0.108 (0.069) | 0.475 (0.036) | 0.723 (0.033) | 0.104 (0.068) | 0.115 (0.076) |
| **Support Vector Machines** | 0.279 (0.039) | 0.545 (0.039) | 0.570 (0.042) | 0.509 (0.107) | 0.193 (0.024) |
| **Random Forest Classifier** | 0.193 (0.075) | 0.512 (0.045) | 0.715 (0.041) | 0.209 (0.088) | 0.186 (0.081) |
| **Gradient Boosting Classifier** | 0.135 (0.097) | 0.510 (0.044) | 0.780 (0.028) | 0.106 (0.079) | 0.195 (0.130) |
| **Ada Boost classifier** | 0.161 (0.076) | 0.512 (0.038) | 0.761 (0.031) | 0.141 (0.071) | 0.197 (0.099) |
| **Gaussian Naieve Bayes** | 0.205 (0.070) | 0.515 (0.036) | 0.684 (0.093) | 0.263 (0.140) | 0.181 (0.049) |
| **Logistic Regression** | 0.285 (0.042) | 0.549 (0.046) | 0.569 (0.049) | 0.519 (0.090) | 0.197 (0.030) |
| **Quadratic Discriminant Analysis** | 0.148 (0.086) | 0.440 (0.058) | 0.527 (0.172) | 0.311 (0.226) | 0.113 (0.066) |
| **Mean Target - SMOTE** | | | | | |
| **K Nearest Neighbours** | 0.209 (0.062) | 0.474 (0.064) | 0.547 (0.051) | 0.365 (0.124) | 0.147 (0.043) |
| **Support Vector Machines** | 0.279 (0.034) | 0.539 (0.040) | 0.529 (0.054) | 0.554 (0.102) | 0.188 (0.023) |
| **Random Forest Classifier** | 0.235 (0.072) | 0.534 (0.047) | 0.708 (0.039) | 0.274 (0.092) | 0.209 (0.063) |
| **Gradient Boosting Classifier** | 0.242 (0.072) | 0.540 (0.047) | 0.715 (0.042) | 0.280 (0.100) | 0.219 (0.063) |
| **Ada Boost classifier** | 0.274 (0.059) | 0.531 (0.065) | 0.538 (0.093) | 0.520 (0.125) | 0.190 (0.050) |
| **Gaussian Naieve Bayes** | 0.271 (0.041) | 0.532 (0.046) | 0.546 (0.050) | 0.511 (0.089) | 0.185 (0.029) |
| **Logistic Regression** | 0.271 (0.046) | 0.539 (0.044) | 0.591 (0.049) | 0.463 (0.094) | 0.193 (0.032) |
| **Quadratic Discriminant Analysis** | 0.275 (0.036) | 0.536 (0.038) | 0.535 (0.067) | 0.537 (0.105) | 0.187 (0.024) |
| **Mean Target - Undersampling** | | | | | |
| **K Nearest Neighbours** | 0.240 (0.059) | 0.504 (0.064) | 0.561 (0.061) | 0.420 (0.115) | 0.169 (0.041) |
| **Support Vector Machines** | 0.266 (0.054) | 0.524 (0.061) | 0.528 (0.065) | 0.519 (0.120) | 0.180 (0.038) |
| **Random Forest Classifier** | 0.265 (0.045) | 0.525 (0.049) | 0.524 (0.068) | 0.526 (0.132) | 0.179 (0.029) |
| **Gradient Boosting Classifier** | 0.258 (0.045) | 0.515 (0.053) | 0.525 (0.056) | 0.500 (0.102) | 0.175 (0.032) |
| **Ada Boost classifier** | 0.264 (0.042) | 0.522 (0.049) | 0.531 (0.056) | 0.509 (0.105) | 0.179 (0.031) |
| **Gaussian Naieve Bayes** | 0.254 (0.056) | 0.520 (0.052) | 0.562 (0.068) | 0.457 (0.126) | 0.178 (0.040) |
| **Logistic Regression** | 0.258 (0.045) | 0.512 (0.052) | 0.504 (0.055) | 0.524 (0.109) | 0.172 (0.029) |
| **Quadratic Discriminant Analysis** | 0.252 (0.073) | 0.519 (0.073) | 0.568 (0.069) | 0.446 (0.153) | 0.177 (0.050) |

Table A.6: Cross-Validation results for each combination of data preparation/label weighting/modelling technique tested during the classification modelling. Figures in brackets are the standard deviations for the relevant statistic.

