



# THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.



# **Data-driven approaches for predicting asthma attacks in adults in primary care**

Holly Tibble

The University of Edinburgh

Primary Supervisor: Prof. Aziz Sheikh

Co-Supervisors: Dr. Athanasios Tsanas & Prof. Robert Horne

Advisors: Dr. Mehrdad Mizani & Prof. Colin Simpson

Thesis submitted in fulfilment of the requirements  
for the research degree of PhD Medical Informatics

Edinburgh University

2021

## Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. Parts of this work have been published in BMJ Open, Scientific Reports, The British Journal of Clinical Pharmacology, and BMC Medical Research Methodology.

A handwritten signature in black ink, appearing to read "Halcyon". The signature is written in a cursive, flowing style with some loops and flourishes.

# Abbreviations

Abbreviation	Term
<b>A&amp;E</b>	Accident and Emergency (Hospital Department)
<b>ACOS</b>	Asthma- Chronic Obstructive Pulmonary Disease Overlap Syndrome
<b>ADRN</b>	Administrative Data Research Network
<b>ALHS</b>	Asthma Learning Healthcare System
<b>ATS</b>	American Thoracic Society
<b>AUC</b>	Area Under the Receiver Operator Curve
<b>AUKCAR</b>	Asthma UK Centre for Applied Research
<b>BI</b>	Bookmaker's Informedness
<b>BMI</b>	Body Mass Index
<b>BNF</b>	British National Formulary
<b>BS</b>	Brier Score
<b>BTS</b>	British Thoracic Society
<b>CART</b>	Classification and Regression Tree (algorithm)
<b>CCI</b>	Charlson Comorbidity Index
<b>CHI</b>	Community Health Index
<b>CI</b>	Confidence Interval
<b>CIC</b>	Class Imbalance Coefficient
<b>CMA</b>	Continuous Multiple-interval measures of medication Availability
<b>COPD</b>	Chronic Obstructive Pulmonary Disease
<b>CPRD</b>	Clinical Practice Research Datalink
<b>CSA</b>	Continuous Single-interval measures of medication Availability
<b>CSG</b>	Continuous Single-interval measures of medication Gaps
<b>CV</b>	Cross-validation
<b>DNA</b>	DeoxyriboNucleic Acid
<b>EBC</b>	Exhaled Breath Condensate
<b>eDRIS</b>	electronic Data Research and Innovation Service
<b>EHR</b>	Electronic Health Record
<b>EIM</b>	Exhaled Inflammatory Markers
<b>EMD</b>	Electronic Monitoring Device
<b>EQUATOR</b>	Enhancing the Quality and Transparency of health Research (guidelines)
<b>ERS</b>	European Respiratory Society
<b>ESCOMP</b>	European Society for Patient Adherence, Compliance, and Persistence
<b>FENO</b>	Fractional Exhaled Nitric Oxide
<b>FEV<sub>1</sub></b>	Forced Expiratory Volume (in one second)



<b>Abbreviation</b>	<b>Term</b>
<b>FN</b>	False Negative
<b>FP</b>	False Positive
<b>FVC</b>	Forced Vital Capacity
<b>GERD</b>	Gastroesophageal Reflux Disease
<b>GINA</b>	Global Initiative for Asthma
<b>GLM</b>	Generalised Linear Regression
<b>GMA</b>	Geometric Mean Accuracy
<b>GP</b>	General Practitioner
<b>HRT</b>	Hormone Replacement Therapy
<b>ICD</b>	International Classification of Diseases
<b>ICS</b>	Inhaled Corticosteroids
<b>IgE</b>	Immunoglobulin E
<b>IL-5</b>	Interleukin-5 (protein)
<b>IQR</b>	Interquartile Range (herein upper (75 <sup>th</sup> ) and lower (25 <sup>th</sup> ) percentile values, rather than their difference)
<b>IRLS</b>	Iteratively Reweighted Least Squares
<b>ISPE</b>	International Society for Pharmacoepidemiology
<b>k-NN</b>	k-Nearest Neighbours (algorithm)
<b>LABA</b>	Long-Acting $\beta$ -2 Agonist
<b>LAMA</b>	Long-Acting Muscarinic Antagonists
<b>LRTI</b>	Lower Respiratory Tract Infection
<b>LTRA</b>	Leukotriene Receptor Antagonists
<b>mAb</b>	Monoclonal Antibodies
<b>MCC</b>	Matthews Correlation Coefficient
<b>MCG</b>	Microgram
<b>MG</b>	Milligram
<b>MPI</b>	Message Passing Interface
<b>MPR</b>	Medication Possession Ratio
<b>NBC</b>	Naïve Bayes Classifier
<b>NHLBI</b>	(American) National Heart, Lung, and Blood Institute
<b>NICE</b>	(English) National Institute of Clinical Excellence
<b>NNT</b>	Number Needed to Treat
<b>NPV</b>	Negative Predictive Value
<b>NUTS-3</b>	Nomenclature of Units for Territorial Statistics Level-3
<b>OCS</b>	Oral Corticosteroids
<b>OP</b>	Optimised Precision
<b>OR</b>	Odds Ratio
<b>PBPP</b>	Public Benefit and Privacy Panel
<b>PEFR</b>	Peak Expiratory Flow Rate

<b>Abbreviation</b>	<b>Term</b>
<b>PIS</b>	Prescribing Information Service
<b>PPI</b>	Patient and Public Involvement
<b>PPV</b>	Positive Predictive Value
<b>PROM</b>	Patient Reported Outcome Measure
<b>QOF</b>	Quality and Outcomes Framework
<b>RF</b>	Random Forest (algorithm)
<b>RI</b>	Relationship Index
<b>ROC</b>	Receiver Operator Curve
<b>RSV</b>	Respiratory Syncytial Virus
<b>SABA</b>	Short-Acting $\beta$ -2 Agonist
<b>SIGN</b>	Scottish Intercollegiate Guidelines Network
<b>SIMD</b>	Scottish Index of Multiple Deprivation
<b>SMOTE</b>	Synthetic Minority Over-Sampling Method
<b>SPT</b>	Skin Prick Test
<b>SURE</b>	Safe Users of Research data Environment (certification)
<b>SVM</b>	Support Vector Machine
<b>TN</b>	True Negative
<b>TP</b>	True Positive
<b>TRIPOD</b>	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (guidelines)
<b>UCI</b>	University of California, Irvine
<b>UK</b>	United Kingdom
<b>UR6</b>	The 6-category Scottish Government Urban Rural Classification Scale
<b>USA</b>	United States of America
<b>VPN</b>	Virtual Private Network

# Frequently Used Notation

The following mathematical notational conventions are used throughout this thesis:

- Scalar values are written in italic lower-case letters, for example  $k$ .
- Vectors are written in bold lower-case letters, for example  $\mathbf{x}$ .
- Matrices are written in bold capital letters, for example  $\mathbf{X}$ .  $\{\cdot\}_{i,j}$  denotes the  $i^{\text{th}}$  row,  $j^{\text{th}}$  column matrix entry. The subscript  $n$  in the form  $x_n$  indicates the  $n^{\text{th}}$  element of a vector.
- Unless otherwise specified, the Euclidean distance is used.
- The set of integers, known as the natural numbers, is denoted  $\mathbb{N}$ , and the strictly positive (not including zero) subset of natural numbers is denoted  $\mathbb{N}^+$ .
- The set of real numbers is denoted  $\mathbb{R}$ , and the strictly positive subset of real numbers is denoted  $\mathbb{R}^+$ .
- Scaled or transformed vectors or scalars are denoted  $\mathbf{x}^*$  and  $k^*$ , respectively.
- The modulus, or absolute value, of a number, here  $x$ , is denoted  $|x|$ .

# Abstract

## Background

Asthma attacks cause approximately 270 hospitalisations and four deaths per day in the United Kingdom (UK). Previous attempts to construct data-driven risk prediction models of asthma attacks have lacked clinical utility: either producing inaccurate predictions or requiring patient data which are not cost-effective to collect on a large scale (such as electronic monitoring device data). Electronic Health Record (EHR) use throughout the UK enables researchers to harness comprehensive and panoramic patient data, but their cleaning and pre-processing requires sophisticated empirical experimentation and data analytics approaches. My objectives were to appraise the previously utilised methods in asthma attack risk prediction modelling for feature extraction, model development, and model selection, and to train and test a model in Scottish EHRs.

## Methods

In this thesis, I used a Scottish longitudinal primary care EHR dataset with linked secondary care records, to investigate the optimisation of an asthma attack risk prediction model. To inform the model, I refined methods for estimation of asthma medication adherence from EHRs, compared model training data enrichment procedures, and evaluated measures for validating model performance. After conducting a critical appraisal of the methods employed in the literature, I trained and tested four statistical learning algorithms for prediction in the next four weeks, i.e. logistic regression, naïve Bayes classification, random forests, and extreme gradient boosting, and validated model performance in an unseen hold-out dataset. Training data enrichment methods were compared across all algorithms to establish whether the sensitivity of estimating relatively uncommon event incidence, such as asthma attacks in the general asthma population, could be improved. Secondary event horizons were also examined, such as prediction in the next six months. Empirical experimentation established the balanced accuracy to be the most appropriate prediction model performance measure, and the calibration between estimated and

observed risk was additionally assessed using the Area Under the Receiver-Operator Curve (AUC).

## **Results**

Data were available for over 670,000 individuals, followed for up to 17 years (177,306 person-years in total). Binary prediction of asthma attacks in the following four-week period resulted in 1,203,476 data samples, of which 1% contained one or more attacks (12,193 total attacks). In the preliminary model selection phase, the random forest algorithm provided the best balance between accuracy in those with asthma attacks (*sensitivity*) and in those predicted to have attacks (*positive predictive value*) in the following four weeks. In an unseen data partition, the final random forest model, with optimised hyper-parameters, achieved an AUC of 0.91, and a balanced accuracy of 73.6% after the application of an optimised decision threshold. Accurate predictions were made for a median of 99.6% of those who did not go on to have attacks (*specificity*). As expected with rare event predictions, the sensitivity was lower at 47.7%, but this was well balanced with the positive predictive value of 48.9%. Furthermore, several of the secondary models, including predicting asthma attacks in the following 12 weeks, achieved state-of-the-art performance and still had high potential clinical utility.

## **Conclusions**

I successfully developed an EHR-based model for predicting asthma attacks in the next four weeks. Accurately predicting asthma attacks occurrence may facilitate closer monitoring to ensure that preventative therapy is adequately managing symptoms, reinforce the need to keep abreast of triggers, and allow rescue treatments to be administered quickly when necessary.

# Lay Summary

My main challenge in this thesis was to find a way to predict asthma attacks that can be used by healthcare professionals to help them improve patient care.

Asthma attacks kill roughly four people every day on average in the United Kingdom. Doctors and nurses can prescribe very high strength oral steroids (such as prednisone) for a few days, to help patients' symptoms improve quickly and make it less likely that they will need to go to the hospital. Repeated use of oral steroids can however lead to very unpleasant side effects like bone weakness, thinned skin, and poor eyesight. Therefore, while we want to identify a high number of the people that might need oral steroids to avoid emergency care for their exacerbated symptoms, we also want to avoid using them when not absolutely necessary.

Other researchers have tried to predict asthma attacks before, but they have come across a lot of problems. They could not work out when people with asthma were not regularly taking their prescribed daily medication, which is very common and a major cause of sudden changes in symptoms. The researchers often did not check their system worked properly, meaning that their results can be hard to trust, and often used data which are hard and expensive to obtain, for example data from personal monitoring machines such as electronic peak flow monitors or smart-inhalers. In the UK, patients have individual *electronic health records*, digital versions of their medical history that their GP maintains. These make it easy for doctors to find important information from your past, and everything is stored more safely than paper copies. These records can also be used in medical research, after removing information that could be used to work out who someone is. A collection of mathematical methods known as *machine learning* often work really well with electronic health records to make predictions. This is because these methods need a huge number of data records to work, but they can find out very specific combinations of traits (like the height, weight, age, and medical history of a person) that changes the risk of asthma attacks.

In this thesis, I show step-by-step how to build an asthma attack prediction system. By using better data and better methods than previous studies to make our prediction system, I believe that I can now predict asthma attacks better than before. However, there is still more work that can be done, and I have highlighted some ideas to make the system work even better.

# Acknowledgements

I would foremost like to thank my supervisors, Prof. Aziz Sheikh, Dr. Athanasios Tsanas and Prof. Robert Horne, for their support and contributions towards my research. Furthermore, the inputs of Dr. Mehrdad Mizani and Prof. Colin Simpson were both constructive and creative, and I am extremely grateful for your time.

I have had the fortunate opportunity to collaborate with many fantastic researchers during my research, who have been very formative in the direction of work and my approach to writing. In particular, I would like to thank Dr. Amy Chan, Dr. Helen Stagg, Dr. Bernard Vrijens, Prof. Sabine de Geest and Dr. Ahmar Shah, for their encouragement and guidance.

I would like to thank my interim examiners, Dr. Ahmar Shah, Prof. Steve Cunningham, and Prof. Steff Lewis, for their constructive feedback and support throughout my research, and Dr. Chris Newby and Prof. Ewen Harrison for agreeing to conduct my final examination.

I would like to thank the members of the DARTH research group for their time, companionship, and support. Both Elsie Horne and Dimitrios Doudesis, in particular, provided valuable feedback and technical assistance on areas of my work.

Of course, thank you to my family – my wonderful mum, Jill Tibble, read every page of this before the first draft was even sent around, as she always does.

My housemate for the vast majority of my PhD, Abigail Elliot, has been a continuous source of inspiration. Her work ethic is both motivating and exasperating, but as much as her infectious drive pushed me to work harder, her affection and kindness helped me keep me sanity. Until we entered pandemic lockdown, and then sanity was a lost cause for either of us.



My partner Nick has had endless patience with my rants about data, and never once expressed regret at moving in with me during my final three months of writing, despite frequent demands for coffee and hugs.

Finally, none of this would have been possible without the patient data, collected by the NHS and entrusted by patients to researchers like myself, to improve health and care for everyone.

# Contents

<b>Declaration</b> .....	<b>i</b>
<b>Abbreviations</b> .....	<b>ii</b>
<b>Frequently Used Notation</b> .....	<b>i</b>
<b>Abstract</b> .....	<b>ii</b>
<b>Lay Summary</b> .....	<b>iv</b>
<b>Acknowledgements</b> .....	<b>vi</b>
<b>Contents</b> .....	<b>viii</b>
<b>List of Figures</b> .....	<b>xiv</b>
<b>List of Tables</b> .....	<b>xviii</b>
<b>List of Appendices</b> .....	<b>i</b>
<b>1 Introduction</b> .....	<b>3</b>
1.1 Background .....	3
1.1.1 Asthma.....	3
1.1.2 Asthma Treatments.....	4
1.1.3 Asthma Attacks .....	6
1.1.4 Machine Learning .....	7
1.1.5 Prediction Modelling with Electronic Health Records.....	8
1.2 Key Technical Terminology .....	9
1.3 Aims and Objectives .....	13
1.4 Scope and Structure of the Thesis.....	14
1.5 Key Contributions.....	16
<b>2 Overview of Electronic Health Records and Dataset for Analyses</b> .....	<b>20</b>

2.1	Electronic Health Record Data for Model Training.....	20
2.1.1	Strengths of Using Electronic Health Records for Health Research....	20
2.1.2	Limitations of Using Electronic Health Records for Health Research..	22
2.2	Asthma Learning Healthcare System Data.....	23
2.2.1	Introduction .....	23
2.2.2	Ethics .....	25
2.2.3	Data Processing.....	27
2.3	Training Data Population.....	37
2.4	Asthma Attack Ascertainment.....	38
2.5	Chapter Summary .....	39
<b>3</b>	<b>A Critical Appraisal of the Predictive Value of Asthma Attack Risk Factors</b>	<b>40</b>
3.1	Feature Selection for Prediction Modelling .....	40
3.2	Introduction to Missing Data.....	41
3.3	Asthma Attack Risk Factor Inclusion and Exclusion Criteria.....	43
3.4	Demographic Risk Factors.....	45
3.4.1	Current Age.....	45
3.4.2	Sex.....	45
3.4.3	Socioeconomic Status .....	46
3.4.4	Ethnicity .....	46
3.5	Risk Factors Relating to Lifestyle.....	47
3.5.1	Smoking .....	47
3.5.2	Obesity.....	48
3.6	Asthma-Related Risk Factors .....	49
3.6.1	Asthma Symptom Control.....	49
3.6.2	Lung Function .....	50
3.6.3	Exhaled Inflammatory Markers .....	51
3.6.4	Controller Treatment Intensity and Severity.....	51

3.6.5	Medication Adherence .....	54
3.6.6	Previous Asthma Attacks and Unscheduled Care .....	55
3.7	Other Comorbidities .....	57
3.7.1	Eosinophilia.....	57
3.7.2	Atopy.....	58
3.7.3	Respiratory Infections .....	60
3.7.4	Other Chronic Comorbidities.....	61
3.8	Other Notable Risk Factors.....	62
3.9	Conclusions.....	65
<b>4</b>	<b>Comparison of Pharmacy-Based Measures of Asthma Controller</b>	
	<b>Medication Adherence in the Asthma Learning Healthcare System Data .....</b>	<b>73</b>
4.1	Background.....	73
4.2	Methods .....	75
4.2.1	Prescription-Based Adherence Measures .....	75
4.2.2	Identifying Asthma Controller Medications.....	80
4.2.3	Controller Medication Cleaning.....	81
4.2.4	Analysis Plan .....	86
4.3	Results .....	88
4.3.1	Asthma Prescription Record Identification .....	88
4.3.2	Asthma Controller Medication Prescription Record Processing .....	90
4.3.3	Single Interval Adherence Measures .....	93
4.3.4	Multiple Interval Adherence Measures.....	95
4.3.5	Correlation between Time-Matched Adherence Measures.....	98
4.4	Adherence Measure Selection .....	101
4.4.1	Principal Findings.....	101
4.4.2	Results in Context.....	103
4.4.3	Limitations and Future Directions .....	104
4.4.4	Adherence Measure Selection for Asthma Attack Risk Prediction Modelling.....	107

<b>5</b>	<b>Machine Learning</b> .....	<b>110</b>
5.1	Introduction to Machine Learning.....	110
5.2	Process Flow for Model Training, Selecting, and Testing.....	112
5.3	Classification Algorithms.....	114
5.3.1	Generalised Logistic Regression.....	114
5.3.2	Naïve Bayes Classifiers.....	115
5.3.3	K-Nearest Neighbours.....	116
5.3.4	Decision Trees.....	118
5.3.5	Support Vector Machines.....	120
5.3.6	Ensemble Learning.....	123
5.4	Evaluating Model Performance.....	127
5.4.1	Probabilistic Performance Measures.....	128
5.4.2	Confusion Matrices and the Data Imbalance Problem.....	130
5.4.3	Confusion Matrix Performance Measures.....	131
5.4.4	Model Calibration.....	135
5.5	Training Data Enrichment.....	137
5.6	Model Interpretability.....	140
5.6.1	Global Model Interpretation.....	140
5.6.2	Local Model Interpretation.....	141
5.7	Summary.....	143
<b>6</b>	<b>Performance Measures for Binary Classification Problems</b> .....	<b>144</b>
6.1	Previous Work.....	144
6.2	Methods.....	146
6.2.1	Simulated Confusion Matrices.....	146
6.2.2	Real Datasets.....	147
6.2.3	Analysis Plan.....	149
6.3	Results.....	150

6.3.1	Experiment 1: The effect of class imbalance on performance measures with set true positive and negative class accuracy.....	150
6.3.2	Experiment 2: The effect of true positive and negative class accuracy on performance measures with set class imbalance.....	153
6.3.3	Empirical Analyses.....	155
6.4	Summarising Findings and Recommendations for Choice of Performance Measure .....	159
6.4.1	Summary of Experimental Investigation .....	159
6.4.2	Results in Context.....	160
6.4.3	Recommendations Dictating Performance Measure Choice .....	161
6.5	Summary remarks.....	165
<b>7</b>	<b>Asthma Attack Risk Prediction Model.....</b>	<b>167</b>
7.1	Previous Work.....	167
7.1.1	Study Setting.....	171
7.1.2	Study Methodology .....	171
7.1.3	Model Performance.....	176
7.1.4	Conclusions .....	176
7.2	Published Guidelines for Developing and Reporting Clinical Risk Prediction Models .....	177
7.3	Methods .....	178
7.3.1	Inclusion and Exclusion Criteria.....	178
7.3.2	Asthma Attacks .....	179
7.3.3	Risk Factors .....	180
7.3.4	Analysis Plan .....	181
7.3.5	Enrichments .....	185
7.3.6	Parallel Programming for Increased Efficiency .....	186
7.4	Results .....	188
7.4.1	Analysis Population.....	188
7.4.2	Outcome Ascertainment .....	195
7.4.3	Algorithm and Enrichment Selection.....	197

7.4.4	Model Performance.....	205
7.4.5	Feature Importance.....	209
7.4.6	Model Calibration .....	212
7.4.7	Model Discrimination in Subgroups .....	213
7.4.8	Secondary Endpoints.....	216
7.5	Conclusions.....	221
<b>8</b>	<b>Discussion .....</b>	<b>222</b>
8.1	Key Findings .....	222
8.2	Strengths.....	224
8.3	Limitations .....	227
8.4	Implementation.....	235
8.5	Future Work .....	241
8.6	Conclusion .....	246
	<b>References .....</b>	<b>247</b>
	<b>Appendices .....</b>	<b>285</b>
	<b>Glossary .....</b>	<b>373</b>

# List of Figures

Figure 1.1: Normal airways, airways of someone with asthma, and airways during an asthma attack (illustration by Tibble, H., 2018) .....	4
Figure 2.1: Linked Analysis Dataset Flow Diagram.....	24
Figure 4.1: Prescription calendar example, with 28 days of supply obtained in a 31-day refill interval.....	76
Figure 4.2: Decision tree illustrating the selection of medication quantity from prescribed and dispensed quantity variables in the Asthma Learning Healthcare System prescribing dataset .....	83
Figure 4.3: Illustration of the medication strength identification natural language pathways .....	85
Figure 4.4: Flowchart of ICS and ICS+LABA prescription record exclusions.....	89
Figure 4.5: Bar chart of the number of prescriptions per individual during their follow-up in the Asthma Learning Healthcare System Dataset.....	93
Figure 4.6: Boxplots (without outliers) of the values of each single interval availability adherence measure.....	94
Figure 4.7: Boxplots of CMA5s and CMA8s for (A) all follow-up time, (B) years of follow-up, and (C) quarters of follow-up.....	97
Figure 4.8: Spearman correlation between single interval adherence measures at each prescription refill.....	99
Figure 4.9: Spearman correlation between multiple interval adherence measures in (A) all of follow-up, (B) years, and (C) quarters .....	100
Figure 4.10: Spearman correlation between single and multiple interval adherence measures in (A) all of follow-up, (B) years, and (C) quarters .....	102



Figure 4.11: Bland-Altman plot of annual CMA8_2 and (year matched) CSA_3 estimates .....	109
Figure 5.1: Using a 5-NN algorithm to estimate the family that an unlabelled child belongs to, based on their height and age .....	117
Figure 5.2: A decision tree to distinguish the Tibble siblings.....	120
Figure 5.3: Convex hull of one class (Setosa) from the Iris dataset.....	121
Figure 5.4: Linear support vector machine example using a modification of the Iris dataset, separating iris's of species Setosa and Versicolor by their petal length and width .....	122
Figure 5.5: A visualisation of the three paradigms of ensemble classification .....	125
Figure 5.6: Density plot of estimated probabilities by observed outcome .....	129
Figure 5.7: Example of a Receiver Operator Curve .....	129
Figure 5.8: A scatterplot showing original samples of versicolor irises, from R.A. Fisher's Iris dataset, alongside SMOTE generated samples.....	138
Figure 6.1: The effect of varying the Class Imbalance Coefficient (CIC; size of the negative class relative to the positive class, which had 100 samples), for set values of the sensitivity and specificity .....	152
Figure 6.2: The effect of varying sensitivity and specificity on performance measures for balanced (50% positive samples) and imbalanced classes (10%, 25%, 75% and 90% positive samples).....	154
Figure 6.3: Decision tree for selecting performance measure in binary classification settings .....	163
Figure 7.1: Model training and validation process .....	183

Figure 7.2: Selected values of $k, z$ for SMOTE enrichment, and their resulting sample size, relating to the original training data sample size .....	186
Figure 7.3: Diagram of a multi-core processor .....	187
Figure 7.4: Asthma attack risk prediction model analysis population flow diagram.	189
Figure 7.5: Rate of asthma attack events per 10,000 person-years, and percentage of total attacks by event type .....	195
Figure 7.6: Percentage of all patients and all asthma attacks in analysis population by maximum British Thoracic Society (BTS) treatment step during follow-up.....	196
Figure 7.7: Venn diagram of patients with one or more asthma attack according to study criteria (green) and Read Codes (orange) .....	197
Figure 7.8: Boxplots of the area under the curve for each algorithm and enrichment method.....	199
Figure 7.9: Boxplots of balanced accuracy for each algorithm, enrichment method, and classification threshold approach .....	200
Figure 7.10: Boxplots of specificity for each algorithm, enrichment method, and classification threshold approach .....	201
Figure 7.11: Boxplots of sensitivity for each algorithm, enrichment method, and classification threshold approach .....	202
Figure 7.12: Boxplots of Positive Predictive Value (PPV) for each algorithm, enrichment method, and classification threshold approach.....	203
Figure 7.13: Sensitivity, Positive Predictive Value (PPV) and Matthews Correlation Coefficient (MCC) for the four RF models, using unenriched training data and optimised classification thresholds .....	204

Figure 7.14: Receiver Operator Curve for model performance in holdout partition, with threshold values indicated by colour .....	208
Figure 7.15: Density plot of logarithm of estimated probabilities by observed outcome in holdout partition .....	209
Figure 7.16: Top and bottom ranked features by Gini importance .....	211
Figure 7.17: Model calibration by risk deciles in the holdout partition .....	212
Figure 7.18: Discrimination by asthma attack history in the holdout partition.....	213
Figure 7.19: Discrimination by asthma severity, according to treatment step, in the holdout partition .....	214
Figure 7.20: Discrimination by asthma attack severity in the holdout partition.....	215
Figure 7.21: Discrimination by smoking status in the holdout partition.....	216
Figure 7.22: Adjusted feature importance of the top ten most important features in primary analysis, across secondary endpoints.....	218
Figure 7.23: Adjusted feature importance of the top ten most important features in primary analysis, across secondary endpoints, for attacks presenting to secondary care only .....	219
Figure 7.24: Median-relative feature importance for consultation month by health care setting and event horizon (weeks).....	220

# List of Tables

Table 2.1: Metadata for clinical data sources in the ALHS data.....	25
Table 2.2: Features present in the ALHS primary care registration dataset.....	28
Table 2.3: Features present in the ALHS primary care encounters dataset.....	29
Table 2.4: Features present in the ALHS primary care prescriptions dataset.....	31
Table 2.5: Features present in the ALHS primary care prescription dose description dataset.....	32
Table 2.6: Features present in the ALHS accident and emergency presentations dataset.....	33
Table 2.7: Features present in the ALHS inpatient hospital admissions dataset .....	35
Table 2.8: Features present in the ALHS mortality dataset.....	36
Table 3.1: Asthma attack risk factor inclusion and exclusion criteria .....	44
Table 3.2: Comparison of Global Initiative for Asthma (GINA) and British Thoracic Society (BTS) 2019 asthma treatment recommendations.....	52
Table 3.3: Asthma attack Read Codes (Version 2) .....	56
Table 3.4: Asthma attack risk prediction model risk factors .....	66
Table 4.1: Start and end of analysis window within observation period for continuous, multiple-interval, measures of medication availability and gaps.....	77
Table 4.2: Corticosteroid asthma therapy exclusion brands.....	80
Table 4.3: Corticosteroid asthma therapy exclusion formulation and indication terms .....	81
Table 4.4: Daily medication dose frequency keywords and observed incidence .....	82
Table 4.5: Asthma medication classifications in the ALHS prescription data (n=4,965,714) .....	90

Table 4.6: Asthma controller medication daily dose frequency and quantity of doses per dose time.....	91
Table 4.7: Summary table of the values of each single interval availability adherence measure.....	94
Table 4.8: Spearman correlation coefficients between single prescription adherence measures for subsequent refills.....	95
Table 4.9: Median and spread of CMA1 across time windows.....	96
Table 4.10: Spearman correlation between multiple prescription adherence measures for subsequent intervals (years and quarters) .....	98
Table 5.1: An annotated binary confusion matrix .....	131
Table 6.1: Confusion matrix cell calculations for experiment 1 with varying class imbalance coefficients .....	146
Table 6.2: Confusion matrix cell calculations for experiment 2 with balanced classes .....	147
Table 6.3: UCI dataset characteristics, before processing .....	148
Table 6.4: Description of empirical analyses of performance measures .....	150
Table 6.5: Results in empirical analyses with varying levels of class imbalance ....	156
Table 6.6: Results in empirical analyses with variations on the third analysis (high imbalance) .....	158
Table 7.1: Characteristics of previous asthma attack risk prediction models .....	168
Table 7.2: Model performance measures reported in asthma attack risk prediction studies .....	174
Table 7.3: Demographics of the ALHS analysis population .....	191
Table 7.4: Prevalence of comorbidities in ALHS analysis population.....	193

Table 7.5: Class sample sizes for enrichment methods across iterations.....	198
Table 7.6: Summary statistics of model performance measures from 100 data partition iterations, and the hold-out data partition .....	206
Table 7.7: Confusion matrix for model performance in holdout partition.....	207
Table 7.8: Performance measures for secondary endpoints.....	217

# List of Appendices

Appendix A: Notable Events and Achievements .....	286
Appendix B: Asthma Attack Risk Factor Read Codes (Version 2) .....	289
Appendix C: Illustration of Algorithm Used to Assign British Thoracic Society/Scottish Intercollegiate Guidelines Networks (2019) Treatment Steps .....	292
Appendix D: UK Asthma Medication Brand and Generic Names, Formulations, and Medication Strength (Adults) .....	295
Appendix E: Asthma Diagnosis and Management Read Codes (Version 2).....	302
Appendix F: Asthma Primary Care Encounter Read Codes (Version 2) .....	305
Appendix G: Comorbidity Read Codes (Version 2) .....	307
Appendix H: Visualisation of CMA Adherence Measures.....	330
Appendix I: Density Plots of Adherence Measures .....	333
Appendix J: Linkage of Primary Care Prescribing Records and Pharmacy Dispensing Records in the Salford Lung Study: Application in Asthma .....	337
Appendix K: Density Plots of Performance Measures in Iterations of Empirical Data Analyses .....	342
Appendix L: Application of Sokolova and Lapalme’s Performance Measure Invariance Properties to Further Performance Measures.....	346
Appendix M: Relevant Risk Prediction Model Guidelines Items and Location within Thesis .....	347
Appendix N: Chronic Obstructive Pulmonary Disease Diagnosis Read Codes (Version 2) .....	349
Appendix O: Summary of Features in Risk Prediction Model.....	352

Appendix P: Machine Learning Classification Algorithms: Functions for Implementation in R, and Hyper-parameter Ranges .....	358
Appendix Q: Deviations between the Final Analysis and the Published Protocol Paper Analysis Plan .....	359
Appendix R: Algorithm and Enrichment Selection: Additional Performance Measure Boxplots.....	367
Appendix S: Feature Importance.....	370



# 1 Introduction

In this chapter, I will provide some preliminary background into the topics discussed, for ease of reading, define the technical vocabulary that will be used throughout, and describe the structure of this thesis.

## 1.1 Background

### 1.1.1 Asthma

#### **DEFINITION: ASTHMA**

*“A disease characterized by recurrent attacks of breathlessness and wheezing, which vary in severity and frequency from person to person. In an individual, they may occur from hour to hour and day to day.*

*This condition is due to inflammation of the air passages in the lungs and affects the sensitivity of the nerve endings in the airways so they become easily irritated. In an attack, the lining of the passages swell causing the airways to narrow and reducing the flow of air in and out of the lungs.”*

World Health Organization <sup>1</sup>

Asthma is a chronic long-term lung disease characterised by inflammation of the airways and sensitivity of the nerve endings in the airways so they become easily irritated (known as *hyper-responsiveness*) by stimuli including allergens <sup>1</sup>. This inflammation obstructs the airways and can result in wheezing, chest tightness, coughing and shortness of breath <sup>2</sup>. An *asthma attack* is the sudden increase of constriction to the airways, as shown in Figure 1.1, which leads to a drastic worsening of symptoms including wheezing and coughing, often further exacerbated by the excessive production of thick mucus <sup>3</sup>. In extreme cases, unless aggressive emergency care is administered immediately, blood oxygen saturation may be reduced until the individual loses consciousness and dies.

In recent years, asthma has been estimated to affect between 235 and 339 million people worldwide <sup>4-8</sup>. Prevalence rates vary greatly between countries <sup>7</sup>, as a result of genetic and sociodemographic population factors <sup>9,10</sup>, and diagnosis ascertainment criteria (such as doctor-diagnosed vs. treated) <sup>7</sup>. Respiratory diseases, including asthma, are the third most common type of chronic illness worldwide <sup>11</sup>, and the United Kingdom (UK) is amongst the countries with the highest asthma prevalence <sup>7</sup>.

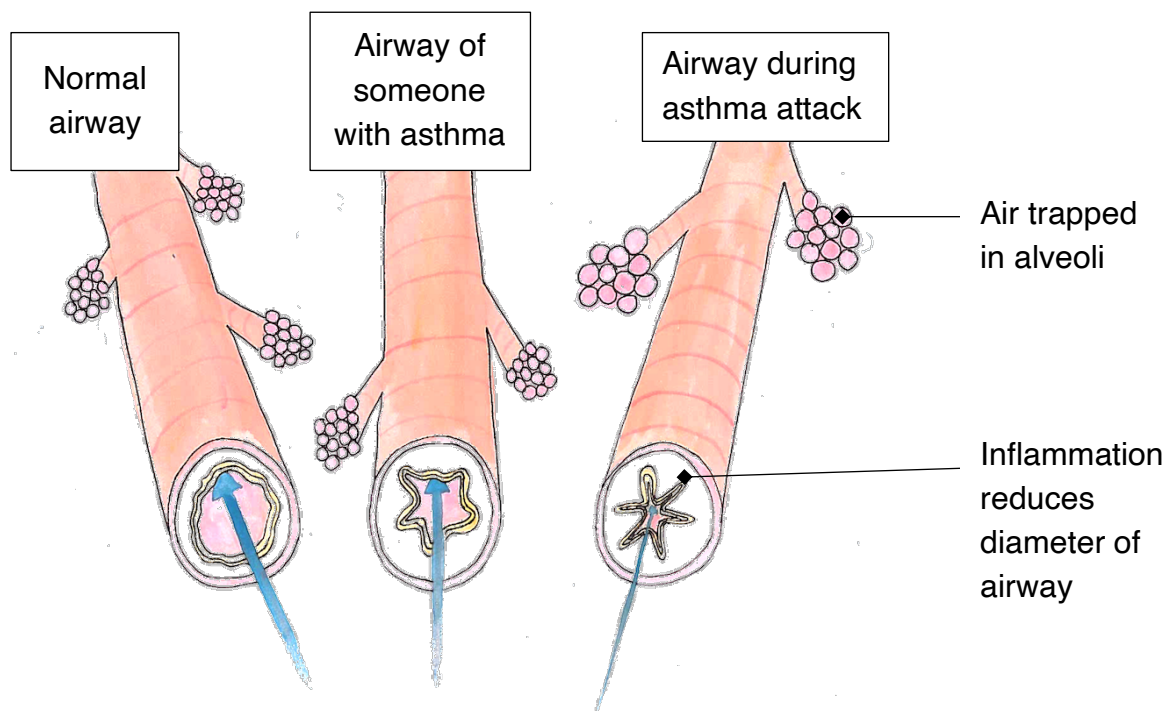


Figure 1.1: Normal airways, airways of someone with asthma, and airways during an asthma attack (illustration by Tibble, H., 2018)

Mukherjee *et al.* <sup>12</sup> estimated in 2016 that 3.8 per 1000 people in the UK have a first-time asthma diagnosis by a General Practitioner (GP) every year (age-standardised; 95% Confidence Intervals (CI) = 3.8–3.9), resulting in an estimated 16% of the population ever having been diagnosed by a clinician (95% CI 14.3-16.9).

### 1.1.2 Asthma Treatments

Asthma preventative treatments (also known as *controller* or *prophylactic* medications), are used to minimise airway hyper-responsiveness, resulting in fewer

daily symptoms and lower risk of an attack. The most common asthma preventer medication is a class of drugs known as Inhaled Corticosteroids (ICS), which are usually taken twice daily (morning and evening) by inhalation. ICS moderate inflammation through a mechanism of deactivating inflammatory genes, reducing airway hyper-responsiveness, and thus controlling asthma symptoms <sup>13,14</sup>.

Other common asthma preventer medications, which are often taken alongside ICS, include Long-Acting  $\beta$ -2 Agonists (LABAs), and Leukotriene Receptor Antagonists (LTRAs) <sup>15</sup>. There are also procedures aimed at controlling asthma symptoms, such as bronchial thermoplasty, and add-on therapies such as allergy treatments (including antihistamines) in cases where symptoms may be exacerbated by common or severe allergies.

Maintenance Oral Corticosteroids (OCS) can also be prescribed for patients whose asthma cannot be controlled even with very high doses of ICS (of which a single dose will be a lower strength than a dose of OCS) and add-on therapies. They are, however, often considered a last resort, because high doses of steroids result in an undesirable safety profile and side-effects, including increased risk of diabetes <sup>16–19</sup>, bone weakness <sup>19–22</sup>, and psychotic and affective disorders <sup>19,22–25</sup> with repeated use. Despite well-documented fears <sup>26,27</sup>, there is no evidence of the same side-effects (or other major safety concerns) for ICS <sup>26,28</sup>.

More recently, there has been substantial development of biological treatments including Monoclonal Antibodies (mAb; externally produced), proteins which the immune system uses to neutralise pathogens in the body. In immunological diseases such as asthma, they usually work by binding to, and inhibiting, *interleukins* (immune response signalling molecules) <sup>29–32</sup> or other antibodies <sup>33</sup>. They are increasingly recommended as a more tolerable alternative to OCS <sup>34</sup>.

The most common reliever (also known as *rescue*) medication is a Short-Acting  $\beta$ -2 Agonist (SABA). Like LABA, SABA opens the airways by relaxing the muscles in the airways, however the effects only last for around three to six hours, unlike LABA which

may be longer than 12 hours. Until recently, those without persistent asthma symptoms may have been prescribed SABA alone, however this is no longer recommended as of the 2019 Global Initiative for Asthma (GINA) guidelines <sup>35,36</sup>.

### 1.1.3 Asthma Attacks

Asthma attacks lead to more than 25 deaths per week on average in the UK <sup>37,38</sup>. If a patient contacts a health professional promptly after a severe exacerbation of symptoms, short courses of high-strength systemic steroids (oral or suspension) can be prescribed on top of preventative therapy to relieve exacerbations and reduce the need for (transfer to or continuation in) emergency care <sup>39–43</sup>. Systematic reviews by the Cochrane Airways Group have found that steroid courses administered in Accident and Emergency (A&E) reduced the probability of both inpatient admission and subsequent relapse of an asthma attack by 60% <sup>44,45</sup>.

Sudden-onset asthma attacks, defined by development over a period of less than six hours, are uncommon (6–20% of cases <sup>43,46</sup>) and typically the exacerbation of symptoms will be evident days before the peak of the attack <sup>47,48</sup>. Sudden-onset asthma attacks may be more commonly triggered by allergens, exercise or stress <sup>46,47</sup>, and less commonly triggered by respiratory infections <sup>49</sup>, compared to longer-onset asthma attacks.

While primary care providers are able to prescribe OCS courses, the first point of contact at symptoms decline for many patients is emergency care, either due to sudden-onset, or a lack of awareness of their symptom change or severity <sup>50</sup>. Indeed, the 2014 National Review of Asthma Deaths <sup>51</sup> reported that almost half of those who died had not sought medical assistance, or did so too late for emergency care to reach them. Even when medical assistance is sought, it is not always easy to tell when symptom decline will lead to an attack. Frequent use of OCS can lead to dangerous side-effects, especially when several courses are taken over a short duration. As such, clinicians need to be able to accurately gauge an individual patient's current risk.

The rationale for the development of this model is the hypothesis that risk classification facilitates efficient intervention. If a clinician was able to establish that a patient was identified as high risk of an asthma attack in the near future (such as, in the next four weeks), but was not immediately presenting with an exacerbation, they could be prompted to conduct certain small interventions to ensure that attacks were managed in a timely manner, and the need for secondary care was reduced. This would reduce patient anxiety, risk of life-threatening attacks, and the burden on the healthcare system of unscheduled care. Such interventions might include reviewing the asthma attack action plan, educating the patient about potentially risk-reducing lifestyle changes, providing the patient with tools and resources for lung function or symptoms self-monitoring, conducting an inhaler technique assessment, reviewing known triggers, and discussing the need for a step-up in asthma treatment. While repeated use of the prediction tool would be unlikely to yield substantially different results unless major lifestyle modifications were made, treatment adherence had changed, treatment had been altered, or the patient was strongly affected by seasonality, the tool is still able to facilitate population risk stratification and be used as a health education tool.

#### 1.1.4 Machine Learning

A primary focus of this thesis is how a methodology known as *machine learning* can be applied to prognostic modelling of health data: in our case for the prediction of future asthma attack risk.

Machine learning is a term with no universally accepted definition<sup>52,53</sup>, but is considered herein to encompass statistical methods using either *parametric* or *non-parametric* computational algorithms to make estimations (continuous numerical value) or predictions (categorical values), or to provide statistical mapping for decision support. *Parametric* algorithms, including logistic regression and linear regression, have fixed parameters, which are estimated from the data. These algorithms are the most commonly used in clinical prediction modelling<sup>54</sup>. *Non-parametric* algorithms have flexible (un-specified) parameters; guided purely by the data, with no enforced assumptions of relationships between variables. Models using parametric algorithms are typically more easily interpreted and generally require less data to build than those

using non-parametric algorithms. Non-parametric algorithm models tend to outperform parametric algorithm models based on predictive accuracy, particularly when domain knowledge is low and the relationships between features are unknown (including interactions between features which cannot easily be expressed *a priori* in a parametric form), but the available relevant historical data are plentiful<sup>55,56</sup>. However, if not handled appropriately, this flexibility can result in random patterns in the data being learned as concepts which do not generalise well to new, unseen, data.

### 1.1.5 Prediction Modelling with Electronic Health Records

One potential source of the required wealth of historical data, discussed in Section 1.1.4, is Electronic Health Records (EHRs). Advances in data storage capacity and processing capabilities have opened doors to new mechanisms and models of health care. In recent years, many countries, including the UK, have strived to digitise their health records<sup>57</sup>. The transition from physical paper-based records to EHRs has been shown to save time, reduce medication errors, and increase adherence to clinical best practices<sup>58-60</sup>. It also generates a rich and wide-covering database, which can be repurposed for medical research, covering large proportions of the population with minimal cost or risk of privacy breach. The limitations of these data, however, are that extracting the required information is at best an arduous task, requiring substantial cleaning, outlier detection, and reformatting. At worst, the task may be impossible: the desired information may not be captured in the data, or the structure of the data may prohibit reliable extraction.

*"The [electronic health record] is not a direct reflection of the patient and physiology, but a reflection of the recording process inherent in healthcare with noise and feedback loops."*

- George Hripcsak & David J. Albers<sup>61</sup>

Furthermore, machine learning algorithms cannot be applied to EHR datasets without substantial pre-processing, and there are many factors that need to be considered when deciding how this should be done. First, we must consider how to define the analysis population, both in terms of asthma diagnosis and further exclusion criteria such as treatment<sup>62–66</sup>, attack history<sup>48,66,67</sup>, age<sup>68–74</sup>, and comorbidities<sup>64,75,76</sup>. Next, we must consider whether to conduct a study *cross-sectionally* (using a single time-point) or *longitudinally* (following people over time). If data permit analysis of multiple time-points per person, it must be considered how the lack of independence between time-points for the same person will be acknowledged, either in the data or in the algorithm itself. For cross-sectional studies, or for each time-point in longitudinal studies, one must consider the duration of the follow-up window in which we are looking for the outcome. Previous studies have mostly assessed whether an asthma attack occurred in the following six to 24 months<sup>66–69,71,73,75,77</sup>. They have also defined asthma attacks differently, some considering only those which resulted in hospital admission to be recorded, and some including any asthma-related unscheduled doctor visits<sup>78</sup>. Different risk factors have been used to predict attacks<sup>54,79</sup>, and different algorithms have been employed to generate the predictions<sup>54</sup>.

When building a clinical prediction model with aspirations for large-scale deployment, it is crucial to balance model performance with feasibility. The model must be able to predict events occurring within a sufficiently short window of time to enable preventative care to be provided at the appropriate time. It should leverage all of the relevant data that are captured in EHRs, and it must provide guidance in an appropriate way respective of the target end-user. Finally, the performance of this model must be estimated and reported in a way such that the user can both be confident in any results produced, while also aware of any potential limitations.

## 1.2 Key Technical Terminology

Herein, I will briefly list a few key terms that will be used when discussing prediction modelling.

### Sample

A single data point, or *observation*. The number of samples, or *sample size*, will be denoted herein as  $N$ . It is important to distinguish the sample size from the analysis population size, the latter of which measures only the number of individuals contributing the samples, rather than the number of samples themselves.

### Feature

A measurable property, or characteristic, of a sample – either comprised of raw data values, or some function of the raw data (such as Body Mass Index, or BMI, calculated using the raw height and weight recorded). Herein, the number of features will be denoted as  $M$ .

### Design Matrix

The design matrix is a matrix of data  $\mathbf{X}$ , comprised of  $N$   $M$ -dimensional row vectors (samples), denoted  $\mathbf{x}_i = (x_{i1} \dots x_{iM})$ , where  $N$  is the number of samples and  $M$  the number of features.

$$\mathbf{X} = \underbrace{\begin{bmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NM} \end{bmatrix}}_{\text{Design matrix}}$$

As the features are also known as dimensions, the term *high dimensional data* can be used to refer to design matrices with a large number of features.

### Feature Space

The  $M$ -dimensional space where each sample in the design matrix is represented by a single point.

### Characteristic

The value of a feature, such as *green* for the feature *eye colour*.



## Outcome

The *response*, or *label*; that which we are attempting to estimate (continuous) or predict (categorical). Each outcome (denoted  $y_i$ ) corresponds to a single sample,  $x_i$ .

$$\mathbf{y} = \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}_{\text{Outcome vector}}$$

## Labelled Data

Data which include a corresponding outcome for each sample.

## Algorithm

A mathematical process which specifies the steps for solving a problem. In machine learning, these steps tend to be iterative and run until a specific criterion is met.

## Supervised Learning

Determining a functional form ( $f$ ) associating a set of features with outcomes.

$$f\left(\begin{bmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NM} \end{bmatrix}\right) = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

## Reinforcement Learning

The machine is able to continuously learn from past errors by reconciling the observed and expected outcomes. A form of supervised learning.

## Unsupervised Learning

Obtaining information from features for a set of samples, with no assigned outcome.

## Training Data

Data used to build a statistical model.

### Query Sample

A sample which was not part of the training data, which is presented in the statistical learning model to estimate the outcome.

### Test Data

Labelled data which are used to test the performance of a constructed statistical model by the comparison of the predicted and observed outcome.

### Class

A categorical outcome. If binary outcomes (e.g. control vs asthma) lead to a binary class classification problem, otherwise (multiple possible outcomes) leads to a multi-class classification problem (see *classification*).

### Classification

A form of supervised learning that assigns query samples on a finite number of classes, as observed in the training data. Binary classification pertains to data with only two possible classes (such as YES and NO), while multi-class classification is for data with three or more possible classes.

### Regression

A form of supervised learning that estimates a continuous or possibly ordered outcome: the latter may also be considered a classification problem in some cases.

### Model

The product of applying a machine learning algorithm to training data. The model then allows estimation or prediction of outcomes (either classification or regression) for unseen test data.

### Over-fitting

A model has learned very well from the training data, but fails to generalise in new, unseen data. Implementing overly complex models which demonstrate low prediction error in the training data but do not generalise well to test data, resulting in considerable deviation between training and testing dataset performance.