## A.3 Free Text Analysis



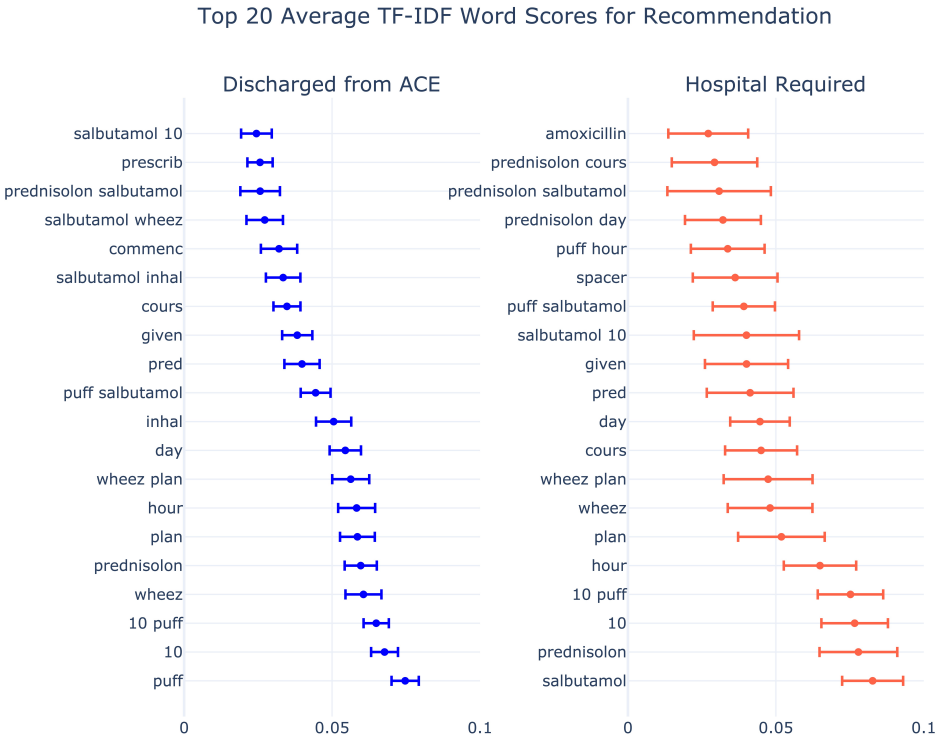
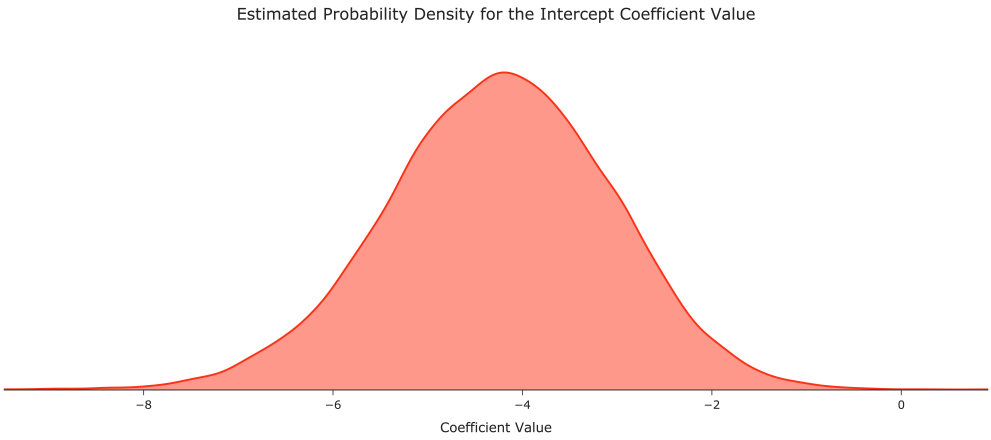