### Validation

The process of establishing the reliability of the model's performance in unseen data. Validation can be *internal* if the testing data are from the same database, and are processed under the same conditions, as the training data (for example, if a random partition of 20% of the data was kept aside for validation) or *external* if it comes from a different database (for example, a similar but distinct study population). Validation can also be classed as *temporal* if it comes from the same source but at a later time-point (for the same exact individuals, or for an overlapping sample of new and previously used individuals) <sup>80</sup>, however we will consider that a subset of external validation methods herein and restrict internal validation to random partitioning.

To some, the phrase *internal validation* may be used to refer to model fit statistics calculated in the same data samples used to train the model, rather than in unseen data from the same distribution, however we will refer to this as *in-sample validation* to avoid confusion. Both internal and external validation as defined herein are *out-of-sample* validation methods.

## **1.3 Aims and Objectives**

In this thesis, I aim to critically review previous literature relating to each aspect of asthma attack risk prediction modelling, to test select methodologies to guide model building, and to construct and validate an asthma attack prediction model for deployment in UK primary care.

My objectives are:

1. To review opportunities and limitations of using EHRs for medical research,
2. To evaluate criteria for selecting the analysis population from primary and secondary care records,
3. To compare criteria for defining asthma attacks in EHRs,
4. To systematically evaluate previously identified asthma attack risk factors, for the utility and feasibility of their integration in a parsimonious statistical learning model,
5. To investigate machine learning methods which can be implemented on the model building platform,
6. To build and validate an asthma attack risk prediction model from mined EHR data and statistical machine learning algorithms.

## **1.4 Scope and Structure of the Thesis**

The second chapter describes the two datasets used in my thesis research, with respect to their generation, format, and contents, as well as noting the ethics approvals in place ensuring that the benefit of my research outweighs any potential harm to the individuals represented in the data. Of note, Section 2.1 investigates the strengths and weaknesses of using EHRs for medical research reported in the literature (Objective One), Section 2.3 reviews the ways in which the EHR population may be restricted for the model training data (Objective Two), and Section 2.4 highlights different approaches to ascertaining the incidence of asthma attacks (Objective Three).

Chapter three comprises a critical appraisal of previously identified potential risk factors for asthma attacks (Objective Four). The review in particular highlights their previously described predictive ability, whether they are reported routinely in EHRs, methods of extracting the pertinent information from EHRs, and the duration for which a measurement of a potentially time-varying risk factor can be reliably maintained (such as age, which cannot be assumed constant for more than one year).

In the fourth chapter, I further explore one particular risk factor identified in chapter three, for which there was insufficient guidance relating to how the information should be extracted from EHRs: medication adherence. I began by detailing the asthma treatment pathway, describing the expert-recommended framework for quantifying adherence, and comparing select measures in the primary dataset. The comparisons between adherence measures included the inter-measure and intra-measure (temporal) correlations between the measures at different timescales.

Chapter five delves into the machine learning methods which underpin this thesis (Objective Five) by detailing the process by which a statistical learning model is trained and tested, the different algorithms which I compared, and the methods by which the performance of the model can be summarised. At the end of this chapter, I also reviewed a problem I anticipated might impact the performance of the model, the relatively low incidence of asthma, and reviewed methods by which negative effects could be circumvented.

Chapter six comprises a series of experiments in both simulated data and real-world datasets to examine how various binary classification performance measures summarise scenarios anticipated as a result of the low incidence of asthma attacks described in Chapter four.

The seventh chapter assimilates the findings from all of the previous chapters to describe the training and testing of the asthma attack risk prediction model (Objective Six).

Finally, in chapter eight I discuss the strengths, limitations, potential impact, and future directions of this body of work, with a summary of the key learnings.

A summary of my milestones is available in Appendix A.

## 1.5 Key Contributions

The four key contributions of this thesis are as follows. First, I conducted a narrative review of literature on the estimation and reporting of asthma medication adherence, highlighting the value of standardised methodology and demonstrating how this could be used to increase the impact of research both within and across medical conditions.

Second, from the findings of this review, I conducted an investigation of the patterns observed in data from electronic inhaler monitoring devices, which provide granular records of exact device actuation times. This multi-dimensional overview contributed to a subsequent study comparing methods of approximating from EHRs the agreement between a prescribed medication regimen and the patient's resulting regimen execution. This analysis provided a thorough review of the methodology as it pertained to a real-world case study, informing asthma research and providing a template for similar investigations in other medical conditions.

Third, I constructed an algorithm for probabilistically linking asthma medication prescription and dispensing records. This allows researchers to identify prescriptions that were not collected, even when the two necessary data sources do not contain a unique identifying link.

In order to facilitate the identification and optimisation of the prediction model which provides the best results, I compared the ability of multiple binary classification model performance measures to detect specific failings in model prediction, specifically pertaining to prediction of rare events. Using these findings, I constructed a decision aid to assist researchers in the selection of a performance measure for general data problems.

Finally, I developed an asthma attack risk prediction model, which utilised data routinely recorded in the UK primary care setting, to provide forecasted risk within various event horizons between one week and one year in the future. The model had improved performance compared to others in the literature and is readily implementable thanks to publicly available R scripts.

These key contributions have resulted in the following publications and presentations:

### **Peer-Reviewed Journal Papers**

**Tibble H**, Chan AHY, Mitchell EA, Horne E, Doudesis D, Horne R, Mizani MA, Sheikh A, Tsanas A. (2021) A Data-Driven Typology of Asthma Medication Adherence using Cluster Analysis. *Scientific Reports*. 10(1). 14999.

**Tibble H**, Lay-Flurrie J, Sheikh A, Horne R, Mizani MA, Tsanas A., The Salford Lung Study Team. (2020) Linkage of Primary Care Prescribing Records and Pharmacy Dispensing Records in the Salford Lung Study: Application in Asthma. *BMC Medical Research Methodology*. 20(1). 303.

**Tibble H**, Flook M, Sheikh A, Tsanas A, Horne R, Vrijens B, De Geest S, Stagg HR. (2020) Measuring and reporting treatment adherence: what can we learn by comparing two respiratory conditions? *British Journal of Clinical Pharmacology*. 87(3). 825-836.

**Tibble H**, Horne E, Horne R, Mizani AM, Simpson CR, Sheikh A, Tsanas A. (2019) Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model. *BMJ Open*. 9(7), e028375.

### **Conference papers**

**Tibble H**, Chan A, Mitchell EA, Horne R, Mizani MA, Sheikh A, Tsanas A. (2019) Heterogeneity in Asthma Medication Adherence Measurement. *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. p. 899-903.

### **Conference abstracts**

**Tibble H**, Sheikh A, Tsanas A, Horne R, Mizani M, Simpson C, Lay-Flurrie, J. (2019) Linkage of Primary Care Prescribing Records and Pharmacy Dispensing

Records in Asthma Controller Medications. *International Journal of Population Data Science*, 4(3).

### **Oral presentations**

Heterogeneity in Asthma Medication Adherence Measurement. IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), 28<sup>th</sup> – 30<sup>th</sup> October 2019, Athens, Greece.

Linkage of Primary Care Prescribing Records and Pharmacy Dispensing Records in the Salford Lung Study: Application in Asthma. *Administrative Data Research*, 9<sup>th</sup> – 11<sup>th</sup> December 2019, Cardiff, UK.

Hormonal contraceptives and clinical outcomes of asthma in reproductive-age women: UK population-based cohort study. Asthma UK Centre for Applied Research (AUKCAR) ASM, 26<sup>th</sup> March 2020, Virtual Event.

### **Poster presentations**

Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model. AUKCAR ASM, 12<sup>th</sup> March 2019, London, UK.

A Data-Driven Typology of Asthma Medication Adherence Subgroups and their Associated Clinical Outcomes. Scottish Informatics and Computer Science Alliance (SICSA), 18<sup>th</sup> – 19<sup>th</sup> June 2019, Aberdeen, UK.

Measuring and Reporting Treatment Non-Adherence: What Can We Learn from the Cross-Comparison of Two Respiratory Conditions? AUKCAR ASM, 26<sup>th</sup> March 2020, Virtual Event.

Linkage of Primary Care Prescribing Records and Pharmacy Dispensing Records in the Salford Lung Study: Application in Asthma Adherence Research. AUKCAR ASM, 26<sup>th</sup> March 2020, Virtual Event.



Linkage of Primary Care Prescribing Records and Pharmacy Dispensing Records in the Salford Lung Study: Application in Asthma Adherence Research. International Society for Pharmacoepidemiology (ISPE) Mid-Year Meeting, 15<sup>th</sup> September 2020, Virtual Event.

## 2 Overview of Electronic Health Records and Dataset for Analyses

In this chapter, I review the strengths and limitations of using EHRs in medical modelling, and describe the study dataset used herein, with respect to format, size, and pre-processing.

### 2.1 Electronic Health Record Data for Model Training

While the determination of whether or not the data source is appropriate for the model's intended use is dependent on the required domains of information, the data structure itself inevitably has some influence on the processing requirements, the model's limitations, and the exact specifications of the model's features. In this section, I review the rationale and limitations of the use of EHRs in prediction modelling.

#### 2.1.1 Strengths of Using Electronic Health Records for Health Research

There are many benefits to using EHRs for medical research over data from other sources. The cost of recruitment for *primary* research studies (using purpose-built datasets) which require direct participant contact increases by the population size and follow-up duration, due to additional research time spent on consent and data collection. As EHRs are collected routinely, *secondary* analyses can be conducted using larger numbers of participants than in primary analyses, without considerable cost increase<sup>81-83</sup>. A recent study of risk prediction models developed with EHRs found that the median population size was over 25,000 individuals<sup>84</sup>. Large, population-representative sample sizes increase the internal validity of research by enabling sufficiently powered subgroup analysis and the identification of less common risk factors<sup>83</sup>.

EHRs collect a panoramic view of patient safety<sup>85</sup>, capturing a wide (many people) but typically shallow (low granularity) net of information about a population. They

typically have a lower risk of *selection bias* (in which the samples are not representative of the population under analysis) than patient-recruiting cohort studies<sup>83,86–88</sup>. This leads to increased generalisability, resulting in more reliable predictive performance in similar but distinct populations.

Many traditional studies are only able to capture data for a finite duration and at set time-points. Although many clinical measurements can only be taken by healthcare professionals, some Patient Reported Outcome Measures (PROMs) and data measurements may be completed by patients self-reporting on their symptom in paper-based or electronic forms. In EHR-based studies, almost any contact with a healthcare professional will result in data being captured, which can be used for analysis. This generates a longitudinal dataset with long follow-up duration. More frequent data collection also means that time-varying risk factors will be more accurately recorded, and that a single individual may be able to contribute multiple samples (for example, stratifying by year) to the analysis, further increasing the number of records available for analysis<sup>83</sup>. We may also be able to use data from the start of a patient's registration at a primary care practice, which means both more data, and access to their documented medical history<sup>82</sup>. Traditional studies on the other hand would rely on self-reported medical (including family) history, which is subject to recall bias<sup>89–91</sup>.

In traditional studies, the physiological, demographic, and other documented clinical data of interest (potential risk factors and confounders) must be predetermined and cannot be changed retrospectively, as the data would not have been captured. The amount of data collected from patients is also limited by resource availability (cost, time, and equipment). In EHR-based studies, there is a wealth of information captured which can be repurposed for analysis<sup>83</sup>. Furthermore, it is possible to discover new risk factors for which the data would not have been collected for use in traditional studies. Tests that are not conducted in routine practice cannot be included in analysis. While this might generally be considered a limitation, it means that EHR-based studies are inherently practicable for primary care-based decision support.

## 2.1.2 Limitations of Using Electronic Health Records for Health Research

As with any data source, there are limitations to the use of EHRs in academic research. The first consideration is that EHRs are not designed for research. Primary data are often of higher quality for research than secondary data, as they are being collected for that express purpose, rather than administrative purposes, for example <sup>92-94</sup>. EHR data may be entered inaccurately by busy clinicians <sup>95</sup>, or by other colleagues who were not present at the consultation (based on the text notes recorded by the clinician). Therefore, data which are not deemed important to the clinician (such as the weight of an individual with a healthy body mass) may not be recorded.

Furthermore, while primary care records may be used to capture data regarding medical recommendations, such as written prescriptions and referrals, in many cases the result of said recommendation is unknown <sup>96</sup>. A future record indicating whether the referral was followed through, or whether the prescription was effective, is only possible when there is a follow-up primary care appointment. As those with co-morbidities are seen more frequently by practitioners <sup>97,98</sup>, and are often excluded in traditional prospective studies, this may introduce selection bias in studies using outcomes ascertained only from primary care records.

Changes in study participation legislation may influence the risk of selection bias, such as changing to an opt-in consent model <sup>83,99</sup>. Previous studies comparing those who have opted in to the organ donation registry in the UK have found that older people and ethnic minorities were less likely to opt-in <sup>100</sup>.

All UK primary care data were recorded using Read Codes as the standard clinical terminology system between April 1986 and March 2018. In England and Wales, Read Codes are currently being replaced by SNOMED-CT, however Scotland currently maintains Read Code (version 2) as the de facto standard. Read Codes are 5-byte (or 4-byte prior to 2010) hierarchical, case-sensitive, and ordered character strings. For example, the code "H33.." is the header for asthma, and the code "H330." for atopic asthma falls underneath, as denoted by their shared first three characters. Much important information, however, is recorded in the unstructured free-text fields.

A 2012 study into health records in the UK Clinical Practice Research Datalink <sup>101</sup> (formerly known as the General Practice Research Dataset) found that cause of death was written solely in the free-text (not in any coded or structured cells) in almost 20% of mortality records. This free-text cannot always be made available to researchers, as it requires extensive anonymising which is often a manual process. Even when they are available, free-text data are much harder to use for quantitative analysis, and often requires natural language processing for the accurate extraction of useable clinical information <sup>102,103</sup>.

## 2.2 Asthma Learning Healthcare System Data

### 2.2.1 Introduction

The Asthma Learning Healthcare System (henceforth ALHS) data were created by University of Edinburgh researchers (led by Dr. Ireneous Soyiri and Dr. Colin Simpson) in order to develop and validate a prototype *learning healthcare system* for asthma patients in Scotland. In a learning healthcare system, patient data are repurposed for a continuous loop of knowledge-generation, evidence-based clinical practice change, and change assessment and validation <sup>104</sup>. The project aimed to increase understanding of variation in asthma outcomes and create benchmarks for clinical practice in order to reduce sub-optimal care. The study recruited over half a million patients from 75 general practices in Scotland, with primary care records linked to national A&E, hospital and mortality datasets using the Scottish health identification number known as the Community Health Index (CHI).

As shown in Figure 2.1, six primary datasets were cleaned and linked to create the ALHS data: a primary care registry containing patient demographics, primary care prescribing records, primary care Read Codes (version 2, see Section 2.1.2), inpatient hospital admissions, A&E records and deaths (Table 2.1).

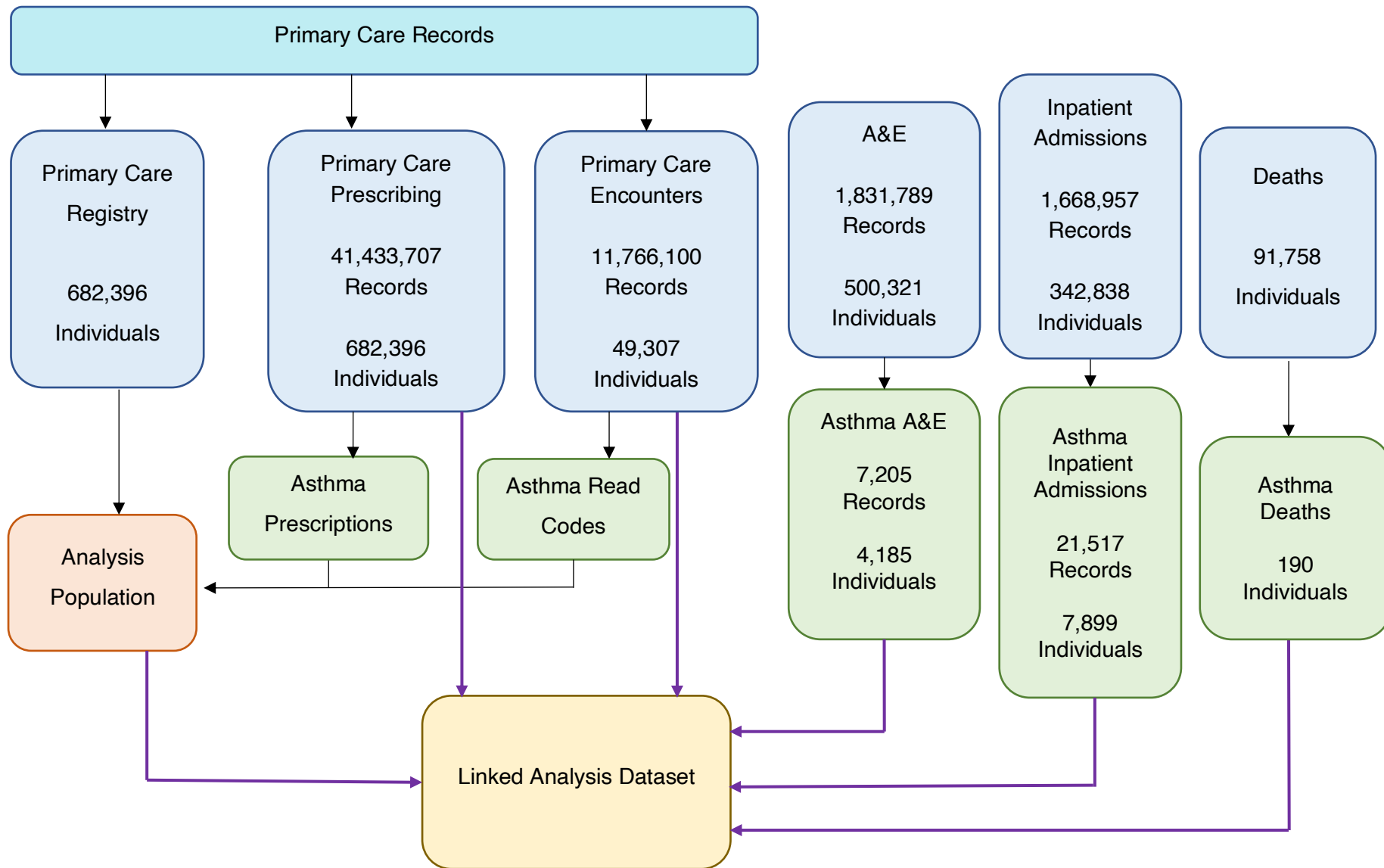


Figure 2.1: Linked Analysis Dataset Flow Diagram

Table 2.1: Metadata for clinical data sources in the ALHS data

<b>Data</b>	<b>Number of Records</b>	<b>Number of Individuals</b>	<b>Valid Date Range</b>
Primary Care Registry	706,546	682,396	N/A
Primary Care Encounters	11,766,100	49,307	Jan 2000 – Nov 2017
Primary Care Prescribing	41,433,707	671,304	Jan 2009 – Apr 2017
Accident & Emergency	1,831,789	500,321	Jun 2007 – Sep 2017
Hospital Inpatient Admissions	1,668,957	342,838	Jan 2000 – Mar 2017
Mortality	NA	91,758	Jan 2000 – Mar 2017

Note: Primary care encounter records available for the subset of the population with asthma diagnosis only

The primary care prescribing records were linked to pharmacy dispensing data, so that only collected prescriptions were included. This data linkage is not a perfect process, however, as prescription will have only a single identification code, regardless of the number of items. There is no item-specific identifier for each medication prescribed. As such, if the items are listed in a different order on the dispensing and prescribing records, additional information relating to a specific item (such as dosing direction notes from the pharmacist) may be assigned to the wrong prescription item.

As described in Table 2.1, the valid date range for records was between January 2000 and November 2017, although there were records dated outside of these limits which could be considered erroneous (such as the year 2099; date of record event can be documented as a distinct value to the date of record creation). To align the datasets, however, records dated before January 2009 or after March 2017 (henceforth referred to as the *study period*) were excluded.

### 2.2.2 Ethics

While there are many benefits in healthcare data analysis for the population under analysis, there are also concerns at the individual level about health data usage, including personal and private information becoming publicly leaked, and the potential for insurance companies to use aggregate level data to calculate risk which may raise premiums even when the individual is in fine health <sup>105</sup>.

Research has shown that the opt-in/opt-out status of a study results in different populations defined by their clinical profiles <sup>106–108</sup>, and demography <sup>108–110</sup>. The General Data Protection Regulation (GDPR), which came into effect on 25<sup>th</sup> May 2018, dictates that all data collected by European Union (EU) or UK organisations must have protection prioritised: unambiguous, informed consent must be received from all individuals for the storage and processing of any personal data, and all subjects have the right to revoke consent at any time.

In a public lecture in 2017, entitled "*How Big Data Can Inform Mental Health Research*" (University of Melbourne), Associate Professor Nicolas Cherbuin of the Australian National University stated that public perception was of the key barriers to modern health informatics, and that in order to gain trust of the public we as researchers need to be more transparent about the benefit-risk ratio. Fundamentally, denying that risks to data security exist is a lie, and instead being upfront about these risks and their likelihood may be more effective.

So, what are the risks to the public? There have been documented cases of researchers using their position of power to look up personal data relating to their peers, as well as the more commonly discussed fear of cyberterrorism. Illegitimate collection of health data relating to a specific individual, however, is unlikely unless that individual's data is especially valuable (a political figure, for example). Perhaps a more likely risk to the public is accidental publication of identifiable data, for example the publication of Australian Medicare data in 2016 which enabled identification of doctors, and even patients in certain rural areas <sup>111</sup>.

In our study, individual patient data was collected at the practice-level, and individual consent was not obtained. Permissions for the ALHS project were obtained from the South East Scotland Research Ethics Committee 02 [16/SS/0130] and the Public Benefit and Privacy Panel (PBPP) for Health and Social Care [1516-0489].



The ALHS data are held by the National Services Scotland electronic Data Research and Innovation Service (eDRIS) in the National Safe Haven. To be able to access the ALHS data, researchers must be added to the study team, and have their analysis plan approved by the PBPP. They must also have passed the Safe Users of Research data Environment (SURE) training, provided by the Administrative Data Research Network (ADRN).

Data were initially accessed only through the Edinburgh Safe Haven portal – a monitored, secure hub based at the Bioquarter, which requires two-factor authentication. After the UK entered lockdown for the COVID-19 pandemic (March 23<sup>rd</sup>, 2020), eDRIS announced that researchers would be able to apply for special access from their home personal computers (via Virtual Private Network, or VPN). On July 6<sup>th</sup>, 2020, my home access for this study was approved, and data were once again accessible.

Any outputs, including *metadata* (information about the specifications of the data, including size and format), are subject to disclosure checking by the eDRIS team, in order to ensure that no identifiable data are released.

## 2.2.3 Data Processing

### 2.2.3.1 Primary Care Registry Data

The primary care registry dataset contained 706,546 records for 682,396 unique individuals. 659,505 individuals (96.6%) had only a single registration record, while 21,720 had two (3.2%), and the remaining 1171 (0.2%) had between three and five. Features of the dataset are described in Table 2.2, including the Scottish Index of Multiple Deprivation (SIMD), a composite geographic-level measure incorporating income, employment, education, health, access to services, crime and housing <sup>112</sup>, and the 6-category Scottish Government Urban Rural Classification Scale <sup>113</sup> (UR6).

1032 records (1022 unique people) were missing registration date, and for 90 of these the deduction date was also missing. The earliest registration date was January 1<sup>st</sup>, 1860 (9937 records), and 10,559 records (1.5%) had registration before the year 1910. The latest registration was August 21<sup>st</sup>, 2017. The earliest deduction date was January 1<sup>st</sup>, 2010, and the latest was August 3<sup>rd</sup>, 2914 (erroneous). Only 10 records had deduction date after 2017 (the end of the study period). Those with missing registration or deduction dates were treated as the individual having registered before the study period and deducted afterwards. No modifications were made to outlier dates outside of the study period. 48,503 records were missing information related to DataZone, SIMD (both 2009 and 2012) and UR6. An additional 3,402 records were solely missing UR6 data. DataZones were linked to higher level area codes using the Nomenclature of Units for Territorial Statistics Level-3 (NUTS-3) <sup>114</sup>. Missing values for NUTS, SIMD, and UR6 were coded as a new category: “missing”,

Table 2.2: Features present in the ALHS primary care registration dataset

Feature Name	Data Type	Description	Example
ID	String	Unique Patient Identifier	“000001”
Sex	String	Sex	“F”
Age	Numeric	Age at 31 <sup>st</sup> , March 2018, or at deduction date if recorded	26
RegDate	Date (YYYY-MM-DD)	Date of registration at practice	“2003-04-02”
DeductionDate	Date (YYYY-MM-DD)	Date of deduction from practice	“2020-07-05”
RegStatus	String	Registration Status	“REG014”
DataZone	String	SNS 2001 DataZone of residence	
SIMD2012quintile	Numeric	SIMD quintile 2012	4
SIMD2009quintile	Numeric	SIMD quintile 2009	5
UR6_Code	Numeric	UR6 Code	5
UR6_Desc	string	Label of UR6 Level	“Accessible Rural”

Notes: SNS = Scottish Neighbourhood Statistics, SIMD = Scottish Index of Multiple Deprivation, UR6 = 6-category Urban-Rurality Scale

### 2.2.3.2 Primary Care Encounters

The Primary care dataset consisted of 11,766,100 Read Code records, for 49,307 unique patients (data from primary care encounters were only provided for those with a diagnosis of asthma), dated between January 1<sup>st</sup>, 2000 and November 1<sup>st</sup>, 9998 (erroneously). The features of the dataset are described in Table 2.3. Records dated after the study period end (31<sup>st</sup> March 2017; Table 2.1) were removed (n=624,393 records). Two further records were excluded because the Read Codes were missing. 860,092 records were removed because they were duplicates of other records on all variables except the encounter ID (subject, date, Read Code, and values associated with the code). Finally, one record was removed that was not in the correct 5-byte format, and thus could not be verified. This left 10,281,612 records, for 48,975 unique individuals.

Table 2.3: Features present in the ALHS primary care encounters dataset

Feature Name	Data Type	Description	Example
Index1	String	Unique Patient Identifier	"000001"
EncounterKey	Numeric	Unique Encounter Identifier	0001
EventDate	Date (DD-MM-YYYY)	Date of Encounter	01-01-2000
ReadCode	String	Read Code	"H33.."
Data1	Numeric	Numeric value associated with Read code (e.g. height in centimetres)	7
Data2	Boolean	Boolean value associated with Read Code (e.g. recent medication review)	TRUE
Data3	String	String value associated with Read Code (e.g. unit of Data1 value)	"kg"

### 2.2.3.3 Primary Care Prescriptions

The prescriptions dataset contained 41,433,707 records for 671,304 individuals. The features of the data are described in Table 2.4. 673 records were removed that were dated outside of the study period, leaving 41,433,034 records for 671,298 individuals. These records were linked by unique record identifier (named *ndx*) to a dataset containing the dose directions (instructions for taking, including the interval of doses; described in Table 2.5). There was one dose direction record for every prescription record in the Prescribing Information Service (PIS) dataset, and no additional records. 39 records were excluded after linking, however, as the dose directions indicated that the record should be deleted due to an error, leaving 41,432,995 records remaining (671,298 individuals).

From manual inspection of the excluded records, it became apparent that the dose directions were not always accurately matched to the prescription record. Upon consulting with the National Services Scotland Information Services Division Principal Pharmacist, Stuart McTaggart, I was informed that soon after initiation of the currently employed system, there were instances reported of the order of medications on the same prescription being different between the prescription and dispensing records, resulting in incorrect matching (correspondence in June 2020). Although feedback and improvement to this system has resulted in improvement over time, the issue still persists. The limitations in analyses resulting from this linkage process are described later in Section 8.3.

### 2.2.3.4 Accident and Emergency Presentations

There were 1,831,789 A&E records in the study period, and the features are described in Table 2.6. Records dated after the right censoring date were excluded (n=100,118). A further 651,465 (36%) records were removed as they contained no presenting complaint text or primary disease code (*Disease1Code*), leaving 1,080,206 records for 360,297 unique individuals.

Table 2.4: Features present in the ALHS primary care prescriptions dataset

Feature Name	Data Type	Description	Example
Index10	String	Unique Patient Identifier	"000001"
PIApprovedName	String	Medication Name	"BECLOMETASONE DIPROPIONATE"
PIPrescribableItemName	String	Brand Name	"BECLOMETASONE"
PIDrugFormulation	String	Formulation	"NASAL SPRAY"
PIItemStrength.UOM	String	Medication Strength	"400MCG "
PrescDate	Date (DD-MM-YYYY)	Date of Prescription	01-01-2000
DispDate	Date (DD-MM-YYYY)	Date of Dispensing	02-01-2000
PIBNFChapterCode	Numeric Categorical	BNF Chapter Code	12
PIBNFSectionCode	Numeric Categorical	BNF Section Code	1202
PIBNFSubsectionCode	Numeric Categorical	BNF SubSection Code	120201
PIBNFParagraphCode	Numeric Categorical	BNF Paragraph Code	1202010
PatientAgeatPaidDate	Numeric	Age of patient at date of prescription	30
ageatDispDate	Numeric	Age of patient at date of dispensing	30
ndx	Numeric	Unique record identifier for linkage to dose description dataset	1
Prescribed_quantity	Numeric	Quantity of medication prescribed	1
Dispensed_quantity	Numeric	Quantity of medication dispensed	1

Note: BNF = British National Formulary

Table 2.5: Features present in the ALHS primary care prescription dose description dataset

Feature Name	Data Type	Description	Example
ndx	Numeric	Unique record identifier for linkage to PIS dataset	1
ePRNativeDoseInstructions	String	Dose instructions	“take once daily”
ePRNDName	String	Full medication name (brand or formulation, and medication strength)	“BECONASE 400mcg”
Amount.min	Numeric	Minimum dose amount	1
Amount.max	Numeric	Maximum dose amount	2
Amount.unit	String	Dose unit (e.g. spray, mcg, nebule)	“spray”
Timing_freq.min	Numeric	Minimum dose timing	1
Timing_freq.max	Numeric	Maximum dose timing	2
Timing_freq.unit	String	Dose timing unit (e.g. hour, day, week)	“day”
Timing_interval.min	Numeric	Minimum inter-dose interval	5
Timing_interval.max	Numeric	Maximum inter-dose interval	8
Timing_interval.unit	String	Dose interval unit (e.g. hour, week, day)	“hours”
As_required	Boolean	To be taken as required, TRUE/FALSE	TRUE
As_directed	Boolean	To be taken as directed, TRUE/FALSE	FALSE

Note: PIS = Prescribing Information Service

Table 2.6: Features present in the ALHS accident and emergency presentations dataset

Feature Name	Data Type	Description	Example
IndexNumber	String	Unique Patient Identifier	“000001”
ArrivalDate	Date (DD-MM-YYYY)	Date of Presentation	01-01-2000
PresentingComplaintText	String	Primary Complaint Description	“hurt knee”
AlcoholInvolvedCode	Numeric Categorical	Flag for Alcohol Involvement	1
TriageCategoryCode	Numeric Categorical	Triage Category (e.g. standard, urgent, etc.)	3
PatManTypeCategoryCode	Numeric Categorical	Patient Management (Minor/Major/Resuscitation)	2
InvestigationType1Code	Boolean	Investigation Type Code	TRUE
InvestigationType2Code	Boolean	Investigation Type Code	TRUE
InvestigationType3Code	Boolean	Investigation Type Code	TRUE
Procedure1Code	Numeric Categorical	Procedure Code	9
Procedure2Code	Numeric Categorical	Procedure Code	9
Procedure3Code	Numeric Categorical	Procedure Code	9
Diagnosis1Code	Numeric Categorical	Diagnosis Code	16
Diagnosis2Code	Numeric Categorical	Diagnosis Code	16
Diagnosis3Code	Numeric Categorical	Diagnosis Code	16
Disease1Code	String	ICD10 Code	“F29”
Disease2Code	String	ICD10 Code	“F29”
Disease3Code	String	ICD10 Code	“F29”
ArrivalModeCode	Numeric Categorical	Arrival mode (e.g. air ambulance)	2

Notes: The National Services Scotland Information Services Division A&E Data Recording Manual specifies that investigation type codes should be numeric, but the features are Boolean in this dataset.

Diagnosis codes are recorded as only the top level (numeric) code, such as ‘9’, rather than the full, more specific, alpha-numeric code, such as 09B.

ICD10 = International Classification of Diseases, version 10

Diagnoses made from presentations to A&E are recorded in (up to) three fields in the data and are coded using the International Classification of Diseases (ICD) medical classification system, developed by the WHO. In the current version (ICD10, to be replaced in 2022), codes beginning with “J” are part of the subclass relating to the respiratory system, and the subdivision “J4” relates to chronic lower respiratory diseases. “J45” generally describes asthma, and “J46” specifically indicates an acute asthma attack.

Both “J45” and “J46” (and child codes belonging to these parent classes) were used to identify asthma-related A&E presentations, as well as the keyword *asthma* in their presenting complaint text. In total, 7,205 (1.1%) A&E presentations were flagged as asthma-related, for 4,185 unique individuals.

#### 2.2.3.5 Inpatient Hospital Admissions

This dataset is also known as SMR01. There were 1,668,957 inpatient admission records in the study period, and the dataset features are described in Table 2.7. Of these, 21,517 (1.3%, 7,899 individuals) were identified as relating to asthma by the presence of ICD code “J45” or “J46” (*Condition\_1*).

#### 2.2.3.6 Mortality

There were 91,758 records in the study period, for which 91,022 (99.2%) were for deaths (rather than stillbirths). The features of the dataset are described in Table 2.8. Of these, 190 (0.2%) had “J45” or “J46” as the primary cause of death.



Table 2.7: Features present in the ALHS inpatient hospital admissions dataset

Feature Name	Data Type	Description	Example
Index8	String	Unique Patient Identifier	"000001"
DATE_OF_BIRTH_YM	Numeric Date (MMYYYY)	Month and Year of Birth	041990
ADMISSION_DATE	Date (DD-MM-YYYY)	Admission Date	01-01-2000
DISCHARGE_DATE	Date (DD-MM-YYYY)	Discharge Date	02-01-2000
AGE_IN_YEARS	Numeric	Age at admission	19
SEX	Numeric Categorical	Sex	1
ADMISSION_TYPE	Numeric Categorical	Type of Admission	32
LENGTH_OF_STAY	Numeric	Length of stay (days)	2
Condition_1	String	ICD10 Code (Primary)	"F29"
Condition_2	String	ICD10 Code	"F29"
Condition_3	String	ICD10 Code	"F29"
Condition_4	String	ICD10 Code	NA
Condition_5	String	ICD10 Code	NA
Condition_6	String	ICD10 Code	NA
MAIN_OPERATION	String	Operation Code	"G459"
OTHER_OPERATION_1	String	Operation Code	"G459"
OTHER_OPERATION_2	String	Operation Code	"G459"
OTHER_OPERATION_3	String	Operation Code	"G459"
TOTAL_NUMBER_OF_EPISODES	Numeric	Total Number of admissions in previous 12 months	3

Note: ICD10 = International Classification of Diseases, version 10

Table 2.8: Features present in the ALHS mortality dataset

Feature Name	Data Type	Description	Example
Index6	String	Unique Patient Identifier	"000001"
DEATH_DATE	Date (DD-MM-YYYY)	Date of Death	01-01-2000
PRIMARY_CAUSE_OF_DEATH_0	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_0	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_1	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_2	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_3	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_4	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_5	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_6	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_7	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_8	String	ICD10 Code	I21
SECONDARY_CAUSE_OF_DEATH_9	String	ICD10 Code	I21
NRSStillbirths	Binary	Binary flag for stillbirth	0
NRSDeaths	Binary	Binary flag for death	1

Note: ICD10 = International Classification of Diseases, version 10

## 2.3 Training Data Population

Using EHRs as our primary data source to train the model allows us to include a broad spectrum of the population; the full study population is thus representative of any Scottish individual registered at a GP clinic. Despite this, we may wish to restrict the analysis population in some ways.

First, we want to include only those with asthma, as that is the target user of this prediction model. Previous studies have identified people with asthma as those with either *clinician-diagnosed* asthma <sup>48,68,70,115–117</sup> or *clinician-diagnosed-and-treated* asthma. Clinician diagnosed asthma can be ascertained in Scottish EHRs from Read Codes, although the codes used by various studies may differ. For example, Papi *et al.* used the Quality and Outcomes Framework (QOF) codes <sup>68</sup>, while Turner *et al.* additionally included codes relating to asthma management <sup>68</sup>, such as Read Code “8B3j.” for *asthma medication review*.

A study by Nissen *et al.* compared the number of people that were identified as having asthma in the UK Clinical Practice Research Datalink (CPRD) according to various criteria, and then estimated the percentage of those identified by clinical review to have a true diagnosis of asthma (a measure known as the Positive Predictive Value, or PPV) <sup>118</sup>. They found that including symptom-based codes (wheeze, breathlessness, chest tightness and cough) in the inclusion criteria, instead of specific diagnosis codes, even in combination with medication and reversibility testing, resulted in very poor PPV (56%). Requiring evidence of airway reversibility drastically reduced the number of identified individuals and did not show any improvement to the PPV. Adding recently prescribed asthma medication to combined diagnosis and management codes did not improve the PPV for asthma diagnosis, but also did not drastically reduce the population size.

Having both a sufficiently large analysis population and having a population of almost exclusively people with true asthma are very important. However, it is also useful to know approximately how many people with true asthma were excluded by these criteria, as excluding them might introduce bias. Unfortunately, Nissen *et al.* did not assess this <sup>118</sup>.

Some studies further stratified their population, such as including only those with mild to moderate asthma <sup>119</sup> or severe asthma <sup>64–66,120</sup>. While this might reduce the analysis population heterogeneity, it also restricts the utility of the resultant prediction model, as asthma attacks occur even in those with mild asthma <sup>75,121,122</sup>. Indeed, the majority of asthma attacks occur in low severity asthma patients (due to the small proportion of people with asthma that are considered moderate-to-severe <sup>123</sup>). Some studies also excluded people with <sup>64,65,67,119,124</sup> recent exacerbations, preventing first-time exacerbators from being detected. In order to maximise the (potential) benefit to patients of the methodology and data available, I did not wish to restrict asthma patients in such a way. However, subgroup analyses should be conducted to appraise the model's discriminatory performance across the strata of asthma severity (stratification process described later in Section 3.6.4).

Multiple studies additionally excluded individuals with Chronic Obstructive Pulmonary Disease (COPD) <sup>64,75,76</sup> or any chronic respiratory disease, excluding asthma <sup>68,117</sup>. Individuals with both asthma and COPD diagnoses may have either been misdiagnosed (most commonly asthma misdiagnosis <sup>125</sup>), or have a condition known as Asthma-COPD Overlap Syndrome (ACOS), which is known to result in higher rates of attacks than asthma alone and may be associated with different risk factors <sup>126,127</sup>.

## 2.4 Asthma Attack Ascertainment

There is no single consensus on the definition of asthma attacks; a joint report by the American Thoracic Society and the European Respiratory Society (2009; ATS/ERS) <sup>78</sup> reviewed asthma attack definitions in the literature, which included use of OCS (sometimes of at least 3 or 5 days in duration <sup>128</sup>), emergency room visits,

hospitalisations, unscheduled doctor visits, and/or decline in peak flow. Similarly, a 2017 review of EHR-specific asthma attack definitions found that prescriptions of OCS (with or without a concurrent primary or secondary care asthma-related encounter), SABA prescriptions, and secondary care events relating either to asthma attacks or to conditions including pneumonia in previously-diagnosed asthma patient, were all used

129.

The ATS/ERS Task Force recommend that for retrospective studies in EHR datasets, *severe exacerbations* (attacks) are defined as either a prescription of OCS, an asthma-related A&E department visit, or an asthma-related hospital admission <sup>78</sup>. The ascertainment of an asthma-related OCS prescription is described in Section 3.6.6.

Another important consideration is the time window in which we will look into the future and attempt to predict events, known as the *event horizon*. In this analysis, I aim to identify people who would benefit from further monitoring and health education interventions, and as such event horizons of between one week and 12 months (the recommended maximum time between asthma reviews) will be considered.

## **2.5 Chapter Summary**

In this chapter, I have reviewed the practical and ethical rationale for the use of EHRs in medical modelling studies and have reviewed the literature regarding population inclusion criteria and outcome ascertainment criteria from related studies.

## 3 A Critical Appraisal of the Predictive Value of Asthma Attack Risk Factors

In this chapter, I review my process for identifying, evaluating, and selecting the risk factors to be used in my asthma attack prediction model. There are a number of important considerations for the inclusion of risk factors in the model, including their effect, their changeability, and the feasibility of measuring them.

### 3.1 Feature Selection for Prediction Modelling

In order to build a classification model, we need to provide a set of features and the *ground truth* (the observed clinical outcome), and the model can learn to estimate how these features map onto the outcome. These features are pieces of information about the patient at a specific point in time (and generally, may also include non-time-varying features, e.g. birthplace). While we could provide any information to the model, the best predictions will come when the features provided all have some effect on the risk of the outcome: in our case, asthma attacks.

As such, it is necessary to construct a list of the risk factors, identified from the research literature and clinical input and rationale, which will be measured and recorded for each patient. While limiting our model to the assessment of known factors means that we are unable to appraise new features which contributes to a patient's risk, it is not feasible to extract every single piece of information about a patient's life. Furthermore, including too many features into the model, especially when there is no biological rationale for any effect, increases the risk of over-fitting (see Section 1.2).

It might seem intuitive that tracking a patient's current symptoms would indicate when the patient was declining and would take into account small changes in environment (like the fact that their neighbour bought a new cat, which they are allergic to). However, the clinical markers of asthma control (such as peak expiratory flow) and the occurrence of asthma attacks are often at a disconnect <sup>130</sup>. The biological

mechanisms responsible for asthma attacks may be different to those causing wheezes, coughs and chest tightness<sup>131</sup>. Indeed, some individuals are more prone to attacks than others, with past attack history being commonly found to be one of the strongest risk predictors for future attacks<sup>48,62,79</sup>.

When narrowing down a list of potential risk factors for inclusion, another consideration is whether that factor is *modifiable*. A *modifiable* risk factor is one which can be purposefully changed in some way, such as smoking status or weight. It is important to distinguish them from *non-modifiable* but *time-varying* risk factors, which change over time but cannot be controlled (such as age). Including enough time varying factors ensures that the model is sensitive to changes within the same person, for example capturing that a patient with severe allergic rhinitis (also known as hay fever) might be at higher risk during pollen season.

A final note is that not all risk factors are possible to extract from EHRs. For example, the results of clinical tests which are not routinely conducted will not be available for the majority of patients. Some data are routinely captured in other sources, such as the weather by the national meteorological office. This information could be incorporated into the training data for model building with relative ease, however it would be a more difficult task for a clinician to incorporate this information when they are evaluating a patient's risk themselves.

## 3.2 Introduction to Missing Data

In a design matrix with  $N$  samples and  $M$  features (as introduced in Section 1.2), cells in which there was no data available are known as missing data. The proportion of samples that had missing data for any one feature is known as the missingness.

Many of the features that are described in the following sections are extracted from primary care records, which are stored in the long format, described in Section 2.2.3.2. Rather than having explicitly left blank the cells of the design matrix, this corresponds to not having recorded an entry relating to that feature at a timepoint. Whether this

results in missing data for a feature depends on how often corresponding Read codes are recorded, as well as how features are derived over multiple previous primary care encounters. For example, a feature might be derived as whether a status has been recorded ever, in the past year, or on the day of the encounter itself.

Many statistical learning algorithms require data to have no missing values<sup>132</sup>. There are different approaches to overcoming missingness, of which the simplest is complete case analysis: excluding any samples with any missing data. As well as reducing the sample size, and thus the power to detect any real and valid associations or patterns, record deletion may introduce bias to the analysis, depending on the type of missingness. Missing data can be categorised into 4 levels: Structurally Missing, Missing Completely At Random (MCAR), Missing At Random (MAR), or Missing Not At Random (MNAR). Structurally missing occurs when the value should not exist. For example, if I record the age at diagnosis of asthma for a group of people, those without a diagnosis of asthma should not have a value here.

When data are MCAR, the fact that the data are missing is independent of both the observed (other features in the design matrix) and unobserved (the value itself, and other missing values) data. When data are MCAR, the missing values cannot be reasonably estimated using statistical methods, or common sense. For example, in a clinical trial, if whether a patient was in the placebo or intervention arm was missing, this should not be possible to ascertain from their baseline characteristics.

MAR data, in contrast, may be possible to predict using the data which was available for a sample. For example, if someone had been smoking for more than five years, we could reasonably infer they were adult rather than paediatric.

Finally, when data are MNAR, the unknown values themselves are associated with the likelihood that a value was missing. For example, if a primary care practitioner had never recorded whether someone had been diagnosed with cancer, it is mostly likely not the case. Due to the multitude of Read codes available, naturally, irrelevant data is likely to be omitted from the record. A diagnostic test or examination which is



costly to conduct (time, resources, financial expense) is unlikely to be conducted unless the practitioner thinks the result will provide some value to the patient's care. Thus, certain untested results may be inferred by the nature of being untested.

As well as the complete case approach described above, it is possible to overcome issues with missing data by replacing the cell with some value, a method known as *imputation*. When data is MCAR, a simple imputation such as the mean value (if normally distributed) across the sample will not introduce bias. When data is MAR, values to be imputed can be estimated using the non-missing data and statistical learning algorithms, such as those discussed in Section 5.3. Multiple imputation creates multiple copies of the dataset containing missing values, imputes the values with some random error included, and combines the results of the imputed datasets<sup>132</sup>.

Finally, in the missing indicator method, missing values are not imputed at all. Instead, a new category is added to categorical data ('missing') or a simple imputation can be conducted in continuous data, alongside a new feature flagging the samples in which imputation was conducted<sup>133</sup>. This method is known to bias the estimates for confounders in parametric methods<sup>133</sup>, but may even strengthen the predictive ability of the model, depending on the mechanism of missingness<sup>134</sup>.

### **3.3 Asthma Attack Risk Factor Inclusion and Exclusion Criteria**

In this section, I define the criteria that I applied to risk factors for asthma attacks identified in the literature, in order to determine which should be included in my model. Studies conducted only in children were included, however the weight of evidence relating to the adult population was carefully reviewed.

Risk factors discussed in two systematic reviews (Loymans *et al.*<sup>54</sup> and Buelo *et al.*<sup>79</sup>), and the 2019 guidelines jointly written by British Thoracic Society (BTS) and the

Scottish Intercollegiate Guidelines Network (SIGN) <sup>135</sup>, were systematically compared against the selection criteria described in Table 3.1, using supplementary literature for the review where necessary. Evidence is presented from the literature of an effect on the incidence of asthma attacks, the feasibility of extraction from EHRs, and the time-varying nature, of each identified risk factor. Furthermore, I detail the process by which the feature should be extracted and processed from EHRs, including missing data handling, Clinical code lists (such as Read Codes), categorisation, and outlier detection. Note that all Read Code lists provided in this thesis were compiled by the ALHS study team.

In line with the recommendations made by Goldstein *et al.* <sup>84</sup> in their review of EHR-based risk prediction models, I made sure to use a large number of time-varying features, and to carefully consider missing data.

Table 3.1: Asthma attack risk factor inclusion and exclusion criteria

Criteria	Inclusion	Exclusion
Feasibility	<ul style="list-style-type: none"> <li>The information is captured in one of the available data sources.</li> <li>The measurement is conducted routinely.</li> </ul>	<ul style="list-style-type: none"> <li>The information requires complex natural language processing to extract.</li> </ul>
Effect	<ul style="list-style-type: none"> <li>There is some evidence in the literature of either a protective or harmful effect on the incidence of asthma attacks (either in adults or by a mechanism that is likely valid for adults and children).</li> </ul>	<ul style="list-style-type: none"> <li>Any effect observed in the literature has been found exclusively in a child population.</li> </ul>
Changeability	<ul style="list-style-type: none"> <li>If time-varying, the time scale used for calculation of the feature has been clearly defined in the literature.</li> </ul>	N/A

## 3.4 Demographic Risk Factors

In this first section, I review four demographic risk factors identified in the previous reviews: age, sex, socioeconomic status, and ethnicity.

### 3.4.1 Current Age

Increasing age in adults (18 and over) may be associated with a lower risk of acute asthma attacks (hospitalisation or A&E presentation) <sup>115,120,136</sup>, but a higher rate of subacute episodes <sup>75</sup>. Age is captured in primary care patient registry data (Section 2.2.3.1) and was included in the analysis as a time-varying feature. Those with missing age were excluded from my analysis.