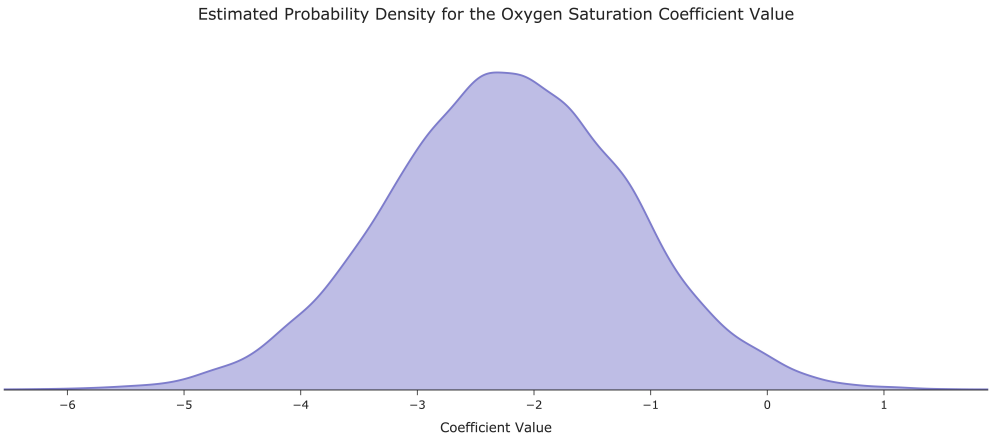

Figure A.2: 20 highest TF-IDF scores for patients successfully treated by ACE or later referred to hospital. Error bars are the standard errors for each value and hence are much larger for the values taken from patients that required hospital treatment, given that these examples are much fewer in number

## A.4 Bayesian Analysis

| | ELPD | HPD Lower | HPD Higher |
|---|---|---|---|
| **Mentions Salbutamol** | -143.704 | 0.336 | 1.630 |
| **Oxygen Saturation** | -145.465 | -4.220 | -0.442 |
| **Referral From GP** | -146.071 | 0.058 | 1.368 |
| **Referral Profession Doctor** | -146.380 | 0.031 | 1.252 |
| **Mentions Asthma** | -146.393 | 0.014 | 1.337 |
| **Referral From CCDA** | -146.648 | -2.605 | 0.002 |
| **Gut Feeling Well** | -146.774 | -1.263 | 0.055 |
| **APLS Resp Rate High** | -146.953 | -0.072 | 1.220 |
| **Illness Severity Moderate** | -147.021 | -0.107 | 1.399 |
| **Activity Level Usual** | -147.066 | -1.109 | 0.118 |
| **Food Allergy** | -147.433 | -0.235 | 1.383 |
| **APLS Heart Rate High** | -147.442 | -0.252 | 1.454 |
| **Referral Date Winter** | -147.451 | -0.176 | 1.048 |
| **APLS Resp Rate Normal** | -147.456 | -1.100 | 0.195 |
| **Age Range Secondary** | -147.483 | -0.369 | 1.971 |
| **Referral Date Spring** | -147.483 | -1.633 | 0.233 |
| **Resp Rate** | -147.497 | -0.528 | 3.063 |
| **Temperature** | -147.512 | -0.702 | 3.216 |
| **Gut Feeling Low Concern** | -147.520 | -0.194 | 1.066 |
| **Referral Time Evening** | -147.527 | -0.436 | 2.193 |
| **Age Range Primary** | -147.534 | -1.090 | 0.189 |
| **Referral Profession Consultant** | -147.586 | -2.281 | 0.329 |
| **APLS Heart Rate Normal** | -147.644 | -1.364 | 0.320 |
| **Heart Rate** | -147.718 | -0.728 | 2.983 |
| **Gut Feeling Unwell** | -147.735 | -0.245 | 5.856 |
| **Referral Profession ANP** | -147.843 | -1.207 | 0.297 |
| **ACE Resp Rate High** | -147.852 | -0.269 | 0.934 |
| **Referral From ED** | -147.993 | -1.103 | 0.336 |
| **Group Ethnicity Other** | -148.028 | -2.223 | 0.454 |
| **Referral Date Autumn** | -148.058 | -0.989 | 0.329 |
| **ACE Resp Rate Normal** | -148.074 | -0.847 | 0.346 |
| **Other Allergy** | -148.233 | -0.651 | 1.197 |
| **Meets ACE Criteria** | -148.250 | -0.806 | 0.394 |
| **Group Ethnicity Asian** | -148.277 | -0.436 | 0.778 |
| **Age** | -148.277 | -0.904 | 1.718 |
| **Age Range Pre School** | -148.284 | -0.420 | 0.806 |
| **Referral Date Summer** | -148.286 | -0.553 | 1.033 |
| **ACE Heart Rate Normal** | -148.295 | -0.528 | 1.018 |
| **Drug Allergy** | -148.300 | -2.017 | 0.642 |
| **ACE Heart Rate High** | -148.310 | -1.063 | 0.531 |
| **Referral Time Afternoon** | -148.342 | -0.728 | 0.482 |
| **Referral Profession Registar** | -148.383 | -0.977 | 0.614 |
| **Sepsis None Noted** | -148.391 | -1.359 | 1.056 |
| **Referral Time Morning** | -148.399 | -0.662 | 0.543 |
| **Gender Male** | -148.403 | -0.524 | 0.716 |
| **Group Ethnicity European** | -148.476 | -0.610 | 0.640 |
| **ACE Resp Rate Low** | -148.490 | -1.173 | 0.744 |
| **Safeguarding Issues** | -148.492 | -0.989 | 0.817 |
| **ACE Heart Rate Low** | -149.909 | -3.316 | 2.049 |

Table A.7: Estimated log pointwise predictive densities and upper/lower 95% Highest Posterior Density coefficient intervals for each of the bayesian logistic regression models trained using an individual feature



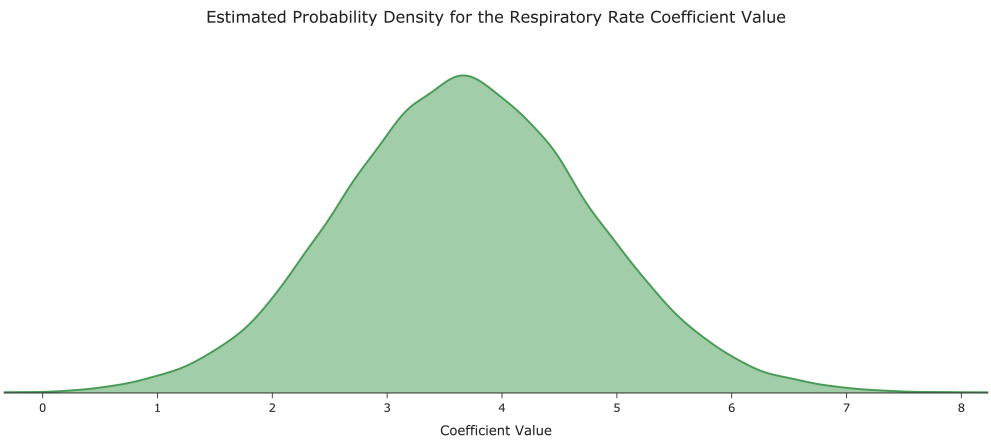




Figure A.3: Estimated probability densities for the coeficient values of the intercept and numerical features as taken from the best Bayesian logistic regression model. Results correspond with the HPD values in Figure 5.3

## A.5 Additional features

Table A.8: Occurrence and effect on hospitalisation of demographic variables . The 'coefficient' and '$p$' columns are the output of binomial GLM with hospitalisation as the response variable, and the variable as the sole explanatory variable. Hospitalisation % refers to the proportion of hospitalised cases that were 'true' for each variable . As such, if this percentage is greater than the occurrence of the condition, then that demographic characteristic is linked to a disproportionately high rate of hospitalisation.