### 3.4.2 Sex

Adult women with asthma are often found to have a higher risk of asthma attacks than men with asthma <sup>75,77,120,136,137</sup>. More adult women than adult men present for asthma to primary care <sup>138</sup> and A&E <sup>139–142</sup>, and a higher proportion of A&E presentations result in hospital admission for women <sup>139,143,144</sup>. Women are also more likely than men to return to A&E for asthma after discharge <sup>140,145</sup>.

A study of prepubescent children with asthma found that boys had a higher risk of asthma attacks than girls <sup>146</sup>, in line with findings that boys have more primary care consultations, A&E presentations, and hospital admissions than girls <sup>138,140,142,146,147</sup>. Significant interactions between age and sex on asthma attack risk have also been identified, with higher incidence in male than female children and adolescents, but higher incidence in adult women than adult men <sup>138,147,148</sup>. One explanation for this is puberty, and the changing levels of sex hormones. This hypothesis is supported by further studies demonstrating changes in exacerbation risk at certain phases of the menstrual cycle, or before and after events such as menarche, pregnancy and menopause <sup>147,149–151</sup>.

Sex, like age, is captured in primary care patient details data, and those with missing data were excluded from the study.

### 3.4.3 Socioeconomic Status

Several studies found that risk factors related to socioeconomic status were predictive of asthma attacks. In the United States of American (USA), higher rates of asthma-related hospitalisation in adults have been associated with lower income <sup>72</sup>, Medicaid health coverage status <sup>77</sup>, and lower levels of education <sup>115</sup>. Recent studies in England and Wales have found that deprivation was associated with higher rates of asthma attacks <sup>121,152,153</sup>, and there is some evidence of a non-linear interaction with age <sup>152</sup>.

Deprivation is captured in the patient registry dataset by the composite measure SIMD <sup>112</sup>, with changes in address represented in updated registration information, and thus time-varying. Missingness for social deprivation was coded into a new category: *missing*.

### 3.4.4 Ethnicity

A study from the USA found that non-White adults with asthma had higher risk of acute attacks than White adults, but without any controlling for socioeconomic factors <sup>120</sup>. Another USA study, however, found that in children with asthma and Medicaid health insurance, Black and Hispanic children were at a higher risk of A&E presentation than White and Asian children <sup>146</sup>. The effect persisted after additional county-level socioeconomic features were included, however no individual-level features were tested. Other USA studies have found similar increases in risk of asthma attacks for Asian, Black, Indigenous, and Hispanic children, compared to White or non-Black children, controlling for confounding factors such as deprivation and parental education <sup>71,79,154–156</sup>.

A Scottish study found that Pakistani, Indian and other South Asian people had 20-50% higher rates of asthma-related hospitalisation than White Scots, and Chinese people had 30-40% lower rates <sup>157</sup>. While disaggregating genetic from sociological effects in the incidence of asthma attacks is very difficult, these findings indicate that including ethnicity where possible may help to improve risk prediction.

Unfortunately, while ethnicity is sometimes recorded in primary care data, it often contains high levels of missingness <sup>157</sup>, and as such was not included in my model.

## 3.5 Risk Factors Relating to Lifestyle

In this section, I review two lifestyle-based risk factors: smoking status and obesity, although the latter may also be caused at least in part by genetics rather than solely lifestyle.

### 3.5.1 Smoking

Three of the highlighted risk prediction model studies identified smoking as a risk factor for asthma attacks <sup>67,75,117</sup>, with between 17% and 77% increased odds compared to non-smokers, although Loymans *et al.* <sup>67</sup> observed some confounding with lung function and Fractional Exhaled Nitric Oxide (FENO). This increased risk is in line with previous findings that smokers have higher risk of inpatient admission for asthma attack than non-smokers and former smokers <sup>67,75,117,158</sup>. There is also some evidence of an increased risk of asthma attacks with exposure to environmental tobacco smoke <sup>79</sup>, such as from a cohabitant.

Smoking status can be recorded in primary care data, using Read Codes. There is no gold standard approach for categorising smoking status from Read Codes, and thus previous studies have often used different code lists <sup>159–161</sup>. Similarly, the process for mapping Read Codes to smoking status category is not always obvious, such as the decision rule for the presence of the code “137X.”: *Cigarette consumption*. Atkinson *et al.* described a complex decision tree for a seven-level categorisation (including distinguishing between smokers and relapsed smokers), however the validity of the categorisation was low when compared to GP manual classifying <sup>161</sup>.

The list of Read Codes used herein, and corresponding 3-level (current, former, and non-smoker) categorisation rules, is presented in Appendix B. They are primarily the codes and simplified decision rules presented by Atkinson *et al.* <sup>161</sup>, however only the Version 2 Read Codes were included. Those with unknown smoking status were

coded as non-smokers <sup>162,163</sup>. Note that if a previous record indicated that an individual was a smoker, and a later record indicated that the individual had never smoked, using the most recent record results in ignoring this inconsistency. Smoking status is not static, and was revised upon a recorded change.

### 3.5.2 Obesity

Compared to being in the normal BMI range, being overweight or obese is associated with an increased risk of asthma attacks <sup>164–167</sup>. Luo *et al.* <sup>116</sup> and Bateman *et al.* <sup>66</sup> used BMI as a continuous feature, with the latter finding a 10% increased risk of a severe exacerbation per increase in kg/m<sup>2</sup>. Blakey *et al.* <sup>75</sup> used categorical, with 16% higher odds of an attack for those with BMI between 25 and 30 (overweight) and 27% increase for those with BMI over 30 (obese), compared to normal BMI (18.5-24.9). There was no significant difference between normal BMI and being underweight (<18.5 BMI) or having unknown BMI. Miller *et al.* <sup>120</sup> used BMI as a binary feature, dichotomising around the higher threshold of 35kg/m<sup>2</sup>.

In my analysis, obesity was a binary flag indicating that an individual's most recent BMI recording was greater than or equal to 30 kg/m<sup>2</sup>, calculated using records of continuous BMI values, dichotomous obesity codes, or BMI calculated from recent height and weight records (Appendix B). If no BMI could be calculated from available records, an individual was considered most-likely non-obese <sup>168</sup>.

A study in the USA <sup>165</sup> compared exacerbation rates by season in obese, overweight and normal weight individuals, and found that there was an interaction between season and BMI such that the increased risk from high BMI had a higher effect in autumn and winter (14-16% in spring and summer, compared to 34-41% in autumn and winter). As such, and also to be discussed further in Section 3.7.2, I will also be including calendar month as a risk factor.

## 3.6 Asthma-Related Risk Factors

In this section, six risk factors relating to asthma presentation and management are reviewed: asthma symptom control, lung function, exhaled inflammatory markers, controller treatment, treatment adherence, and history of asthma attacks.

### 3.6.1 Asthma Symptom Control

Many of the studies include some marker of current symptoms and management, including the number of  $\beta$ -2 agonists prescribed (SABA and LABA) <sup>72,73,116</sup>, the mean daily reliever (self-reported) reliever actuations <sup>66,75</sup>, the ratio of reliever to controller medications prescribed <sup>146,169</sup>, and asthma control questionnaire scores <sup>66,67,70,124,170–172</sup>. Osborne *et al.* <sup>115</sup> also used more explicit survey data to create binary features flagging those with nightly nocturnal symptoms and symptoms impacting school/work attendance, which had risk ratios of 1.99 (95% CI = 1.40-2.80) and 1.45 (95% CI = 1.16-1.80), respectively. The presence of nocturnal symptoms was no longer statistically significant after controlling for lung function, but symptoms impacting school/work attendance remained in the model.

Haselkorn *et al.* <sup>71</sup> and Zeiger *et al.* <sup>173</sup> also used survey data, with the USA's National Heart, Lung, and Blood Institute (NHLBI) guidelines for asthma control categorisation <sup>174</sup>: a composite measure of symptoms, night-time awakenings, interference with normal activity, and SABA use.

Like the studies by Blakey *et al.* <sup>75</sup> and Bateman *et al.* <sup>66</sup>, I evaluated asthma control using the mean SABA (in micrograms) used per day during the last SABA prescription refill interval, estimated from the strength and volume prescribed and the number of days between the most recent and the preceding prescription. Additionally, a variable was included to flag any prescriptions of nebulised SABA in the last 90 days, indicating a probable short-term increase in symptoms.

### 3.6.2 Lung Function

*Spirometry* is a pulmonary function test, conducted in specialist care, that can measure Forced Vital Capacity (FVC; volume of exhalation) and Forced Expiratory Volume (FEV<sub>1</sub>; volume of the first second of exhalation). In patients with obstructed airflows, asthma can be diagnosed by measuring *bronchodilator reversibility* - change in spirometry results before and after inhaling 400µg of a bronchodilator such as salbutamol<sup>175</sup>. Pre-bronchodilator spirometry results can also be used as a measure of lung function. Peak Expiratory Flow Rate (PEFR) is an alternative lung function test, which can be measured using a simple, cheap to manufacture plastic device. For this reason, it is often distributed to asthma patients for home monitoring purposes, as well as being tested routinely in the primary care setting<sup>176</sup>. FEV<sub>1</sub> was used in several of the examined risk models<sup>66,67,115,124,172</sup>, and FVC in two<sup>120,171</sup>. Like UK studies by Blakey *et al.*<sup>75</sup> and Turner *et al.*<sup>68</sup>, I decided to use PEFR as it is recorded more frequently in UK primary care practice, where my prediction model is designed to be used, than other lung function measures. The Read Codes used to identify PEFR measurements are listed in Appendix B.

PEFR, FEV and FVC are all routinely standardised by comparing results to the expected values based on age, height and sex<sup>177</sup>, or the best recorded historical value for that individual. The expected values to use as a reference vary depending on the source<sup>178</sup>, and many do not account for ethnic variation<sup>176</sup>. As such, the BTS/SIGN Guidelines<sup>135</sup> promote using an individual's personal best PEFR value as the reference.

On the date of any asthma consultation, the most recent PEFR measurement within seven days was converted to a categorical percentage of their previously recorded (any historical measurement) maximum (>90%, 80-90%, 70-80%, or less than 70%), or missing if there were no recorded values in the last seven days. Typically, this measurement will have taken place during the reference consultation itself, however if a measurement had been taken at a recently occurring consultation, the evaluation may not be repeated.



### 3.6.3 Exhaled Inflammatory Markers

Exhaled Inflammatory Markers (EIMs) are signals of an inflammatory immune response which can be detected from the analysis of Exhaled Breath Condensate (EBC). Robroeks *et al.*<sup>70</sup> found that concentration of the protein Interleukin-5 (IL-5) and the acidity of EBC were both significant predictors of asthma attacks in univariate analyses, and that IL-5 remained significant in multivariate analysis. Furthermore, Robroeks *et al.*<sup>70</sup> demonstrated that increased IL-5 was detectable a month prior to exacerbation, unlike FENO, which only raised at the start of the attack and within 3 days had on average returned to its value three months prior to the attack. When Van Vliet *et al.*<sup>69</sup> attempted to replicate the findings of this study in a further dataset, they found that none of the EIMs (in isolation or together) were able to predict asthma attacks. While Sato *et al.*<sup>124</sup> included FENO in their decision tree to improve the specificity of the model, its inclusion decreased the overall performance, as measured by the Area Under the Receiver Operator Curve (AUC; detailed later in Section 5.4.1).

Further studies included in the BTS/SIGN Guidelines evidence base demonstrate similarly unclear effects<sup>179–181</sup>. EIMs are also not routinely measured in primary care, and so were unlikely to be available for a sufficiently large proportion of the population to add any predictive value to the model. Indeed, the BTS/SIGN guidelines<sup>135</sup> discourage the use of FENO testing except in specialist asthma clinics. As such, I decided not to include any EIMs as risk factors in the model.

### 3.6.4 Controller Treatment Intensity and Severity

Asthma severity can be estimated by examining the prescribed medications used to control symptoms. There are several established treatment classifications used in the UK, including the GINA<sup>35</sup> and BTS Treatment Steps<sup>135</sup>. The differences between the two classifications of treatment steps (2019 editions) are detailed in Table 3.2. Note that the GINA guidelines<sup>35</sup> also state that the preferred reliever option across all treatment steps is a low-strength as-needed combination ICS+LABA inhaler (specifically formoterol). Both Turner *et al.*<sup>68</sup> and Bateman *et al.*<sup>66</sup> used the GINA treatment management steps in their analysis, while Price *et al.*<sup>117</sup> used the BTS treatment steps.

Table 3.2: Comparison of Global Initiative for Asthma (GINA) and British Thoracic Society (BTS) 2019 asthma treatment recommendations

Step	GINA Guidelines <sup>35</sup>	BTS Guidelines <sup>135</sup>
1	As-needed Low-Strength ICS + LABA, As needed Low-Strength ICS	As-needed Low-Strength ICS
2	Low-Strength ICS, As-needed Low-Strength ICS + LABA, LTRA, As needed Low-Strength ICS	Low-Strength ICS
3	Low-Strength ICS + LABA, Medium-Strength ICS, Low-Strength ICS + LTRA	Low-Strength ICS + LABA
4	Medium-Strength ICS + LABA, High-Strength ICS, Medium-Strength ICS + LTRA, Medium-Strength ICS + add-on therapy	Medium-Strength ICS, Low-Strength ICS + LTRA, Medium-Strength ICS + LABA
5	High-Strength ICS + LABA, High-Strength ICS + LABA + add-on therapy, Medium-Strength ICS + LABA + OCS, High-Strength ICS + OCS, Medium-Strength ICS + LTRA + OCS, Medium-Strength ICS + add-on therapy + OCS	Medium-Strength ICS + LTRA, Medium-Strength ICS + LTRA + add-on therapy, Medium-Strength ICS + LTRA + LABA, Medium-Strength ICS + LTRA + LABA + add-on therapy, High-Strength ICS, High-Strength ICS + LTRA, High-Strength ICS + LTRA + add-on therapy, High-Strength ICS + LABA, High-Strength ICS + LABA + add-on therapy, High-Strength ICS + LTRA + LABA, High-Strength ICS + LTRA + LABA + add-on therapy

Other studies used custom severity indicators in their studies. In their study of asthma attacks in those with severe asthma, Miller *et al.* <sup>120</sup> identified users of nebulised ipratropium bromide as a proxy for any nebuliser use, and found that it was a mild but significant indicator of later hospitalisation or A&E presentation. They additionally included current diagnoses of cataracts, which may have been used as an indicator of prolonged use of high-strength steroids <sup>182</sup>.

Schatz *et al.* <sup>136</sup> found that individuals on a sustained *high intensity* treatment regimen (high-strength ICS+LABA) had an increased risk of asthma attacks over those who were only prescribed the high intensity regimen in the index year, but no higher risk of hospitalisation.

While treatment intensity is often considered a proxy for asthma severity, the treatment regimen in itself also independently has some predictive value. Price *et al.* <sup>117</sup> observed a non-linear effect, with BTS treatment steps 0 and 1 (low-strength ICS only) resulting in higher odds of asthma attack than with Step 2 (in which LABA is added). Samuels-Kalow *et al.* <sup>183</sup> found that those without controller medication prescribed had 4.43 times higher odds of high emergency department utilisation. Similarly, in the USA, Grana *et al.* <sup>77</sup> defined treatment severity using dispensed, rather than prescribed, medications: combining both the quantity claimed and the medications themselves in their categorisation. They found that the lack of any pharmacy plan had 20% higher odds of an asthma attack requiring hospitalisation than even those requiring a single course of OCS (or multiple bursts with duration under 28 days).

In my analysis I used the 2019 BTS 5-step treatment classification, updated at any change in regimen. The BTS classification was chosen over the GINA classification because there were fewer treatment options at the lower steps, and some of the same treatment options at higher BTS step were considered lower in the GINA guidelines. I hypothesised that more distinction between the lower steps, at which the majority of patients are treated, would improve the predictive ability of this feature. The ascertainment of the patient treatment steps is described fully in Appendix C, as prescribers may deviate from the recommended treatment steps when appropriate.

Note that treatment intermissions of longer than 120 days were coded as treatment step *zero*, in order to prevent discontinued treatments carrying over. Additionally, the number of asthma controller medications dispensed in the previous calendar year was recorded.

### 3.6.5 Medication Adherence

#### **DEFINITION: ADHERENCE**

*“The extent to which a person’s behaviour taking medication, following a diet, and/or executing lifestyle changes, corresponds with agreed recommendations from a health care provider”*

- World Health Organization <sup>184</sup>

Non-adherence to asthma controller treatments has been repeatedly highlighted as a major contributing factor to excess mortality and life-threatening asthma attacks. A 2015 systematic review and meta-analysis by Engelkes *et al.* <sup>185</sup> found that, in both adults and children, poor adherence was associated with higher rates of exacerbations, across varying study designs, adherence definitions, and asthma attack definitions. For example, a study of almost 100,000 individuals <sup>186</sup> found a 14% reduction in odds of asthma emergency department presentation or hospitalisation between those above and below the 75<sup>th</sup> percentile of adherence, as measured by the Medication Possession Ratio (MPR). Williams *et al.* <sup>76</sup> estimated that one in four reported asthma attacks could have been attributed to poor adherence. A recent study by Chongmelaxme *et al.* <sup>187</sup> also found that risk was reduced when comparing moderate and poor adherence, as well as good and moderate adherence.

Four of the studies examined in this section found significant associations using some measure to approximate medication adherence. Luo *et al.* <sup>116</sup>, Schatz *et al.* <sup>72</sup>, and Lieu *et al.* <sup>169</sup> included simple counts of the number of controller medications dispensed during the study, but without any observation of the number of units that were *prescribed*. As such, it is not possible to disambiguate lower-strength treatments (likely associated with lower risk) and poor adherence (higher risk). Blakey *et al.* <sup>75</sup>

used the MPR but found that lower adherence was associated with a lower risk of attack. They speculated that this may have been a result on confounding with asthma severity (individuals with milder asthma self-managing their treatment successfully), but this was not investigated, as asthma severity was not included as a covariate.

Adherence is rarely explicitly captured in EHRs, so its estimation requires careful consideration. When it is captured it is reported directly by PROMs, such as standardised questionnaires and psychometric scales <sup>188</sup>. More commonly, however, adherence is estimated from prescribing and dispensing records, using some function of the expected and observed time between subsequent prescriptions, based on the quantity of medication prescribed. Many methods of measuring adherence from EHRs have been defined, however no gold standard has been proposed, and very few studies have attempted to critique the measures <sup>189–191</sup>. More information was required in order to determine the most appropriate methodology, and so a more in-depth review of the literature, as well as experiments in my data, was conducted (Chapter 4).

### 3.6.6 Previous Asthma Attacks and Unscheduled Care

As discussed in Section 3.1, previous asthma attacks are consistently identified as the strongest risk factor for subsequent attacks. Almost all of the examined risk model studies found a significant effect of this nature, but have quantified past history in different ways, including previous unscheduled asthma care (primary and/or emergency <sup>73,115,136,170,183</sup>), oral steroid bursts <sup>120,169</sup>, and combinations thereof <sup>67,68,72,74,75,116,117,148,192</sup> (including the ATS/ERS Taskforce attack ascertainment criteria <sup>78</sup>, introduced in Section 2.4), often matching the criteria they used to ascertain the outcome.

Miller *et al.* <sup>120</sup> found that in those with asthma attacks during the study, 16% had required intubation in the past, and less than 5% had not required any steroid bursts in the three months prior to baseline (23% had required three or more). Similarly, Loymans *et al.* <sup>67</sup> found 3.8 times higher odds of a severe asthma attack in those who had required oral steroids in the previous year. Grana *et al.* <sup>77</sup> looked at both the

timing, frequency, and location of unscheduled asthma care, and found that more frequent and more recent episodes, as well as those requiring hospital admission as opposed to A&E or primary care presentation, all increased the risk of subsequent attacks.

Episodes of secondary care for asthma (A&E presentations or inpatient admissions) are not automatically recorded in primary care data, although letters may be sent to the patient's GP in order to inform them (indicated in the Read Codes in Table 3.3). As such, EHR primary care data may or may not include Read Codes referring to secondary care encounters but will record any primary care prescriptions of OCS.

As this model will be deployed in the primary care setting, it is important to use only the data that GPs have access too. As such, if a previous secondary care encounter has not been coded in the patient notes, it cannot be used as a predictor of future risk. One limitation of this approach is that previous history of secondary care asthma encounters may be recorded in the free text medical notes, rather than in Read Codes, and are thus not able to be identified for this model. In practice, however, this means that the model may perform better in real life (when information is more certain) than in the testing data.

Table 3.3: Asthma attack Read Codes (Version 2)

Read Code (V2)	Term
H3301	Extrinsic asthma with asthma attack
H3311	Intrinsic asthma with asthma attack
H333.	Acute exacerbation of asthma
H33z000	Status asthmaticus NOS
H33z011	Severe asthma attack
H33z1	Asthma attack
633d.	Emergency asthma admission since last appointment
663m.	Asthma accident and emergency attendance since last visit
8H2P.	Emergency admission, asthma
663y.	Number of asthma exacerbations in past year

Note: NOS = Not Otherwise Specified

In my analysis, I included a binary flag for whether there had been more than one inhaled steroid prescription in either the previous or current (to date) calendar year. Prescriptions of prednisolone oral steroids (brand names listed in Appendix D) were identified as related to an asthma attack if they met the following conditions: 1) They were prescribed to someone with a diagnosis of asthma or receiving asthma treatment, 2) They were prescribed on the same day as an asthma-related consultation (identified by the presence of any Read Code listed in Appendices Appendix E or Appendix F on the same day), and 3) The total prescribed dosage was between 200 and 1000 mg (based on the British National Formulary Version 80 (BNF80) recommendation that 40-50mg daily be prescribed for asthma attacks, for at least 5 days <sup>193</sup>).

Additionally, the time since the last asthma attack was recorded (either steroid prescription or Read Code), categorised as 'one to two years', 'six months to one year', 'three to six months', 'one up to three months', 'in the last month', or 'none in the last two years'.

### **3.7 Other Comorbidities**

Finally, other comorbidities which may interact with asthma to increase risk of adverse outcomes are reviewed: eosinophilia, atopy, respiratory infections, and other chronic comorbidities.

#### **3.7.1 Eosinophilia**

Eosinophilia is defined as elevated counts of eosinophil white blood cells. A common threshold for defining an elevated count is  $\geq 400$  cells per  $\mu\text{L}$ , used by Blakey *et al.* <sup>75</sup>, Turner *et al.* <sup>68</sup>, and Price *et al.* <sup>117</sup>. Forno *et al.* <sup>74</sup> used the raw laboratory data, and thus had eosinophil count as a continuous feature (with a median of approximately 2.6  $\log_{10}\text{cells/mm}$ ).

Blakey *et al.*<sup>75</sup> found that eosinophil counts over  $\geq 400$  cells per  $\mu\text{L}$  resulted in 21% higher odds of an asthma attack., while Turner *et al.*<sup>68</sup> found 46% higher odds of attack and Price *et al.*<sup>117</sup> found a very similar estimate of 48% increased odds. Forno *et al.*<sup>74</sup> found a 2.7 times odds increase for each unit increase in  $\log_{10}$ cells/mm.

Blakey *et al.*<sup>75</sup> and Turner *et al.*<sup>68</sup> found that 66% and 61% (respectively) of their UK EHR study populations had at least one recorded eosinophil reading. Blakey *et al.*<sup>75</sup> also found that the *missing* group had significantly lower odds of an attack than those with  $< 400$  cells per  $\mu\text{L}$ , demonstrating that this information should not be considered missing at random.

McGrath *et al.*<sup>194</sup> have noted that eosinophilia can be either persistent or intermittent, even in the absence of ICS treatment, which typically targets eosinophil-specific inflammation<sup>195,196</sup> and thus may reduce eosinophil levels when used regularly. Eosinophilia should thus be considered time-varying.

Records of blood eosinophil counts were extracted from continuous Read Codes values (Appendix B). I dichotomised the recorded value at  $\geq 400$  cells per  $\mu\text{L}$ , or missing for those without any recorded measurements, using the most recent Read Code record at any time.

### 3.7.2 Atopy

#### **DEFINITION: ATOPY**

*“The genetic tendency to develop allergic diseases such as allergic rhinitis, asthma and atopic dermatitis (eczema). Atopy is typically associated with heightened immune responses to common allergens, especially inhaled allergens and food allergens.”*

- American Academy of Allergy, Asthma, and Immunology<sup>197</sup>



Atopy can be quantified by measuring Immunoglobulin E (IgE) levels from blood samples after exposure to individual allergens, from total IgE levels after exposure to a range of allergens, or by reaction to a Skin Prick Test (SPT). Diagnoses of allergic comorbidities, such as eczema and allergic rhinitis, may also be used as a proxy.

Blakey *et al.*<sup>75</sup> identified diagnoses of nasal polyps, anaphylaxis, active eczema and active rhinitis as risk factors for subsequent asthma attack incidence. All four diagnoses were significant in the multivariable model, signifying that such comorbidities may have distinct mechanisms to general atopy for increased risk of asthma attacks.

Loymans *et al.*<sup>67</sup> investigated several measures of atopy, including total IgE of over 100 kU/mL, chronic sinusitis (which they postulated may be related to nasal polyps), self-reported food allergy, or exposure to sensitised allergens (defined as positive-specific IgE titres to any of house dust mite, grass pollens, or birch pollens and/or IgE positivity to cat or dog, combined with ownership)<sup>67</sup>. The cut-off for labelling atopy from IgE levels is subjective, with Westerhof *et al.*<sup>198</sup> using a substantially higher value of total IgE over 350 kU/mL, for example. Sinusitis was the only significant predictor in the final model, controlling for spirometry and FENO, with an OR of 2.39 (95% CI = 1.11-5.14).

Luo *et al.*<sup>116</sup> counted the number of distinct recorded allergies, binary flags for food allergy, drug or material allergy, or environmental allergy, and prescriptions for allergies or nasal steroid sprays (treatment for rhinitis). Haselkorn *et al.*<sup>71</sup> also counted the number of allergic triggers, and found a consistent trend that more allergies increased risk of asthma attack.

Engelkes *et al.*<sup>148</sup> defined atopy as the diagnosis of either rhinitis or eczema (significant predictors in all but one cohort, with ORs between 1.09 and 2.07), but also included nasal polyps as a covariate (although it was found to be very rare in all of the cohorts they investigated, possibly contributing to its non-significance). Price *et al.*<sup>117</sup> used rhinitis and eczema separately, with 10% and 8% increased odds of attacks with diagnosis, respectively.

Finally, Forno *et al.*'s study <sup>74</sup> in children looked at family history of allergic conditions, as the data were taken from a genetic study with detailed family medical histories, however in primary care this is rarely recorded, especially for adults.

I used four separate allergic comorbidities as markers of atopy: diagnoses of rhinitis or eczema, and any history of nasal polyps or anaphylaxis (Read Codes listed in Appendix G). Additionally, I included prescriptions of a corticosteroid nasal sprays. The four allergic comorbidities and nasal sprays were categorised as 'never', 'in the past year', 'in the past 5 years' (not including in the past year), or 'longer than five years ago'.

The current month was included as an additional feature, allowing possible interactions between season and seasonal allergies to be detected in non-parametric analyses, as well as the interaction between season and BMI as discussed in Section 3.5.2. The allergic comorbidities and nasal sprays were categorised as 'never', 'in the past year', 'in the past five years' (not including in the past year), or 'longer than five years ago'.

### 3.7.3 Respiratory Infections

Respiratory infections have been found to be associated with higher odds of frequent exacerbations <sup>62</sup>. In adults presenting at hospital for asthma attacks, approximately 35-75% have a respiratory virus detected <sup>199–201</sup>. Rhinovirus is the most common agent involved in acute episodes of wheeze in adults, followed by coronaviruses <sup>200,202–205</sup>. One study in Singapore <sup>206</sup> found that in those presenting to hospital for asthma attacks, a higher percentage of the near fatal attacks (requiring ventilatory support) had a concurrent infection of a picornavirus (which includes rhinovirus) or adenovirus, although no difference was identified for influenza infections.

Luo *et al.* <sup>116</sup> only included diagnoses of bronchiolitis in the baseline year in their final model, a Lower Respiratory Tract Infection (LRTI) affecting the bronchioles. Bronchiolitis is commonly caused by Respiratory Syncytial Virus (RSV) or rhinovirus,

but almost exclusively occurs infants and children under the age of two <sup>207</sup>. Miller *et al.* <sup>120</sup> also only included one infection-related feature in their final model: any history of pneumonia, lung inflammation primarily affecting the alveoli.

Turner *et al.* <sup>68</sup> found that any LRTI diagnosis in the baseline year increased odds of asthma attack by 48%, but with very wide confidence intervals demonstrating low precision (95% CI = 4 - 214% odds increase). Price *et al.* <sup>117</sup> only considered LRTI diagnoses in the baseline year that resulted in antibiotic prescription, and also found a modest increase in risk (18% for one LRTI, and 28% for two or more).

LRTIs were flagged in the primary care dataset using the Read Codes listed in Appendix G. The maximum of the number of infections in the previous or current calendar year was used to estimate susceptibility to infection, and the time since the most recent infection was used to identify periods of recovery, categorised as 'In the past two weeks', 'Between two weeks and up to two months ago', 'Between two months and up to six months ago', 'Between six months and up to twelve months ago', 'Between one year and up to two years ago', or 'None in the last two years'.

In addition, although not relevant to my analysis due to the time-period of the datasets used herein, there is some conflicting evidence on whether infection from the recent pandemic novel coronavirus (SARS-CoV-2) disease (known as COVID-19) provoked worse clinical outcomes in asthma patients. Early studies found high prevalence of asthma in those hospitalised with COVID-19 in the UK <sup>208,209</sup>, however the evidence pertaining to elevated mortality risk remains inconclusive <sup>209-211</sup>.

### 3.7.4 Other Chronic Comorbidities

For many disease areas, including respiratory diseases, musculoskeletal diseases, and cardiac diseases, higher rates of comorbidity are seen in people with asthma than in the general population <sup>212-215</sup>. Some of these conditions interact with asthma in ways which increases risk of asthma attacks.

Luo *et al.*<sup>116</sup> identified diagnoses of diabetes without chronic complications and COPD as a risk factor for asthma attacks. Miller *et al.*<sup>120</sup> also found active diabetes to be a significant predictor of asthma attacks, in addition to cataracts (discussed in Section 3.6.4). Grana *et al.*<sup>77</sup> found significant increases in risk with diagnoses of ischaemic heart disease (OR = 1.64) and COPD (OR = 1.75).

Price *et al.*<sup>117</sup> found that diabetes and ischaemic heart disease were risk factors for hospitalisation (ORs of 1.64 and 1.53, respectively), and that anxiety/depression (associated with impaired immune response<sup>216</sup>) and diabetes were associated with increased odds of two or more attacks of any type (ORs 1.09 and 1.11, respectively). Finally, Engelkes *et al.*<sup>148</sup> identified Gastroesophageal Reflux Disease (GERD) as a significant risk factor in some, but not all, of the cohorts in their multicentre study.

As well as the aforementioned studies, Schatz *et al.*<sup>136</sup> also identified COPD as a risk factor for asthma attacks. However, as discussed in Section 2.3, those with COPD were excluded from our study in case the risk factors and mechanisms in those with Asthma-COPD Overlap Syndrome are distinct. As such, it could not be included as a risk factor.

In my risk prediction model, I included as binary variables (flags) the 17 diagnostic categories of the adapted Charlson Comorbidity Index (CCI; any history)<sup>217,218</sup>, and diagnoses of anxiety/depression or GERD. The diagnostic Read Codes are listed in Appendix G. Anxiety/depression and GERD features were categorised as ‘never’, ‘in the past year’, ‘in the past 5 years’ (not including in the past year), or ‘longer than five years ago’.

### **3.8 Other Notable Risk Factors**

There are established differences between the sexes in the severity of asthma (Section 3.4.2), which it has been hypothesised are related to sex hormones. This hypothesis is supported by observed changes in asthma severity around various female reproductive events such as menarche and menopause<sup>147,149–151</sup>. While sex

chromosomes cannot be changed, sex hormones can be modified by hormone therapies, for indications such as hormonal contraception, Hormone Replacement Therapy (HRT; menopause symptom relief), and sex reassignment therapy. A study in Scottish women taking oral contraceptives <sup>219</sup> found lower rates of asthma onset, but no difference in the incidence of wheezing attacks. Our recently published UK-wide study found that combined oestrogen/progestogen hormonal contraceptive (but not progestogen only contraceptive) use was associated with a lower risk of asthma attacks, increasing with duration of use <sup>220</sup>. I was not able to include hormonal therapies as a risk factor in my analysis, as the relevant Read Codes were not available in the ALHS dataset, and identification from prescribing records was beyond the scope of this body of work.

Various air pollutants, including carbon monoxide, ozone, PM<sub>2.5</sub> (atmosphere Particulate Matter of less than 2.5 micrometres diameter, less than 1/20<sup>th</sup> of the diameter of a human hair), and nitrogen dioxide, have been linked to increased risk of asthma attacks <sup>221–223</sup>. Stratification by age shows that older children and adolescents may be more vulnerable than adults <sup>221–223</sup>. While this information is not explicitly captured in EHRs, I will be including the local area identifier, as well as a measure of rurality, in the analysis which will account for some confounding due to pollution <sup>224</sup>.

Two studies found that the number of prescribers for an individual was associated with their risk of asthma attacks, however they were both conducted in children <sup>72,73</sup>. One of the two studies also explored the effect in adults, however they did not identify any significant relationship <sup>72</sup>. Additionally, the number of GPs an individual had seen about their asthma was evaluated by another adult study, however it did not contribute to the final decision tree <sup>169</sup>. As such, I did not include this information as a risk factor in my model.

Occupational exposures such as cleaning products, exhaust fumes, and animal products may increase daily symptoms and asthma attack risk <sup>225–227</sup>. Osborne *et al.* <sup>115</sup> also identified increased caffeine consumption as a mild risk factor for asthma

attacks, however this is not captured routinely in primary care. Neither exposure is recorded routinely in primary care data and thus was not included in the model.

There is strong evidence that respiratory infections are a serious risk factor for subsequent asthma attacks, as discussed in Section 3.7.3. As such, people with severe asthma are often encouraged to have the seasonal influenza vaccination <sup>228,229</sup>. While many studies have reported lower rates of asthma attacks in those who have had the seasonal influenza vaccination <sup>230</sup>, I have seen no evidence for any preventative mechanism other than reducing the rate of influenza infection. Indeed, only one study was identified that included a subgroup analysis of the effect of vaccination in those who did not have any influenza vaccination, and they found no significant change in asthma attack rates <sup>231</sup>. As such, the inclusion of respiratory infections as a predictive feature should be a more informative and reliably predictive feature than the vaccination itself, and thus the latter was not included in analyses. Another argument for the inclusion of influenza vaccination status is that it may be considered in part a proxy for higher health engagement, which was already captured in some measure through the feature counting the number of asthma consultations in the previous calendar year.

Luo *et al.* <sup>116</sup> also looked at a number of asthma attack risk factors which were not used in other studies' models, such as religion, primary language, marital status, vital signs, and time since diagnosis. Unfortunately, none of these risk factors were possible to ascertain in the ALHS dataset, as the relevant Read Codes were not included in the data extract. Guidance sought informally from clinical colleagues informed me that the low incidence of recording in primary care records meant that there would likely be high missingness if the additional codes were requested, and so I decided not to include them in my analysis. While time since diagnosis is not explicitly coded in primary care datasets, it can be determined using the first diagnosis code date for an individual. Unfortunately, this information is not always available (such as if the diagnosis occurred prior to the study start, or when diagnosis is not clearly and explicitly recorded) and as such this was also not included in the analysis.

### **3.9 Conclusions**

In this section, I have reviewed the evidence of effect on asthma attack incidence, feasibility of extraction from EHRs, and time-varying nature of identified potential risk factors, for inclusion in my risk prediction model.

In Table 3.4 I have listed the risk factors which were included in my risk prediction model, the method of extracting them, their duration of effect, and the format of the feature.

Table 3.4: Asthma attack risk prediction model risk factors

<b>Risk Factor</b>	<b>Extraction Method</b>	<b>Time</b>	<b>Feature Format</b>
Age	Difference in integer years between date of birth, from primary care patient registration dataset, and date of event	Time-varying, per record	Positive integer, no missing values allowed
Sex	Recorded in Primary Care, from primary care patient registration dataset	Constant	Categorical {'F', 'M'}, no missing values allowed
Socioeconomic Status	SIMD, from primary care patient registration dataset	Time-varying, updated at changes to patient registration	Categorical {1:5, missing}
Local Area Code	Nomenclature of Units for Territorial Statistics Level-3 (NUTS-3) codes, linked from the data zone (2001 version) the person was residing in at registration	Constant	Categorical {Not Listed, including missing}
Rurality	Scottish Government Urban Rural Classification Scale, from primary care patient registration dataset	Time-varying, updated at changes to patient registration	Categories {1:6, missing}
Smoking Status	Primary care Read Codes listed in Appendix B	Time-varying, most recent category.	Categories {'current', 'former', 'non-smoker'}
Reliever Medication Usage	SABA dosage prescribed divided by SABA refill interval, from primary care prescriptions	Time varying, most recent closed refill interval	Continuous positive real value, or zero if no previous (closed) SABA refill interval



<b>Risk Factor</b>	<b>Extraction Method</b>	<b>Time</b>	<b>Feature Format</b>
Peak Expiratory Flow	Primary care Read Codes listed in Appendix B. Standardised as the percentage of the best measurement to date (including that measurement)	Time-varying, most recent measurement in the previous 7 days	Categorical {'>90%', '80-90%', '70-80%', 'less than 70%', 'missing'}
BTS Step	Prescribed asthma controller medications, processed as detailed in Appendix C (not directly aligned with BTS Steps as presented in Table 3.2)	Time-varying, most recent treatment step estimated from prescriptions in the last 120 days	Positive integer in range [0,4]
Recent LRTI	Read Primary care Read Codes listed in Appendix G, more than one distinct event in previous calendar year or current calendar year	Time-varying, most recent event	Binary (1=Multiple recent LRTIs, 0 = One or fewer recent LRTIs)
Recent Asthma Encounters	Primary care Read Codes listed in Appendix E and Appendix F, more than one distinct event in previous calendar year or current calendar year	Time-varying, most recent event	Binary (1= Multiple recent Asthma Encounters, 0 = One or fewer recent Asthma Encounters)
Recent Steroid Prescriptions	Steroid treatments (identification described in Section 3.6.6), more than one distinct event in previous calendar year or current calendar year	Time-varying, most recent event	Binary (1= Multiple recent Steroid Prescriptions, 0 = One or fewer recent Steroid Prescriptions)
Number of asthma controller medications	Number of asthma controller medications (identification process discussed later in Section 4.2.2) in the previous calendar year	Time-varying, annual	Positive integer, no missing values allowed

<b>Risk Factor</b>	<b>Extraction Method</b>	<b>Time</b>	<b>Feature Format</b>
Time Since Last Asthma Attack (Recorded in Primary Care)	Primary care Read Codes listed in Table 3.3, difference between date of last event and current date	Time-varying, most recent event	Categorical {'one to two years', 'six months up to one year', 'three up to six months', 'one up to three months', 'in the last month' or 'none in the last two years'}
Adherence	Prescribed asthma controller medications, processing method to be determined	To be determined	To be determined
Eosinophilia	Read Primary care Read Codes listed in Appendix B, dichotomised the recorded value at $\geq 400$ cells per $\mu\text{L}$	Time-varying, most recent measurement	Categorical {' $\geq 400$ ', '<400', 'missing'}
Month	Calendar month of Event	Time-varying, per record	Categorical {'January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'December'}
Rhinitis Diagnosis	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code	Time-varying, per record	Categorical {'Never', 'In the past year', 'One up to five years ago', 'Longer than five years ago'}
Eczema Diagnosis	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code	Time-varying, per record	Categorical {'Never', 'In the past year', 'One up to five years ago', 'Longer than five years ago'}

<b>Risk Factor</b>	<b>Extraction Method</b>	<b>Time</b>	<b>Feature Format</b>
Anxiety/Depression Diagnosis	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code	Time-varying, per record	Categorical {'Never', 'In the past year', 'One up to five years ago', 'Longer than five years ago'}
Nasal Polyps Diagnosis	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code	Time-varying, per record	Categorical {'Never', 'In the past year', 'One up to five years ago', 'Longer than five years ago'}
Anaphylaxis Diagnosis	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code	Time-varying, per record	Categorical {'Never', 'In the past year', 'One up to five years ago', 'Longer than five years ago'}
GERD Diagnosis	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code	Time-varying, per record	Categorical {'Never', 'In the past year', 'One up to five years ago', 'Longer than five years ago'}
Corticosteroid Nasal Sprays	Spray formulations with drug name "mometasone", "fluticasone", "beclometasone" or "budesonide"	Time-varying, annual	Categorical {'Never', 'In the past year', 'One up to five years ago', 'Longer than five years ago'}

Risk Factor	Extraction Method	Time	Feature Format
Time since last LRTI	Read Primary care Read Codes listed in Appendix G	Time-varying, annual	Categorical {'In the past two weeks', 'Between two weeks and up to two months ago', 'Between two months and up to six months ago', 'Between six months and up to twelve months ago', 'Between one year and up to two years ago', 'None in the last two years'}
Nebulised SABA	Prescriptions for any nebulised SABA in the last 90 days	Time-varying, most recent measurement	Binary (1=Prescription in the last 90 days, 0 = No prescription in the last 90 days)
Obesity	Read Primary care Read Codes listed in Appendix B. Continuous BMI values dichotomised at 30 kg/m <sup>2</sup> . Categorical BMI values dichotomised as obese or non-obese. Continuous BMI calculated from height and weight	Time-varying, most recent measurement	Binary (1=Obese, 0 = Non-Obese)
AIDS	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= AIDS, 0 = No AIDS)
Cancer	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Cancer, 0 = No Cancer)

<b>Risk Factor</b>	<b>Extraction Method</b>	<b>Time</b>	<b>Feature Format</b>
Cerebrovascular disease	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Cerebrovascular disease, 0 = No Cerebrovascular disease)
Chronic pulmonary disease	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Chronic pulmonary disease, 0 = No Chronic pulmonary disease)
Congestive heart disease	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Congestive heart disease, 0 = No Congestive heart disease)
Dementia	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Dementia, 0 = No Dementia)
Diabetes (without complications)	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Diabetes, 0 = No Diabetes)
Diabetes with complications	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Diabetes with complications, 0 = No Diabetes with complications)
Hemiplegia	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Hemiplegia, 0 = No Hemiplegia)
Metastatic tumour	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Metastatic tumour, 0 = No Metastatic tumour)

<b>Risk Factor</b>	<b>Extraction Method</b>	<b>Time</b>	<b>Feature Format</b>
Mild liver disease	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Mild liver disease, 0 = No Mild liver disease)
Moderate liver disease	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Moderate liver disease, 0 = No Moderate liver disease)
Myocardial infarction	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Myocardial infarction, 0 = No Myocardial infarction)
Peptic ulcer disease	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Peptic ulcer disease, 0 = No Peptic ulcer disease)
Peripheral vascular disease	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Peripheral vascular disease, 0 = No Peripheral vascular disease)
Renal disease	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Renal disease, 0 = No Renal disease)
Rheumatological disease	Read Primary care Read Codes listed in Appendix G, most recent diagnosis code within the last year	Time-varying, updated at first observed diagnosis code	Binary (1= Rheumatological disease, 0 = No Rheumatological disease)

## **4 Comparison of Pharmacy-Based Measures of Asthma Controller Medication Adherence in the Asthma Learning Healthcare System Data**

As discussed in Section 3.6.5, adherence has been identified before as an important risk factor in the prediction of asthma attacks. Adherence is not assessed routinely in primary care, and while it can be estimated using prescribing records, the methods require careful consideration. In this chapter, I conducted an in-depth review of the methods described in the literature and conducted experiments in the ALHS data to determine the most appropriate approach for my purposes.

### **4.1 Background**

Medication adherence is defined as the process by which a patient takes their medication, in accordance to the regimen agreed to with their healthcare provider <sup>184</sup>. Non-adherence to treatment for chronic diseases is high <sup>184</sup>, and is a substantial impediment to treatment effectiveness <sup>63,64,185,232–234</sup>. Furthermore, subsequent poor clinical outcomes may lead to unnecessary dose escalation and/or additional treatment to control symptoms, itself resulting in the onset of avoidable side-effects <sup>17,19,22,235,236</sup>.

Estimates of non-adherence incidence are crucial for approximating associated costs (both financial and quality of life) <sup>237–241</sup>, identifying the most at-risk patients <sup>242–244</sup>, and accurately appraising the effectiveness of new treatments <sup>245–247</sup>. Prescription records provide opportunity to estimate adherence cost-effectively and at scale, although of course not all aspects of adherence can be measured (such as whether the medication is being taken once collected). While many methods have been proposed and tested <sup>248</sup>, there is currently no consensus on what should be considered the gold standard, both in terms of clinical relevance and utility in statistical modelling <sup>191,249–251</sup>.

Boissel *et al.*<sup>252</sup> summarised the challenges regarding medication adherence neatly: *“One challenge in studying varying [adherence] is that no single feature can express it.”* Despite this, there are very few studies that have reported more than one measure of adherence, by which to compare and critique approaches. For example, Engelkes *et al.*<sup>185</sup> reported that in studies that investigated adherence and risk of asthma attacks using EHRs, only 1 in 17 computed more than one adherence measure.

Crucially, studies which have evaluated multiple measures have often demonstrated their non-equivalence in associations with clinical outcomes. Williams *et al.*<sup>253</sup> compared two measures, one of which measured the amount of medication prescribed over the duration, and one of which measured the amount of time without medication available, and found that the latter was consistently more strongly correlated with clinical outcomes, including hospitalisations and use of emergency treatment. Similarly, Ismaila *et al.* found that the proportion of time with medication in supply was less strongly associated with severe outcomes (such as intensive care unit admission and intubation) than the continuous renewal of prescriptions without a gap of more than 30 days<sup>254</sup>.

Electronic monitoring devices (EMDs) enable the real-time tracking of inhaler use, by means of a small electronic chip fitted onto an inhaler, which records the date and time of each dose taken<sup>255</sup>. In a recent investigation, outwith my thesis, of patterns of adherence within an EMD dataset, I compared five approaches to summarising the longitudinal data over a six-month period<sup>256,257</sup>. I demonstrated that the simple measure of percentage of doses taken was the best single adherence measure at capturing the diversity of medication taking patterns. However, it failed to distinguish between those with long intermissions of treatment and those with frequently missed single doses, the latter of which has less severe consequences on asthma control. As Alleman *et al.*<sup>258</sup> noted:

*“Some temporal sequences of deviations from the prescribed regimen may be more detrimental to treatment effectiveness and safety compared to others.”*



While this EMD dataset provided extremely granular medication taking data, the observed limitations of using a single aggregate measure are only exemplified in EHR data, where adherence measures are far less precise approximations of the underlying behaviour.

Collectively, these studies highlight that the method of measuring medication non-adherence is not trivial, and that further work is needed in order to guide best practice. As highlighted in my recent review of adherence measurement and reporting in two respiratory conditions (asthma and tuberculosis), it is essential to understand how longitudinal data aggregation may mask clinically relevant changes in medication taking behaviour<sup>255</sup>. In this chapter, I critically appraise a variety of different measures of medication adherence in a UK EHR dataset, in order to guide the selection of the most appropriate measure for my risk prediction model. As a chronic condition with high prevalence<sup>7,8</sup> and high rates of non-adherence<sup>259–265</sup>, asthma is in many respects an ideal case study to highlight some of the specific aspects that researchers must consider in other diseases.

## 4.2 Methods

### 4.2.1 Prescription-Based Adherence Measures

Two single refill interval measures were used herein; the Continuous Single interval measures of medication Availability (CSA; Equation (4.1) and Gaps (CSG; Equation (4.2)<sup>189</sup>. Both measures use the *supply days obtained at refill* (how many days the prescription should last for if taken as prescribed; 28 days in example Figure 4.1) and the *refill interval duration* (the time between this prescription and the next; 31 days in example Figure 4.1).