| Variable | Occurrence | Hospitalisation (%) | Coefficient | $p$ |
|---|---|---|---|---|
| Male Gender | 267 (61.2%) | 51 (65.4%) | 0.217 | 0.407 |
| Ethnic group | | | | |
| Pakistani | 205 (47%) | 37 (47.4%) | 0.020 | 0.935 |
| White | 137 (31.4%) | 25 (32.1%) | 0.035 | 0.895 |
| Other | 94 (21.6%) | 16 (20.5%) | -0.076 | 0.804 |

Table A.9: List of co-morbidities extracted from Connected Bradford. GORD = Gastro-oesophageal reflux disease.

| Condition | Occurrence |
|---|---|
| Asthma | 217 (48.7%) |
| Eczema | 209 (46.9%) |
| Bronchitis | 74 (16.6%) |
| Pneumonia | 59 (13.2%) |
| Common cold | 46 (10.3%) |
| Rhinitis | 43 (9.6%) |
| Croup | 22 (4.9%) |
| GORD | 22 (4.9%) |
| Total | 446 |

Table A.10: List of prescription data extracted from Connected Bradford

| Drug category | Occurrence |
|---|---|
| Fast bronchodilators | 348 (77.9%) |
| Prednisolone | 232 (51.9%) |
| Slow bronchodilators | 141 (31.5%) |
| Antihistamines | 138 (30.9%) |

Table A.11: List of visit data extracted from Connected Bradford.

| Visit type | $n$ (%) |
|---|---|
| Family Practice Visits | 20,558 (81.7%) |
| Emergency Room Visit | 2,767 (11.0%) |
| Inpatient Visit | 1,671 (6.6%) |
| Outpatient Visit | 159 (0.6%) |

Table A.12: Background air pollution concentration estimates grouped by admission to hospital. Concentrations are presented in $\mu g \cdot m^3$, and reported as medians with upper and lower quartiles.

| Variable | Discharged from ACE | Admitted to Hospital |
|---|---|---|
| $NO_2$ | 15.99 [14.13, 17.97] | 16.65 [14.48, 19.51] |
| $PM_{10}$ | 13.16 [12.11, 13.84] | 13.25 [12.38, 13.86] |
| $PM_{2.5}$ | 9.1 [8.31, 9.63] | 9.15 [8.57, 9.64] |

Table A.13: Indices of multiple deprivation (IMD) grouped by admission to hospital.

| Variable | Description | Discharged from ACE | Admitted to Hospital |
|---|---|---|---|
| BHSScore | Barriers to housing and services | 17.87 [14.42, 21.37] | 17.02 [14.67, 21.46] |
| CriScore | Crime | 1.23 [0.89, 1.66] | 1.35 [0.79, 1.72] |
| EduScore | Education, skills and training | 50.46 [28.84, 64.01] | 56.51 [33.97, 65.63] |
| EmpScore | Employment | 0.17 [0.12, 0.2] | 0.18 [0.13, 0.21] |
| EnvScore | Living environment | 38.49 [24.42, 52.43] | 38.37 [24.72, 53.73] |
| HDDScore | Health deprivation and disability | 0.82 [0.5, 1.13] | 0.93 [0.59, 1.13] |
| IDCScore | Income deprivation affecting children | 0.26 [0.19, 0.32] | 0.29 [0.21, 0.36] |
| IDOScore | Income deprivation affecting older people | 0.32 [0.19, 0.5] | 0.33 [0.21, 0.52] |
| IMDScore | Index of multiple deprivation | 43.77 [31.91, 54.02] | 46.67 [32.58, 55.44] |
| IncScore | Income | 0.26 [0.17, 0.32] | 0.28 [0.18, 0.33] |
| ASScore | Adult skills sub-domain | 0.48 [0.38, 0.56] | 0.48 [0.4, 0.57] |
| CYPScore | Children and young people sub-domain | 0.71 [0.37, 1.17] | 0.92 [0.55, 1.25] |
| GBScore | Geographical barriers sub-domain | -0.56 [-0.94, -0.14] | -0.51 [-0.86, -0.14] |
| IndScore | Indoors sub-domain | 0.92 [0.47, 1.28] | 0.89 [0.24, 1.26] |
| OutScore | Outdoors sub-domain | 0.5 [0.13, 0.84] | 0.59 [0.28, 0.92] |
| WBScore | Wider barriers sub-domain | 1.04 [-0.27, 2.01] | 1.03 [-0.15, 2.01] |

Table A.14: Healthcare distance metrics (in meters or log meters) grouped by admission to hospital.

| name | Discharged from ACE | Admitted to Hospital |
|---|---|---|
| Distance to surgery | 917.01 [554.52, 1789.65] | 1153.24 [604.38, 1681.72] |
| Distance to hospital | 2265.56 [1334.02, 3628.76] | 2277.96 [1527.18, 3290.22] |
| log(Distance to surgery) | 6.82 [6.32, 7.49] | 7.05 [6.4, 7.43] |
| log(Distance to hospital) | 7.73 [7.2, 8.2] | 7.73 [7.33, 8.1] |

Table A.15: Occurrence and effect on hospitalisation of co-morbidity variables of interest. The 'coefficient' and '$p$' columns are the output of binomial GLM with hospitalisation as the response variable, and the variable of interest as the sole explanatory variable. Hospitalisation % refers to the proportion of hospitalised cases that were 'true' for each variable of interest. As such, if this percentage is greater than the occurrence of the condition, then that condition is linked to a disproportionately high rate of hospitalisation. 'Other resp' = Other respiratory conditions, comprising influenza, common cold, hay fever, sinusitis, croup and streptococcal pharyngitis.

| Condition | Time | Occurrence | Hospitalisation (%) | Coefficient | $p$ |
|---|---|---|---|---|---|
| Eczema | Any | 209 (46.9%) | 49 (62.8%) | 0.691 | 0.003 |
| Other resp. | Any | 155 (34.8%) | 33 (42.3%) | 0.478 | 0.039 |
| Bronchitis | Any | 74 (16.6%) | 20 (25.6%) | 0.589 | 0.033 |
| Pneumonia | Any | 59 (13.2%) | 19 (24.4%) | 0.907 | 0.002 |
| Eczema | Year | 39 (8.7%) | 10 (12.8%) | 0.703 | 0.042 |
| Pneumonia | Year | 22 (4.9%) | 8 (10.3%) | 1.012 | 0.024 |

Table A.16: Occurrence and effect on hospitalisation of prescription variables of interest. 'Time' refers to prescription of a medication of the indicated group within a fixed period before ACE acceptance. The 'coefficient' and '$p$' columns are the output of binomial GLM with hospitalisation as the response variable, and the variable of interest as the sole explanatory variable. Hospitalisation % refers to the proportion of hospitalised cases that were 'true' for each variable of interest. As such, if this percentage is greater than the occurrence of the condition, then that condition is linked to a disproportionately high rate of hospitalisation.

| Prescription | Time | Occurrence | Hospitalisation (%) | Coefficient | $p$ |
|---|---|---|---|---|---|
| >12 Inhalers | Any | 27 (6.2%) | 8 (10.3%) | 0.863 | 0.007 |
| >4 Prednisolone courses | 6 months | 43 (9.9%) | 12 (15.4%) | 0.765 | 0.028 |
| >3 Prednisolone courses | Year | 53 (12.2%) | 13 (16.7%) | 0.732 | 0.018 |
| >4 Prednisolone courses | Any | 44 (10.1%) | 13 (16.7%) | 1.22 | 0.037 |
| Prednisolone | 6 months | 100 (22.9%) | 22 (28.2%) | 0.476 | 0.066 |
| Prednisolone | Year | 151 (34.6%) | 31 (39.7%) | 0.431 | 0.065 |
| Slow-acting bronchodilators | 1 month | 42 (9.6%) | 12 (15.4%) | 0.797 | 0.023 |
| Slow-acting bronchodilators | 6 months | 89 (20.4%) | 25 (32.1%) | 0.887 | <0.001 |
| Slow-acting bronchodilators | Year | 107 (24.5%) | 29 (37.2%) | 0.766 | 0.002 |
| Slow-acting bronchodilators | Any | 140 (32.1%) | 35 (44.9%) | 0.647 | 0.006 |