Continuous single-interval measure of medication acquisition (CSA):

$$\frac{\text{Supply days obtained at refill}}{\text{Refill interval duration}} \quad (4.1)$$

Continuous single-interval measure of medication gaps (CSG):

$$\frac{\text{Refill interval duration} - \text{Supply days obtained at refill}}{\text{Refill interval duration}} \quad (4.2)$$

M	T	W	T	F	S	S
{28}						
			{28}			

Figure 4.1: Prescription calendar example, with 28 days of supply obtained in a 31-day refill interval

Note: {28} denotes 28 days of supply obtained

Vollmer *et al.*<sup>190</sup> defined eight adherence measures which used multiple refills (known as the Continuous Multiple-interval measures of medication Availability, or CMAs), labelled CMA1 to CMA8 (summarised in Table 4.1, with illustrated examples provided in Appendix H). CMA measures 1 through 4 are explicitly measures of medication acquisition rather than medication taking, as they use the amount of medication obtained over a period in the numerator, rather than any calculations requiring acknowledgement of the spacing and gaps in availability. This makes them relatively simply to calculate but results in an overly simplified reflection of the observed time series. In contrast, CMA5 to CMA8 incorporate the timing of the prescriptions (all at once, or evenly spaced) within the observation period to better detect gaps in medication availability. They can accordingly be considered Continuous Multiple-interval measures of medication Gaps, or CMGs. They are inhibited from detecting over-supply of medications, which mark that a patient is using their medication at more regular intervals or dosages than they had been instructed<sup>266</sup>.

Table 4.1: Start and end of analysis window within observation period for continuous, multiple-interval, measures of medication availability and gaps

Measure	Start of Window	End of Window	Derivation
CMA1	Day of first prescription in observation period	Day before final prescription in observation period	$\frac{\text{Supply days obtained in window}}{\text{Window duration}} \quad (4.3)$
CMA2	Day of first prescription	End of observation period	$\frac{\text{Supply days obtained in window}}{\text{Window duration}} \quad (4.4)$
CMA3	Day of first prescription	Day before final prescription	$\frac{\text{Supply days obtained in window}}{\text{Window duration}} \quad (4.5)$
CMA4	Day of first prescription	End of observation period	$\frac{\text{Supply days obtained in window}}{\text{Window duration}} \quad (4.6)$
CMA5	Day of first prescription in observation period	Day before final prescription in observation period	$\frac{\text{Days with medication available in window}}{\text{Window duration}} \quad (4.7)$
CMA6	Day of first prescription	End of observation period	$\frac{\text{Supply days obtained in window}}{\text{Window duration}} \quad (4.8)$

CMA7	Start of observation period	End of observation period	$\frac{\text{Supply days obtained in window}}{\text{Window duration}}$	(4.9)
CMA8	Day that supply that was available at the start of observation period theoretically exhausted	End of observation period	$\frac{\text{Days with medication available in window}}{\text{Window duration}}$	(4.10)

Of the first four CMA measures, only two (CMA1 and CMA3) were designed for patient-level analysis. Capping an individual's adherence estimate at 1 is used when averaging adherence across a population, as it ensures that one patient's over-supply does not balance out another's under-supply. CMA2 is calculated in the same manner as CMA1, but with the terminal gap included. As such, observation periods containing only a single prescription can still produce an estimate of adherence, unlike CMA1. However, as both CMA1 and CMA2 sum the amount of medication, rather than the number of days with medication, obtaining a large amount of medication near the end of the follow-up will result in inflated estimates. It is primarily used in cases where a single prescription during the observation period is likely, and thus enables adherence estimates for more of the population. As Vollmer et al. themselves note, in a chronic condition such as asthma this is not expected so long as the observation period is long enough<sup>190</sup>, relative to the expected duration of supply dispensed (typically 1-2 months).

Of the gap-related measures, CMA6 was excluded from analyses as it relates to CMA5 in the same way that CMA2 relates to CMA1; it extends the calculation period past the final prescription to the end of the person's follow-up. CMA7 is a simplified version of CMA8, which does not exclude the period in which the remaining supply from the last dispensing prior to the start of observation was being used. The two measures are expected to be markedly different only when an intervention of some variety would have changed adherence at the beginning of the observation period, such as in a clinical trial. Although no such interventions were deployed in our analysis, we might wish to differentiate periods such as calendar years more distinctly so that associations between clinical events and changes to adherence might be identified.

As such, the three measures selected as the most appropriate for my study were CMA1, CMA5, and CMA8.

## 4.2.2 Identifying Asthma Controller Medications

Asthma can be effectively managed in the majority of individuals through regular use of ICS <sup>14,267,268</sup>, although additional therapies may also be used in those with poor control. One such additional therapy is LABA, which may be prescribed in a stand-alone, or combination ICS+LABA inhaler.

To identify asthma medications, the medication's name (a concatenation of the *PIApprovedName* and *PIPrescribableItemName* variables) was searched for various keywords relating to the medication ingredients and brand names listed in Appendix D, an update of the classification used previously by Mukherjee et al. <sup>160</sup>, with the addition of formulations and dosages approved for asthma treatment in adults (extracted from the British National Formulary Version 80; September 2020 <sup>269</sup>). The medications were then labelled with their corresponding drug class (such as LABA).

Many corticosteroids are also used in other dosages and formulations for conditions such as rhinitis (e.g. nasal sprays) and Crohn's disease (e.g. foam enemas). All medications with the formulation listed as 'sprays' or 'drops' were excluded. The brand names of the medications were also checked to exclude brands relating to treatments for inhaled medications for related conditions, such as COPD, or for nasal sprays with missing formulation variable (Table 4.2). Finally, the dose directions (*ePRNativeDoseInstructions*) and medication name (*ePRNDName*) were searched for keywords listed in Table 4.3 to further exclude other formulations.

Table 4.2: Corticosteroid asthma therapy exclusion brands

<b>Corticosteroid Asthma Therapy Exclusion Brands</b>
"NASONEX", "FLIXONASE", "ANORO ELLIPTA", "SUMATRIPTAN", "AVAMYS", "RHINOCORT", "NASOBEC", "NASOFAN", "RYNACROM", "PIRINASE", "SPIOLTO", "DYMISTA", "POLLENASE", "VIVIDRIN", "DUAKLIR", "SEEBRI", "ULTIBRO", "PRED FORTE", "TRELEGY", "TRIMBOW", "BRALTUS", "RINATEC", "ENTOCORT", "BENACORT", "AIRCORT", "BUDEFLAM", "BUDENOFALK",

"CORTIMENT", "JORVEZA", "AZELASTINE", "CUTIVATE", "ELOCON", "NALCROM", "CATACROM", "ASPIRE", "OPTICROM", "OPTREX", "BECONASE", "MURINE", "ACLIDINIUM", "GENUAIR", "OLADATEROL", "YANIMO"
--

Table 4.3: Corticosteroid asthma therapy exclusion formulation and indication terms

<b>Corticosteroid Asthma Therapy Exclusion Formulation and Indication Terms</b>
"NASAL", "NOSE", "NOSTRIL", "NASULE", "HAYFEVER", "EYE", "EAR", "DROP", "TONGUE", "FOAM", "ENEMA", "RECTAL", "SUPPOSITOR", "CREAM", "OINTMENT", "ULCER", "SKIN", "PATCH", "APPLY"

The designated class of medication for each remaining record is reported in Appendix D, with corticosteroid solutions distinguished from inhaled formulations by listed formulation (“SOL”, “CAPS”, or “TABS”) or by the presence of any of the following keywords in the dose directions or *ePRNDName*:

"SACHET", "RESPULE", "NEB", "VIAL", "AMPOULE"

#### 4.2.3 Controller Medication Cleaning

In order to estimate adherence, the date when a medication supply should be exhausted must be calculated, if used according to the dose directions: a function of the amount that should be taken every day, and the volume of the prescription. The amount that should be taken every day is the product of the number of times a day in which medicine should be taken, and the specified dosage each time.

First, the number of daily doses (*dose frequency*) that should be taken each day were extracted from prescriptions, using the keywords (and combinations of keywords) listed in Table 4.4. For example, if one puff of an inhaler twice per day is prescribed, the dose frequency is twice daily. Missing dose frequencies were imputed as the most common (mode) by *medication type* (such as beclometasone, rather than by brand).

Secondly, the number of doses (inhalations, tablets, and so forth) of medicine that should be taken at each dose time (*dose quantity*; for example TWO PUFFS) was estimated by searching for the numbers one, two, three, or four (in numerals and written out) preceded by “take\_” or “inhale\_”, or followed by “\_to\_be\_taken\_”, “\_at\_”, “\_daily”, “\_puf” (with a single ‘f’ to allow for typographical errors), “\_p “ or “p\_” (‘p’ being commonly used shorthand for puffs). For all of the above, an underscore is used here to denote a space. When this information could not be extracted, the mode by medication type was imputed.

Table 4.4: Daily medication dose frequency keywords and observed incidence

Daily Dose Frequency	Key Words		
Once	"ONCE", "O-D", "O.D"		
	"DAILY", "EVERY DAY", "EACH DAY"	WITHOUT	"TWICE", "TWO TIMES", "2 TIMES", "TD", "TID", "BID", "BD", "B-D", "B.D", "FOUR TIMES", "4 TIMES", "QID"
	"MORN"		"NIGHT", "EVE", "BEDTIME"
	"MANE"		"NOCTE"
	"NOCTE"		"MANE"
	"AM"		"PM"
	"PM"		"AM"
	"A.M"		"P.M"
	"P.M"		"A.M"
Twice	"TWICE", "TWO TIMES", "2 TIMES", "TD", "TID", "BID", "BD", "B-D", "B.D"		
	"MORN"	WITH	"NIGHT", "EVE", "BEDTIME"
	"AM"		"PM"
	"A.M"		"P.M"
	"MANE"		"NOCTE"
Four Times	"QID", "FOUR TIMES", "4 TIMES"		
Unknown	N/A		



Next, the number of medication units (inhalers, boxes of tablets, and so forth) prescribed was extracted (*unit quantity*). The ALHS dataset contained two variables relating to medication quantity – the prescribed and dispensed amount. A final quantity variable was derived as shown in Figure 4.2.

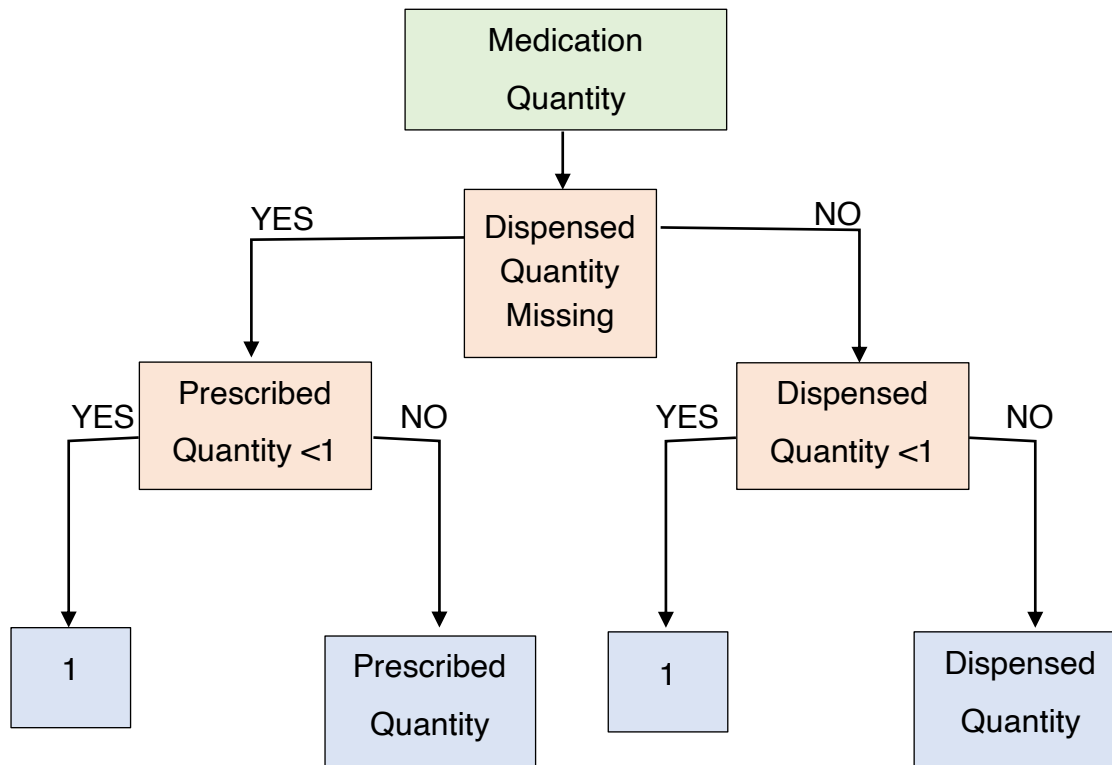


Figure 4.2: Decision tree illustrating the selection of medication quantity from prescribed and dispensed quantity variables in the Asthma Learning Healthcare System prescribing dataset

In order to estimate the number of *prescribed doses*, I multiplied the number of medications units by the number of doses per unit (*unit volume*), extracted from the free-text prescription information. To do this, I searched for any of the values [200, 120, 112, 100, 60, 56, 50, 40, 30, 28, 24, 20, 14, 5] followed by any of “DOSE” (with a space), “-DOSE”, or “ X ”. Additionally, records with quantity of 14 or over were included as extracted values of prescribed doses. The next step was to impute unit volume values for the prescriptions where information could not be extracted. Medications are frequently available in different pack sizes depending on the strength: lower strengths are often available in larger volumes. As such, the prescribed strength was also extracted from the data. Note that herein strength refers to the amount of

medication that is taken in a specified dose. Strength is sometimes used interchangeably in the literature with dose, however the latter refers herein to the unit of medication taken at a single point of ingestion (such as two puffs on an inhaler, or one tablet).

First, I searched through the free-text prescription information (*ePRNDName*) for any of the following medication strengths in micrograms [10000, 5000, 4000, 2000, 1000, 500, 400, 320, 250, 200, 184, 160, 125, 100, 92, 80, 65, 50], followed by “MCG” or “MICROGRAM”, with and without spaces between the value and phrase. Additionally, for ICS+LABA medications, which have medication strengths for the ICS and LABA components separately, the values could proceed “/” (without a space). By searching through the values in descending order I ensure that “250 MCG” is not extracted as “50 MCG”, for example. Following that, I searched for the following medication strengths in milligrams [0.5, 20, 10, 5, 4, 2, 1] followed by “MG” or “MILLIGRAM” (again, with or without a space between). Similarly to the micrograms, 0.5 is searched prior to the integer values such that “0.5MG” is not extracted as “5MG”. Finally, the microgram values previously specified could also be followed by “CLICKHALER”, “ACCUHALER”, “EVOHALER”, or “TURBOHALER”, or preceded by “QVAR”, “SERETIDE”, “SERETIDE MDI”, “INHAL”, or “ALVESCO”. This process is illustrated in Figure 4.3. The extracted milligram value is multiplied by 1000 to convert all extracted values into micrograms.

Missing medication strengths were investigated to identify the most appropriate level for mode imputation: brand, medication type, or drug class.

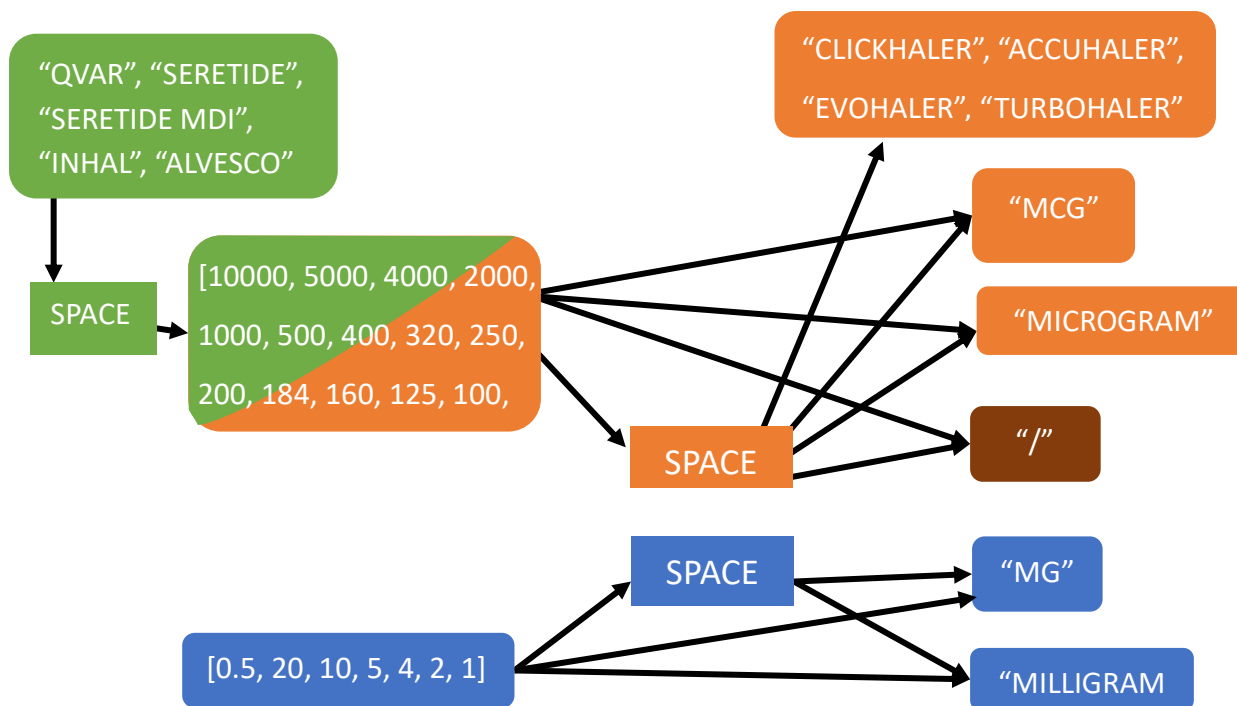


Figure 4.3: Illustration of the medication strength identification natural language pathways

The extracted medication strength for a prescription was compared to a look-up table of available medication strengths by brand (presented in Appendix D, pressurised inhaler or inhalation powder only for ICS or ICS+LABA), in order to flag values that were outside of the range of medication strengths prescribed for that medication specifically for asthma, indicating it may have been prescribed for another condition and should be excluded. The range of included medication strengths were only matched to their medication type (such as beclometasone), rather than specific brand, to account for generic substitution dispensing.

Returning to the unit volume imputation, the modal value by strength and medication type and brand was imputed for missing values. If there were no records with extractable unit volume (and thus no mode could be calculated) the value was imputed as the smallest unit volume listed for that brand (or the most common brand for generic medications) and medication strength from Electronic Medicines Compendium (EMC) website, medicine.org.uk, which hosts information on all medicines licensed for use the UK.

Finally, the prescribed doses were calculated as the number of units prescribed multiplied by the unit volume. For some values of prescribed quantity, it is not always clear whether it relates to the number of doses or the number of units (such as 20, which is high for a number of inhalers, but low for a number of doses). The prescribed quantity was manually reviewed in order to guide data-driven thresholds for classing quantity as number of units or number of doses.

An algorithm was developed by McTaggart *et al.* <sup>270</sup> for the extraction of prescription data from the free-text prescribing fields, which is applied automatically to all research datasets extracted from the Scottish PIS. The accuracy of their algorithm has not been tested in data unseen in the derivation process, however in the subset of the derivation data pertaining to respiratory therapy, data was extracted for 95.3% of records <sup>270</sup>. In the PIS data, however, there was a high amount of missing data in the asthma medications, and so I developed my own methods for use herein, as described. When both methods had managed to extract values for the number of doses per day, the amount to take at each dose, and the strength of the medication, the agreement was between 99.6 and 99.9%. My methods consistently resulted in lower missingness (before imputation): 13% versus 10% for daily dose frequency, 13% versus 11% for dose quantity, and 62% versus 8% for dose strength. The most common phrases which were not translated (no information extracted) were “Morning and night” (equalling two daily dose times), “[*n*] inhalations” or “[*n*] inspirations” (equalling *n* units of inhaled medication to be taken at each dose time), and ICS+LABA medications such as Seretide and Symbicort which were commonly listed without the unit (i.e. “SERETIDE 250”).

#### 4.2.4 Analysis Plan

The CSA and CSG were calculated for each refill in the observation period except the individual’s last, as the adherence measures require a subsequent prescription to cap the duration. Rolling means of the CSA for the last 3, 5, and 10 refills were also calculated, denoted CSA\_3, CSA\_5, and CSA\_10. The means were not weighted by either the chronology of the intervals, or the length of each interval.

The definition of each CMA provided by Vollmer *et al.*<sup>190</sup> states that in CMA5 and CMA8 the number of days of *theoretical use* should assume “medications taken as directed and new medications banked until needed”. In line with work by Galozy *et al.*<sup>266</sup>, I additionally included two variations of this approach to estimating medication supply. Three measures of supply were used to estimate whether there was medication available on each day, with supply estimation method henceforth denoted by the numbers below following an underscore:

1. Assuming all medication was lost or disposed of at a new prescription (ignoring leftovers) and calculated using only the time since the last dispensing, and how much was dispensed.
2. Assuming the maximum amount available after a dispensing was double the amount dispensed (capping the leftovers)
3. Assuming all leftovers were available, and no medication was ever lost, disposed of, or went out of date (as utilised by Vollmer *et al.*<sup>190</sup>).

For example, CMA5\_1 denotes CMA5 with over-supply discarded.

For all measures, multiple prescriptions obtained on the same day were condensed into a single record by summing the supply obtained and removing the first record (which would have the refill interval duration calculated as zero days). For the multiple refill measures, I processed separately each individual’s entire observation period, and sub-periods of single calendar years and three-month blocks (quarter-years).

CMA8 could only be calculated when some prior history of medication was known, such that the supply quantity at the start of the observation period can be calculated. For this analysis, the calculation of CMA8 in sub-periods which start at the beginning of follow-up will assume no carryover (equivalent to CMA7).

Changes in medication (therapy type, strength, brand, etc.) were disregarded in this analysis, such that it was assumed carryover was not discarded when a new treatment began, however changes to the number of doses to be taken each day were assumed to come into effect immediately, even in cases when the carryover supply was for a different medication as well as daily dosing regimen.

First, described the results of the asthma prescription identification process, and reported on the number of records excluded at each stage, as well as the proportions of each characteristic before and after imputation was conducted. For each measure, I provided summary statistics, and Spearman correlation coefficients between one interval (refill, quarter, or year) and the next. Correlation coefficients were considered to denote strong ( $|r| > 0.7$ ), moderate ( $0.3 < |r| < 0.7$ ), or weak ( $|r| < 0.3$ ) statistical associations. Spearman correlation coefficients measure the strength (and direction) of monotonic relationships between two variables.

The Spearman correlation coefficients between different measures for the same time period (all of follow-up, years, and quarters) were also calculated. There is no perfect method to map the single interval measures (including the rolling averages) to the multiple interval measures, such that they can be compared. My approach for this analysis was to match each single interval measure time-point to the period it matched up to.

## **4.3 Results**

### **4.3.1 Asthma Prescription Record Identification**

A negligible (fewer than 10) number of prescriptions for mAb treatments were identified, which were removed from further analysis to prevent accidental disclosure. After these exclusions, there were 5,684,338 potential asthma medications, identified by brand names and active ingredients, of which 2,342,339 (41.2%) were either ICS or combination ICS+LABA medications (Figure 4.4). ICS Records with formulation listed as a spray ( $n=687,511$ ) or a drop ( $n=22,938$ ) were excluded, leaving 1,631,890 were ICS or ICS+LABA medications (32.8%).

A further 1332 records were excluded based on the identification of non-asthma indication brand names (of which none were ICS or ICS+LABA), leaving 4,970,983 records. Finally, 5269 records (1407 ICS or ICS+LABA records) were excluded as they contained one or more formulation exclusion keywords, leaving 4,965,714 records for 187,487 unique individuals (1,630,483 ICS or ICS+LABA). 2675 ICS records were reclassified as steroid solutions (Table 4.5), leaving 1,627,808 ICS and ICS+LABA prescriptions.

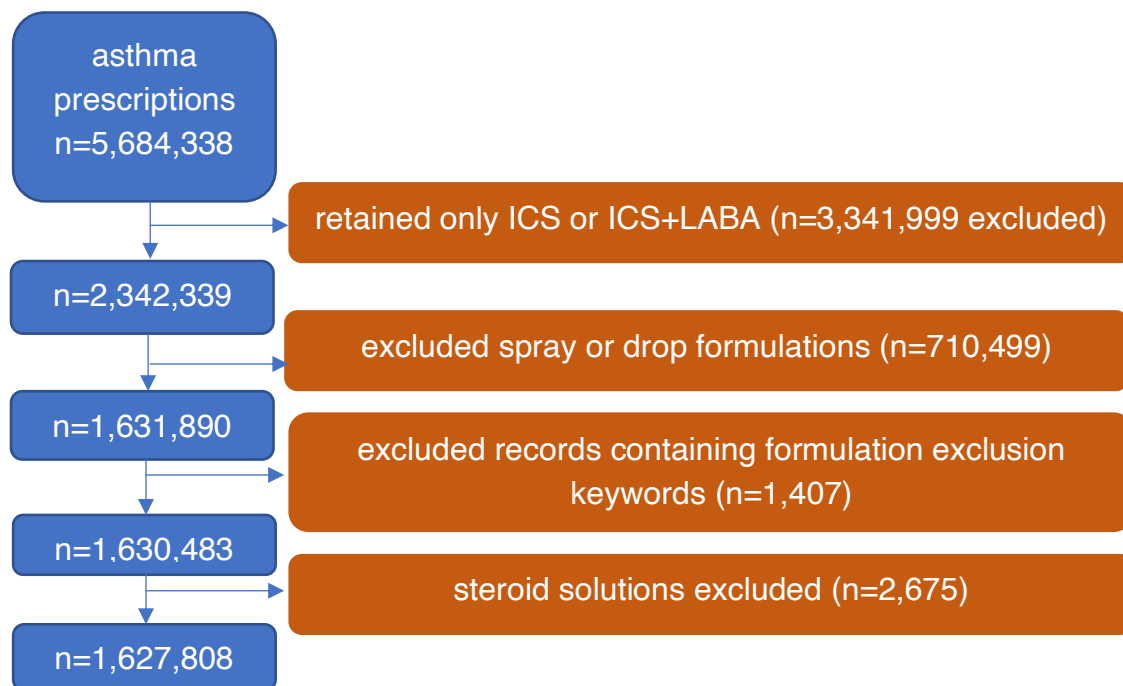


Figure 4.4: Flowchart of ICS and ICS+LABA prescription record exclusions

Notes: ICS = Inhaled Corticosteroids, LABA = Long Acting  $\beta$ -2 Agonist

Overall, 44.5% of prescriptions had identified brand names prescriptions, of which 20.6% had a generic drug substituted at dispensing.

Table 4.5: Asthma medication classifications in the ALHS prescription data (n=4,965,714)

<b>Drug Class</b>	<b>Number of Prescriptions (%)</b>	<b>Percentage of class branded at prescription</b>
ICS	644,907 (13.0)	88.3
CS Solution	2675 (0.1)	27.7
ICS+LABA	982,901 (19.8)	>99.9
LABA	567,916 (11.4)	21.8
SABA	1,976,932 (39.8)	21.1
Other	790,383 (15.9%)	14.3

Notes: 'other' category included LTRA (3.2%), Long-Acting Muscarinic Antagonists (LAMA; 1.0%), Theophylline (1.2%), and OCS (10.2%) medications.

#### 4.3.2 Asthma Controller Medication Prescription Record Processing

The modal dose frequency by medication (ingredients) was imputed when a value could not be extracted (Table 4.6): once daily for Ciclesonide and Fluticasone Vilanterol (n=1210), else twice daily (n=162,102). The modal dose quantity by medication was also imputed when a value could not be extracted: one dose at each daily dose time for Budesonide, Ciclesonide, Fluticasone Vilanterol, Fluticasone Salmeterol, and Mometasone (n=71,880), else two (n=103,621).

181 prescriptions had missing dispensed quantity (0.1%), and 50,556 (3.1%) had distinct prescribed and dispensed quantities. Of these, 50,490 had prescribed quantity of zero (99.9%). 59.5% of prescriptions had final quantity value 1, 39.7% had 2, 0.8% had higher. The maximum recorded quantity was 480.



Table 4.6: Asthma controller medication daily dose frequency and quantity of doses per dose time

(N=1,627,808)		Before Imputation	After Imputation
		Percentage of Prescriptions	
<b>Dose Frequency</b>			
	Once	2.3	2.4
	Twice	87.4	97.4
	Four Times	0.2	0.2
	Unknown	10.0	-
<b>Dose Quantity</b>			
	One	35.8	40.2
	Two	53.0	59.3
	Three	0.2	0.2
	Four	0.3	0.3
	Unknown	10.8	-

221 prescriptions had extracted medication strength values which were not listed on the lookup table presented in Appendix D, and were thus excluded. This left 1,627,587 prescriptions for 91,332 unique individuals. No strength value could be extracted for 8.4% of prescriptions (n=136,151), and the modal medication strength by medication was imputed.

The unit volume could be extracted for only 15.2% of records, and the modal value by medication strength and medication type and brand was imputed. Of the 60 combinations of medication, brand, and medication strength, 42 (70%) had at least 80% *confidence*: the imputed modal unit volume leading by a majority of at least 80% of the samples with extracted values. 55 of the combinations (92%) had confidence over 60%, translating to 82.6% of the imputed prescriptions. The most common combinations also tended to have higher confidence, leading to a median confidence in imputed records of 99.97% (interquartile range, or IQR, between 99.8 and 100%). There was only one combination confidence lower than 50% (Relvar Ellipta 184mcg), but this represented only 0.3% of the imputed records (n=3792). The next lowest

confidence combination was Seretide 250mcg (n=233,068) for which 59% of the extractable records stated 120 doses per pack and 41% said 60. For the 1.2% of prescriptions for which no modal value could be calculated for imputation, values were manually imputed. From manual review, the threshold above which quantity was assumed to quantity doses rather than units was 15.

The expected duration of the medication supply was calculated as the quantity dispensed divided by the daily number of doses. The median duration was 60 days, with an interquartile range of 30 to 60 (range 3.5 to 1100). The prescription interval duration was calculated as the number of days between the date of that prescription, and the next chronological prescription for that person. The range was 0 to 2964 days (median 53, and IQR 30 to 88), with 91,332 records (one per person) with no calculated interval duration, as they were either the only, or the final prescription in the study period for that person. If multiple prescriptions were issued on the same day, their quantities (and expected supply durations) were summed, thus assuming that the medications would be taken sequentially and not simultaneously. This resulted in 1,600,419 records being retained.

15,329 people (16.8%) had only a single prescription during their follow-up (Figure 4.5); their single-interval measures (CSA and CSG), CMA1, and CMA5s could not be calculated for any interval length.

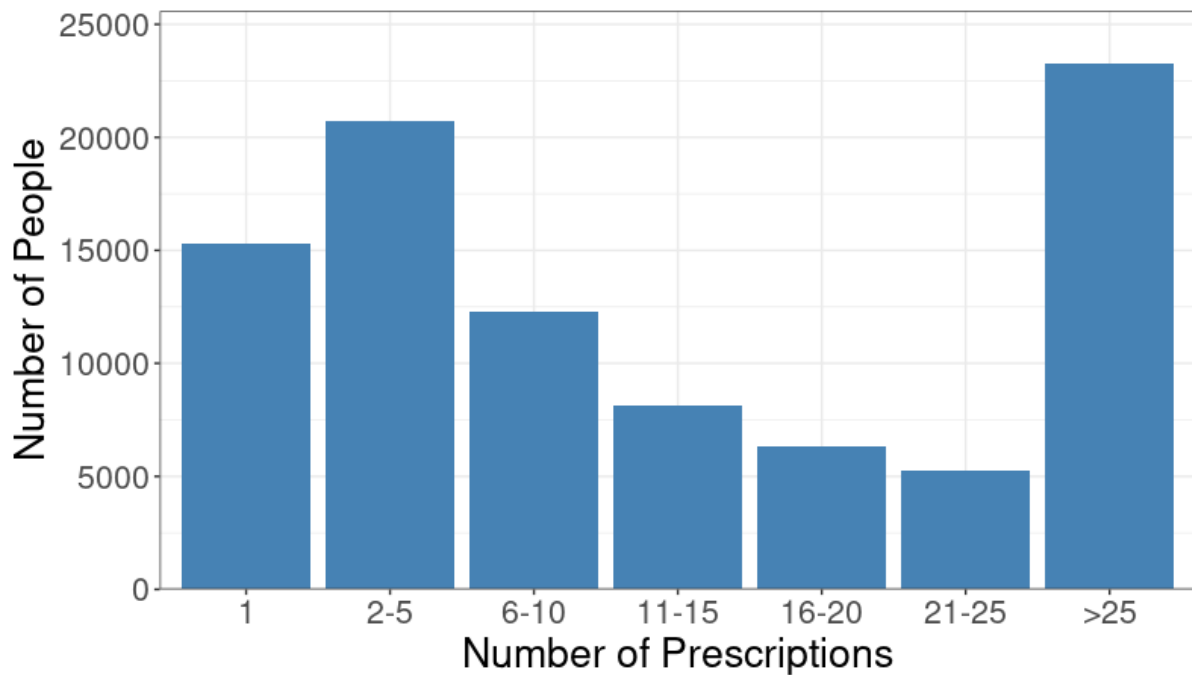


Figure 4.5: Bar chart of the number of prescriptions per individual during their follow-up in the Asthma Learning Healthcare System Dataset

### 4.3.3 Single Interval Adherence Measures

Of the single interval measures, the CSG (Equation (4.2)) is the only one with a finite range  $[0, 1)$ . Across all people, the median prescription interval gap was  $<0.001$  (mean 0.21), and the upper quartile was 0.41. The within person median CSGs ranged between 0 and 0.99, with a median of medians of 0.24.

The CSA and cumulative variations were all unbounded, and the CSA had a maximum value of 480 (median 1.00). Longer windows (more refills) for rolling averages of CSA resulted in higher values (Figure 4.6), likely due to some combination of survivor bias (only 49% of people had 10 or more prescriptions during their follow-up) and the reduced impact of a single poor interval (regardless of its length or chronology). The interquartile range width was similar between window sizes, however (between 1.01 for CSA\_3 (0.79 to 1.80) and 1.08 (0.57 to 1.67) for CSA; Table 4.7). Density plots of the single interval adherence measures are presented in Appendix I.

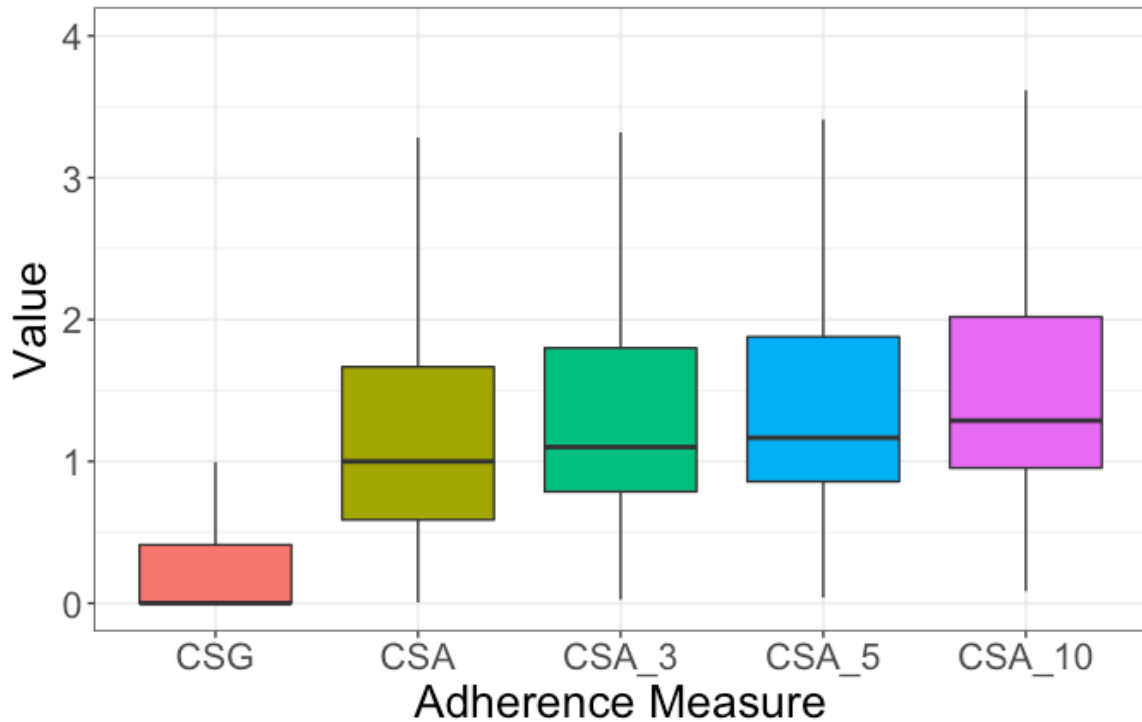


Figure 4.6: Boxplots (without outliers) of the values of each single interval availability adherence measure

Notes: CSA = Continuous Single interval measures of medication Availability

The number following the underscore denotes the number of previous prescriptions the estimate is averaged over.

Outliers (extending above 50 for the rolling average CSAs, and up to 480 for CSA) have been excluded from this plot to aid readability.

Table 4.7: Summary table of the values of each single interval availability adherence measure

	<b>CSA</b>	<b>CSA_3</b>	<b>CSA_5</b>	<b>CSA_10</b>
25 <sup>th</sup> Percentile	0.59	0.79	0.86	0.95
Median	1.00	1.10	1.17	1.29
75 <sup>th</sup> Percentile	1.67	1.80	1.88	2.02
Interquartile Width	1.08	1.01	1.02	1.07

Note: The number following the underscore denotes the number of previous prescriptions the estimate is averaged over.

Table 4.8 summarises the correlation coefficients between one refill and the following two refills (respectively). The correlation for the next refill was strong for the rolling window measures and moderate for the single interval measures. For the refill after next, the correlation remained strong for only CSA\_5 and CSA\_10 (R 0.817 and 0.914, respectively).

Table 4.8: Spearman correlation coefficients between single prescription adherence measures for subsequent refills.

Measure	Spearman Correlation	
	Compared to one refill later	Compared to two refills later
	Correlation Coefficient (Number of Samples)	
CSA	0.422 (n=1,433,084)	0.419 (n=1,364,718)
CSG	0.393 (n=1,433,084)	0.386 (n=1,364,718)
CSA_3	0.836 (n=1,242,777)	0.682 (n=1,187,508)
CSA_5	0.908 (n=1,135,211)	0.817 (n=1,085,638)
CSA_10	0.958 (n=909,304)	0.914 (n=869,819)

Notes: All correlation coefficients were statistically significant, with  $p < 0.001$

The number following the underscore denotes the number of previous prescription refills the estimate is averaged over.

#### 4.3.4 Multiple Interval Adherence Measures

There were 647,585 person-years, of which 232,251 (35.9%) contained no prescriptions and 114,890 (17.7%) contained only one. Therefore, CMA1 and CMA5s could not be calculated for 53.6% of person-years. Similarly, there were 2,230,480 person-quarters, of which 1,183,672 (53.1%) contained no prescriptions and 643,809 (28.9%) contained only one. Therefore, CMA1 and CMA5s could not be calculated for 81.9% of person-quarters.

Table 4.9 shows that as the interval decreases in length, the values increased for the decreasing number of intervals in which a value could be calculated. Quarters in which there were multiple prescriptions (and a value could thus be calculated) tended to be instances of more medication being collected than is required for that period. Density plots of the single interval adherence measures are presented in Appendix I.

Table 4.9: Median and spread of CMA1 across time windows

<b>Time Window</b>	<b>Median</b>	<b>Interquartile Range</b>	<b>Range</b>	<b>Number Not Calculable</b>
All of follow-up	0.609	0.318-0.979	0.007-120.000	15,329 (16.8%)
Years	0.929	0.600-1.333	0.040-480.000	347,141 (53.1%)
Quarters	1.225	0.896-2.000	0.078-480.00	3,414,152 (81.9%)

Like the CMA1, the three CMA5 measures require there to be at least two prescriptions in each analysis interval, such that there is at least one with a known end date. The CMA8 measures have no such requirement. In Figure 4.7, both the CMA5 and CMA8 measures (for all supply calculation methods) also increase on average when the interval is shorter, even for the single prescription cases with the CMA8s. While the CMA8s were always markedly lower than their CMA5 counterpart (by supply calculation method), the difference decreased with the length of the interval.

There is also a consistent trend with higher values when over-supply is allowed, but also note that the capped oversupply closely resembles the uncapped oversupply in distribution, implying that especially high (outlier) quantities of over-supply were uncommon.

The Spearman correlation between the subsequent years was highest for the CMA8s, and substantially lower for CMA1 than any other measure (Table 4.10). In both the CMA5s and CMA8s, the supply calculation method with no oversupply had the highest correlation to the next period. Across all period comparisons, the CMA8s had the strongest correlation. This difference in the continuity between the measures highlights the effect of excluding the incomplete prescription intervals in the time-period; the retrospective CMA8 is a much more appropriate proxy for current adherence than either CMA5 or CMA1.

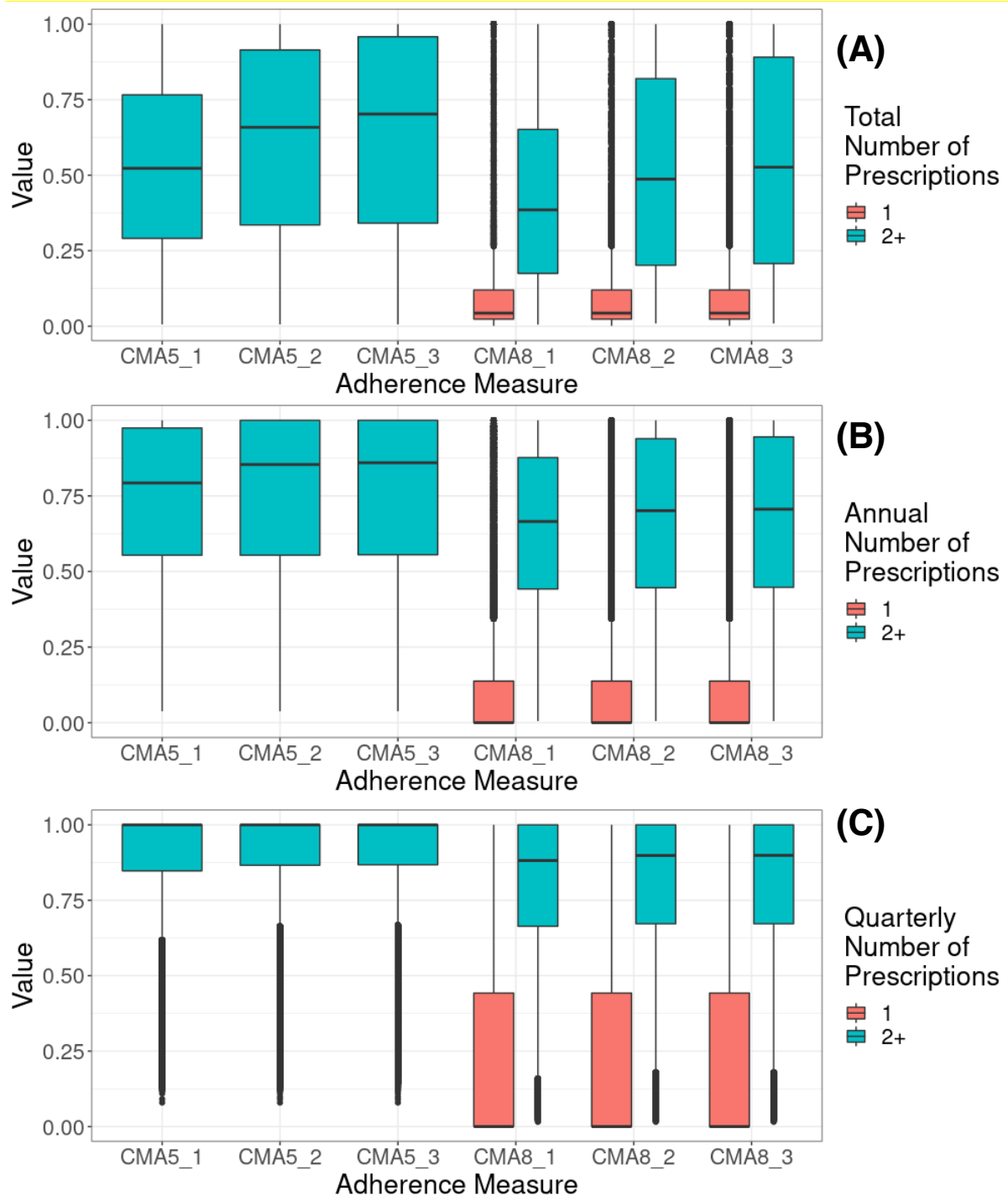


Figure 4.7: Boxplots of CMA5s and CMA8s for (A) all follow-up time, (B) years of follow-up, and (C) quarters of follow-up

Note: Number following underscore denotes the supply estimation approach: 1) Assuming all medication was lost or disposed of at a new prescription (ignoring leftovers) and calculated using only the time since the last dispensing, and how much was dispensed, 2) Assuming the maximum amount available after a dispensing was double the amount dispensed (capping the leftovers), 3) Assuming all leftovers were available, and no medication was ever lost, disposed of, or went out of date.

Table 4.10: Spearman correlation between multiple prescription adherence measures for subsequent intervals (years and quarters)

Measure	Spearman Correlation			
	Compared to one year later	Compared to two years later	Compared to one quarter later	Compared to two quarters later
CMA1	0.512 (n=231,536)	0.428 (n=177,540)	0.459 (n=345,126)	0.418 (n=300,438)
CMA5_1	0.463 (n=231,536)	0.387 (n=177,540)	0.387 (n=345,126)	0.346 (n=300,438)
CMA5_2	0.405 (n=231,536)	0.323 (n=177,540)	0.370 (n=345,126)	0.331 (n=300,438)
CMA5_3	0.397 (n=231,536)	0.317 (n=177,540)	0.369 (n=345,126)	0.330 (n=300,438)
CMA8_1	0.710 (n=556,253)	0.612 (n=466,263)	0.617 (n=2,139,148)	0.550 (n=2,049,158)
CMA8_2	0.692 (n=556,253)	0.595 (n=466,263)	0.616 (n=2,139,148)	0.549 (n=2,049,158)
CMA8_3	0.691 (n=556,253)	0.594 (n=466,263)	0.616 (n=2,139,148)	0.549 (n=2,049,158)

Notes: All correlation coefficients were statistically significant, with  $p < 0.001$

Number following underscore denotes the supply estimation approach: 1) Assuming all medication was lost or disposed of at a new prescription (ignoring leftovers) and calculated using only the time since the last dispensing, and how much was dispensed, 2) Assuming the maximum amount available after a dispensing was double the amount dispensed (capping the leftovers), 3) Assuming all leftovers were available, and no medication was ever lost, disposed of, or went out of date.

#### 4.3.5 Correlation between Time-Matched Adherence Measures

The single interval CSA was strongly (negatively) correlated with the CSG (-0.91; Figure 4.8), and shorter window rolling averages of CSA were more correlated with the single interval CSA (0.74, 0.66, and 0.57, in order of window length). The correlations between the 3-refill and 5-refill measures, the 3-refill and 10-refill measures, and the 5-refill and 10-refill measures were both strongly positive (0.89, 0.76 and 0.85, respectively).



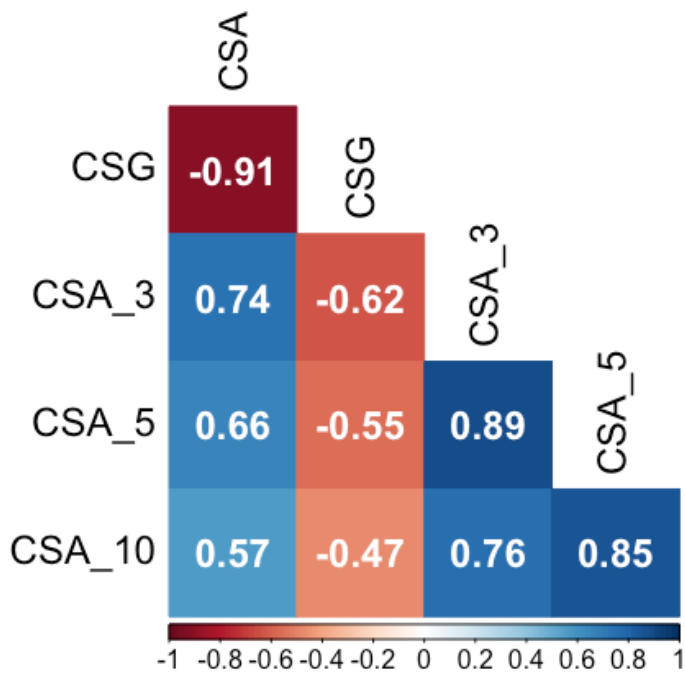


Figure 4.8: Spearman correlation between single interval adherence measures at each prescription refill

Note: The number following the underscore denotes the number of previous prescriptions the estimate is averaged over.