Table A.17: Occurrence and effect on hospitalisation of visit variables of interest. The 'coefficient' and '$p$' columns are the output of binomial GLM with hospitalisation as the response variable, and the variable of interest as the sole explanatory variable. Hospitalisation % refers to the proportion of hospitalised cases that were 'true' for each variable of interest. As such, if this percentage is greater than the occurrence of the condition, then that condition is linked to a disproportionately high rate of hospitalisation.

| Visit type (Threshold) | Time | Occurrence | Hospitalisation (%) | Coefficient | $p$ |
|---|---|---|---|---|---|
| Emergency Room Visits (>4) | 1 year | 23 (5.3%) | 8 (10.3%) | 1.079 | 0.012 |
| Family Practice Visits (>4) | 6 months | 251 (57.6%) | 52 (66.7%) | 0.504 | 0.031 |
| Family Practice Visits (>6) | 1 year | 323 (74.1%) | 62 (79.5%) | 0.550 | 0.039 |
| Inpatient Visits (>1) | 6 months | 57 (13.1%) | 14 (17.9%) | 0.641 | 0.037 |
| Inpatient Visits (>2) | 1 year | 59 (13.5%) | 16 (20.5%) | 0.824 | 0.006 |
| Inpatient Visits (>4) | 3 years | 71 (16.3%) | 18 (23.1%) 0.613 | 0.029 | |
| Inpatient Visits (>9) | Total | 37 (8.5%) | 11 (14.1%) | 0.765 | 0.028 |
| Any Visits (>6) | 1 month | 23 (5.3%) | 8 (10.3%) | 1.012 | 0.024 |
| Any Visits (>4) | 6 months | 274 (62.8%) | 55 (70.5%) | 0.533 | 0.027 |
| Any Visits (>8) | 1 year | 309 (70.9%) | 58 (74.4%) | 0.541 | 0.037 |

Table A.18: Occurrence and effect on hospitalisation of air pollution variables of interest. The 'coefficient' and '$p$' columns are the output of binomial GLM with hospitalisation as the response variable, and the variable of interest as the sole explanatory variable. Hospitalisation % refers to the proportion of hospitalised cases that were 'true' for each variable of interest. As such, if this percentage is greater than the occurrence of the condition, then that condition is linked to a disproportionately high rate of hospitalisation.

| Variable (Threshold) | Occurrence | Hospitalisation (%) | Coefficient | $p$ |
|---|---|---|---|---|
| $NO_2 > 18.75\ \mu g{\cdot}m^3$ | 80 (17.9%) | 22 (28.2%) | 0.713 | 0.008 |
| $PM_{10} > 13.5\ \mu g{\cdot}m^3$ | 147 (33%) | 33 (42.3%) | 0.502 | 0.032 |

Table A.19: Occurrence and effect on hospitalisation of indices of multiple deprivation (IMD) variables of interest. Descriptions for abbreviations are found in Table A.13. The 'coefficient' and '$p$' columns are the output of binomial GLM with hospitalisation as the response variable, and the variable of interest as the sole explanatory variable. Hospitalisation % refers to the proportion of hospitalised cases that were 'true' for each variable of interest. As such, if this percentage is greater than the occurrence of the condition, then that condition is linked to a disproportionately high rate of hospitalisation.

| Variable (Threshold) | Occurrence | Hospitalisation (%) | Coefficient | $p$ |
|---|---|---|---|---|
| CYPScore >0.85 | 191 (42.8%) | 41 (52.6%) | 0.492 | 0.032 |
| EduScore >45.75 | 252 (56.5%) | 54 (69.2%) | 0.524 | 0.028 |
| IDCScore >0.25 | 187 (41.9%) | 42 (53.8%) | 0.482 | 0.035 |