Over the entire study period, for people with at least two prescriptions, the correlation between all multiple-interval measures was mostly strong ( $R > 0.65$ ). The correlations were above 0.9 within the CMA5s and CMA8s, and also between the CMA5s and CMA1 (Figure 4.9A). The same was true within years and quarters, with the only substantial difference being the declining correlation between CMA1 and the other CMA measures as the interval decreased (Figure 4.9B and Figure 4.9C).

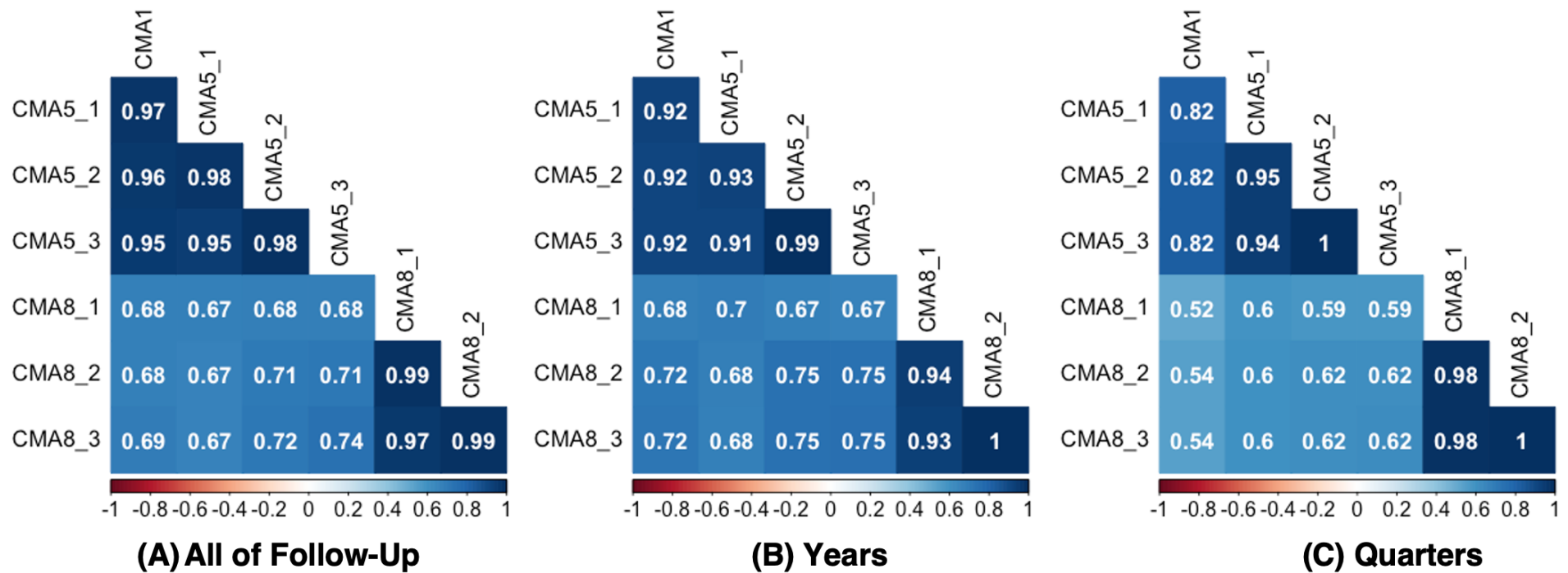


Figure 4.9: Spearman correlation between multiple interval adherence measures in (A) all of follow-up, (B) years, and (C) quarters

Notes: Number following underscore denotes the supply estimation approach: 1) Assuming all medication was lost or disposed of at a new prescription (ignoring leftovers) and calculated using only the time since the last dispensing, and how much was dispensed, 2) Assuming the maximum amount available after a dispensing was double the amount dispensed (capping the leftovers), 3) Assuming all leftovers were available, and no medication was ever lost, disposed of, or went out of date.

The correlation between the CMA8 measures when there was only a single prescription was always 1; these cases were excluded so as not to skew the results.

There is no perfect method to map the single interval measures (including the rolling averages) to the multiple interval measures, such that they can be compared. Figure 4.10A shows that over an individual's full follow-up (each individual refill is matched to the person's full follow-up adherence estimates), all single interval measures are moderately well correlated with all CMAs. The strongest correlation is between CMA1 and the 10-refill CSA, but this is only available for people with at least 10 refills, which likely introduces some confounding. When the individual refills were matched to their annual or quarterly adherence estimates (Figure 4.10B and Figure 4.10C) similar trends were observed, but had strong correlation with the shorter window rolling measures (3- and 5-interval, and 10-interval at the annual level only).

## **4.4 Adherence Measure Selection**

### **4.4.1 Principal Findings**

Appropriate selection of an adherence measure is crucial to ensure that the nuance in prescribing records is captured. Rolling average windows covering higher numbers of refills are more susceptible to survivor bias, and neglect recent gaps in prescriptions, while single interval measures have massive variance. Across all individuals with asthma controller medications, 17% had only a single prescription, and thus no values for CMA1 or the CMA5s could be calculated. Periods of oversupply were common, but the minimal difference between the capped and uncapped oversupply estimates demonstrate that they rarely skewed estimates substantially in annual and quarterly estimates (more so in the full follow-up). The CMA8 had stronger correlation between subsequent years and quarters than either the CMA1 or CMA5s, which meant that retrospective estimates (such as for the previous year) were likely a better proxy for current use.

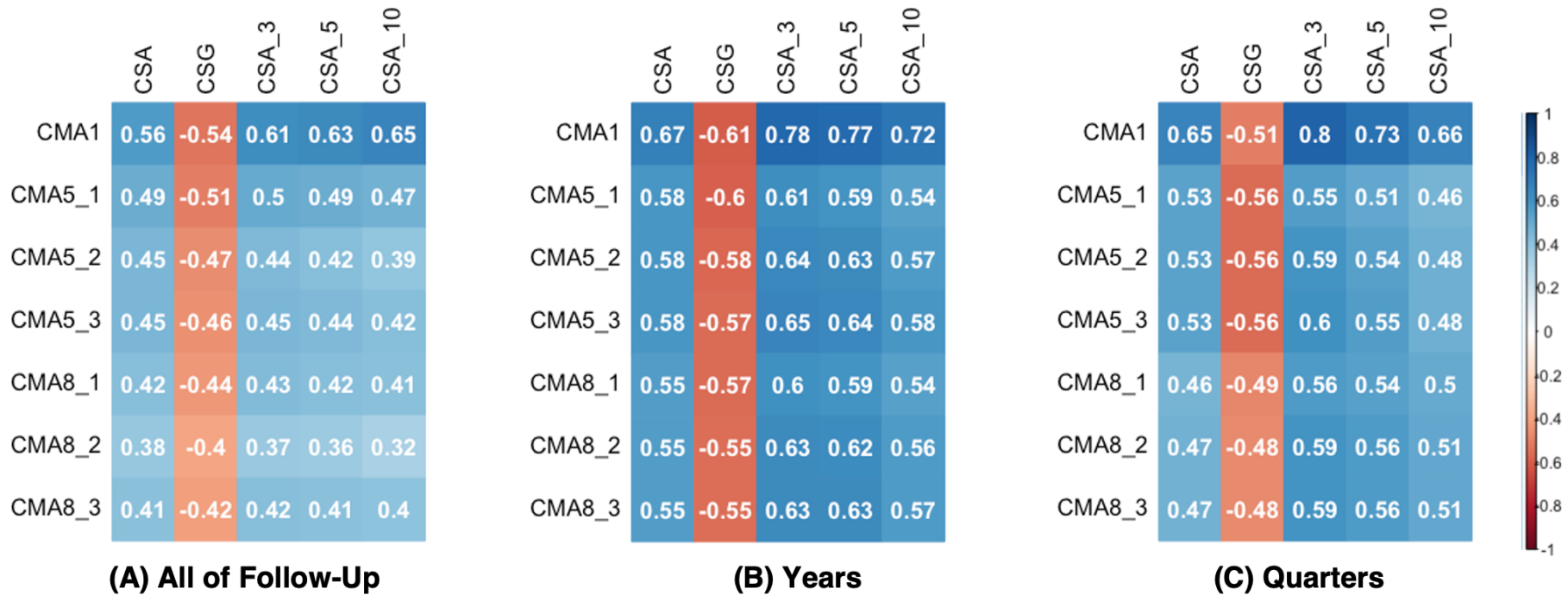


Figure 4.10: Spearman correlation between single and multiple interval adherence measures in (A) all of follow-up, (B) years, and (C) quarters

Notes: The number following the underscore in the CMA measures denotes the supply estimation approach: 1) Assuming all medication was lost or disposed of at a new prescription (ignoring leftovers) and calculated using only the time since the last dispensing, and how much was dispensed, 2) Assuming the maximum amount available after a dispensing was double the amount dispensed (capping the leftovers), 3) Assuming all leftovers were available, and no medication was ever lost, disposed of, or went out of date.

The number following the underscore in the single interval measures denotes the number of previous prescriptions the estimate is averaged over.

The correlation between the CMA8 measures when there was only a single prescription was always 1; these cases were excluded so as not to skew the results.

#### 4.4.2 Results in Context

A recent review by LeClerq and Choi <sup>271</sup> compared four variations on the CMA5 in pharmacy dispensing records for multiple sclerosis patients, in which the end of the observation window is censored in different ways. All variations used the equivalent of the uncapped medication supply estimation approach defined herein. The window end decision rule for the first version was as defined for the CMA5: the date on which the last refill in the full observation window was made (referred to as '*last fill*'). In the second, the observation window continued until some predefined date, much like the CMA8 ('*fixed*'). In the third, the window ended when the last fill would be exhausted, if used compliantly ('*last fill plus*'). Finally, the fourth approach adds up to 30 days of non-adherence to the end of the last refill's exhaustion date ('*last fill plus plus*'), with the rationale that one cannot be certain whether the treatment had been discontinued by the medication professional, but wish to compromise between the *last fill plus* and *fixed* approaches. Across all participants, the mean adherence was lowest for the *fixed* variation (0.87, with 23% under 0.8) and highest for the *last fill plus* variation (0.92, with 14% under 0.8). In subsets of participants with higher numbers of refills during follow-up, the difference between the variations became smaller. This is well aligned with our findings in Figure 4.7, i.e. that the CMA8 was substantially lower for the full follow-up than the CMA5, but the difference decreased in magnitude when annual or quarterly periods were considered, as full periods without a single prescription were removed. The pertinence of the possibility of authorised medication discontinuations effecting estimates depends largely on the study design and population.

Similarly, Bjarnadottir *et al.* challenged some of the assumptions and parameters used in their calculation of an adherence measure equivalent in its primary state to the CMA1 as defined herein <sup>272</sup>. In one comparison, they noted, similarly to LeClerq and Choi <sup>271</sup>, that using fixed follow-up end dates led to lower mean adherence than when an individual's follow-up was right-censored at the theoretical end of their last refill's supply. Similarly to this study, they also reported that shorter follow-up periods resulted in higher mean adherence, that people with fewer prescriptions tended to have lower adherence when a fixed end-point was used, and that disposing of any

over-supply drastically reduced mean adherence (54% of the study participants had at least one overlap in prescriptions during follow-up).

A recent comparison of adherence measures in a South African pharmacy claims database demonstrated using Bland-Altman plots (scatterplots of the difference between two measures against the mean of the two measures) that even similarly defined measures may show poor agreement <sup>273</sup>. Evidently, measures for estimating adherence from EHRs are very sensitive to key underlying assumptions. Buono *et al.*'s recent review of EHR adherence measurement methodology <sup>274</sup> highlights the importance of matching the measure to its intended purpose in the analysis. For example, it is important to consider the most meaningful timescale for a specific purpose. Averaging adherence across a long period means that it would not be possible to evaluate how changes to adherence affect the likelihood of an event occurring in a short period, for example adherence in last month might be more pertinent than in the previous calendar year for predicting whether an attack is likely to occur in the next week. On the other hand, only using adherence measured over a short duration to extrapolate over a longer study might introduce some bias from seasonal variations in adherence. Similarly, the optimal approach for estimating medication supply may vary depending on the condition (liquid solutions, for example, are more prone to volume loss by spillage than tablets) and the population (such as children being perhaps more likely to have multiple inhalers at one time than adults).

#### 4.4.3 Limitations and Future Directions

A fundamental limitation of measuring adherence from EHRs is that prescription recording systems cannot record whether a medication is actually taken, and indeed whether it is taken appropriately (including with good or poor technique). Unlike treatments which are taken orally, poor inhaler technique limits the ingestion of inhaled asthma medication <sup>244</sup>. As such, EHRs cannot be considered a good estimator of the *implementation* of a treatment regimen <sup>188</sup>.

If prescribing and dispensing records are both available (and linked), then it is also possible to identify when an individual has failed to *initiate* a prescribed regimen. In

datasets such as the one used herein, in which entries correspond to pharmacy claims, only dispensed medications are recorded and thus regimens which were never initiated cannot be detected. Asthma non-initiation rate estimates from the USA and Canada range between 8-20%, as ascertained from prescriptions claims datasets <sup>259–263</sup>.

The primary value in estimating adherence from EHRs is in evaluating an individual's *persistence*, including the duration and incidence of unscheduled treatment intermissions (an extended duration of consecutively missed doses, with the minimum duration varying by treatment and condition <sup>275–279</sup>). Intermissions may occur many times in the unbounded duration of asthma treatment, particularly as 30-50% of asthma patients in western Europe are classed as having intermittent asthma <sup>280–282</sup>, according to the GINA guidelines <sup>35</sup>. Additionally, the most common reasons for a sanctioned treatment discontinuation are possible to identify in the EHRs, by searching for changes in prescriptions <sup>250</sup> or Read Codes relating to revised diagnosis, a change in regimen, or asthma resolution (common in childhood asthma <sup>283</sup> and occupational asthma <sup>284</sup>).

Specifically relating to the methods employed herein, the primary limitations of this study relate to the extreme complexity of EHRs, and the procedures that were implemented for data extraction. First, data extraction from the free-text fields of the drug description and instruction was handled using very basic approaches. In the following example dose instructions, the bold text highlights words (and segments) which will result in the exclusion of the corresponding records, according to my process:

“TAKE 8 A DAY IF PEAK FLOW **DROPS** BELOW 220 IN ACCORDANCE WITH  
PERSONAL ASTHMA PLAN”

“MAKE APPOINTMENT FOR REVIEW PLEASE”

“USE AFTER **NASAL SPRAY**”

This was unfortunately unavoidable without conducting a full manual review, or using more complex natural language processing techniques, which was outside of the scope of this study.

Secondly, the data linkage between prescribing and dispensing records in Scottish EHRs (conducted by National Services Scotland Information Services Division) is not a perfect process, as prescriptions containing multiple items have only a single identifier, rather than an item-specific identifier. As such, if the items are listed in a different order on the dispensing and prescribing records, additional information relating to a specific item (such as dosing direction notes from the pharmacist) may be assigned to the wrong prescription item. Although feedback and improvement to this system has resulted in improvement over time, the issue still persists, and the incidence of such mis-matching (and subsequent erroneous exclusions) is hard to estimate. From a manual review of a sample of 1000 asthma medications included herein, less than 1% were obviously incorrect (either named a different medication or described a method of ingestion inherent to a different formulation, such as 'inject'). Although rare, this mismatch is likely to have led to a small number asthma-related records being erroneously excluded on the basis of indication, as they contained exclusion keywords. Motivated by this observation, I designed an algorithm which could be used to probabilistically link prescribing and dispensing records for asthma controller medications <sup>265</sup>, utilising the information recorded in free-text fields; further details are provided in Appendix J.

There is some evidence that the use of EHRs to estimate adherence is more appropriate in adults than in children: the latter may result in substantial overestimation, as seen by Jentzsch *et al.* <sup>264</sup> in their study of children with asthma (population average 70% vs 52%), as the refills are likely coordinated by their parents, regardless of the child's medication taking. Furthermore, the impact of not being able to assess implementation in EHRs is thought to be low in adults. In the general adult population, it has been estimated that (across multiple conditions) only 10% of adults could be classed as engaged (not discontinued), but poorly implementing their treatment regimen <sup>285</sup>.



A final note is that it is becoming increasingly common to recommend patients self-manage their treatment to some extent, and use their inhaler only *as needed* <sup>188,286,287</sup>. Such patients can be flagged using dosage instructions recorded in prescription records. For those patients, adherence is not a meaningful measure of their exacerbation risk, although medical usage patterns, measured in the same way, may still have some predictive value.

#### 4.4.4 Adherence Measure Selection for Asthma Attack Risk Prediction Modelling

First, consider the most meaningful timescale for measurement for my aim: short-term prediction of asthma attacks. Averaging adherence across the entirety of follow-up (up to a maximum of 7 years in ALHS, see Section 2.2.1) will not be as sufficiently sensitive to predicting asthma attacks at smaller resolutions, such as in the following four weeks for example. Furthermore, in this analysis the upper interquartile range of expected prescribed medication supply was 60 days. As such, one would expect the majority of patients to have at least one refill every quarter; however this stipulation may bias estimation such that individuals with overlapping prescription supplies or longer duration supplies (more common when asthma is stable) are more likely to have quarters with no prescriptions, and thus no estimable adherence. As such, I decided that years were the optimal period lengths for CMA-based adherence estimation. In the single interval availability measures, the 3-refill CSA rolling average was selected as the optimal balance between the risk of discounting previously accumulated oversupply and the risk of a recent interval of poor adherence being masked by prior intervals of good adherence. Similarly, the availability measures were preferable to the complementary medication gaps measure, as averaging the gaps will not enable oversupply periods to balance out longer intervals.

Secondly, the optimal CMA approach was selected. Both CMA5 and CMA8, which use the estimated medication supply, have some ability to cap outliers, while CMA1 has an infinite range. Both CMA1 and CMA5 cannot be calculated for those with only one prescription during a year, whereas CMA8 is available for anyone with at least one prescription during the year, which was found to be common in this data. When

treatment is discontinued, CMA5 and CMA1 will discard the time from the last prescription in the year onwards, whereas CMA8 will estimate poor adherence. The side-effect that authorised discontinuation will result in low estimates of adherence may be less pertinent in the main analysis because those with asthma resolved flags were excluded, assuming they are well utilised codes. As such, CMA8 was selected as the optimal CMA for this analysis, using the capped over-supply approach with one-year intervals.

Next, the approach for estimating medication supply must be determined. While Figure 4.7 shows that the population distribution of the measures does not vary greatly between supply estimation approaches at the annual and quarterly level, Table 4.9 shows that at an individual level there is a noticeable effect resulting in allowance of oversupply consistently reducing correlation between subsequent periods. Due to the finding by Bjarnadottir *et al.*<sup>272</sup> of high prevalence of overlapping prescriptions, the capped oversupply (second variation) approach was selected as an appropriate compromise between the risk of inflating supply due to medication switching or loss, and the risk of ignoring genuine overlaps.

The final comparison is between the multiple interval measure (CMA8\_2 annual) and the single interval measure (CSA\_3). As demonstrated in Figure 4.10, the correlation between CSA\_3 measurements in the same year as the corresponding CMA8\_2 values was moderate (correlation coefficient 0.63). To better understand the relationship between the two measures I created a Bland-Altman plot, which plots the mean of the two values against the difference between them (Figure 4.11).

There is a fairly consistent average difference of approximately zero until the mean of the two values exceeds one, with greater heterogeneity as the mean increases. When the CMA8\_2 is poor, it is either due to consistently poor implementation or discontinuation (with or without poor implementation prior to discontinuation). Recalling that CSA\_3 right-censors at the date of the most recent prescription, when discontinuation did occur, the difference between the CSA\_3 and CMA8\_2 is entirely dependent on the implementation prior to discontinuation, leading to increased

heterogeneity of deviation between the measures. Ultimately, I decided to use both CMA8\_2 at the annual level (for the previous calendar year) and CSA\_3 in my risk prediction model. The two measures enable me to capture different aspects of adherence without concern for their collinearity.

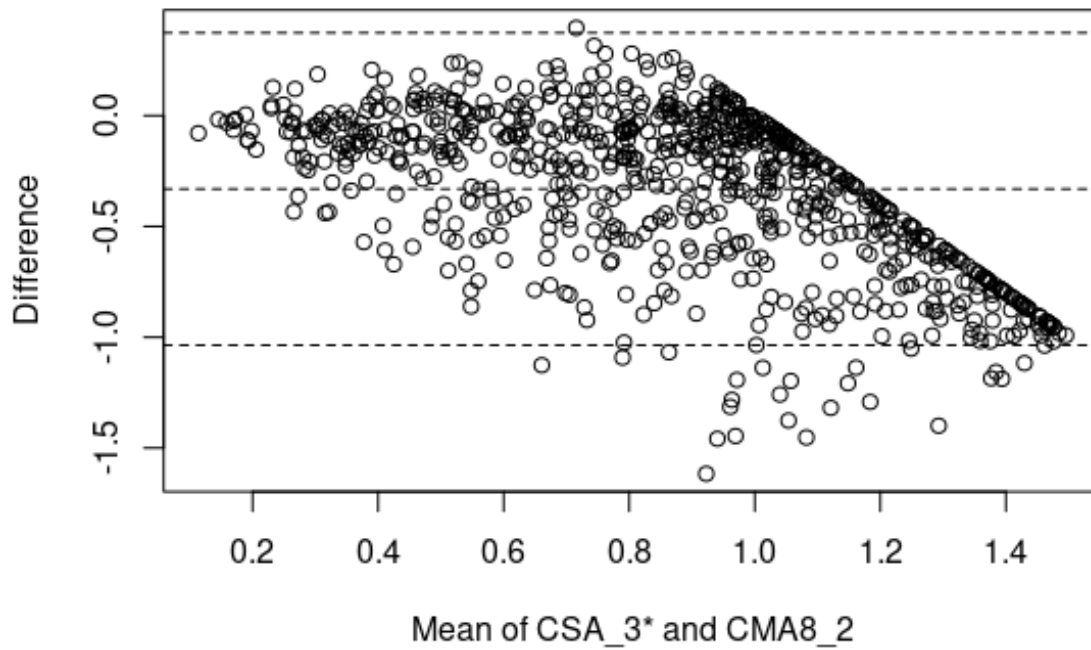


Figure 4.11: Bland-Altman plot of annual CMA8\_2 and (year matched) CSA\_3 estimates

## 5 Machine Learning

In this chapter, I introduce in greater detail the core concepts of machine learning and the process of building a machine learning model. Following that, I describe the classification algorithms and model performance measures considered for this analysis and introduce the concepts of training data enrichment and model interpretability.

### 5.1 Introduction to Machine Learning

As stated in Section 1.1.4, machine learning is a term with no universally accepted definition<sup>52,53</sup>. Herein, I use the term to describe the set of statistical methods which use computational algorithms to make estimations or predictions, or to provide statistical mapping for decision support. These estimations are more agile than rule-based approaches in cases where the statements are extremely complex to program manually (like predicting whether an email is spam by the content), but they can also be more accurate in cases where the true underlying physiological mechanism is unknown (for example, the immediate occurrence of an asthma attack).

Supervised learning models utilise algorithms applied to *labelled* training data, in which a corresponding outcome is known for each sample, and determine a functional form associating a set of features with the outcomes. Unsupervised learning, in contrast, does not require known outcomes, and obtains information about the data structure based on the features alone. For both learning paradigms, although most commonly in supervised learning, it is possible to evaluate the performance of the model by comparing the estimations against the *ground truth*; observed outcomes withheld from the model training process. In supervised learning, this can be done by querying completely unseen data with known outcomes. In unsupervised learning, it is good practice to evaluate how well the model distinguishes samples included in the training data as controls (deliberate outliers).

As previously introduced in Section 1.1.4, parametric algorithms are defined in terms of a finite number of unknown values, known as *parameters*, that are estimated from the data. Generalised linear algorithms, including logistic regression and linear regression, are the most commonly used parametric statistical mapping algorithms. The functional form ( $f$ ) of the design matrix  $\mathbf{X}$  defines the expression of random variables and parameters mapping the samples to their outcomes. A random variable is a variable whose value depends on the outcome of a random phenomenon. For example, the random variable 'age' can be mathematically interpreted as a function which maps any randomly selected individual from a population to their years of age. The realisations of each random variable can be presented using a probability distribution, for example the probability of a random person's age being over 105 years old is very small. A random variable can be composed of single features, or *interactions* between multiple features. Their specification is based on scientific domain knowledge, and intuition. The data-driven estimations of the parameters may consist of the *intercept* ( $\alpha$ ), the expected mean outcome value when the realisation of all random variables is equal to zero, and the random variable *coefficients* ( $\beta_1, \beta_2 \dots \beta_k$ ), which quantify the average contribution of a unit change in the realisation of a random variable, all else being equal, to the outcome.

*Non-parametric* algorithms, on the other hand, do not require the specification of the parameters or random variable formulation, and the functional form is instead inferred as part of the statistical learning process. They are thus more flexible when relationships between features are non-linear<sup>55,56</sup>. This does, however, make their interpretation more complicated<sup>288,289</sup>.

There are, of course, advantages to both parametric and non-parametric models. Parametric models are more easily interpreted, as the coefficients can be used to quantify the relative effect of the features onto the outcome. This can help identify important risk factors in prediction models, for example. They are also able to test hypothetical relationships, with non-zero coefficients often indicating that a term was useful in the model, assuming there is no collinearity. Finally, they require less data to build than non-parametric models, as the structure is pre-determined rather than

inferred from the samples. Non-parametric models, however, often outperform parametric models based on predictive accuracy, particularly when domain knowledge is limited, but there is a vast quantity of available relevant historical data, as is the case when using EHRs.

## 5.2 Process Flow for Model Training, Selecting, and Testing

As introduced in Section 1.2, the development of any statistical learning model is conducted in three stages: (1) a selection of algorithms (and hyper-parameters) are chosen, and trained on the partition of the data known as the *training* set, (2) the trained models are compared in the *validation* set in order to find the optimal model selection (including the choice of algorithm and hyper-parameter values), (3) the chosen model is evaluated using the *testing* set.

When a model is trained, the model fit in the model training data (*in-sample model validation*) can be evaluated. If the model overfits to the training data, however, it may not generalise well to out-of-sample data. The balance between avoiding overfitting and capturing the pertinent relationships between features in the data is known as the *bias-variance trade-off*.

Model validation in out-of-sample data can confirm whether overfitting has occurred to a degree such that the model's performance might suffer in a new testing dataset, and hence we cannot be confident about how well the model will generalise. As introduced in Section 1.2, models can be validated *internally* (from a random partition of the same dataset, and thus taken from the same feature distribution) or *externally* (in a distinct dataset, with potentially different feature distributions).

External validation is often a better confirmation that over-fitting has not occurred, as the differences in distributions of uncommon feature values may highlight weaknesses in the model, however it may not always be possible to obtain sufficiently similar data from another source.

One method of internal validation is to set aside a random subset of the data for validation purposes, however this both a) reduces the sample size that is available to train the model, and b) incurs the risk that the partitioning itself randomly produced especially good or bad performance. The former is particularly problematic if the training data sample size was small, and thus you risk not capturing the full diversity of the applicable population. One way to overcome both problems, however, is to use *k*-fold cross-validation (CV), in which one  $k^{\text{th}}$  of the data are used for testing (and the remainder for training) in a process repeated for a total of *k* times, ensuring that each sample is in the testing partition in exactly one of the *k* folds. Using a higher number of folds is often desirable, but decreases the quantity of data that are used for testing and thus increases the variation between folds.

Other variations of cross-validation include the *leave-one-out* CV, in which each sample is in turn used as the single query sample for a model trained on all remaining data (equal to *k*-fold CV when *k* is equal to the number of samples). This is known as an exhaustive method, such that every combination of testing and training samples (within the parameters that the testing set size must be equal to one sample) are permuted. The generalised form, *leave-p-out* CV, takes all possible ways of dividing the data such that the testing partition comprises *p* samples. This results in  $\frac{n!}{p!(n-p)!}$  permutations, in which *n* is the sample size, and *x*! denotes the factorial of *x*.

In *stratified* CV, the *k* folds are selected such that the class balance (or mean response for regression) is approximately equal for all partitions. Finally, in *repeated* CV, the data partitioning in *k*-fold CV is repeated multiple times. This allows for smaller values of *k* to be used when the sample size is low, but with additional confidence in the average performance.

## 5.3 Classification Algorithms

Classification, as introduced in Section 1.2, is the estimation of outcomes from a finite set, also known as a categorical outcome, or the *class*, of a sample. The classification algorithms which will be tested in my risk prediction model are described in detail in the following sections.

### 5.3.1 Generalised Logistic Regression

Generalised Logistic Regression (GLM) is a statistical model, based on the logistic function, for modelling binary classes. In the logistic model, the log-odds (the logarithm of the odds, denoted  $l$ ) for the event/class is a linear combination of  $k$  continuous or binary features ( $\beta$ ) and their respective parameters ( $\alpha$ ):

$$l = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (5.1)$$

The logistic function can then be used to convert log-odds ( $l$ ) to the probability ( $p$ ) of positive event, given the observed data. The standard logistic function, with values in the range  $[0,1]$  and a logistic growth rate of 1, is defined as follows:

$$p = \frac{1}{1 + e^{-l}} \quad (5.2)$$

The estimated class is then assigned by comparing the probability of the positive class to some threshold, typically with default 0.5.

In general terms, regression models estimate the parameters of the model by finding the set of parameter values which maximise some function. In logistic regression, maximum likelihood estimation is most commonly used: the parameter values which make the derivative of the log-likelihood equal to zero (the stationary point, in this case the maxima) are found using the Newton-Raphson optimisation algorithm (expressed as the Iteratively Reweighted Least Squares, or IRLS) based on the gradient descent method. The exponent of the coefficients from a logistic regression model, known as the Odds Ratio (OR), are a common way of reporting the estimated strength of



association. Increased odds (OR>1) mean that higher values of that feature (or the presence, for binary features) are associated (correlated) with higher odds of the outcome. An OR of 1 implies no correlation. For more information, and the formulation of the IRLS, refer to Hastie et al. (Book chapter 4) <sup>290</sup>.

### 5.3.2 Naïve Bayes Classifiers

Naïve Bayes Classifiers (NBCs) are in fact a whole family of parametric classifiers, which use Bayes' theorem to evaluate the probability of each class, given the distribution of the classes in the training data. The 'naïve' term comes from the assumption that features are all independent, however the classifier often performs well even when this assumption is violated <sup>291</sup>. Bayes' theorem is as follows:

$$p(C_k | x) = \frac{p(C_k) p(x | C_k)}{p(x)} \quad (5.3)$$

It can be thought of as the probability of a sample having class 'k' is equal to *the proportion of labelled samples with class 'k' (the prior)* multiplied by *the proportion of the labelled samples with class 'k' that have the same characteristics as our query sample (the likelihood)*, divided by *the proportion of labelled samples with the same characteristics as our query sample (the evidence)*.

Extensions of the classifier are often centred around the characteristics of the data; Bernoulli naïve Bayes has binary features (e.g. yes, is it raining), and Gaussian naïve Bayes has continuous features (e.g. height of the subject's father), for example <sup>292</sup>. In order to calculate probability of a characteristic which is continuous, we must make an assumption about the distribution of this feature within our dataset. For example, if we assume that the height of our subjects was normally distributed then we can calculate the *evidence* by calculating the mean and standard deviation of heights within the training dataset and applying a Gaussian function. In practise, the distributions are selected individually for each feature, based on its class <sup>293</sup>.

### 5.3.3 K-Nearest Neighbours

For dataset with two features, such as height and age, it would be possible to plot each sample as a point in a 2-dimensional figure. Similarly, if we added an additional feature, we could make an appropriate 3-dimensional plot, and so on. This is a simple way of explaining what we call a (Euclidean) *feature space* – a collection of vectors of information, such as our dataset’s features. For a dataset with  $M$  features, we have an  $M$ -dimensional feature space.

The k-Nearest Neighbours ( $k$ -NN) algorithm entails finding the  $k$  training data samples in the  $M$ -dimensional feature space with the smallest distance between them if you were drawing a straight line. From these identified closest samples, one can either select their modal label (classification) or their mean value (regression).

In Figure 5.1, I demonstrate a simple 5-NN example for three classes (families) in a 2-dimensional (height and age) feature space. The denoted query sample, which appears as red dot (‘Unlabelled person’), needs to be assigned to one of the three families on the basis of its proximity to its closest neighbours. We can see that the majority (three) of the five closest neighbours to our unknown child are in family B, and so we place them in that class. The estimated probability of each class is the proportion of neighbours that are in that class (estimated probability of being in class B = 0.6). Although there is no limit to the number of features in a feature space, higher dimensions evidently become harder to visualise.

The example in Figure 5.1 uses distance in Euclidean space (known as the Euclidean distance). It is also the most commonly used distance measure, however there are many alternatives, each with their own strengths and weaknesses. For example, the Hamming distance is calculated as the proportion of features with different values between two samples <sup>294</sup>, and is therefore used for purely categorical (usually binary) features. There are also composite measures designed for mixed data, such as the Gower distance <sup>295</sup>, which calculate a distance metric for each feature type, and then combine them into a single distance value.

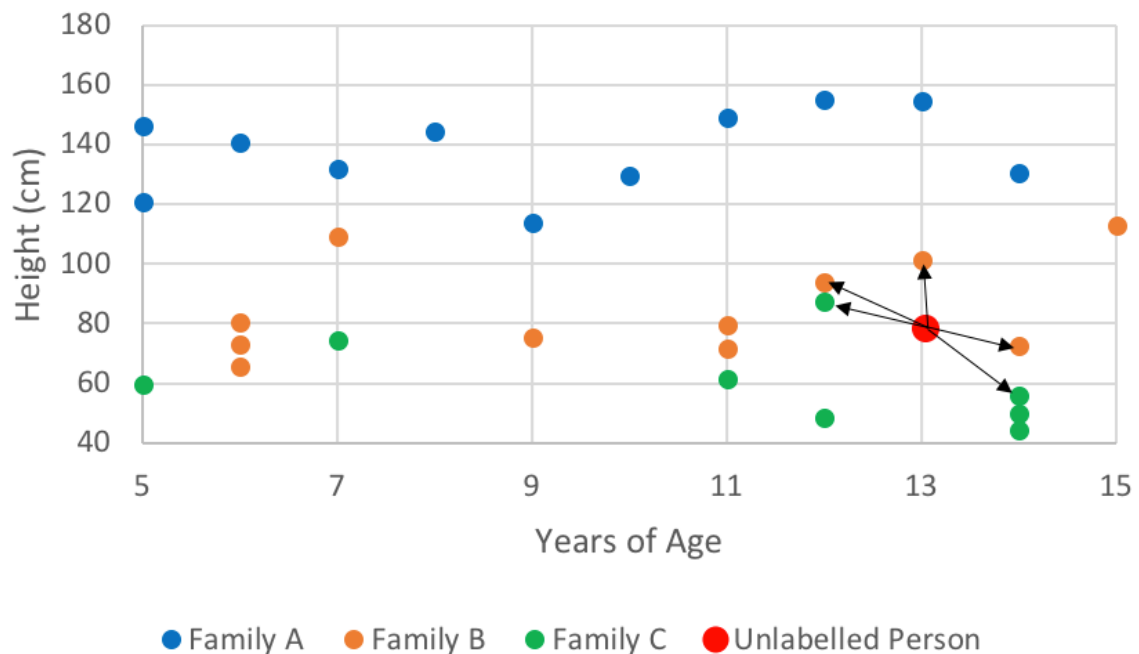


Figure 5.1: Using a 5-NN algorithm to estimate the family that an unlabelled child belongs to, based on their height and age

The Euclidean distance is best used with numerical (non-binary) data, but in practice it is usually the default choice. When using the Euclidean distance for any distance-based algorithm (such as  $k$ -NN), the scale of the features is of great importance. As such, a sample which was identical in all regards to another except one feature which had a much wider scale than the others may have a higher distance than a sample which was mildly different for all other features. Features should therefore be rescaled to have the same range, or a range that is meaningful relative to the feature importance.

$k$ -NN is an example of a *lazy learning* algorithm, which means it does not conduct any generalisations until a query is made (a new data sample requires assessing), as opposed to *eager learning*, where the system will calculate the outcome of any given query before they are made. As such,  $k$ -NN does not produce a trained model which can then be used for new query samples without disclosing any potentially identifiable information about the individuals in the training dataset. As our training data cannot

be shared for patient confidentiality reasons, the use of  $k$ -NN is not feasible for this analysis.

### 5.3.4 Decision Trees

A decision tree is a basic non-parametric algorithm containing a series of decision statements (such as ‘height > 3’, with products true or false) with each decision leading to a different set of subsequent statements until it reaches the terminus: the predicted class. As an example, a single terminal node may have the associated decision rule *if condition1 and condition2 then class1*.

The tree is constructed starting from the root. The algorithm assesses which of the features can be used to split the data (and where for continuous features), in order to maximise some measure of the data variability. Generally, this performance is related to the homogeneity of each terminal node: having as many of the training samples be in the same class as possible. Different implementations of the decision tree algorithm use different measures. The CART (Classification and Regression Tree) implementation<sup>296</sup>, for example, uses the Gini impurity ( $G$ ): the probability of a new random sample (at a specific node in the tree) being incorrectly classified, if all samples at that node were randomly classified according to the distribution of classes observed in the training data at that node<sup>297</sup>. It can thus be calculated as follows, where  $J$  is the number of distinct classes, and  $P(i)$  is the proportion of samples with class  $i$ :

$$G = 1 - \sum_{i=1}^J P(i)^2 \quad (5.4)$$

The quality of a potential split can be assessed by summing the impurity at each branch, with weighting for the proportion of the samples at the top of the branch that filter in each direction. For example, if a split of 10 training samples gave 0.5 impurity on one branch for 8 training samples, and the other branch gave 0.0 impurity for the remaining 2 training samples, the overall impurity would be  $0.5 \cdot 0.8 + 0.0 \cdot 0.2 = 0.4$ .

Decision tree algorithms are *recursive* (the same steps repeating until a stop command is issued) and *greedy* (makes the best choice at the time, regardless of the impact further down the line). They will continue to find further splits until a stopping criterion is met, including any of the following:

- every *terminal* (end of the branch) node's dataset contains only one class (for classification trees), or only five samples (regression trees),
- subsequent splits would result in nodes below a minimum size threshold,
- the maximum tree height (distance from root to terminal node) is reached,
- the maximum number of nodes is reached.

The predicted class probability is estimated as the proportion of training samples that were in each class at the terminal node a query sample reaches. As such, if the tree does not stop growing until every terminal node contains only a single class, the tree will optimistically estimate the probability to be 100% for the larger class and 0% for the smaller class.

As well as improving the quality of the estimated probability estimates, the setting of stopping criteria prevents the tree from becoming overly complex and *over-fitting* to the data. Another method to prevent this is to *prune* the tree: using *cross-validation* (partitioning the data into complementary subsets; see Section 5.2) to identify splits that are more likely to result from over-fitting and reduce the predictive accuracy of the model in out-of-sample data. These branches are then pruned, and the node becomes a terminal node.

Shallow decision trees are very easy to interpret, and have value even with very few samples, as demonstrated by the example in Figure 5.2 (which shows a set of rules that can be used to tell me and my siblings apart).

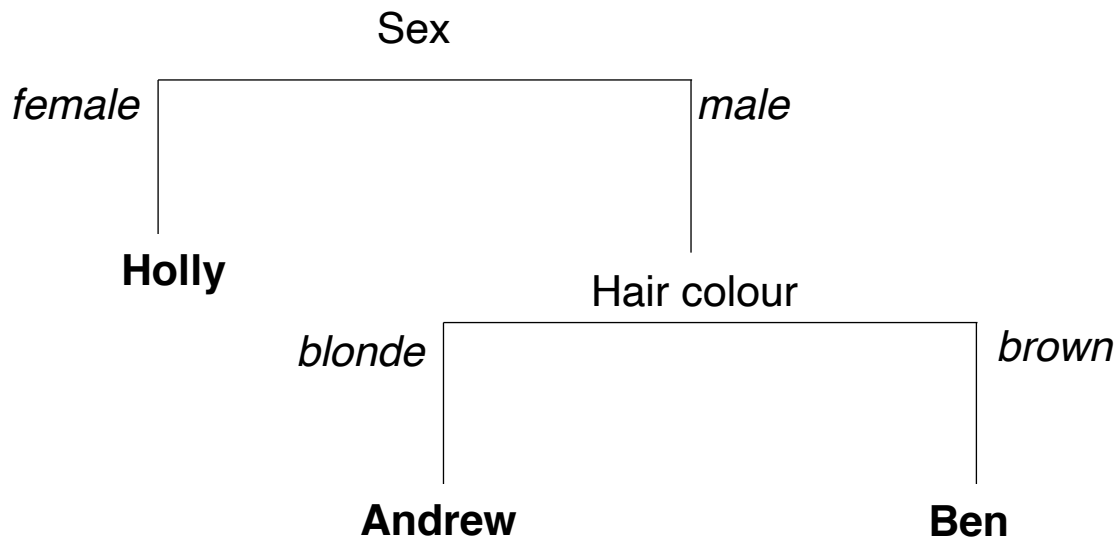


Figure 5.2: A decision tree to distinguish the Tibble siblings

### 5.3.5 Support Vector Machines

A Support Vector Machine (SVM) is a non-parametric model which creates a hyperplane that best defines the areas in  $M$ -dimensional feature space inhabited by samples belonging to one of two classes. To put this more simply, imagine a dataset with two features, and binary classes. An SVM creates a boundary which best separates the samples when plotted.

The calculation of the location of the boundary, known as the *hyperplane*, is an optimisation problem. In the simple *linearly separable* 2-dimensional case, there is a straight line which can be drawn between the two classes and separate them perfectly. To find the optimal hyperplane we can simply iterate between combinations of the samples on the class's convex hull (Figure 5.3, using R.A. Fisher's Iris dataset <sup>298</sup>) and rotate the angle of the two parallel lines which separate them to find the largest margin.

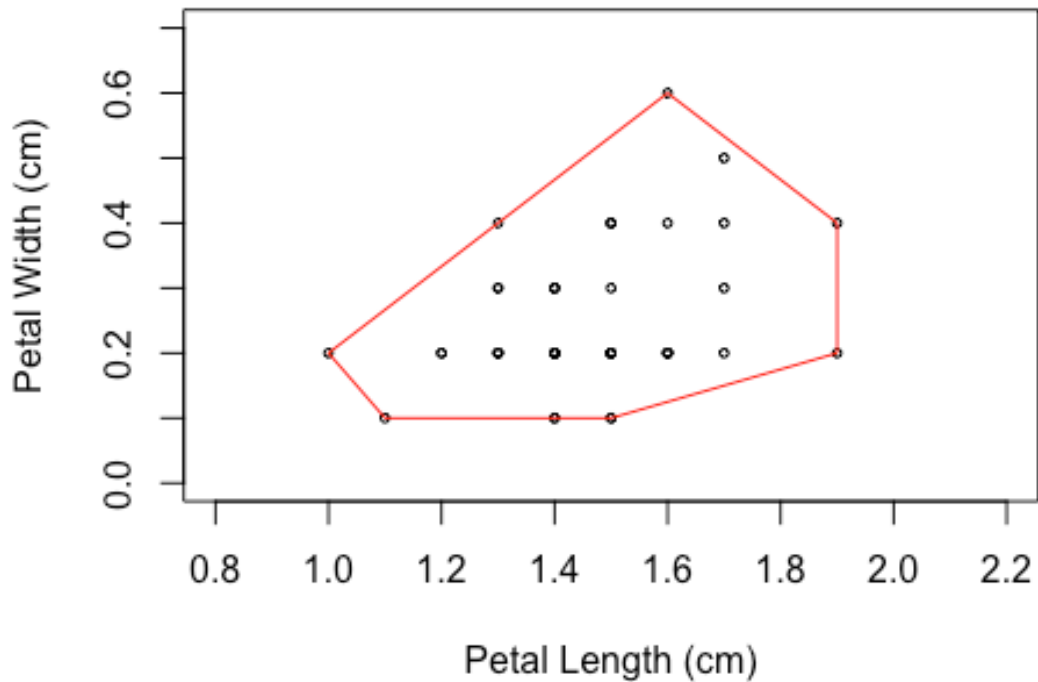


Figure 5.3: Convex hull of one class (Setosa) from the Iris dataset

We can see in Figure 5.4 a hyperplane which divides two classes of Iris flower, based on two features: the width and length of the petals. In the figure, regular samples are represented with a hollow circles and triangles, while the samples on the convex hull which were used to calculate the hyperplane are represented by a filled shape; these are the *support vectors*. The dotted lines are the boundaries for the maximal *margin* between the classes, around the hyperplane.

The example here is very simple, but there will be cases in which the classes overlap in feature space. One way in which non-linearly separable cases can be handled is the implementation of a *soft margin* <sup>299</sup>. Herein, samples are allowed to cross the margin, however they incur a penalty based on the distance they cross it.

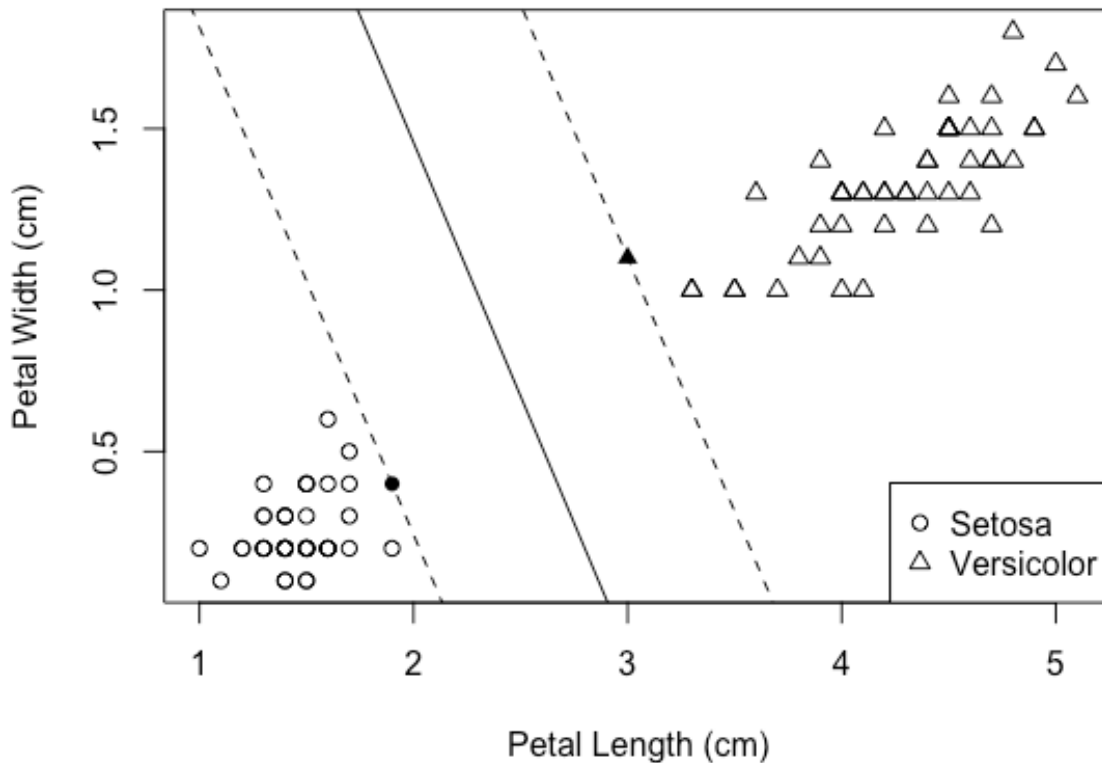


Figure 5.4: Linear support vector machine example using a modification of the Iris dataset, separating iris's of species Setosa and Versicolor by their petal length and width

Non-linear hyperplanes can also be calculated by applying a transformation which expands the feature space and finding some linear boundary in this new higher-dimensional space. The transformation itself is not usually explicitly specified, but we define it by describing the kernel function <sup>290</sup>; that which computes the inner products of two samples (a scalar value from the multiplication of two row vectors in the design matrix) in our new feature space, using the inner products of the same samples in the original feature space.

A back-transformation is required to convert the SVM native output to the approximate class probabilities, commonly using the method of Platt scaling <sup>300</sup>, and implemented by training a cross-validated logistic regression model on top of the SVM's class estimations. This process is thus very computationally intensive in large datasets.



Additionally, while SVMs are strong at predicting with relatively few samples, their complexity (and thus training time) increases drastically with larger training sample size<sup>301</sup>. As such, the computing power required to train an SVM model in such large data as used herein, and the requirements for extensive tuning, meant that it was not feasible to include SVMs in this analysis.

### 5.3.6 Ensemble Learning

Ensemble learning is the method of combining multiple *base models* (also known as *weak learners*), either in parallel or in sequence, in order to improve out-of-sample performance. In order to maximise the performance of an ensemble method, base models should be as diverse as possible, so that they have different regions of competence. Diverse base models can be created in multiple ways, such as altering prediction model parameters<sup>302,303</sup>, using multiple random subsets of the data samples<sup>303–305</sup>, combining base models created using different algorithms (method known as a *heterogeneous ensemble*)<sup>302,303,305</sup>, and using subsets of the available variables (method known as the *random subspace method*)<sup>303,306</sup>. Diversity of base models can be evaluated by calculating the range of pairwise base model correlations<sup>307</sup>, such as the Q-statistic<sup>303</sup> and the double-fault measure<sup>308</sup>, and non-pairwise measures such as the Kohavi-Wolpert variance<sup>309</sup>.

There are three common approaches which describe both the generation of the base models and the consensus function which aggregates them (Figure 5.5). *Bagging* (shortened form of bootstrap aggregating) trains base models on a subset of the full sample, using random sampling with replacement, known as a bootstrap sample. These base models are typically *homogeneous*; they use the same learning algorithm. Using bootstrapping gives us multiple independently sampled subsets of the underlying analysis population, with variations between subsets in the distributions of less common characteristics helping to overcome problems with variance. They also enable probabilistic exploration of the different properties of the feature space, and they can be combined using some deterministic algorithm.

For example, we may take the mean estimation of a regression model, or the modal prediction of a classification model. The aim is that the resulting ensemble has lower variance (reducing the risk of over-fitting) than both the constituent parts and the model trained on the full sample simultaneously. The most common bagging algorithm is the random forest, an extension of decision trees, which is explained in Section 5.3.6.1.

*Boosting* is a sequential method (unlike bagging, in which base models can be generated in parallel) of selecting targeted training data samples in order to improve performance in cases where previous base models have had lower performance. Like bagging, it typically uses homogeneous base models, but instead of focussing on reducing variance, boosting focuses on reducing the ensemble model's *bias*. Reducing bias means avoiding having certain query samples with inaccurate prediction, as a result of the base models not picking up on the nuances captured in the data. High bias results in the opposite of over-fitting, known as *under-fitting*, when the model fails to capture the trends observed in the training data. Two boosting methods, adaptive boosting and gradient boosting, are described in Sections 5.3.6.2.

Finally, *stacking* is an approach for combining heterogeneous base models using a meta-learner, with both of the base learners (collectively) and meta-learner trained on separate data partitions. Meta-learners include the *Behavior Knowledge Space* <sup>310</sup>, by which commonly occurring combinations of characteristics in the training data set are noted along with their true classification, the *Borda Count* <sup>303</sup>, which uses the likelihood of each response from each classifier, rather than simply submitting the best choice to the committee (similar to the alternative voting system), and the *Dynamic Classifier Selection* <sup>311</sup>, which highlights the class suggested by the base model which performed best on the training data on similar samples.

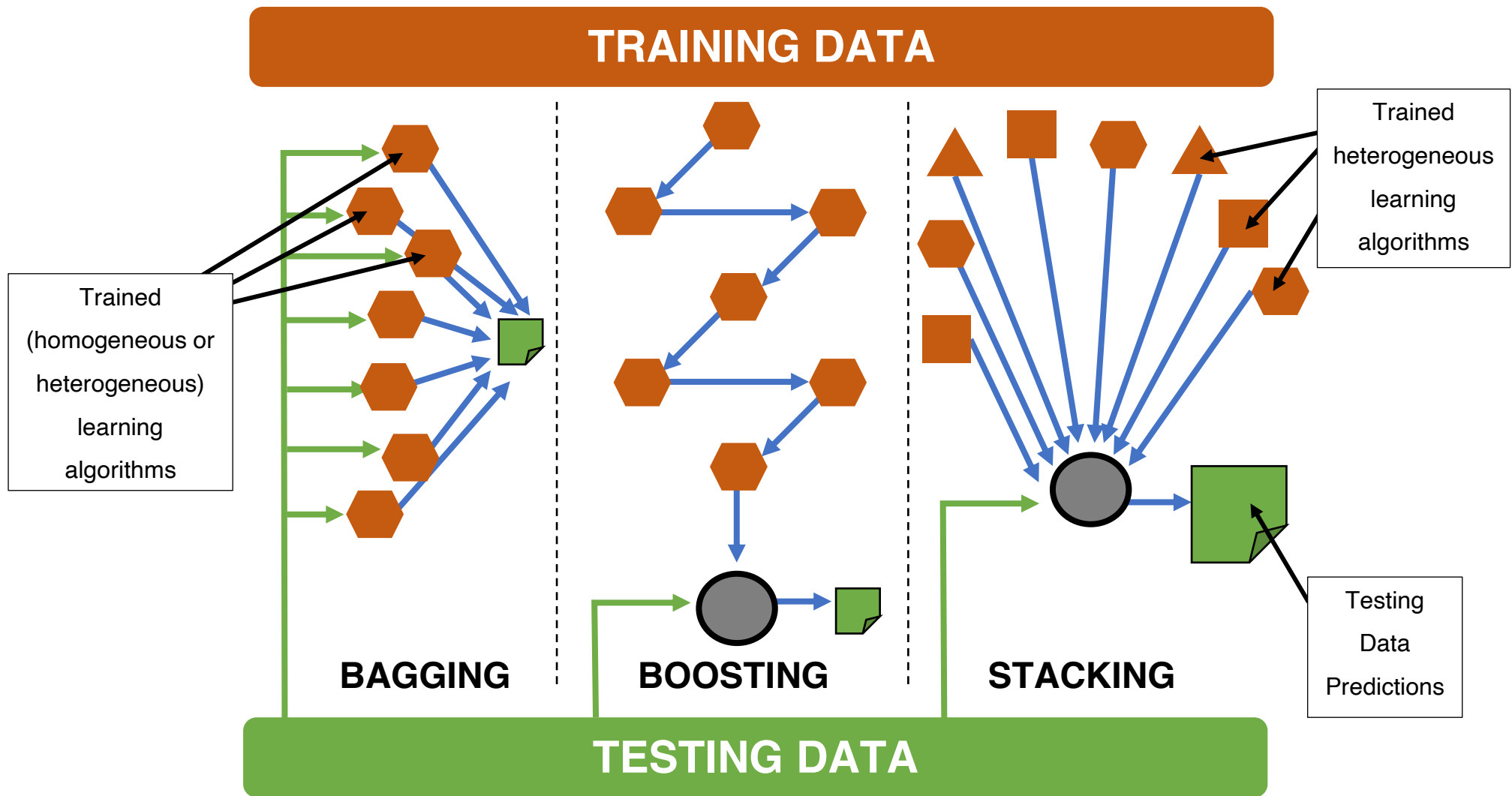


Figure 5.5: A visualisation of the three paradigms of ensemble classification

Note: Testing data are presented into the trained (and optimised) classifiers to provide a best estimate of out-of-sample performance.

There exists no one method (ensemble or consensus building) that consistently outperforms competing learning algorithms across all datasets <sup>312,313</sup> (in machine learning parlance known as the *no-free lunch theorem*), nor are they guaranteed to outperform all of the base models. Furthermore, they are computationally considerably more intensive than single models. They can also be challenging to interpret, because of the internal complexity of the ensemble method and its integral base learners.

#### 5.3.6.1 Random Forests

The Random Forest (RF) algorithm uses multiple (500 by default in most implementations) decision trees as base learners, each constructed using a bootstrapped sample of the data, and selecting the splitting feature at each node from a random subset, to increase the diversity between the trees <sup>314</sup>. Random split selection can also be employed, which may allow a slightly suboptimal feature split to be made <sup>315</sup>.

The predicted class probability output from a random forest is the mean of the predicted class (not the estimated class probabilities) from each of the trees in the forest.

#### 5.3.6.2 Adaptive Boosting and Gradient Boosting

Adaptive boosting (also known as *AdaBoost*) uses sample weights to identify cases in which the model needs improving <sup>316</sup>. To start, AdaBoost produces a base model with equally weighted samples. The base model weight, in the range  $(-\infty, \infty)$ , is calculated using the sum of the weights of the incorrectly classified samples. The sample weights are then updated based on whether the sample was correctly classified in the previous model, and the weight of that model. The next base model is then constructed using these new sample weights. This process iterates until some stopping criteria are met.

Gradient boosting is a generalisation of the AdaBoost algorithm, which allows for a wider variety of loss functions (in classification, a function that maps the design matrix

to a real number intuitively representing some "cost" associated with the misclassification between the observed and predicted outputs). Gradient boosting constructs sequential base models that minimise the designated differentiable loss function, using an iterative optimisation method known as gradient descent: progressing along the loss function in the direction of negative gradient (lower cost) to find a local minimum <sup>317</sup>. Further extensions to the method, utilising the sub-gradient descent procedure <sup>318</sup>, also allow non-differentiable loss functions to be used <sup>319</sup>.

We calculate the pseudo-residuals (error) for each sample as the difference between the estimated probability of the class (calculated by applying the logistic function to the log odds) and the observed class (with values 0 or 1). We then construct a base model to estimate the pseudo-residuals and transform them from probabilities to log odds values. These estimated log odds values are used to update our predicted class for each sample by multiplying them by the learning rate, a value between 0 and 1, and adding this to the previous log odds. This process is iterated until some stopping criteria are met. The learning rate affects how much information we take on from the estimated residuals; a value of 0 means that the predicted classes do not update at all, and a value of 1 means that the predicted classes are completely replaced at every iteration. Smaller values will take longer to converge but will result in lower variance; 0.3 is the default value in the R implementation of gradient boosting (package XGBoost <sup>320</sup>), which is on the higher end of the typically used values (most commonly between 0.1 and 0.3). Like the random forest, by default in most implementations 500 trees are grown.

## 5.4 Evaluating Model Performance

In order to choose the best performing model for a classification task (a process known as model selection), and to evaluate the final product, we must be able to assess how well the model is able to classify new 'query' samples, by comparing the predicted outcome of the model with the *ground truth* (the observed outcome). The ground truth might not always be known, and thus the current state of the art prediction (known as the gold standard) may be used instead.

In this section I will review some common performance measures used in binary classification problems. Later, in Chapter 6, I will further explore the strengths and limitations of each performance measure, in order to determine the best choice to use for my analysis.

### 5.4.1 Probabilistic Performance Measures

Model performance can be quantified by comparing the ground truth to the estimated probabilistic outcome of the classifier (the class probabilities), given a set of query samples. Using the class probabilities allows a robust assessment of two fundamental components of model fit: discrimination and calibration<sup>288,321,322</sup>. Good discrimination means the model can distinguish between the classes well, at some optimal threshold of the estimated class probabilities. For example, in Figure 5.6, one may assign some threshold in the range of 0.2 to 0.5 (for which value there is overlap between the class estimated probabilities) in order to separate the two classes.

Good calibration means that there is strong alignment between the estimated probability and the observed rate of events, which can be evaluated by binning the samples into risk groups, such as deciles (see Section 5.4.4). As such, it is a measure of the precision of the forecast. All performance measures penalise poor calibration to some extent, but in practice, some compromise is necessary between a model's discrimination and calibration.

The most well-known probabilistic performance measure is the AUC, the Area Under the Curve (known as the Receiver Operator Curve, or ROC) formed by plotting the sensitivity (the true positive rate) and specificity (the true negative rate), as shown by the shaded area in Figure 5.7. The AUC is equal to the probability that a randomly chosen negative-class sample will have a lower probability of belonging to the positive class than a randomly chosen positive-class sample<sup>323</sup>.

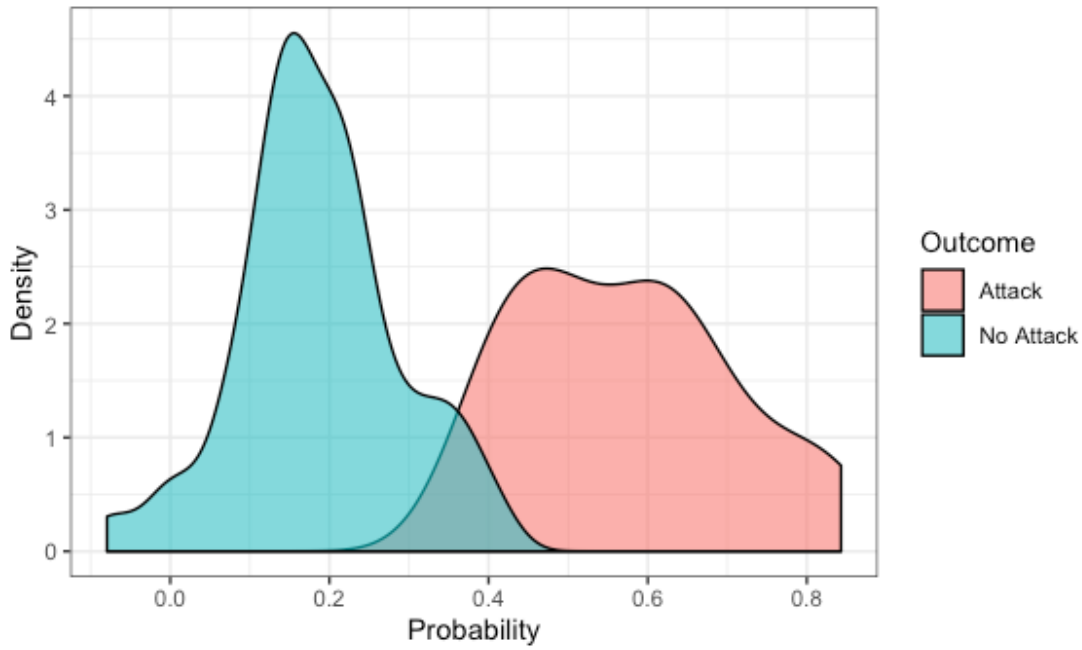


Figure 5.6: Density plot of estimated probabilities by observed outcome

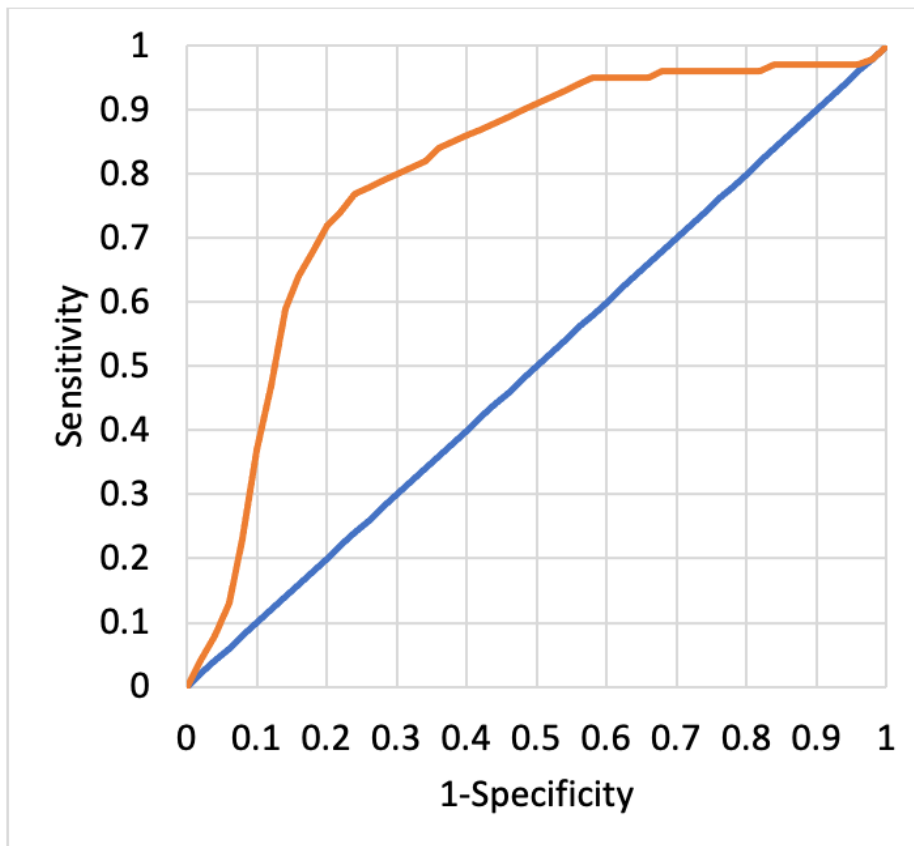


Figure 5.7: Example of a Receiver Operator Curve

A limitation of the AUC is that it is not possible to prioritise minimising false-positives over false-negatives (or vice-versa), which have different misclassification costs. Additionally, it encapsulates performance in regions of the ROC space in which a threshold would never practically be selected, such as the area with either exceptionally low sensitivity or specificity (bottom-left and top-right corners of Figure 5.7) <sup>324</sup>.

Another common probabilistic performance measure is the Brier Score (BS): the mean squared error of the estimated probabilities ( $\hat{y}_i$ ) and the observed outcomes ( $o_i$ ) <sup>325</sup>:

$$BS = \frac{\sum_{i=1}^n (\hat{y}_i - o_i)^2}{n} \quad (5.5)$$

There are many cases when assessment of the raw probabilities is not appropriate. Physician statistical literacy has been found in many studies to be insufficient to ensure that probabilistic outputs will be interpreted effectively <sup>326–330</sup>, and thus we instead assess performance after some decision rule has been defined. In the binary case, this decision rule depends on a single classification threshold value above which the outcome is predicted to occur. This value may be determined by identifying the threshold which optimises the performance, according to non-probabilistic performance measures. The classification threshold might also be pre-defined and not possible to optimise, such as when comparing a model to previous literature.

#### 5.4.2 Confusion Matrices and the Data Imbalance Problem

The performance of a model can be evaluated based on the predicted versus observed class of each query sample, described using the cells of the confusion matrix. The confusion matrix, also known as a contingency table, is a 2\*2 table (or more generally m\*m for multi-class classification problems) of the true and predicted classes. An example of a binary confusion matrix is shown in Table 5.1, including the shorthand notation for the value of each cell.



Table 5.1: An annotated binary confusion matrix

		Observed Outcome (ground truth)	
		Positive	Negative
Predicted Outcome	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Before the confusion matrix performance measures are introduced, there is an important factor relevant to asthma attack risk modelling which must be introduced: Data Imbalance. Consider a dataset in which 99% of the sample belonged to a certain class, denoted as the *majority* (or *major*) class. This is very common when predicting rare events, such as asthma attacks. A model could predict the class of an unlabelled, unseen sample and make the correct prediction for 99% of the samples by simply assigning it to the majority class – without using the features of that sample at all. By some measures, the resulting model would look to be performing very well, but looking at the people who were observed to have the event we would see that none of them were detected by the model.

As such, when there are substantially fewer samples in one class than the other, the performance measure must be selected carefully to ensure that poor performance in one class is detected and penalised appropriately.

### 5.4.3 Confusion Matrix Performance Measures

As shown in Table 5.1, all information about a binary classification model performance can be captured in the four cells (or  $m^2$  cells for an  $m$ -class problem) of the confusion matrix. Despite this, it is frequently useful to be able to summarise all of this

information in a single statistic, for example, when trying to develop an optimisation function or when objectively ranking models for the purposes of model selection. This has sparked interest in the development of classifier performance measures, many of which are summarised in the following sections.

#### 5.4.3.1 Sensitivity and Specificity

The sensitivity and specificity (introduced in Section 5.4.1; also known as the True Positive Rate and True Negative Rate) are *paired measures*, meaning that they only each describe half of the confusion matrix and are rarely considered in isolation. They show the proportion of the samples from the true positive and negative classes, respectively, that are correctly estimated. Using the confusion matrix notation, they are formally expressed as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (5.6)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (5.7)$$

#### 5.4.3.2 Accuracy

The accuracy is the number of correct predictions, both positive and negative, as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.8)$$

#### 5.4.3.3 Balanced Accuracy

The balanced accuracy is an extension of the accuracy specifically for cases with imbalanced classes. It is calculated as the average of sensitivity and specificity as follows:

$$\text{Balanced Accuracy} = 0.5 * (\text{sensitivity} + \text{specificity}) \quad (5.9)$$

The balanced accuracy is also known as the Bookmaker's Informedness (BI) when the latter is scaled to the [0,1] range: the BI otherwise exists in the [-1,1] range.

#### 5.4.3.4 Positive Predictive Value and Negative Predictive Value

The Positive and Negative Predictive values (PPV and NPV) are paired measures (see Section 5.4.3.1) which describe the proportion of samples in the predicted positive and negative classes, respectively, which were correctly predicted. Using the confusion matrix notation, they are formally expressed as follows:

$$PPV = \frac{TP}{TP+FP} \quad (5.10)$$

$$NPV = \frac{TN}{TN+FN} \quad (5.11)$$

#### 5.4.3.5 Markedness

Markedness is a *single* measure (summarises the entire confusion matrix) which uses the PPV and NPV, as follows:

$$\text{Markedness} = PPV + NPV - 1 \quad (5.12)$$

Markedness takes values in the range [-1,1]. To aid comparison between the measures, we have rescaled the markedness to the [0,1] range using min-max normalisation, as follows:

$$x^* = a + \frac{(x - \min(x)) * (b - a)}{\max(x) - \min(x)} \quad (5.13)$$

Where  $x$  is the value pre-scaling, and the desired scale is  $[a,b]$ . The minimum and maximum values of  $x$  are taken from the range of the measure ([-1,1]). The scaled version of the measure,  $x^*$  is equivalent to the average of the sensitivity and specificity.

#### 5.4.3.6 $F_1$ and $F_\beta$ Measures

The  $F_1$  Measure, also known as the  $F_1$  Score, is the harmonic mean of the PPV and the sensitivity. The  $F_1$  Measure is thus defined as follows:

$$F_1 \text{ Measure} = 2 * \frac{TP}{2*TP+FN+FP} \quad (5.14)$$

The generalisation of the  $F_1$  Measure, known as the  $F_\beta$ , enables the user to weight the PPV and the sensitivity, as follows:

$$F_\beta = (1 + \beta^2) * \left( \frac{\text{sensitivity} * \text{specificity}}{(\beta^2 * \text{sensitivity}) + \text{specificity}} \right) \quad (5.15)$$

The value  $\beta$  represents the relative weighting of the sensitivity (accuracy of the true positives) to the PPV (accuracy of the predicted positives). While altering the  $\beta$  value does not change the fact that the positive class is prioritised over the negative class, it allows the user to weight the measure based on the costs of each misclassification respectively. For example, if a false negative is worse than a false positive (such as the risks to the unborn baby if a mother is told they are not pregnant, compared to the distress of mistakenly being told that they are pregnant) then selecting a high  $\beta$  value is preferable. In this analysis, I will include the  $F_{1.1}$  Measure alongside the  $F_1$  Measure.

#### 5.4.3.7 Geometric Mean Accuracy

In mathematics, the geometric mean is an average which uses the product of a set values, as opposed to the arithmetic mean which uses the sum (like in the Balanced Accuracy). In classifier performance, the Geometric Mean Accuracy (GMA) is thus the square-root of sensitivity and specificity, as follows:

$$GMA = \sqrt{\text{sensitivity} * \text{specificity}} \quad (5.16)$$

The GMA is sometimes also referred to as the G-Measure <sup>331</sup>.

#### 5.4.3.8 Matthews Correlation Coefficient

The Matthews Correlation Coefficient (MCC), also known as the Phi coefficient, calculates the correlation between predicted and observed outcomes <sup>332</sup>. It is computed as follows:

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad (5.17)$$

When every sample in the test set is either observed or predicted to be in the same class the MCC fails to compute, due to zero sums on the denominator. The MCC, like markedness, has range [-1,1], and so is scaled to the [0,1] range using the minmax formula (Equation (5.13)).

#### 5.4.3.9 Optimized Precision

Optimized Precision (OP) is a hybrid performance measure, equal to the accuracy minus the Relationship Index (RI) <sup>333</sup>, defined as follows:

$$RI = \frac{|\text{sensitivity} - \text{specificity}|}{\text{sensitivity} + \text{specificity}} \quad (5.18)$$

$$OP = \text{accuracy} - RI \quad (5.19)$$

A lower RI (closer to zero) is best, and occurs when the sensitivity and specificity are higher, but the difference between them is small. The RI has the disadvantage, however, that it equals zero whenever the sensitivity and specificity are equal, regardless of their value. The OP negates this, by incorporating in the accuracy. The OP, like markedness and MCC, is scaled using the minmax formula defined previously (Equation (5.13), from the range [-1,1] to [0,1]).

#### 5.4.4 Model Calibration

Model calibration evaluation is often conducted as a post-hoc analysis of a trained model, to assess how well the predicted risk of an outcome corresponds to the observed outcome on an individual level, as opposed to across the whole population.

A well calibrated model affords additional flexibility of being able to use the estimated class probabilities, rather than just the predicted class, in order to add more nuance to the interpretation of a model's performance.

Two common calibration measures are known as the calibration-in-the-large and the calibration slope, respectively <sup>334</sup>. The calibration-in-the-large quantifies the difference between the means of the observed ( $y$ ) and estimated ( $\hat{y}$ ) probabilities, and can be calculated as the intercept,  $\alpha$ , of the logistic regression shown in Equation (5.20) <sup>335</sup>.

$$y = \alpha + \text{offset}(\hat{y}) \quad (5.20)$$

The calibration slope is estimated by the  $\beta$  term in Equation (5.21). It reflects whether the estimated risks are appropriately scaled, with  $\beta > 1$  indicating that the estimated probabilities do not vary enough; if  $\alpha = 0$  then the estimated probabilities would be too low overall. On the other hand,  $\beta < 1$  indicates that the estimated probabilities are too extreme.

$$y = \alpha + \beta \hat{y} \quad (5.21)$$

It is also possible to recalibrate estimated probabilities <sup>336,337</sup>, such as by Platt scaling (using a logistic sigmoid function; see Section 5.3.5) or isotonic regression <sup>338</sup>. The quality of a recalibration should be assessed in the same manner as any hyperparameter tuning (see Section 5.2): the parameters should be estimated in a training data partition and evaluated when applied to an unseen testing partition. Furthermore, applying a model which is well-calibrated in its derivation dataset to a new external dataset does not guarantee good calibration there, even if the discriminative ability of the model is similar between datasets <sup>339</sup>. As such, it is sometimes desirable to recalibrate models when they are being applied to a new setting.

## 5.5 Training Data Enrichment

Data used in the training of machine learning models can be modified to improve efficiency and accuracy in prediction modelling, known as *data enrichment*. One common reason for employing data enrichment, and indeed the reason it will be used herein, is towards prediction with imbalanced classes, as introduced in Section 5.4.2. With an anticipated asthma attack incidence rate of approximately 0.16 attacks per patient per year <sup>340</sup> (estimated in an unselected UK asthma population, using the ATS/ERS outcome definition <sup>78</sup>, introduced in Section 2.4), imbalanced data are likely to pose a significant barrier to model performance without preventative measures.

Training data enrichment methods can either increase or decrease the size of the training dataset, by adding or removing samples in order to create a more *balanced* dataset, with roughly equal sample representation in each class <sup>341</sup>. *Over-sampling* is an additive method of adding weight to samples of the minority class by duplicating them a specified number of times. This duplication of samples can lead to over-fitting in the minority class, which will result in high predictive performance in the training data but generally lower performance in the unseen data <sup>342,343</sup>. *Random over-sampling* specifically replicates a random subset (with replacement; class-specific targeted bootstrapping) of the minority class, rather than all samples; when conducted multiple times, this reduces the likelihood of over-fitting by preventing individual samples from being too influential. Sample synthesis is a sub-class of additive methods, in which new samples are *synthesised* (artificially generated) rather than replicating already existing samples.

Restrictive methods reduce the number of samples in the training dataset by subsampling the majority class. *Random under-sampling* <sup>342,343</sup> removes samples in the majority class at random. Restrictive methods are useful in cases where there is sufficiently large number of samples belonging to the smaller class, and sample size in the majority class which is larger than required for efficient prediction, although one drawback is the risk that nuances in the majority class determination may be missed by the removal of pertinent samples <sup>343</sup>.

Finally, there are methods which combine restrictive and additive sample-based enrichment methods in order to reduce the negative effects associated with each method class respectively. *Informed sample-based enrichment methods* make use of the underlying structure of the data, such as hierarchy of samples and the distribution of features, in order to prevent the loss of pertinent data or, conversely, the duplication of particularly specific records likely to result in over-fitting. In *Synthetic Minority Over-Sampling TEchnique* (SMOTE) <sup>341,344</sup>, each minor class sample is paired with another from its K-nearest minority class neighbours, and feature values are generated from a uniform distribution within the range of the example sample pair (demonstrated in Figure 5.8, using R.A. Fisher's Iris dataset <sup>298</sup>). This is repeated a specified number of times for each minor class sample. SMOTEing can also use random under-sampling.

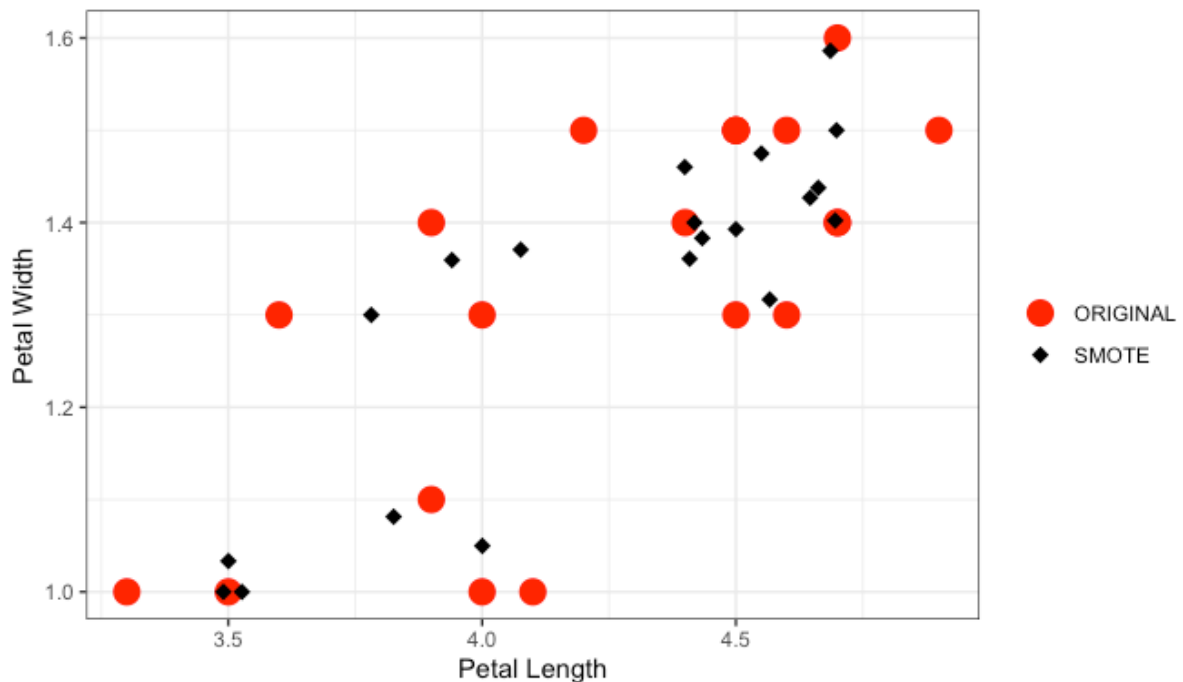


Figure 5.8: A scatterplot showing original samples of versicolor irises, from R.A. Fisher's Iris dataset, alongside SMOTE generated samples



There are two main parameters in the SMOTEing function, for the over-sampling and under-sampling rates. These parameters, and the effect on the training data sample size, will be described using the following notation:

$a$  = the number of training samples in minor class A

$b$  = the number of training samples in major class B

The SMOTEd training data generates  $k$  new synthetic samples for each sample in the minor class, resulting in  $(k+1)*a$  samples in the minor class, and retaining  $z*k*a$  major class samples.

If  $z*k*a$  is in fact larger than  $b$ , then we instead generate  $b-z*k*a$  new samples in the major class. However, as class B already contains more samples than class A, we do not need to generate more samples in this class. In fact, we want to explicitly avoid it. As such, we limit  $z$  to be any positive number:

$$z \leq \frac{b}{a*k}, z \in \mathbb{R}^+ \quad (5.22)$$

Furthermore, we do not want to produce more samples in class A than there are in class B. As such, we limit  $k$  to be any non-negative integer:

$$k \leq \frac{b}{a} - 1, k \in \mathbb{N} \quad (5.23)$$

The user can also specify the number of neighbouring samples (closest distance in Euclidean space) to draw from in the creation of new synthetic samples, commonly 5. In their 2017 asthma attack risk prediction study, Finkelstein and Jeong<sup>345</sup> compared the results from three algorithms (NBC, SVM, and adaptive Bayesian networks), and demonstrated that under-sampling improved sensitivity with mild (or no) decline in specificity, compared to the original training data. Zhang *et al.*<sup>346</sup> compared varying levels of under-sampling, over-sampling, and SMOTEing, for logistic regression, and similarly found that all tested variation improved sensitivity from 0% to at least 55%

(maximum 87%) with a maximum reduction in specificity from 100% to 84%. No changes in the AUC (see Section 5.4.1) were observed.

## 5.6 Model Interpretability

A model can be considered interpretable if the reasons behind a prediction are intuitively understandable to a human. A simple decision tree, like the one shown in Figure 5.2, is completely interpretable. For a deeper decision tree, it may be possible to see in specific (*local*) cases the ruleset that a patient met in order to be classed as they were, but harder to see how all possible rulesets relate to all possible outcomes (*global interpretability*). In the case of a random forest, for example, neither may be easily inferred.

While the attribute of interpretability itself is hard to quantify<sup>347</sup>, more pertinent perhaps is describing the methods by which interpretability might be improved, such that some explanation can be provided to the model's user. This might serve to facilitate trust in the system, or to highlight clinically irrelevant patterns that the training data might have provided<sup>347,348</sup>.

### 5.6.1 Global Model Interpretation

Global interpretability is the trait that the logic and reasoning behind the entire model can be easily understood by a human. Few algorithms inherently provide global interpretability, but for many algorithms the contribution of individual features to the model can be used to aid interpretation of the results.

For linear methods, the feature weight model coefficient can be simply interpreted relative to the other features and to the model intercept. For non-linear methods, the predictive value of each feature (known as the *feature importance*) can be computed by comparing the absolute difference in the performance of the model (see Section 5.4) to the performance when values for each of the features in turn are randomly permuted across testing samples<sup>314</sup>. This estimate can be validated further by including a single random feature and comparing its importance to other features, but

is nonetheless liable to inflate importance of correlated variables <sup>349</sup>. Additionally, for tree-based methods, feature importance can be quantified by averaging the mean decrease in impurity (or the corresponding performance measure used to grow the tree) that would be achieved by using each feature as the splitting criterion for each parent node in each tree <sup>314</sup>. These values should be considered relative to each other, rather than in terms of the computed magnitude.

Another way of gaining global model insight is the application of an *interpreter*, or *explanation*, model over the final prediction model. For example, *single-tree approximation* is the application of a single decision tree to demonstrate the main drivers of a black-box model's predictions <sup>348</sup>. The interpretation model can be appraised on the basis of interpretability and accuracy, but also *fidelity*: how well the interpretation model predicts the outcome predicted by the primary prediction model <sup>348</sup>.

## 5.6.2 Local Model Interpretation

When presented with a prediction that differs greatly from the expected outcome, understanding the model's *local* predictions may be more efficient than the global interpretation.

In a linear model, such as logistic regression, the local effect of each feature is simply the product of the feature weight (coefficient) and the value itself. For non-linear models, we can approximate the effect using Shapley values. Shapley values originated in the field of game theory, used to determine an individual player's contribution to the pay-out in a collaborative game. They can be applied to predictive modelling to estimate how much an individual feature (*player*) contributes towards the prediction (*pay-out*) of a single query sample (*the game*), compared to the average prediction <sup>350</sup>. For example, if the average estimated risk of an asthma attack is 10% in the training data, and one patient has an estimated risk of 60%, we want to estimate the marginal contributions of each characteristic of that patient to this difference (50% increase).

To simplify the explanation of how the estimation of the marginal contributions work, envision a simple prediction model with only four features ( $F=4$ ), of which two are binary (A and B) and two are continuous (X and Y). For our query sample (patient) in question, A is true, B is false, C is 100, and D is 5. To estimate the Shapley value for feature D, we must first define all possible *coalitions* ( $C$ ): unordered sets containing between 0 and  $n-1$  of features. There are  $k=8$  coalitions excluding the feature Y:  $C_Y = \{\text{no features, A, B, X, A+B, A+X, B+X, A+B+X}\}$ .

The first coalition we will look at is A+X, which we will arbitrarily label  $C_{Y,6}$  as it was the sixth in the list above. We note that there are three coalitions of the same size as  $C_{Y,6}$ :  $\mu(C_{Y,6})=3$ . We simulate a set of samples for this coalition by finding  $m$  random training samples with A=TRUE and X=100, and using their values of the missing feature B. We then calculate the estimated event probability for these  $m$  samples: (1) when  $Y=5$  (from the patient's data), and (2) using the samples' Y values. The mean difference between (1) and (2) is the estimated contribution in this coalition (herein denoted  $\omega(C_{Y,6})$ ). This process is repeated for all coalitions, and the Shapley value ( $\varphi_Y$ ) is the weighted average of the marginal contributions across the coalitions:

$$\varphi_i = \frac{1}{F} \sum_{j=1}^k \frac{\omega(C_{i,j})}{\mu(C_{i,j})}$$

The sum of  $\varphi_i$  for each feature ( $i$  between 1 and  $F$ ) is equal to the difference between the estimated event probability for this patient and the estimated event probability for the average patient in the training data.

## 5.7 Summary

In this chapter, I have described the technical methods which have been considered for use in my asthma attack risk prediction model. These included machine learning classification algorithms, performance measures, training data enrichment methods, and interpretability aid methods.

In chapter 7, I will be utilising logistic regression, naïve Bayes classifiers, RF, and extreme gradient boosting algorithms, with a selection of training data enrichments. There is no gold standard approach to performance measurement in classification modelling, as the most appropriate performance measure is dependent on the data (including the class imbalance) and the priorities with regards to minimising errors. In the next chapter, I conduct a detailed review of the performance measures under various empirical and experimental conditions, in order to evaluate their utility for my model selection process.

# 6 Performance Measures for Binary Classification Problems

There are a number of classification performance measures which can be computed directly from a confusion-matrix, some of which have been described in Section 5.4.3. In this section, I conduct a theoretical exploration of binary classification performance measures, in order to guide my selection in the testing of my asthma attack risk prediction model. A particular focus of this body of work is the impact of class imbalance (introduced in Section 5.4.2).

## 6.1 Previous Work

To designate our primary performance measure for our risk prediction model selection, it is essential to understand how performance measures are affected by data idiosyncrasies. There is no gold-standard performance measure which should be used primarily in all cases, or that can capture how well a model performs without losing some of the information captured within the confusion matrix. The use of multiple performance measures retains more of the nuance of the confusion matrix<sup>351</sup>, but there are benefits to using a single performance measure. Firstly, it facilitates the objective ranking of models for the purposes of model selection. Secondly, it is easier to define an objective loss function with a specific measure as part of the optimisation algorithms, towards selecting the best-performing hyper-parameters.

There have been numerous attempts previously to pragmatically explain the differences between various classification performance measures, often aiming to examine the conditions under which a performance measure may produce undesirable estimates<sup>352–355</sup>. These studies each provided a useful perspective on the problems faced when choosing a classification performance measure, and specific use cases in which one performance measure is more illustrative of the model's shortcomings than another. Lacking, however, was pragmatic guidance on the

generalisable implications of their findings, and a pathway for performance measure recommendation.

For example, Chicco and Jurman <sup>353</sup> compared three performance measures (Accuracy, F<sub>1</sub> Measure, and the MCC) in six simulated binary confusion matrices representing use cases defined by levels and direction of class imbalance. They demonstrated that MCC was more informative than the accuracy when classes were imbalanced, and more informative than the F<sub>1</sub> Measure when performance was poor in the negative class. Despite highlighting the cases in which the F<sub>1</sub> Measure is inherently more informative, such as in the prediction of DeoxyriboNucleic Acid (DNA) sequence variants, as discussed by Brown <sup>352</sup>, the authors close by stating their (contradictory) belief that the MCC should be used preferentially over the F<sub>1</sub> Measure and the accuracy in all binary classification task evaluations. Caution should be taken before recommending any such ‘one-performance-measure-fits-all’ approach, and a key priority in this chapter is to investigate the most appropriate measure for specific use-cases.

Sokolova and Lapalme <sup>355</sup> also conducted an investigation of a small selection of measures, but covering a broad spectrum of cases (including multi-class and multi-label classification performance measures). The measures were compared according to eight properties, however they failed to provide any mapping between these properties and pragmatic cases, and many of the measures were not possible to distinguish.

Studies such as Alaiz-Rodriguez *et al.* <sup>356</sup> and Kouznetsov and Japkowicz <sup>357</sup> have also demonstrated methods of aggregating multiple performance metrics into one single metric, using a consensus approach for model selection in a similar way that ensemble models use consensus for classification itself, however this can make interpretation of results much more complicated. Therefore, it is useful to explore a single performance measure for our purposes.

In this chapter, I aimed to use simulated confusion matrices to explore the various methods by which predicted-class based performance measures have been compared and contrasted, and to conduct further empirical analyses to facilitate the selection of the performance measure that will be used in optimisation processes in the development of my asthma attack risk prediction model.

## 6.2 Methods

### 6.2.1 Simulated Confusion Matrices

I conducted two experiments using simulated confusion matrices to illustrate the effects of error imbalance and class imbalance in specific cases of their performance measures.

The first experiment used confusion matrices with varying degrees of class imbalance; changing the size of the negative class relative to the size of the positive class (with 100 samples). Confusion matrices were generated for each combination of sensitivity and specificity in the range 0.1,0.5,0.9, and the Class Imbalance Coefficient (CIC) with values  $\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4,$  and 5. The construction of the confusion matrices is shown in Table 6.1. This resulted in 81 confusion matrices.

Table 6.1: Confusion matrix cell calculations for experiment 1 with varying class imbalance coefficients

	<b>Observed Positive</b>	<b>Observed Negative</b>
<b>Predicted Positive</b>	sensitivity*100	(1-specificity)*100*CIC
<b>Predicted Negative</b>	(1-sensitivity)*100	specificity*100*CIC

Note: CIC = Class Imbalance Coefficient, in range ( $\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4, 5$ )



The second experiment varied the sensitivity and specificity more widely for five cases: 10% positive samples, 25% positive samples, 50% positive samples (balanced classes), 75% positive samples and 90% positive samples. For every case and combination of the sensitivity and specificity between 0.005 and 0.995, at intervals of 0.005, a confusion matrix was constructed with 1000 total samples, as shown in Table 6.2, with 39,601 confusion matrices for each case, where  $z$  equals the proportion of the samples that are in the positive class.

Table 6.2: Confusion matrix cell calculations for experiment 2 with balanced classes

	<b>True Positive</b>	<b>True Negative</b>
<b>Predicted Positive</b>	$1000 * z * \text{sensitivity}$	$1000 * (1 - z) * (1 - \text{specificity})$
<b>Predicted Negative</b>	$1000 * z * (1 - \text{sensitivity})$	$1000 * (1 - z) * \text{specificity}$

Note:  $z$  = the proportion of the samples that are in the positive class, with range (0.1, 0.25, 0.6, 0.75, 0.9)

## 6.2.2 Real Datasets

Two datasets from the University of California, Irvine (UCI) machine learning repository were used to further describe the nuances of the performance measures. The datasets were chosen on the basis of their large size (>25,000 samples), small number of features (<25), and diverse class imbalance proportions.

The first dataset I used in these empirical demonstrations is the UCI dataset ‘default of credit card clients’ (henceforth default) <sup>358</sup>. The dataset contained 24 features (3 nominal, 1 binary, 20 numeric) and 30,000 samples, of which 22% were in the positive class and 78% were in the negative class (Table 6.3). One of the nominal features was a case identifier, which was removed. The two remaining nominal features had 4 and 7 levels respectively. These features were *one-hot encoded*: a data transformation in which each unique level becomes a new binary feature <sup>359</sup>. For example, one-hot encoding a categorical feature ‘eye colour’ might result in three new binary features ‘blue’, ‘brown’, and ‘other’, of which exactly one will have the value 1

for each sample (else 0). The resulting dataset had 20 numerical features, and 12 binary features (1 plus 11 from one-hot encoding).

The second UCI dataset was ‘poker-hand-training-true’ (henceforth poker) <sup>360</sup>, which contained 10 nominal features for 25,010 samples (Table 6.3). Five of the features indicate the rank of card in the player’s hand. While the rank of a card in a poker hand can be considered ordinal, a value of 1 (an ace) is simultaneously both lower than 2, and higher than 13 (a king). As such, I treated the rank as nominal in this dataset. The outcome in this dataset is ordinal and multiclass, with range 0-9, indicating the best hand-rank achievable with the player’s hand.

The poker dataset was used in multiple analyses, as the range of outcomes allowed flexibility in the dichotomisation to a binary outcome to create a range of class imbalance proportions that were not available from other UCI datasets. First, a variation was created by dichotomising on whether or not the player’s hand was ‘bust’ (fitted no ranked category; rank 0 in this dataset). This variation, known henceforth denoted *poker1*, had perfectly balanced classes. The second variation, *poker2*, was created by dichotomising above rank 1 (indicating that the hand could beat a pair of aces, assuming ace-high) and had a minor class proportion of 8%. Finally, the third variation, *poker3*, dichotomised at whether the player’s hand could beat a three-of-a-kind on aces, with 1% probability.

Table 6.3: UCI dataset characteristics, before processing

Dataset Name	Default	Poker1	Poker2	Poker3
<b>Samples</b>	30,000	25,010		
<b>Classification</b>	Binary	Multiclass (ordinal)		
<b>Class Balance</b>	22% positive, 78% negative	50% positive, 50% negative	8% positive, 92% negative	1% positive, 99% negative
<b>Features</b>	24	10		
<b>Feature Types</b>	3 nominal, 1 binary, 20 numerical	Nominal		

### 6.2.3 Analysis Plan

The effects of the varying parameters in the experiments with simulated confusion matrices were visualised with heatmaps, in order to compare the patterns observed in different performance measures.

The empirical analyses consisted of seven datasets and variations (Table 6.4), each randomly partitioned iteratively 100 times, such that 80% of the data were used for training and 20% was used for testing. In the fifth analysis, a training data enrichment method was applied (see Section 5.5), a modification to only data used for model training aiming to improve model performance in the testing partition. Specifically, the method employed herein, known as (random) under-sampling, works by removing a random selection of samples from the majority class in order to retain a subset of the original data samples and ensure the classifier is presented with a balanced dataset. Under-sampling may improve the performance of the model, but tends to increase the variance of the performance measure across random samples, as occasionally important samples will be discarded<sup>361</sup>; this balance is commonly known as the *bias-variance trade-off*<sup>362</sup>. This procedure is only conducted on the training data partition, and the testing data partition was identical (for each iteration) to that of the third analysis.

In the sixth analysis, only the testing partition was modified, with slight under-sampling to increase the positive class percentage from 8% to approximately 19% (deviation caused by under-sampling post random partitioning, with standard deviation of 0.8).

The classification algorithm used for the empirical data analysis was an RF, using Breiman's implementation from the R package `randomForest` with default settings and hyper-parameters: 500 trees were 'grown' (hyper-parameter *ntree*), the number of features evaluated as candidates at each split was equal to the square root of the number of features (rounding to the lower integer number, if needed; hyper-parameter *mtry*), and there was no requirement for a minimum number (greater than one) of training samples in each terminal node (hyper-parameter *nodesize*). Thus the tree will finish growing when all samples in each terminal node are of the same class.

Table 6.4: Description of empirical analyses of performance measures

Analysis	Dataset/Variation	Analysis Description
1	Poker1	Balanced Data
2	Default	Mild Positive Imbalance
3	Poker2	Moderate Positive Imbalance
4	Poker3	Extreme Positive Imbalance
5	Poker2 + Training Enrichment	Moderate Positive Imbalance with Under-sampled Training Data
6	Poker2 + Testing Modification	Moderate Positive Imbalance with Mildly Imbalanced Testing Data

For each dataset, the RF algorithm was trained on the (randomly sampled) training partition, and the resulting model was tested on the testing partition. For each of the 100 iterations, the number of true and false predicted negative and positive testing samples were recorded (the cells of the confusion matrix). Each of the performance measures was then calculated from the confusion matrix of each iteration and dataset (with primary and enriched training data).