Table A.20: Occurrence and effect on hospitalisation of distance variables of interest. The 'coefficient' and '$p$' columns are the output of binomial GLM with hospitalisation as the response variable, and the variable of interest as the sole explanatory variable. Hospitalisation % refers to the proportion of hospitalised cases that were 'true' for each variable of interest. As such, if this percentage is greater than the occurrence of the condition, then that condition is linked to a disproportionately high rate of hospitalisation.

| Variable (Threshold) | Occurrence | Hospitalisation (%) | Coefficient | $p$ |
|---|---|---|---|---|
| Distance to surgery >1200 m | 187 (41.9%) | 39 (50%) | 0.508 | 0.026 |
| Distance to hospital >2090 m | 245 (54.9%) | 50 (64.1%) | 0.516 | 0.031 |
| log(Distance to surgery) >7.1 | 190 (42.6%) | 39 (50%) | 0.482 | 0.035 |
| log(Distance to hospital) >7.6 | 242 (54.3%) | 50 (64.1%) | 0.550 | 0.022 |

# Bibliography


[1] Joni Jabbal and Matthew Lewis. Approaches to better value in the nhs: Improving quality and cost. https://www.kingsfund.org.uk/sites/default/files/2018-10/approaches-to-better-value-october2018_0.pdf, 10 2018.

[2] Care Quality Comission (CQC). Bradford teaching hospitals nhs foundation trust: Inspection report. https://api.cqc.org.uk/public/v1/reports/edcfb304-14c0-4e3e-8557-de663e8533f0?20210113203413, 09 2020.

[3] Health Services Journal (HSJ). Hsj awards 2018: Improvement in emergency and urgent care. https://www.hsj.co.uk/the-hsj-awards/hsj-awards-2018-improvement-in-emergency-and-urgent-care/7023834.article, 11 2018.

[4] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

[5] Sara Bersche Golas, Takuma Shibahara, Stephen Agboola, Hiroko Otaki, Jumpei Sato, Tatsuya Nakae, Toru Hisamitsu, Go Kojima, Jennifer Felsted, Sujay Kakarmath, and et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making*, 18(1), Jun 2018.

[6] Fatemeh Rahimian, Gholamreza Salimi-Khorshidi, Amir H. Payberah, Jenny Tran, Roberto Ayala Solares, Francesca Raimondi, Milad Nazarzadeh, Dexter Canoy, and Kazem Rahimi. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLOS Medicine*, 15(11), Nov 2018.

[7] Thomas Desautels, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, Uli Chettipally, Mitchell D Feldman, Chris Barton, and et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Medical Informatics*, 4(3), Sep 2016.

[8] Royal College of Nursing. Standards for assessing, measuring and monitoring vital signs in infants, children and young people. https://www.rcn.org.uk/-/media/royal-college-of-nursing/documents/publications/2017/may/pub-005942.pdf, 06 2017.

[9] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *Introduction To Statistical Learning: with applications in r*. Springer, 2021.

[10] UCLA: Statistical Consulting Group. Contrast coding systems for categorical variables. https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/, January 2011.

[11] Gregory Carey. Coding categorical variables. http://psych.colorado.edu/~carey/Courses/PSYC5741/handouts/Coding%20Categorical%20Variables%202006-03-03.pdf, March 2003.

[12] R Barandela, J.s Sanchez, V Garcia, and E Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851, 2003.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[14] 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 Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[16] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[17] Andrew Gelman, Christian Robert, Nicholas Chopin, and Judith Rousseau. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2013.

[18] L-P. Boulet. Influence of comorbid conditions on asthma. *European Respiratory Journal*, 33(4):897–906, April 2009.

[19] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, November 2011.

[20] P C Austin and E W Steyerberg. Events per variable (epv) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research*, 26(2):796–662, 2017.

[21] Torsten Hothorn and Berthold Lausen. Double-bagging: combining classifiers by bootstrap aggregation. *Pattern Recognition*, 36(6):1303–1309, June 2003.

[22] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, March 1977.