The distribution of each performance measure in each dataset was evaluated (see Appendix K) and, given the non-normal distribution in many of the performance measures, the across-iteration performance was summarised using the median and interquartile range. Some performance measures fail to compute when specific cells of the confusion matrix are zero. The number of iterations for which this was the case was recorded, and no values were imputed, such that these iterations did not contribute towards the averages.

## 6.3 Results

### 6.3.1 Experiment 1: The effect of class imbalance on performance measures with set true positive and negative class accuracy

In the first experiment, I changed the size of the negative class relative to the constant positive class size and recorded the performance for 9 combinations of the true

positive and negative rate. As the sensitivity and specificity are only affected by performance in the positive and negative classes, respectively, there is no change in performance as the class balance changes. The performance in the positive class (sensitivity) has negligible effect on the accuracy when the CIC is high, indicating that the negative class is larger than the positive class (Figure 6.1), and vice versa. When the class performances are equal, the accuracy is unaffected by the class imbalance. The same is true for the OP, however unlike the accuracy when the performance in one class is very low, even good performance in a larger class has a minimal effect on the overall performance.

The Balanced Accuracy, GMA and  $F_{1.1}$  Measure are all unaffected by the change in class balance when the class performances are held constant. They vary in how much they penalise poor calibration, with the  $F_{1.1}$  Measure showing the greatest effect, and the balanced accuracy showing the least. The MCC shows only a mild effect from class imbalance but demonstrates that balanced classes only result in better performance when both classes are performing well. When both classes are performing badly, unbalanced classes actually result in better performance. The same, but more extreme, is shown for markedness.

Finally, the  $F_1$  Measure is the only performance measure to discriminate between which class is the positive and which is the negative: positive class performance is far more important than performance in the negative class. There are cases when this would be useful, such as when the costs associated with misclassifying a true positive case (someone with the outcome) are substantially higher than for misclassifying a true negative case. For example, incorrect diagnosis of a dangerous and contagious disease leads to undesirable quarantine, whereas incorrectly not diagnosing someone could lead to a deadly outbreak.

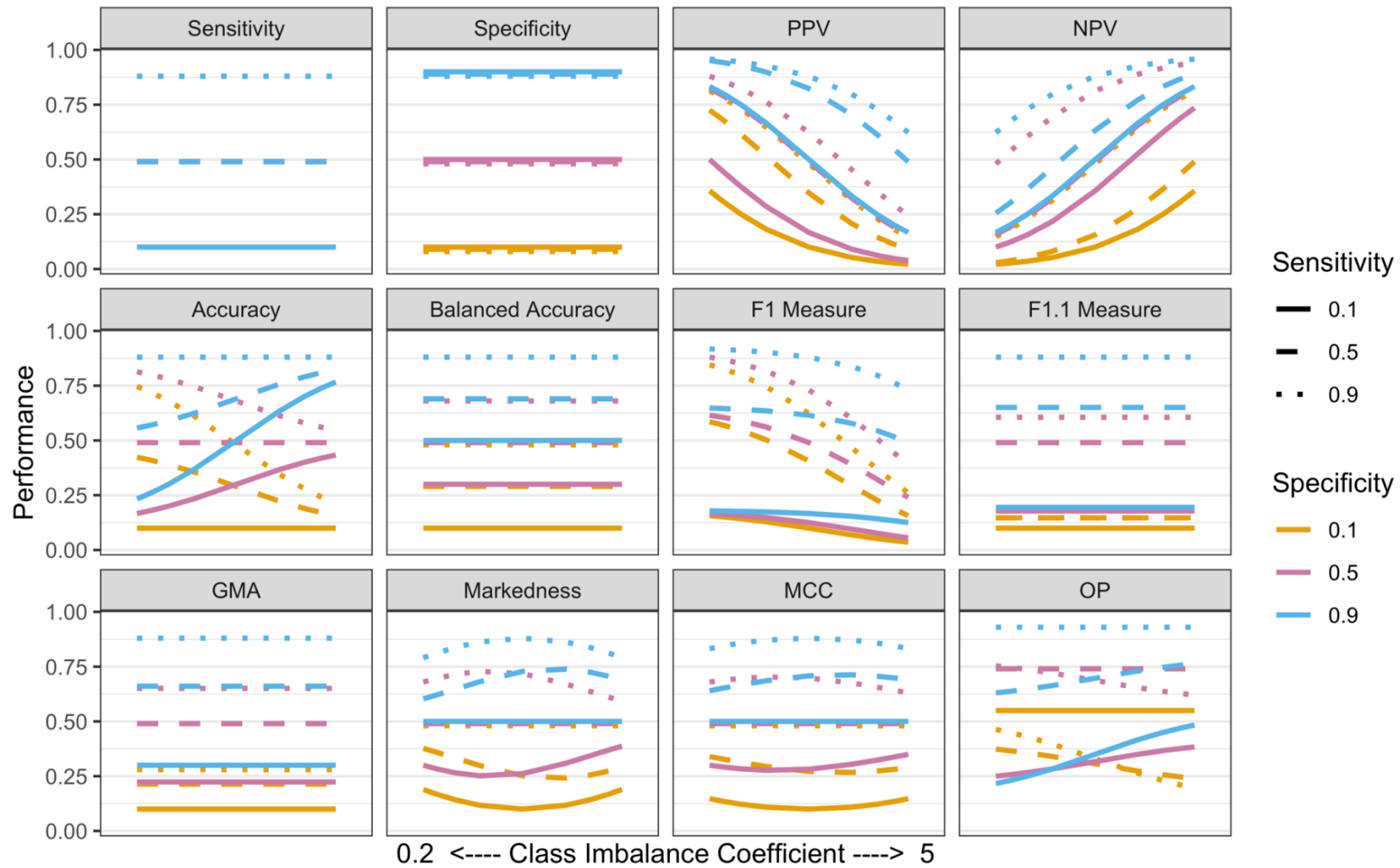


Figure 6.1: The effect of varying the Class Imbalance Coefficient (CIC; size of the negative class relative to the positive class, which had 100 samples), for set values of the sensitivity and specificity

There are also, however, cases when this is not true. It could be that the correct classification of a negative case is more important. In this case, however, you could still use the  $F_1$  Measure by switching the positive and negative classes. Another instance in which the true positive rate is not the more important is when the classes are both important, such as when the treatment regimen for a condition is dangerous, expensive or distressing. In the case of asthma attacks, failing to identify and treat a patient who subsequently experiences an attack is worse than unnecessarily treating a patient with a course of OCS, despite the risk of consequential side-effects. A model may repeatedly classify an individual as high-risk, however, and recurrent OCS courses accumulate a higher risk of adverse outcomes.

### 6.3.2 Experiment 2: The effect of true positive and negative class accuracy on performance measures with set class imbalance

In practice, imbalanced classes can have a substantial effect on the individual class performances. In this section, I explore how the performance measures vary in empirical analyses. Figure 6.2 highlights the effect that interactions between the sensitivity and the specificity have on performance for set values of class imbalance.

Accuracy has parallel contour lines of performance, which means the two classes both contribute linearly to the overall performance. The slope of the contour lines demonstrates the weighting of each class according to the class balance. When balanced, the class accuracies contribute equally to the accuracy. The same is true for the balanced accuracy, both when classes are balanced (in which case it is equal to the accuracy) or imbalanced.

The  $F_1$  Measure has only minimal interaction between the classes, with performance in the positive class being the dominating factor (close to the sensitivity). As such, it is possible to have good overall performance according to the  $F_1$  Measure despite poor performance in the negative class, when the positive class is small, or even when classes are balanced. On the other hand, not even perfect sensitivity and specificity will result in very high performance.

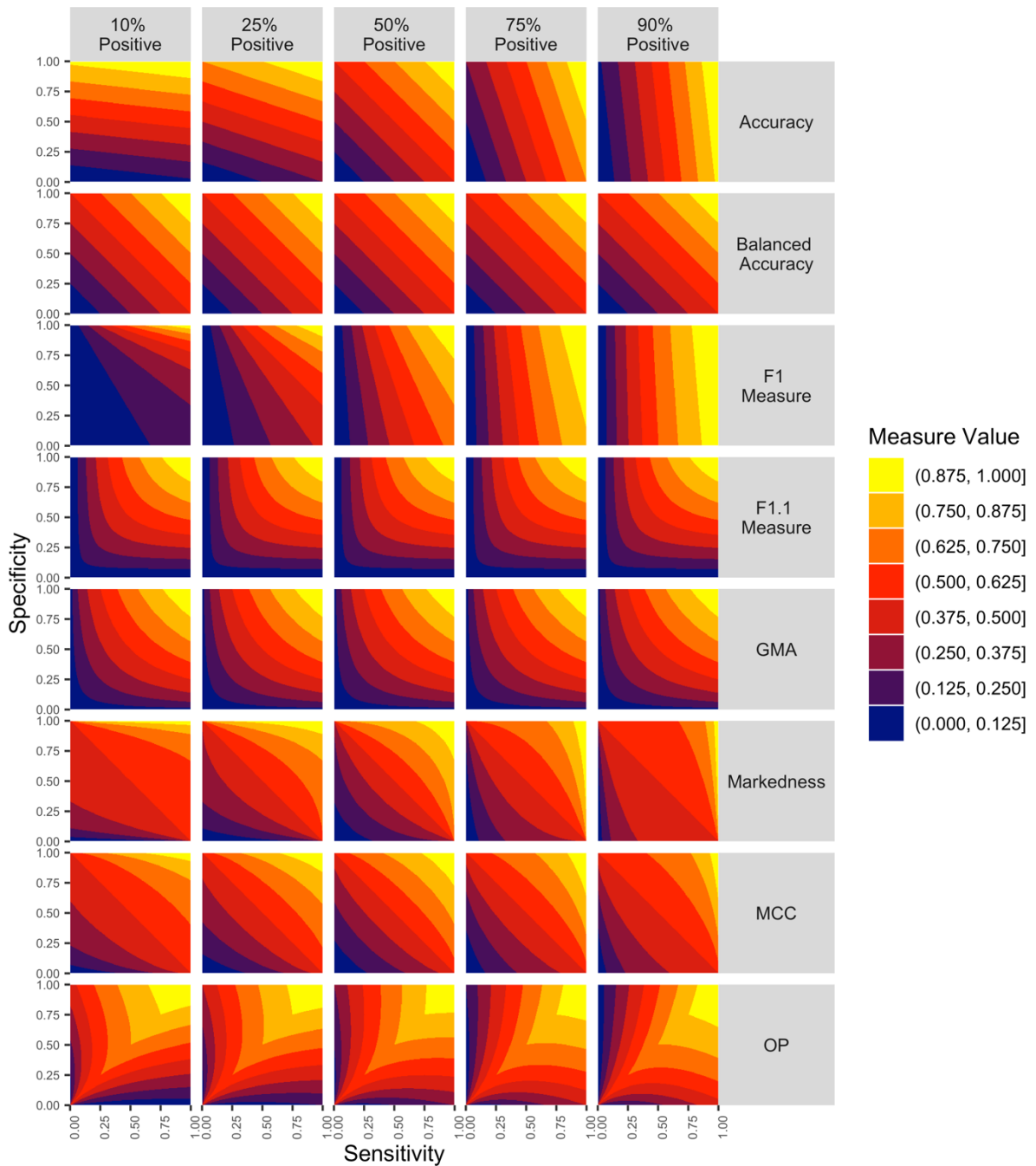


Figure 6.2: The effect of varying sensitivity and specificity on performance measures for balanced (50% positive samples) and imbalanced classes (10%, 25%, 75% and 90% positive samples)



However, using a  $\beta$  value of 1.1 instead of 1 has a big difference, and there is a much more substantial interaction between the classes. The  $F_{1.1}$  Measure and the GMA are similar, and are both sensitive to poorly calibrated models, such that even substantial improvements to performance in one class can have very little effect on the overall performance when performance in the other class is low. The OP is even more sensitive to poorly calibrated models and can provide misleading performance evaluations when performance in either class is low. Like the  $F_{1.1}$  Measure and the GMA, there is very little change across the three cases.

The MCC and markedness are similar to the balanced accuracy but penalise poor calibration slightly less. For both of these performance measures, high imbalance means that smaller changes in the major class performance are weighted more than larger changes in the minor class performance. For example, when 90% of samples are positive, the MCC is higher (in the yellow zone in Figure 6.2) when the specificity is 65% and the sensitivity is over 99% (MCC = 0.880) than when the specificity is 90% and the sensitivity is 95% (orange zone: MCC = 0.873).

### 6.3.3 Empirical Analyses

In the first empirical four analyses, I compared varying levels of imbalance between the classes (Table 6.4). Analysis 1, 3 and 4 use the poker dataset, with 50%, 8% and 1% of samples being in the positive class, respectively. The second analysis uses the default dataset, with 22% of samples in the positive class.

When the classes were balanced (analysis 1), the sensitivity and specificity were similar (0.66 and 0.73, respectively; Table 6.5). All performance measures except the OP (which was the highest at 0.82) fell in the range between the sensitivity and specificity and had similarly low interquartile widths (the difference between the upper and lower interquartiles; 0.008 to 0.012 compared to 0.013 and 0.014 for the sensitivity and specificity, respectively).

Table 6.5: Results in empirical analyses with varying levels of class imbalance

Performance Measure	Analysis 1: 50% Positive	Analysis 2: 22% Positive	Analysis 3: 8% Positive	Analysis 4: 1% Positive
	Median (interquartile width) [% failed to compute]			
AUC	0.766 (0.009) [0]	0.767 (0.009) [0]	0.779 (0.016) [0]	<b>0.824 (0.042) [0]</b>
Sensitivity	<b>0.660 (0.014) [0]</b>	0.367 (0.015) [0]	0.010 (0.006) [0]	0.000 (0.000) [0]
Specificity	0.734 (0.013) [0]	0.946 (0.004) [0]	<b>1.000 (0.000) [0]</b>	<b>1.000 (0.000) [0]</b>
PPV	0.714 (0.012) [0]	0.660 (0.023) [0]	<b>0.750 (0.175) [0]</b>	0.000 (0.000) [96]
NPV	0.683 (0.011) [0]	0.841 (0.006) [0]	0.924 (0.004) [0]	<b>0.992 (0.001) [0]</b>
Accuracy	0.697 (0.008) [0]	0.819 (0.005) [0]	0.924 (0.004) [0]	<b>0.992 (0.001) [0]</b>
Balanced Accuracy	<b>0.697 (0.008) [0]</b>	0.657 (0.008) [0]	0.505 (0.003) [0]	0.500 (0.000) [0]
F <sub>1</sub> Measure	<b>0.686 (0.010) [0]</b>	0.472 (0.014) [0]	0.020 (0.011) [0]	0.000 (0.000) [0]
F <sub>1.1</sub> Measure	<b>0.699 (0.008) [0]</b>	0.553 (0.015) [0]	0.022 (0.012) [0]	0.000 (0.000) [0]
GMA	<b>0.696 (0.008) [0]</b>	0.590 (0.012) [0]	0.101 (0.028) [0]	0.000 (0.000) [0]
Markedness	0.698 (0.008) [0]	0.750 (0.012) [0]	<b>0.836 (0.090) [0]</b>	0.496 (0.000) [96]
MCC	<b>0.698 (0.008) [0]</b>	<b>0.698 (0.010) [0]</b>	0.541 (0.018) [0]	0.499 (0.000) [96]
OP	<b>0.822 (0.009) [0]</b>	0.689 (0.009) [0]	0.472 (0.007) [0]	0.496 (0.001) [0]

Notes: The five reference measures are included at the top (shaded darker) of the table, above the measures compared herein, for reference.

Analysis with highest median performance by performance measure highlighted in bold.

The percentage 'failed to compute' is the percentage of iterations for which no measure value could be calculated due to an empty cell in the confusion matrix (i.e. all query samples predicted to be in the same class).

In the second analysis, the imbalance resulted in high specificity (0.95) and low sensitivity (0.37). Consequently, the accuracy was very high (0.82), but also remained highly precise (interquartile width 0.005). Like the accuracy, the markedness was higher in the second analysis than the first (0.75 versus 0.70), and the MCC remained the same (0.7). All other performance measures were lower in the second analysis, driven by the low sensitivity (as a result of the classifier output being skewed towards the dominating class).

Consistent trends were seen in the third analysis, such that higher levels of imbalance resulted in higher accuracy (0.92) and markedness (0.84), while other performance measures showed lower performance (0.02 to 0.54). The balanced accuracy and MCC showed the least change (0.51 and 0.54 in analysis 3, compared to 0.70 and 0.70 in analysis 1, respectively). The markedness also became more imprecise (as indicated by the interquartile width; from 0.008 and 0.012 in analyses 1 and 2, respectively, to 0.090) than the other performance measures, while the accuracy and balanced accuracy stayed the most precise (0.004 and 0.003). Finally, in the fourth analysis, for the first time that the imbalance has resulted in no test samples being predicted to be in the minor class on 96% of the iterations, meaning that the PPV, MCC, and markedness all failed to compute (see last column in Table 6.5). Table 6.5 also lists the AUC for each analysis for completeness; more imbalance resulted in a higher AUC precision (0.82 versus 0.77 in analysis 1) but with lower precision (0.042 versus 0.009).

In Table 6.6, the results of the fifth (enriched training data) and sixth (modified testing data) analyses are compared to the third analysis. In analysis 5, enrichment has improved the sensitivity from 0.01 to 0.65, while maintaining an acceptable specificity (0.70, reduced from 1.00). However, the decline in prevalence between the training and testing data naturally results in a lower PPV, reduced from 0.75 to 0.15. The accuracy also declined, from 0.92 to 0.69, as the major class is no longer performing exceptionally.

Table 6.6: Results in empirical analyses with variations on the third analysis (high imbalance)

Performance Measure	Analysis 3: 8% Positive Training and Testing	Analysis 5: 50% Positive Training 8% Positive Testing	Analysis 6: 8% Positive Training 19% Positive Testing
	Median (interquartile range) [% failed to compute]		
AUC	<b>0.779 (0.016) [0]</b>	0.735 (0.017) [0]	<b>0.779 (0.012) [0]</b>
Sensitivity	0.010 (0.006) [0]	<b>0.649 (0.031) [0]</b>	0.010 (0.005) [0]
Specificity	<b>1.000 (0.000) [0]</b>	0.697 (0.015) [0]	<b>1.000 (0.001) [0]</b>
PPV	0.750 (0.175) [0]	0.151 (0.011) [0]	<b>1.000 (0.200) [1]</b>
NPV	0.924 (0.004) [0]	<b>0.960 (0.004) [0]</b>	0.809 (0.010) [0]
Accuracy	<b>0.924 (0.004) [0]</b>	0.693 (0.015) [0]	0.810 (0.010) [0]
Balanced Accuracy	0.505 (0.003) [0]	<b>0.673 (0.014) [0]</b>	0.505 (0.003) [0]
F <sub>1</sub> Measure	0.020 (0.011) [0]	<b>0.245 (0.016) [0]</b>	0.020 (0.011) [0]
F <sub>1.1</sub> Measure	0.022 (0.012) [0]	<b>0.674 (0.014) [0]</b>	0.022 (0.012) [0]
GMA	0.101 (0.028) [0]	<b>0.672 (0.015) [0]</b>	0.101 (0.027) [0]
Markedness	0.836 (0.090) [0]	0.555 (0.006) [0]	<b>0.902 (0.100) [1]</b>
MCC	0.541 (0.018) [0]	<b>0.598 (0.010) [0]</b>	0.543 (0.018) [1]
OP	0.472 (0.007) [0]	<b>0.829 (0.012) [0]</b>	0.415 (0.008) [0]

Notes: The five reference measures are included at the top (shaded darker) of the table, above the measures compared herein, for reference.

Analysis with highest median performance by performance measure highlighted in bold

Analysis three was designed to have high levels of class imbalance (92:8 ratio) in both training and testing sets. The two variations (analyses 5 and 6) increase the number of samples in the positive class in the training only and testing only sets, respectively.

Similarly, the balanced accuracy,  $F_1$  Measure,  $F_{1.1}$  Measure, GMA, and OP have all improved as a result of the improved sensitivity. The markedness, which is driven by the PPV and NPV, declined from 0.84 to 0.56. The MCC, however, remained similar between analyses 3 (0.54) and 5 (0.60) as both poor sensitivity/specificity balance and poor NPV/PPV balance result in lower performance.

In the sixth analysis, the only performance measures which were affected by the change in testing outcome distribution were the accuracy and OP (worse when the testing data were more balanced; 0.81 and 0.42 versus 0.92 and 0.47, respectively) and the markedness (better when the testing data were more balanced; 0.90 versus 0.84).

The AUC was similar between the three analyses presented in Table 6.6, but slightly lower on the fifth analysis for which the training data had been enriched (0.74 versus 0.78 in analyses 3 and 6).

## **6.4 Summarising Findings and Recommendations for Choice of Performance Measure**

### **6.4.1 Summary of Experimental Investigation**

I assessed experimental (controlled) and empirical binary classification outputs, to compare performance measures under certain conditions, with a focus on class imbalance.

For accuracy and the  $F_1$  Measure, performance is strongly affected by the proportion of the classes (i.e. how much one class dominates). As such, poor performance in the non-dominant class often goes unnoticed. The balanced accuracy, markedness, and the MCC may select as the optimal model one which could be improved by recalibration, which would improve the performance in the under-performing class, but not without sacrificing some of the performance in the better class. Additionally, this is only desirable when the model or scenario allows recalibration (not true, for

example, when using domain-expert decision rules) and when sufficient post-hoc validation can be conducted. Between these three measures, the balanced accuracy remains unaffected by class imbalance, while the markedness and (to a lesser extent) the MCC still weights performance more in the major class than the minor class. The GMA,  $F_{1.1}$  Measure and OP penalise poor calibration strongly, and as such it is unlikely that models which could benefit from recalibration methods <sup>337,363</sup> would be discounted.

Across low and moderate levels of class imbalance, the accuracy and markedness both increased as the major class performance increased, even when the minor class performance rapidly declined simultaneously. The same was true of the MCC, but only when imbalance was high. The other performance measures all penalised the decline in minor class performance significantly, despite its small size.

#### 6.4.2 Results in Context

As introduced in Section 6.1, Sokolova and Lapalme <sup>355</sup> defined a taxonomy of eight performance measure invariance properties as the preservation of performance evaluation after certain elementary matrix operations are made to the confusion matrix (such as increasing the number of samples in one class while preserving the class-specific performance). They highlighted that such invariance can be beneficial or adverse, depending on the objective of the classification task, but failed to map these invariance properties to pragmatic use cases. For completeness, the invariance properties of the performance measures used herein are described in further detail in Appendix L. While several of the measures used therein could not be distinguished at all, they did highlight the main difference the GMA and  $F_{1.1}$  Measure: the GMA is invariant to the swapping of the positive and negative columns. This property is characteristic of the robustness of a measure to study specific definitions (asthma attack as the positive outcome) but has little significance in most practical settings. Furthermore, the  $\beta$  value of 1.1 was chosen as a comparator in this analysis because it is a close approximation of the value which gives true invariance, but the  $F_{1.1}$  Measure actually results in lower variance than the GMA in many practical settings, as demonstrated in Tables Table 6.5 and Table 6.6.

To the best of my knowledge, Luque *et al.* <sup>354</sup> is the only study which attempted to translate their findings into guidance. They used cluster analysis to group performance measures according to their handling of class imbalance, quantified as the bias of the performance measure to the direction and magnitude of imbalance. They define the property *null-biased* as when the degree of imbalance, independent of the sensitivity and specificity, changes the resulting performance value (illustrated as parallel lines in Figure 6.1). They concluded that in imbalanced data classification, the GMA and Balanced Accuracy (more accurately, they refer to the Bookmaker's Informedness, discussed in the Section 5.4.3.3) are the best null-biased measures "*if their focus on successes (dismissing the errors) presents no limitation for the specific application where they are used*". In other words, if you do not want the imbalance to explicitly affect the relative weight of the true positive and negative class accuracies, but good discrimination is the overall aim. This is the case in my asthma attack risk prediction modelling: any weighting between class inaccuracies should be chosen on the basis of medical rationale rather than observed class imbalance.

On the other hand, the authors highlight how the MCC is the best choice when the NPV/PPV balance must be considered as well as the sensitivity/specificity balance. Such is the case when the prevalence is expected to be different in the testing data than the model was trained on. This is not expected to be the case in my analyses, as the training and testing data are random partitions of the same original dataset.

### 6.4.3 Recommendations Dictating Performance Measure Choice

There are two primary contexts in which performance measures are used: comparing multiple models in the same data (validation) and comparing the same model in multiple datasets (testing). In both contexts, reporting as many performance measures as possible (as well as the confusion matrix itself) increases transparency and aids comparison with other research. However, the model selection process itself requires a single, primary performance measure so that models can be objectively compared and ranked. As each performance measure is a differently derived

summary of the model's performance, it is important to select a measure in a way that is meaningful for the given data in that specific application and the investigated research question.

Figure 6.3 shows a decision tree I have constructed to provide a generic roadmap for the comparative recommended use of the performance measures in binary classification settings. I have not included the  $F_1$  Measure in this decision tree, as it has a very distinct and unique use case.

At the first branch, we split on whether the major class comprises more than 60% of samples in the training data. When data are substantially imbalanced (major class representing over 60% of samples), the Accuracy and OP are poor choices of performance measure. Accuracy weighs the class-specific accuracies according to the size of the class, and a model with poorer performance in the minor class (which is common as there are fewer samples to learn from) may not have a much lower accuracy than a model with excellent performance in the minor class. The OP is not appropriate when the class-accuracies are poorly calibrated, as when the sensitivity is low, a model with poor specificity may be appraised to perform better than a model with good specificity (and vice versa).

If the data are fairly balanced, most algorithms will attempt to roughly equate the observed and predicted class balance, resulting in poor discrimination if either class is hard to predict. An important distinction between performance measures is whether they prioritise model calibration or model discrimination. Good calibration ensures high performance across an entire population whereas good discrimination means a higher chance of good performance for an individual. If calibration is the priority, then the OP is a strong candidate measure which will select the best model as the one in which the estimated class probabilities are well aligned with the observed risk for all individuals.



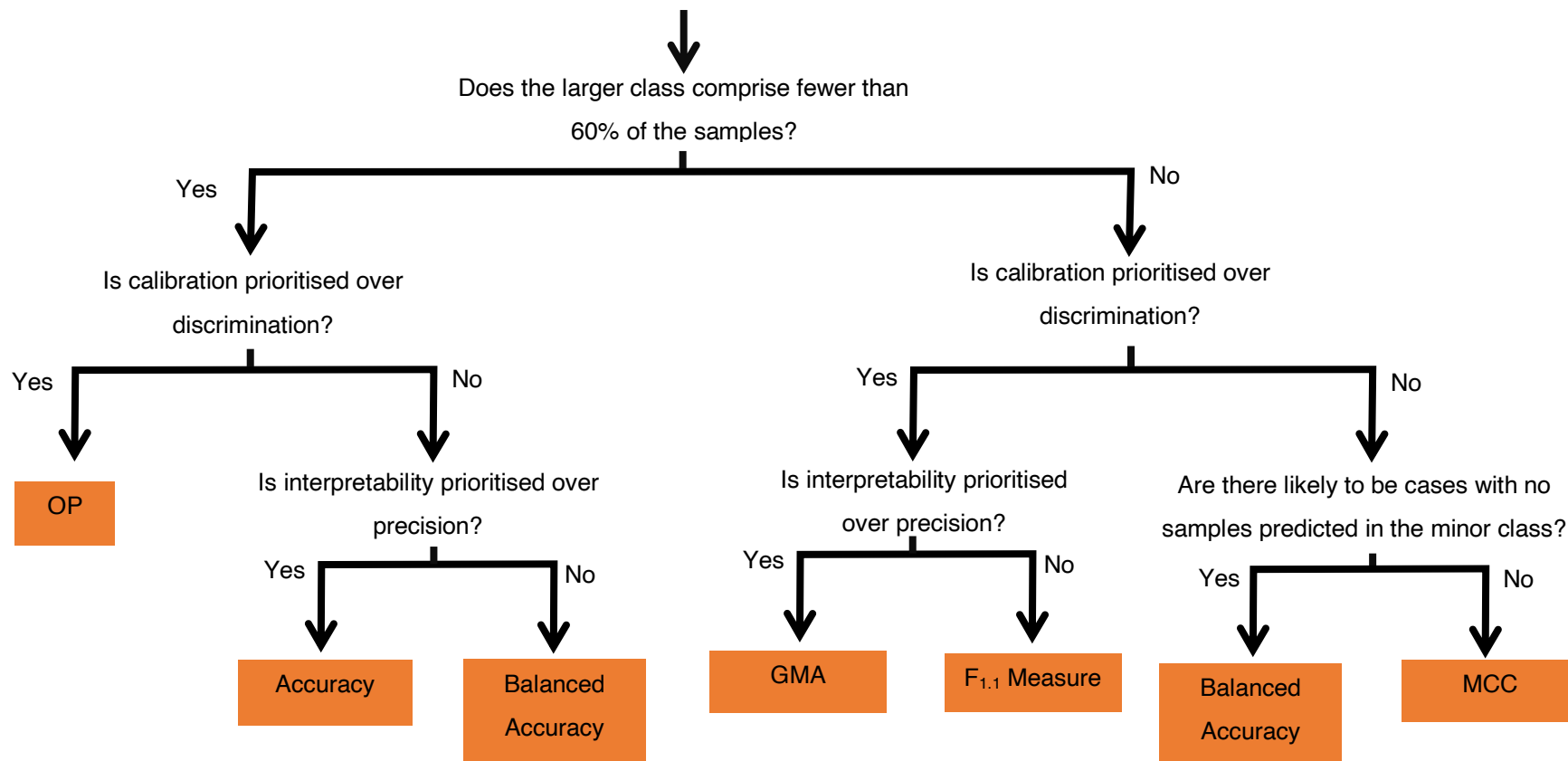


Figure 6.3: Decision tree for selecting performance measure in binary classification settings

If discrimination is more important, then the final question is whether the model should prioritise interpretability or precision. The accuracy is simpler to explain and to understand, whereas the balanced accuracy is the indicated choice for detecting models with poor class performance, especially if the model may be applied to data with a differing prevalence.

Although not presented in this side of the decision tree, the GMA and  $F_{1.1}$  Measure are good middle ground performance measures for lower levels of imbalance when calibration is desired, but not to the extreme of the OP (As shown in Figure 6.2). Both measures will be slightly more favourable than the OP to models in which there is more variance in the observed to estimated risk, and so better able to distinguish between models when calibration is less easily achieved. The GMA and  $F_{1.1}$  Measure will provide very similar results to each other, but the  $F_{1.1}$  Measure is slightly more precise than the GMA, at the expense of being an unconventional measure. Furthermore, the  $F_{1.1}$  Measure penalises models with very low performance in one class quite extremely, so a mean-based model selection algorithm would detect better if one class was underperforming. The markedness, which is also not included in the decision tree, penalises poor calibration even less than the balanced accuracy or accuracy. An optimal use case for this measure was not identified in this analysis, perhaps suggesting why its use has not been so widespread as the other measures (as discussed by Chicco *et al.* <sup>353</sup>).

When data are imbalanced (returning to the top of the decision tree and continuing down the right branch) and one wishes to prioritise calibration over discrimination, the GMA and  $F_{1.1}$  Measure are both suggested, with the same distinctions between the measures described previously applying in this case. When good discrimination is more important, our two prime choices are the MCC and the balanced accuracy. The most substantial difference between the two is that the balanced accuracy only observed the true class accuracies, whereas the MCC also observes the predicted class accuracies. The most important case where this distinction becomes pertinent is when there is a risk of no (or very few) samples being predicted to the minor class. As seen in Figure 6.2, the MCC does not give a balanced overview of the two classes

when the imbalance is extreme and one class is substantially outperforming the other, such that small improvements in the major class affect the performance more than large improvement in the minor class. This finding was also observed by Luque *et al.*<sup>354</sup> (see their Figure 7). As such, when the imbalance is over 90%, the MCC should only be considered when cases with very poor minor class performance have been identified, or when the classification threshold has already been optimised to balance the class performance.

In this case, or when the imbalance is between 60-80%, the benefits of the MCC become more pertinent: it does not reward improved minor class performance when it came at the expense of the class predictive value. The MCC's ability to balance all cells of the confusion matrix in its calculation makes it very stable when comparing across datasets with different prevalence (in external validation, or with a very small testing set, for example) and identify cases where too many samples are predicted as belonging to the minor class (low PPV) as poorly performing.

## 6.5 Summary remarks

In accordance with the well-known machine learning mantra "*there is no unique single best algorithm*" for any problem, there is no single best performance measure for all applications. By thoroughly investigating the results of multiple binary classification performance measures under certain conditions, I have determined the factors which need to be considered in order to facilitate the identification of the likely most suitable performance measure for a given application taking into account the imbalance of the data, the application of the model, and the audience.

In my analyses, a single performance measure is required for two steps of the model selection process: 1) classification threshold optimisation within a model, and 2) model selection itself. For the first step, the risk prediction model could still achieve good discrimination by predicting that no query samples will have an asthma attack (setting the classification threshold to a very high value, such as samples with over 99% predicted risk). As such, even though the model needs to be accurate at the

population level, the calibration is crucial to ensure that poor sensitivity is penalised. The balanced accuracy was chosen accordingly as the most appropriate performance measure for this step and will be used for classification threshold optimisation in Chapter 7. For the second step, once models have been optimised to balance the class accuracies, the MCC is the selected measure for model ranking and selection.

## 7 Asthma Attack Risk Prediction Model

Primary care consultations provide the opportunity for patients and clinicians to assess changes to asthma attack risk. Accurate prediction of risk can instigate timely primary care intervention, prompt more frequency primary care visits, promote risk-reducing lifestyle choices, and encourage patients to seek emergency care following symptom deterioration. Furthermore, highlighting periods when risk is lower can reduce lifetime steroid use and patient anxiety. In this section, I conduct a detailed review of previous asthma attack risk prediction models, in order to establish the criteria for benchmarking my model's performance. I then review the relevant literature regarding reporting guidelines and best practices for prediction model development, describe the final methodology for my prediction model, and finally, present my results.

### 7.1 Previous Work

To critically appraise and benchmark the performance of my own model, it is important to determine the current state-of-the-art, in the context of the various study settings and outcome definitions used in the literature. A 2018 systematic review by Loymans *et al.*<sup>54</sup> had previously identified adult asthma attack risk models, of which eight of the twelve studies identified therein reported some estimate of model performance, and were included in my review. An additional five studies matching these criteria were identified since 2017 from the Google Scholar results of the search terms listed below, for a total of 13 included in this section (Table 7.1). No risk of bias assessment was conducted on these studies, as the purpose of this review was simply to describe methods previously used and to provide guidance on the model benchmarking process.

("asthma attack" OR "asthma exacerbation") AND model AND ("sensitivity" OR "accuracy" OR "C-statistic" OR "AUC")
--

Table 7.1: Characteristics of previous asthma attack risk prediction models

Study Authors	Data Source	Population (Training Sample Size)	Primary Outcome(s)	Event Horizon (Months)	Incidence Risk
Eisner <i>et al.</i> <sup>65</sup> , 2012	Secondary analysis of observational study (EXCELS; NCT00252135)	Diagnosed asthma (N=2878)	One or more serious adverse event	12	Not Reported
			One or more inpatient admission		
			One or more A&E presentation		
			One or more OCS prescription		
			One or more unscheduled doctor visit		
Grana <i>et al.</i> <sup>77</sup> , 1997	US EHRs	Diagnosed or treated asthma (N=54,573)	One or more inpatient admission	12	1.8%
Lieu <i>et al.</i> <sup>169</sup> , 1999	US EHRs	One or more asthma attack in previous two years (N=7141)	One or more A&E presentation or inpatient admission	12	6.9%
Loymans <i>et al.</i> <sup>67</sup> , 2016	Secondary analysis of RCT (ACCURATE; NTR1756)	Diagnosed and treated asthma (N=611)	One or more ATS/ERS defined severe exacerbation	12	13.1%

Notes: Sample size and incidence risk reported from model training set only, except Eisner *et al.* which was an external validation study of a previous model  
Eisner *et al.* also looked at the components of this composite score in isolation, but only the outcome most closely resembling my own has been included herein

Lieu *et al.* also looked at inpatient admissions only

Study Authors	Data Source	Population (Training Sample Size)	Primary Outcome(s)	Event Horizon (Months)	Incidence Risk
Luo <i>et al.</i> <sup>116</sup> , 2020	US EHRs	Diagnosed asthma (164,320 person-years)	One or more A&E presentation or inpatient admission	12	3.4%
Martin <i>et al.</i> <sup>364</sup> , 2020	US EHRs	Diagnosed and treated asthma (N=1787)	One or more ATS/ERS defined severe exacerbation	12	54.8%
Miller <i>et al.</i> <sup>120</sup> , 2006	Secondary analysis of observational study (TENOR; NCT00091767)	Severe asthma (N=2821)	One or more A&E presentation or inpatient admission	6	8.5%
Sato <i>et al.</i> <sup>124</sup> , 2009	Retrospective Observational study	Diagnosed and treated asthma (N=78)	One or more OCS prescription, A&E presentation, inpatient admission, or two (or more) consecutive days of a PEFR of sub-70% of baseline	12	20.5%
Schatz <i>et al.</i> <sup>72</sup> , 2003	US EHRs	Diagnosed <i>or</i> treated asthma (N=6904)	One or more inpatient admission	12	1.2%
Xiang <i>et al.</i> <sup>365</sup> , 2020	US EHRs	Diagnosed and treated asthma (N=31,433)	One or more A&E presentation, inpatient admission or OCS prescription	12	7.8%

Notes: Sample size and incidence risk reported from model training set only

PEFR = Peak Expiratory Flow Rate

Schatz *et al.* conducted a stratified analysis on both children and adults, but only the adult strata is included herein

Study Authors	Data Source	Population (Training Sample Size)	Primary Outcome(s)	Event Horizon (Months)	Incidence Risk
Yurk <i>et al.</i> <sup>366</sup> , 2004	Prospective Observational study	Diagnosed asthma (N=4895)	One or more inpatient admission	12	9%
Zein <i>et al.</i> <sup>367</sup> , 2021	US EHRs	Diagnosed and treated asthma (N=60,302)	One or more OCS prescriptions	12	32.8%
			One or more inpatient admission		1.5%
			One or more A&E presentation		2.9%
Zhang <i>et al.</i> <sup>346</sup> , 2020	Secondary analysis of observational study (SAKURA; NCT00839800)	Diagnosed and treated asthma, and a recent history of asthma attacks (N=2010)	One or more A&E presentation, inpatient admission or OCS prescription initiation lasting for at least 3 days (multiple events per person possible, samples comprised 728,535 follow-up days)	~ 0.1 (3 days)	<0.1%

Notes: Sample size and incidence risk reported from model training set only

Yurk *et al.* did not report the number of people with positive outcomes, and thus a more precise incidence estimate could not be calculated, but also looked at an alternative composite outcome including reduced activity (not presented here).

Zein *et al.* also looked at A&E presentations and inpatient admissions



### 7.1.1 Study Setting

Secondary data sources were used for 11 of the 13 studies identified, including EHRs (all US), RCTs and observational studies. Only two of the studies (Miller <sup>120</sup> and Zhang <sup>346</sup>) used an event horizon of less than 12 months. While identifying patients at high risk of an attack in the next 12 months might highlight an appropriate population for a health education program, or a home monitoring-based intervention, it is not specific enough to assist in the prescribing of a short-term pharmacological intervention (OCS).

Three studies (Miller <sup>120</sup>, Lieu <sup>169</sup>, and Zhang <sup>346</sup>) restricted analysis to patients with severe asthma, however the other studies used the more general criteria of diagnosed and/or treated asthma for the primary population.

As discussed in Section 2.3, while those with mild asthma have a relatively low incidence rate of asthma attacks, they comprise a large proportion of asthma attacks by virtue of being the most populous severity group. As such, limiting my study to those with severe, or even moderate-to-severe asthma, limits the potential benefit of my prediction model.

The definition of an asthma attack (used as outcome) varied considerably across studies, as discussed in Section 2.4. Eisner *et al.* <sup>65</sup> used multiple outcomes in their study, as listed in Table 7.1, one of which was a study defined serious adverse event. Sato *et al.* <sup>124</sup> used peak flow in their (composite) outcome, while the other studies used some combination of A&E presentations, inpatient admissions, and OCS prescriptions (including the outcome definition defined by ATS/ERS).

### 7.1.2 Study Methodology

The most common modelling approach used in the reviewed asthma attack risk prediction studies was logistic regression (11 of 13 studies). Only four of the studies investigated any algorithms other than logistic regression and decision trees (Luo <sup>116</sup>, Xiang <sup>365</sup>, Zein <sup>367</sup>, and Zhang <sup>346</sup>), and only five compared multiple algorithms (Eisner <sup>65</sup>, Luo <sup>116</sup>, Xiang <sup>365</sup>, Zein <sup>367</sup>, and Zhang <sup>346</sup>). Five of the studies only tested a single

model (logistic regression; Grana <sup>77</sup>, Martin <sup>364</sup>, Miller <sup>120</sup>, Schatz <sup>72</sup>, and Yurk <sup>366</sup>), but others tested various algorithm hyper-parameters (Lieu <sup>169</sup>, Luo <sup>116</sup>, Sato <sup>124</sup>, Xiang <sup>365</sup>, Zein <sup>367</sup>, and Zhang <sup>346</sup>), feature sets (Eisner <sup>65</sup>, Lieu <sup>169</sup>, Loymans <sup>67</sup>, Luo <sup>116</sup>, and Xiang <sup>365</sup>), or training data enrichment methods (Luo <sup>116</sup> and Zhang <sup>346</sup>). As introduced in Section 5.3.6, there is no *free lunch*; no algorithm or model guaranteed to produce the ‘universally best’ performance across every dataset. As such, without comparing multiple models it is very unlikely that the best performing model will be identified.

Luo *et al.*’s final model used the extreme gradient boosting algorithm with the positive and negative classes weight hyper-parameter optimised to 0.02 <sup>116</sup>. They note that while this improved the AUC, it resulted in poorly calibrated estimated probabilities. Zhang *et al.* <sup>346</sup> compared under-sampling, over-sampling, and SMOTEing (see Section 5.5), for varying levels of resulting class balance, on a logistic regression model, and found that all methods performed equally, with respect to sensitivity and specificity, and that the performance was best when the classes were equally balanced <sup>346</sup>. Under-sampling was used as their final method due to simplicity, increasing the sensitivity from 0 to 87 with a decline in specificity from 100 to 84.

Only one of the studies reported their performance in an external dataset (Loymans <sup>67</sup>), three used a random split-sample (also known as a hold-out set; Zein <sup>367</sup>, Xiang <sup>365</sup>, Lieu <sup>169</sup>), three used a temporal split-sample (such as the last year of data; Grana <sup>77</sup>, Luo <sup>116</sup>, and Miller <sup>120</sup>), and two used cross-validation (Zhang <sup>346</sup>) or bootstrapping (Schatz <sup>72</sup>). Four of the studies only reported their performance in either the training data (Eisner <sup>65</sup>, Sato <sup>124</sup>, Yurk <sup>366</sup>) or a pooled training and testing set (Martin <sup>364</sup>). The result of this is the performance reported is likely higher than its comparators (and to the performance expected in a practical application of the model) as predictive models tend to perform better on the training data than on new data, as they have learned the patterns present in that data and have fit the model to it <sup>368</sup>.

Five studies (Eisner <sup>65</sup>, Loymans <sup>67</sup>, Miller <sup>120</sup>, Xiang <sup>365</sup>, and Yurk <sup>366</sup>) evaluated the performance of their models using only the AUC (Table 7.2). In contrast, both Grana *et al.* <sup>77</sup> and Lieu *et al.* <sup>169</sup> reported only the predicted classes of their test samples.

The other six studies all reported on both the probabilistic outputs and the predicted classes assigned after a threshold was applied to the probabilistic outputs of the classifiers to obtain a deterministic class estimate. As described in Section 5.4.1, the AUC has many limitations such as measuring performance at thresholds that would never be applied in the real world (such as a false positive rate of 99%) and it is not intuitive to interpret. It can be very useful to supplement deterministic performance measures, however, as it can provide some overview of the sensitivity of the performance depending on the threshold chosen.

Note that multiple models for the same outcome were presented equally by Lieu <sup>169</sup>, Loymans <sup>67</sup>, and Miller <sup>120</sup>. For Miller and Loymans, who each reported only the AUC, the best performing model is reported in Table 7.2. For Lieu *et al.*, who reported multiple performance measures, both models are reported here (models A and B).

Similarly, multiple models with different outcomes were presented equally by Eisner *et al.* <sup>65</sup>, and so have all been reported here with a three-character code denoting the different outcomes.

Five studies (Eisner <sup>65</sup>, Loymans <sup>67</sup>, Miller <sup>120</sup>, Xiang <sup>365</sup>, and Yurk <sup>366</sup>) evaluated the performance of their models using only the AUC. In contrast, both Grana *et al.* <sup>77</sup> and Lieu *et al.* <sup>169</sup> reported only the deterministic predicted classes of their test samples. The other six studies all reported on both the probabilistic outputs and the predicted classes assigned after a threshold was applied to the probabilistic outputs of the classifiers to obtain a deterministic class estimate. As described in Section 5.4.1, the AUC has many limitations such as measuring performance at thresholds that would never be applied in the real world (such as a false positive rate of 99%) and it is not intuitive to interpret. It can be very useful to supplement deterministic performance measures, however, as it can provide some overview of the sensitivity of the performance depending on the threshold chosen.

Table 7.2: Model performance measures reported in asthma attack risk prediction studies

Study	AUC	Sensitivity	Specificity	PPV	NPV	Performance Validation Approach
Eisner <i>et al.</i> <sup>65</sup> , 2012 (SAE)	0.78	-	-	-	-	Performance reported only in the data seen in model training
Eisner <i>et al.</i> <sup>65</sup> , 2012 (INP)	0.69	-	-	-	-	
Eisner <i>et al.</i> <sup>65</sup> , 2012 (A&E)	0.75	-	-	-	-	
Eisner <i>et al.</i> <sup>65</sup> , 2012 (OCS)	0.69	-	-	-	-	
Eisner <i>et al.</i> <sup>65</sup> , 2012 (UDV)	0.68	-	-	-	-	
Loymans <i>et al.</i> <sup>67</sup> , 2016	0.72	-	-	-	-	External
Miller <i>et al.</i> <sup>120</sup> , 2006	0.81	-	-	-	-	Internal (Temporal)
Xiang <i>et al.</i> <sup>365</sup> , 2020	0.70	-	-	-	-	Internal (20% hold-out)
Yurk <i>et al.</i> <sup>366</sup> , 2004	0.71	-	-	-	-	Training set
Grana <i>et al.</i> <sup>77</sup> , 1997	-	70	71	-	-	Internal (Temporal)
Lieu <i>et al.</i> <sup>169</sup> , 1999 (A)	-	49	85	20	-	Internal (50% hold-out)
Lieu <i>et al.</i> <sup>169</sup> , 1999 (B)	-	36	92	25	-	
Luo <i>et al.</i> <sup>116</sup> , 2020	0.86	54	92	23	98	Internal (Temporal)
Martin <i>et al.</i> <sup>364</sup> , 2020	0.67	48	80	74	56	Combined training and validation set
Sato <i>et al.</i> <sup>124</sup> , 2009	0.68	44	92	-	-	Training set

Notes: SAE = serious adverse event, INP = inpatient admission, A&E = accident and emergency presentation, OCS = oral corticosteroid prescription, UDV = unscheduled doctor visit. For the performance validation approach, internal refers to using a single database for training and testing (usually done in some standard way of model validation, e.g. CV), and external refers to having an external database. A hyphen denotes that a measure was not reported.

Study	AUC	Sensitivity	Specificity	PPV	NPV	Performance Validation Approach
Schatz <i>et al.</i> <sup>72</sup> , 2003	0.71	45	87	4	99	Internal (bootstrap & jack-knife estimates)
Zein <i>et al.</i> <sup>367</sup> , 2021 (OCS)	0.71	64	67	51	78	Internal (20% hold-out)
Zein <i>et al.</i> <sup>367</sup> , 2021 (INP)	0.85	86	73	5	100	
Zein <i>et al.</i> <sup>367</sup> , 2021 (A&E)	0.88	84	76	12	99	
Zhang <i>et al.</i> <sup>346</sup> , 2020	0.85	90	83	-	-	Internal (10-rep 5-fold CV)

Notes: SAE = serious adverse event, INP = inpatient admission, A&E = accident and emergency presentation, OCS = oral corticosteroid prescription, UDV = unscheduled doctor visit, CV = cross validation. For the performance validation approach, internal refers to using a single database for training and testing (usually done in some standard way of model validation, e.g. CV), and external refers to having an external database.

### 7.1.3 Model Performance

The AUC values reported ranged between 0.67 and 0.88. Four models achieved an AUC of 0.8 or greater: Luo <sup>116</sup>, Zein: Inpatient <sup>367</sup>, Zein: A&E <sup>367</sup>, and Zhang <sup>346</sup>. All four models achieved greater than 70% specificity, and over 50% sensitivity. In fact, Zhang *et al.* <sup>346</sup> and both Zein *et al.* models <sup>367</sup> achieved over 80% sensitivity. Unfortunately, this performance in the minor class came at the expense of the PPV, which was under 15 for both Zein <sup>367</sup> models and only 23% for Luo *et al.* <sup>116</sup>. No PPV was reported for Zhang *et al.* <sup>346</sup>. The model by Grana *et al.* <sup>77</sup> also achieved over 70% for both sensitivity and specificity, but failed to report PPV. Zein *et al.*'s OCS prediction model <sup>367</sup> was the only to achieve greater than 50% for sensitivity, specificity, PPV and NPV, however with a very low AUC.

### 7.1.4 Conclusions

Defining a benchmark is no simple task, due to the multitude of performance measures used, and the different settings of the models evaluated. Overall, I decided on a composite benchmark, with six criteria which if met a model could be considered conclusively the highest performing:

- Sensitivity and PPV greater than 50% (median sensitivity in Table 7.2)
- Specificity and NPV greater than 70% (median specificity in Table 7.2)
- Balanced accuracy greater than 70% (median balanced accuracy calculated from studies reporting both sensitivity and specificity in Table 7.2)
- AUC greater than 0.70 (median AUC in Table 7.2).

Reviewing the previous models has also reinforced the importance of appraising training data enrichment methods, ensuring that data (including both samples and parameters) from model training are not used in the performance evaluation, and reporting on a wide array of performance measures. Furthermore, providing the confusion matrix allows further performance measures to be calculated post-hoc.

## 7.2 Published Guidelines for Developing and Reporting Clinical Risk Prediction Models

In this section I will review published guidelines for the development and reporting of clinical risk prediction models, and for observational studies using routinely collected data. Items on any of the discussed guidelines which were pertinent to secondary data analyses and to a thesis, rather than a journal publication, will be listed in Appendix M, along with the section in which the item will be covered.

Within the EQUATOR (Enhancing the Quality and Transparency of health Research) network, three guidelines were identified which pertained to risk prediction or prognostic modelling. First, RiGoR (Reporting Guidelines to address common sources of bias in Risk model development, by Kerr *et al.* <sup>368</sup>) primarily aims to improve practices relating to two common causes of bias, which are sometimes combined into *optimism bias*: leaking of information (either data samples, or parameters estimated from said data) from the training set to the testing set (named *resubstitution bias* therein, but often known as *data-leakage*) and reporting on the performance of the model which performs best in the training or validation set, but not the final performance in new unseen data (*model-selection bias*). Such considerations are of great importance to ensuring model validity but are very often missing from less technical guidelines.

Luo *et al.* <sup>53</sup> published guidelines specifically for machine learning predictive models. Some points related strictly to specific algorithms (including recommended hyperparameter fine tuning), but they also stressed the importance of some quantification of clinical benefit from this model over standard practice, such as calculating the Number Needed to Treat (NNT) and/or providing a health economics evaluation.

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guidelines by Collins *et al.* <sup>369</sup> provides well-rounded general guidance for both prognostic and diagnostic models. Few points were raised which were not covered by either of the two previously described guidelines, however they

also comment on the value of publishing a study protocol prior to the start of analysis, and the importance of highlighting any changes to the originally described methodology that were undertaken after the data was explored.

Finally, the RECORD (Reporting of studies Conducted using Observational Routinely-collected health Data) guidelines by Benchimol *et al.*<sup>370</sup> were used to supplement the prediction model-centred guidelines with mandates specifically pertaining to the use of EHRs. These included describing the extent to which the investigators had access to the database population, and including flow charts to demonstrate the data linkage process.

## 7.3 Methods

In this section I review the conclusions I have made in the previous chapters which comprise the methods for my risk prediction model.

### 7.3.1 Inclusion and Exclusion Criteria

The final study population is adults (aged 18 and over) with clinician-diagnosed-and-treated asthma (identified by Read Codes in Appendix E and at least one controller medication as identified in Section 4.2.2). This includes those with *inactive* asthma, as there was no enforced time limit on the recency of asthma controller medication prescriptions.

I chose to focus this analysis on the adult population as there is evidence that use of EHRs to estimate adherence is more appropriate in adults than in children. Jentsch *et al.*<sup>264</sup> saw a substantial overestimation of adherence in their study of children with asthma, possibly as a result of parents coordinating the refills, regardless of the child's medication taking. In the general adult population, it has been estimated that (across multiple conditions) only 10% of adults could be classed as engaged (not discontinued) but poorly implementing<sup>285</sup>. As such, they may be the most appropriate age group for such an assessment. If age was missing from the primary care registry,



it was not possible to ascertain if an individual was an adult, and so they were excluded.

COPD is a condition with a similar presentation to asthma, but a different biological mechanism and different risk factors <sup>126,127</sup>. In addition, a previous UK study by Nissen *et al.* found that over half (52%) of individuals with validated COPD diagnoses also had a previous asthma diagnosis, however the majority (72%) of these were asthma misdiagnoses <sup>125</sup>. I decided to exclude anyone with a COPD diagnosis, as ascertained by the presence of Read Codes listed in Appendix N).

The ALHS datasets were *left-censored* on January 2009 (all records prior to this date were discarded) in order to align with the primary care prescribing data and *right-censored* on March 2017 (all records after to this date were discarded) in order to align with the primary care, inpatient hospital admission, and mortality records (as presented in Table 2.1). Each person's follow-up time was further left-censored at their first treatment event (whichever came first) and right-censored at their date of death or asthma resolution (the cessation of symptoms; Read Code 212G.).

In the final analysis dataset, only primary care encounters for asthma or respiratory infections (days on which at least one asthma-related code was recorded), on days with no steroid prescriptions or secondary care asthma encounters, were retained as samples for training and validation. As such, to be included in the analyses, individuals were required to have at least one such event during their follow-up. Read Codes for asthma encounters other than diagnoses are listed in Appendix F. Finally, those with missing age or sex (in the primary care registry) were excluded from analyses.

### 7.3.2 Asthma Attacks

As discussed in Section 2.4, I used the ATS/ERS Task Force definition of a severe exacerbation <sup>78</sup> to define asthma attacks in my dataset: a prescription of OCS, an asthma-related A&E visit, or an asthma-related hospital admission. In addition, deaths with asthma as the primary cause were considered asthma attacks, and cases of multiple attack-identifying records occurring within a 14-day period were coded as a

single incident. A&E presentations, inpatient admissions, and deaths due to asthma were all identified by the ICD10 codes J45 or J46. Additionally, A&E presentations were flagged as asthma attacks if ‘asthma’ was identified in the free text field for presenting complaint in A&E admissions, (manually checked to confirm no negation, such as ‘not asthma’).

As described in Section 3.6.6, prescriptions of prednisolone oral steroids (brand names listed in Appendix D) were considered indicative of an asthma attack if all of the following conditions were also met: 1) they were prescribed to someone with a diagnosis of asthma or receiving asthma treatment, 2) they were prescribed on the same day as an asthma-related consultation (identified by the presence of any Read Code listed in Appendix E and Appendix F on the same day), and 3) the total prescribed dose was between 200 and 1000 mg (based on the British National Formulary Version 80 (BNF80) recommendation that 40-50mg daily be prescribed for asthma attacks, for at least 5 days <sup>193</sup>).

Medication strength was extracted in ALHS by searching the *ePRNDName* (and *PIItemStrength.UOM* if no value could be identified from *ePRNDName*) for any of 500, 125, 120, 100, 80, 40, 25, 20, 16, 15, 10, 5, 4, 2, 1 or 2.5, followed by “MG”, or “MILLIGRAM” (with or without a preceding space), or 1 followed by “G” (with or without space, and subsequently converted from grams to milligrams). The total dose was then calculated as the quantity (see Section 4.2.3) multiplied by the medication strength.

### 7.3.3 Risk Factors

Table 3.4 describes the full set of risk factors which were included in the analysis. In this section, I describe the feature extraction method, where previously undetermined, and any subsequent data processing not described in Chapter 3.

Following from the findings in Chapter 4, prescribing records were used to estimate two measures of adherence: a rolling average of the three most recent (closed-ended) prescription intervals (known as *CSA\_3*), and the percentage of days in the previous

calendar year for which there was medication supply available, assuming that supply from overlapping intervals is not discarded (CMA8\_2).

The mean inhaled SABA dosage per day and a binary flag for the recent prescription (in the last 90 days) of nebulised SABA were used as proxies for asthma control. The mean inhaled SABA dosage per day was estimated from the dosage and the dates between prescriptions. Non-inhaled reliever medications were identified by “ML” in the *PIItemstrength.UOM*, or any of the following phrases in the *ePRNDName* or *PIDrugformulation*:

“NEB”, “ORAL”, “CAP”, “TAB”, “SYRUP”, “SOL”, “SOLN”, “INJ”

The medication strength was then extracted by searching for any of the values 95, 100, or 200, followed by “MICRO” or “MCG” (with or without a preceding space) in *ePRNDName*, or *PIItemstrength.UOM* if no value could be extracted from the *ePRNDName*. The volume of the inhaler was estimated by searching for 60, 100, 120, or 200, following by “DOSE” (with or without a preceding space or hyphen) in *ePRNDName*. The mean SABA dosage daily was then estimated as the prescribed quantity of inhalers multiplied by the medication strength in micrograms and the volume of the inhaler unit and divided by the number of days until the next prescription.

Of the 45 features (the 43 listed in Table 3.4, plus CSA\_3 and CMA8\_2), 22 were binary, six were continuous, and 17 were categorical (with more than two categories). The distributions and summary statistics of each continuous feature, and the proportions for each value of the categorical features, are presented in Appendix O. The categorical variables were one-hot encoded (described in Section 6.2.2), resulting in a design matrix with 125 columns and 1,154,048 rows.

#### 7.3.4 Analysis Plan

In this analysis, a repeated, random split-sample approach is used to train and validate the models. A 10% partition (115,404 samples) of the ALHS dataset (Section 2.2) was kept aside for holdout testing, and the remaining 90% (1,038,644 samples) was used for model selection and initial performance reporting (henceforth the derivation dataset).

For model selection, the derivation dataset was further randomly partitioned 10 times such that 90% of the samples were used for training, and 10% for validation (Figure 7.1). Feature scaling was conducted on the continuous features by min-max normalisation, introduced previously in Section 5.4.3.5 (Equation (5.13), independently in the training/validation and testing partitions. Feature scaling speeds up gradient descent-based methods (such as logistic regression). This can be illustrated by a simple example: in a dataset with two features, of which one has range [0,1] and the other has range [0,1000000], the learning rate required to complement the scale of the first feature will result in very inefficient descent relative to the second feature. Feature scaling is also necessary for SMOTEing, as it is based on the Euclidean distance between minority samples. As such, a sample which was identical in all regards to another except one feature which had a much wider scale than the others may have a higher distance than a sample which was mildly different for all features. Minmax normalisation was chosen specifically because the range of values for each feature are the only required parameters, rather than the full distribution of the feature. As such, the minimum and maximum observed in this derivation data can be easily shared and compared against the range observed in an external dataset, in order to determine the range to use for the external data scaling.

Outliers were removed from the continuous features before scaling by right-censoring values at the value closest to one significant figure (such as 3000, or 40) to the median multiplied by four. 12 models (Appendix P) were then trained on four variations of the same scaled training data partition, to predict whether an asthma attack would occur in the following four weeks (primary endpoint). The 12 models used the following algorithms: (1) logistic regression, (2) naïve Bayes classifier, (3-6) random forests (with four hyper-parameter values trialled) and (7-12) extreme gradient boosting (two hyper-parameter value sets, of size three and two, for a total of six models trialled). The logistic regression model, employed without any feature selection or polynomial terms, serves as a naïve benchmark to demonstrate the flexibility of the non-parametric machine learning models.

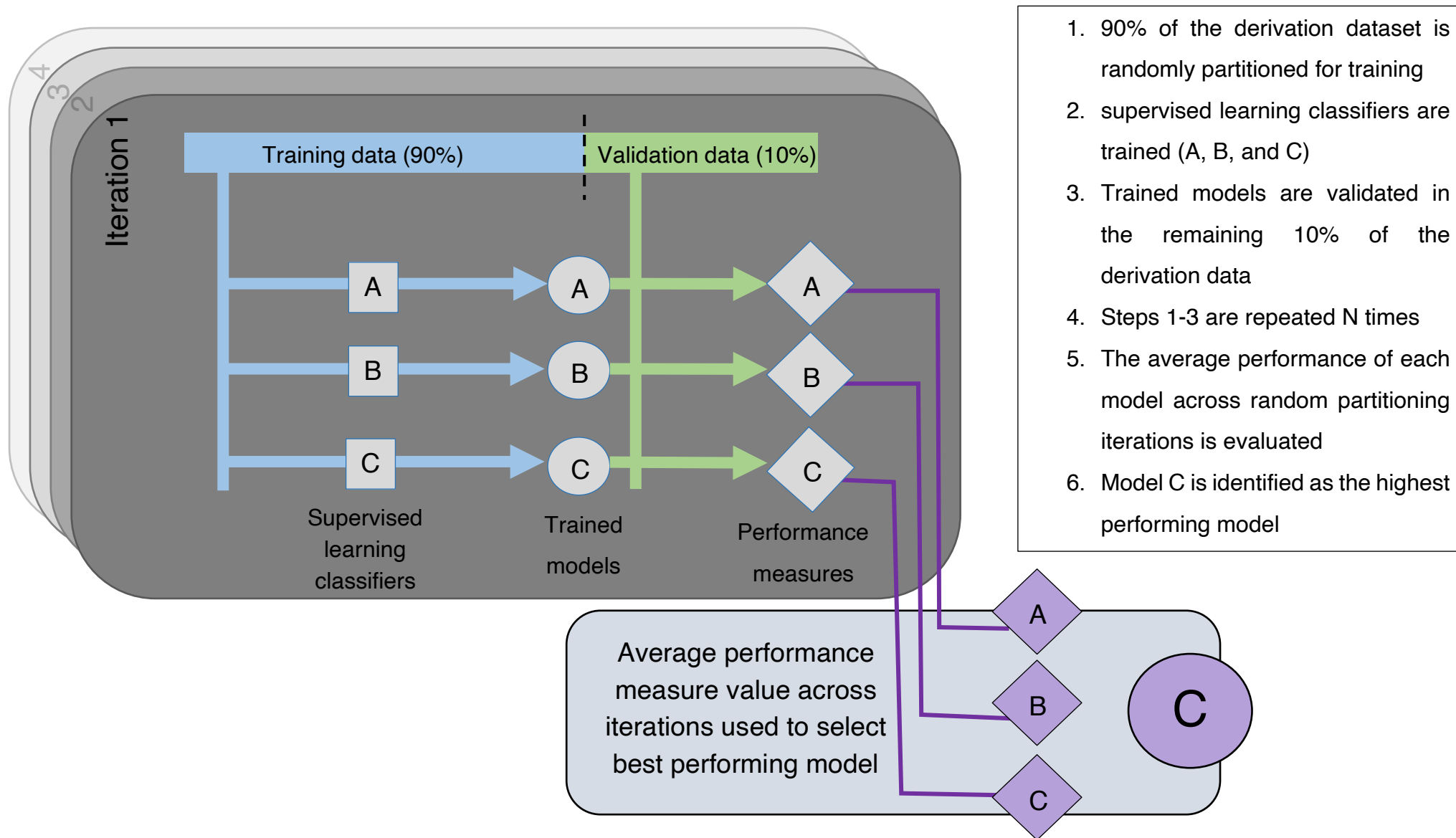


Figure 7.1: Model training and validation process

Variations of the training data employed different data enrichment methods, assessing how to best overcome problems in model performance as a result of low outcome prevalence (see Section 5.5). The variations were: the original training data partition, and the original training data partition with three SMOTEing parameter sets applied (described in Section 7.3.5). The stability of the performance measure estimates across the 10 iterations of the training/validation partitioning was evaluated to ensure that there is sufficient confidence in the selection of the model to proceed to the second model validation phase, else further iterations will be conducted at this stage.

For each iteration, model, and enrichment method, the AUC and Brier Score were calculated, and the confusion matrix was recorded for both the default (probability greater than 0.5) classification threshold, or the threshold that optimised the balanced accuracy in predictions made on the training data. The optimised threshold was identified using golden-section search optimisation<sup>371</sup>; an iterative technique which assumes that the function mapping the threshold values to the resulting balanced accuracy is monotonically increasing until the optimal value, and then monotonically decreasing above it. Across iterations, summary statistics were calculated for each performance measure to provide some estimate of the average performance, and the certainty around that estimate. The best-performing model was selected on the basis of the sensitivity and PPV in the validation partitions for each of the 10 split-sample iterations.

For performance reporting, the derivation dataset was randomly partitioned another 100 times, again with 90% of the data used for training the final selected algorithm, hyper-parameters and enrichment method, and 10% for validation. The model performance was reported aggregated over the 100 iterations of the split-sample process.

Finally, the model was then retrained on the full derivation dataset and tested on the as-of-yet unseen holdout test partition. Retraining the model allows the final testing in the holdout partition to make use of the full wealth of the derivation dataset, as well as selecting a classification threshold informed by the performance in the validation

stage. Each performance measure estimated in this holdout partition was compared to the range of estimates in the 100 previous iterations to ensure that the crossover in samples between the model selection and performance reporting subsets did not bias the results.

The Gini importance <sup>314</sup> of each feature in the full derivation dataset was evaluated (see Section 5.6.1), and the top ten and bottom five features were included in a bar chart for illustration of the relative importance.

Post-hoc analyses of model calibration were conducted on the holdout partition using the calibration slope and calibration-in-the-large (Section 5.4.4). Discrimination in subgroups was evaluated by assessing the classification errors according to the following factors: prior history of asthma attacks, asthma severity, asthma attack severity, and smoking status.

As an unvalidated comparison, the final selection of algorithm, hyper-parameters and enrichment method was retrained (in the derivation dataset) on nine alternative endpoints, tested in the holdout partition. Four alternate event horizons were tested, compared to prediction in the next four weeks: one week, 12 weeks, 26 weeks, and 52 weeks. Five endpoints used the same five event horizons but only for predicting asthma attacks that presented in secondary care. The hyper-parameters for these models have not been fine-tuned, nor the model performance robustly tested, but it serves as a simple indicator of possible utility for varying clinical settings.

The protocol for this analysis was published in *BMJ Open* in 2019 <sup>372</sup>, and any deviations from that protocol are detailed in Appendix Q.

### 7.3.5 Enrichments

The primary SMOTEing parameters, as described in Section 5.5, relate to the degree of over-sampling and under-sampling conducted. In our case, with an expected minor class proportion of 0.01,  $k$  can take any integer value in the range 1 to 98 (Equation (5.22)), and  $z$  any real value in the range 1 to 99 (Equation (5.23)).

As shown in Figure 7.2, the enriched training dataset sample size is at its largest for each value of  $k$  when  $z$  is maximised. I used three scenarios, selected at the 5<sup>th</sup>, 15<sup>th</sup>, and 25<sup>th</sup> percentiles (rounded to the nearest integer) of  $k$  (6, 16 and 25), and the corresponding value of  $z$  (to the nearest 1 decimal place) which minimised the deviation in sample size from the original (unenriched) sample (15.5, 5.2, and 3); denoted by the diamonds. The neighbourhood radius parameter was set to the default value of 5 from the function SMOTE in the package *DMwR*<sup>373</sup>.

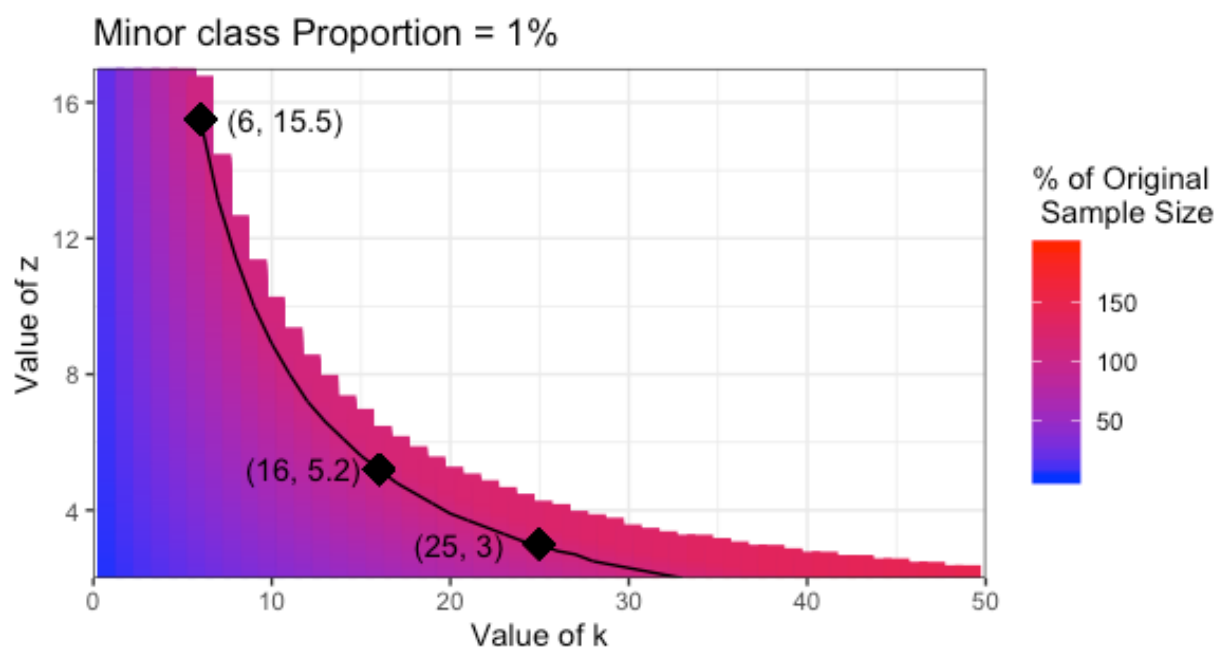


Figure 7.2: Selected values of  $k, z$  for SMOTE enrichment, and their resulting sample size, relating to the original training data sample size

Note: The SMOTEd training data generates  $k$  new synthetic samples for each sample in the minor class, resulting in  $(k+1)*a$  samples in the minor class, and retaining  $z*k*a$  major class samples.

### 7.3.6 Parallel Programming for Increased Efficiency

Training machine learning models, and generating predictions in new data, is a computationally intensive and time-consuming process. One way to improve the efficiency of running such programs is to move from *serial* execution (completing tasks, such as training models, one at a time) to *parallel* execution (running several



tasks *simultaneously*). When programs are being run on a processor with multiple *cores* (a computation unit, capable of running a single task), known as a *multi-core processor* (Figure 7.3), it may be possible to assign independent processes (or *workers*) to compute, and pass their results back to the *master* process. The total volume of work (including data) given to a worker to complete is known as a *chunk*. It can also be possible to use multiple cores spread across multiple networked computers (or *nodes*), using a Message Passing Interface (MPI).

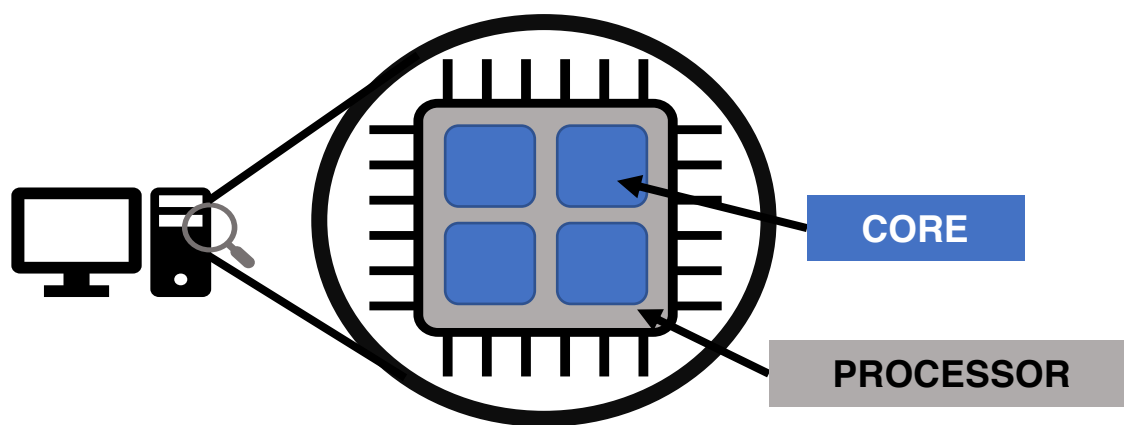


Figure 7.3: Diagram of a multi-core processor

In my main program, described in Section 7.3.4, there were 100 iterations of the model building and validating that could be conducted in parallel. The initial data set up was conducted on a single core, and then the 100 iterations were divided between 5 workers. Each worker trained the models, made predictions about each model in the validation partition, assessed the performance of each model's predictions, and passed these evaluations back to the master process. The master then combined these evaluations into one dataset, so that the performance of each model across all iterations could be reviewed.

## 7.4 Results

### 7.4.1 Analysis Population

The ALHS dataset identified those with a diagnosis of asthma as determined by the presence of one or more diagnostic or asthma management Read Codes, presented in Appendix E <sup>104</sup>. As described in Section 2.2, the primary care encounters dataset had already been restricted to this population, leaving 49,307 individuals. After cleaning (described in Section 2.2.3.2), records for 48,975 remained. A further 1,702 individuals were excluded due to their diagnosis of COPD, leaving 47,273 individuals (Figure 7.4). Finally, only individuals with at least one primary care encounter relating to asthma or a respiratory infection (else they did not have any data to input), excluding a further 522 individuals (46,897 remaining).

I additionally specified that individuals must have had at least one ICS asthma medication prescribed during their follow-up (*clinician-diagnosed-and-treated asthma*), although I did not specify that it had to have been a recent prescription, in line with the findings of Nissen *et al.* <sup>118</sup>. Asthma treatment, and the identification of asthma medications from EHRs, is described in the Section 4.2.2. From the 671,298 individuals remaining after data cleaning was conducted in the primary care prescriptions dataset (Section 2.2.3.3), 91,327 individuals had at least one ICS asthma medication (Section 4.2.2). The intersect between the eligible patients in the primary care encounters dataset and the primary care prescriptions dataset was 31,463 patients.

Follow-up time was calculated from the first eligible prescription record in the primary care prescriptions (asthma medications other than SABA) until the resolution of their asthma, their death, or the end of the study period (March 2017). 1590 patients had Read Codes for asthma resolution, of which 133 patients occurred before the start of their follow-up, and as such they were excluded from the analyses (leaving 31,330 patients).

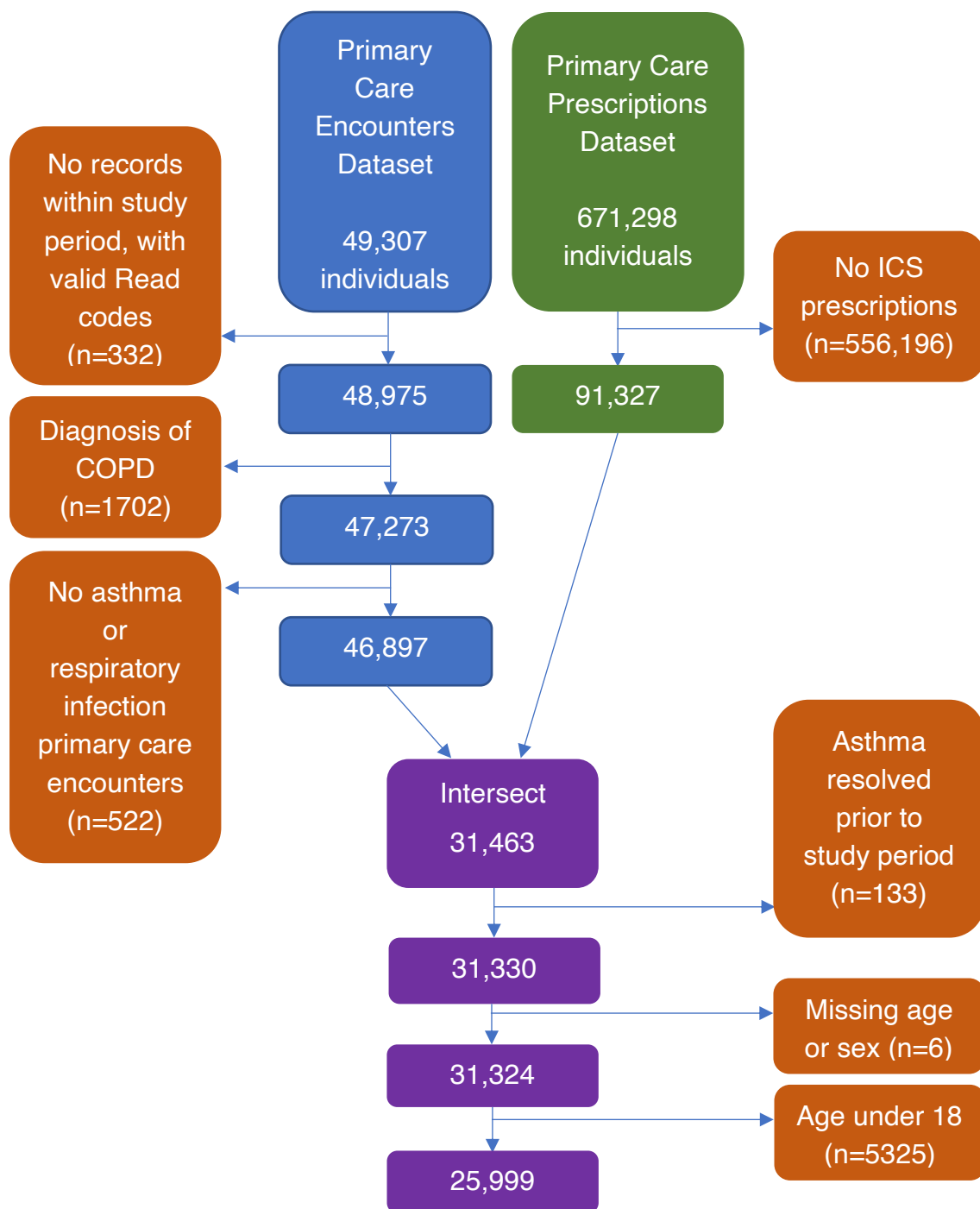


Figure 7.4: Asthma attack risk prediction model analysis population flow diagram

Patients with missing sex and/or date of birth (n=6) or age under 18 (n=5,325) were excluded, leaving 25,999 viable patients. No other demographic exclusion criteria were implemented. This resulted in a total of 177,299 person-years of data, with a median of 7.8 years per person (interquartile range 6.3 to 8.1, range <1 to 8.2).

Patients in the analysis sample were mostly males, aged 18 to 35, female, and living within urban areas (Table 7.3). The vast majority of all patients had no indication of smoking status (or had an explicit non-smoking code) in their baseline year (all records in which they were their baseline age in years). The most prevalent comorbidity in the analysis population was chronic pulmonary disease. This did not include COPD, which had already been excluded, but did include asthma diagnosis codes in addition to pulmonary fibrosis, asbestosis, and others, as listed in Appendix G. As such, while everyone in the study had either an asthma diagnosis or management code, not everyone had a chronic pulmonary disease diagnosis.

Other common comorbidities were anxiety/depression, eczema, GERD, and rhinitis (Table 7.4). Additionally, 9,777 individuals (37.6%) had at least one nasal spray prescription during the study period, of whom 21% had at least one asthma attack (unadjusted OR = 1.34,  $p < 0.001$ ).

Table 7.3: Demographics of the ALHS analysis population

Characteristics		All Patients (N=25,999)	Patients with no attacks during follow-up (N=21,234)	Patients with one or more attacks during follow-up (N=4,765)
<b>Baseline Age</b>				
	18 to 35	10721 (41.2%)	8881 (41.8%)	1840 (38.6%)
	36 to 45	4825 (18.6%)	3859 (18.2%)	966 (20.3%)
	46 to 60	5957 (22.9%)	4779 (22.5%)	1178 (24.7%)
	61 to 75	3385 (13.0%)	2764 (13.0%)	621 (13.0%)
	76 to 99	1111 (4.3%)	951 (4.5%)	160 (3.4%)
<b>Sex</b>				
	Male	10544 (40.6%)	8904 (41.9%)	1640 (34.4%)
	Female	15455 (59.4%)	12330 (58.1%)	3125 (65.6%)
<b>Obesity</b>				
	Not Obese	24187 (93.03%)	4318 (90.62%)	19869 (93.57%)
	Obese	1812 (6.97%)	447 (9.38%)	1365 (6.43%)
<b>Baseline Maximum BTS Step <sup>a</sup></b>				
	0	6281 (24.2%)	5135 (24.2%)	1146 (24.1%)
	1	8394 (32.3%)	7216 (34.0%)	1178 (24.7%)
	2	2518 (9.7%)	2012 (9.5%)	506 (10.6%)
	3	7951 (30.6%)	6360 (30.0%)	1591 (33.4%)
	4	855 (3.3%)	511 (2.4%)	344 (7.2%)

Note: Baseline period for maximum BTS step was all records for which the patient was at the same age as when they entered the study, rather than in the chronological year following the start of the study.

Characteristics		All Patients (N=25,999)	Patients with no attacks during follow-up (N=21,234)	Patients with one or more attacks during follow-up (N=4,765)
Baseline Smoking Status <sup>b</sup>				
	Current	1604 (6.2%)	1127 (5.3%)	477 (10.0%)
	Former	1513 (5.8%)	1114 (5.2%)	399 (8.4%)
	Never	22882 (88.0%)	18993 (89.4%)	3889 (81.6%)
Baseline Scottish Index of Multiple Deprivation				
	1 (Highest Deprivation)	5555 (21.4%)	4350 (20.5%)	1205 (25.3%)
	2	5070 (19.5%)	3992 (18.8%)	1078 (22.6%)
	3	4261 (16.4%)	3476 (16.4%)	785 (16.5%)
	4	5778 (22.2%)	4869 (22.9%)	909 (19.1%)
	5 (Lowest Deprivation)	4664 (17.9%)	3987 (18.8%)	677 (14.2%)
	Missing	671 (2.6%)	560 (2.6%)	111 (2.3%)
Baseline Scottish Urban Rural Classification				
	1 (Large Urban)	8912 (34.3%)	7409 (34.9%)	1503 (31.5%)
	2 (Other Urban Area)	8974 (34.5%)	7169 (33.8%)	1805 (37.9%)
	3 (Accessible Small Towns)	2160 (8.3%)	1642 (7.7%)	518 (10.9%)
	4 (Remote Small Towns)	910 (3.5%)	782 (3.7%)	128 (2.7%)
	5 (Accessible Rural)	2711 (10.4%)	2259 (10.6%)	452 (9.5%)
	6 (Remote Rural)	1512 (5.8%)	1292 (6.1%)	220 (4.6%)
	Missing	820 (3.2%)	681 (3.2%)	139 (2.9%)

Note: Baseline period for most recent smoking status was all records for which the patient was at the same age as when they entered the study, rather than in the chronological year following the start of the study.

Table 7.4: Prevalence of comorbidities in ALHS analysis population

Comorbidity	All Patients (N=25,999)	Patients with no attacks during follow-up (N=21,234)	Patients with one or more attacks during follow-up (N=4,765)	Odds Ratio (95% Confidence Interval)
	Number (Percent)			
AIDS	NR	NR (<0.1%)	NR (<0.5%)	NR
Hemiplegia	NR	NR (<0.1%)	NR (<0.5%)	NR
Anaphylaxis	NR	NR (<0.1%)	NR (<0.5%)	NR
Moderate liver disease	NR	27 (0.2%)	NR (<0.5%)	NR
Metastatic tumour	NR	39 (0.2%)	NR (<0.5%)	NR
Mild liver disease	NR	50 (0.2%)	NR (<0.5%)	NR
Peptic ulcer disease	NR	50 (0.2%)	NR (<0.5%)	NR
Peripheral vascular disease	93 (0.4%)	63 (0.3%)	30 (0.6%)	2.13 (1.38 – 3.29)
Rheumatological disease	126 (0.5%)	64 (0.3%)	62 (1.3%)	4.36 (3.07 – 6.19)
Nasal Polyps	141 (0.5%)	84 (0.4%)	57 (1.2%)	3.05 (2.18 – 4.28)
Dementia	142 (0.5%)	106 (0.5%)	36 (0.8%)	1.52 (1.04 – 2.22)
Congestive heart disease	147 (0.6%)	96 (0.4%)	51 (1.1%)	2.38 (1.69 – 3.35)
Myocardial infarction	162 (0.6%)	108 (0.5%)	54 (1.1%)	2.24 (1.61 – 3.11)
Diabetes with complications	308 (1.2%)	221 (1.0%)	87 (1.8%)	1.77 (1.38 – 2.27)
Cerebrovascular disease	318 (1.2%)	229 (1.1%)	89 (1.9%)	1.75 (1.36 – 2.23)
Renal disease	344 (1.3%)	259 (1.2%)	85 (1.8%)	1.47 (1.15 – 1.88)
Cancer	522 (2.0%)	401 (1.9%)	121 (2.5%)	1.35 (1.10 – 1.66)
Diabetes	607 (2.3%)	419 (2.0%)	188 (3.9%)	2.04 (1.71 – 2.43)
GERD	963 (3.7%)	645 (3.0%)	318 (6.7%)	2.28 (1.99 – 2.62)

Note: Values with percentages under 0.5% for attacks and 0.1% for no attacks have been redacted due to small numbers.

Comorbidity	All Patients (N=25,999)	Patients with no attacks during follow-up (N=21,234)	Patients with one or more attacks during follow-up (N=4,765)	Odds Ratio (95% Confidence Interval)
	Number (Percent)			
Rhinitis	987 (3.8%)	712 (3.4%)	275 (5.8%)	1.77 (1.53 – 2.04)
Eczema	1326 (5.1%)	952 (4.5%)	374 (7.8%)	1.81 (1.60 – 2.05)
Anxiety/Depression	3683 (14.2%)	2523 (11.9%)	1160 (24.3%)	2.39 (2.21 – 2.58)
Chronic pulmonary disease	8822 (33.9%)	7011 (33.0%)	1811 (38.0%)	1.24 (1.17 – 1.33)



## 7.4.2 Outcome Ascertainment

514,785 steroid prescriptions were identified, of which only 17 had a missing or negative value for quantity. The median estimated total dose was 200mg (interquartile range 140-280 mg, and range 1-168,000 mg). 307,369 prescriptions were retained with total dose between 200 and 1000mg, of which 7061 occurred on the dates of asthma consultations and were thus included as steroid bursts.

A total of 12,193 asthma attack events occurred within the study period in our analysis population: 3221 inpatient admissions, 2405 A&E presentations, 6533 primary care OCS courses, and 34 deaths (0.3% of identified severe attacks were fatal). Attack events occurring within 14 days of an initial event were not counted herein as separate attacks. As such, if a patient presented to their GP and were prescribed OCS, but then required subsequent secondary care within 2 weeks, the attack would be labelled herein as OCS. Overall, the rate of asthma attacks in the analysis population was 687.7 per 10,000 person-years (95% CI = 675.5-699.9). As shown in Figure 7.5, the majority of this constituted of OCS prescriptions (368.5 per 10,000 person-years, 95% CI = 359.5-277.4). There were approximately 1.9 deaths per 10,000 person-years.

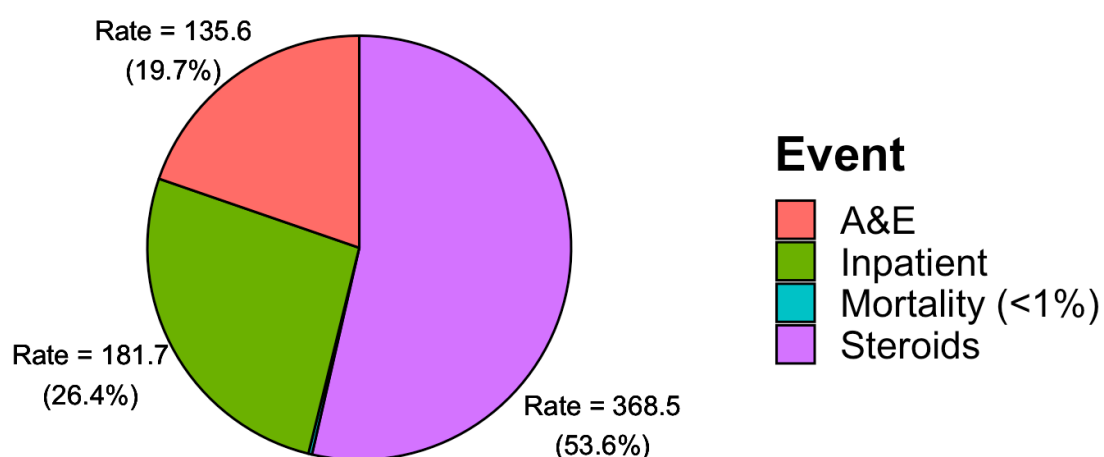


Figure 7.5: Rate of asthma attack events per 10,000 person-years, and percentage of total attacks by event type

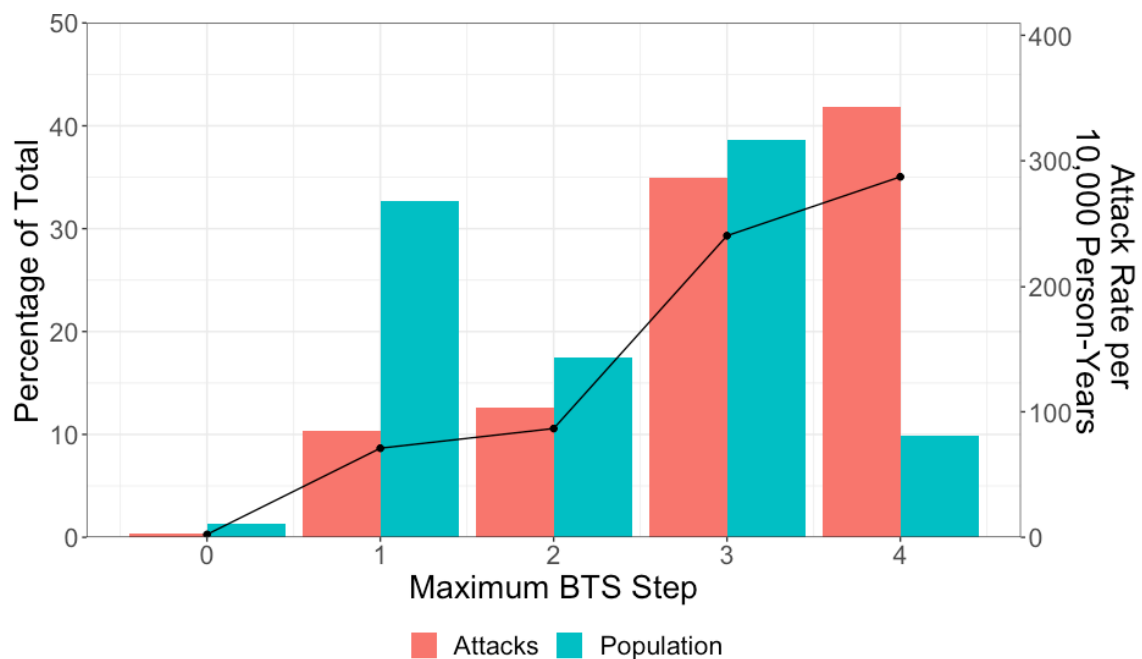


Figure 7.6: Percentage of all patients and all asthma attacks in analysis population by maximum British Thoracic Society (BTS) treatment step during follow-up

Overall, 18.3% of the population had at least one identified attack, but only 14.4% of patients had some record of asthma attacks in their Read Codes. As shown in Figure 7.7, 66.3% of patients with at least one identified asthma attack had no record of the event in their Read Codes. Additionally, 10.1% of people with no identified asthma attacks *did* have a Read Code of an asthma attack (2140/21,234).

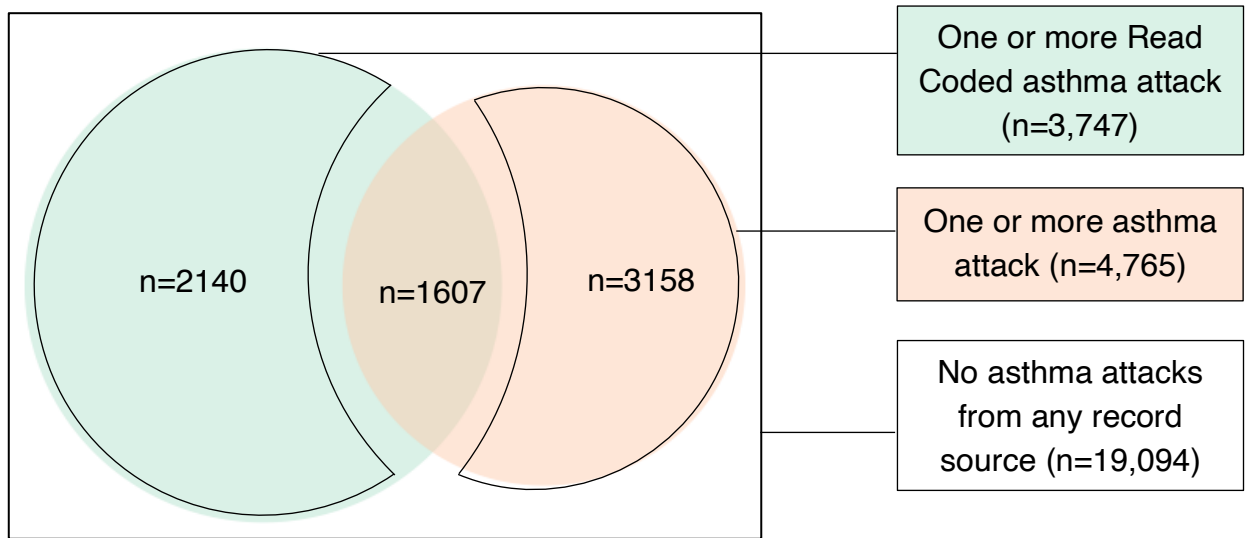


Figure 7.7: Venn diagram of patients with one or more asthma attack according to study criteria (green) and Read Codes (orange)

### 7.4.3 Algorithm and Enrichment Selection

Ten iterations of partitioning, training, and validation were conducted in the derivation dataset to inform the selection of the algorithm and enrichment method to be carried through the next phase of analysis.

Table 7.5 summarises the distribution of the class samples sizes and the minor class proportion. The minor class proportion was actually slightly below our 0.1% estimate (0.08%), which meant that the SMOTE parameters did not result in equal total sample size across enrichment methods. Enrichment method four, for example, had a mean total sample size of only 66% of that of the unenriched data (method one).

Table 7.5: Class sample sizes for enrichment methods across iterations

Enrichment Method	Minor Class Proportion	Class	Mean Sample Size	Minimum Sample Size	Maximum Sample Size
1	0.1%	Negative	926947.5	926913	926997
		Positive	7831.5	7782	7866
		Total	934779	934779	934779
2	15.1%	Negative	704835	700380	707940
		Positive	125304	124512	125856
		Total	830139	824892	833796
3	7.0%	Negative	626520	622560	629280
		Positive	46989	46692	47196
		Total	673509	669252	676476
4	5.1%	Negative	587362.5	583650	589950
		Positive	31326	31128	31464
		Total	618688	614778	621414

Notes: The minor class proportion was estimated from the mean sample size across iterations.

Enrichment methods: (1) unenriched data, (2) high up-sampling SMOTE, (3) medium up-sampling SMOTE, (4) low up-sampling SMOTE.

The boxplots presented in Figure 7.8 show that, across enrichment methods, the RF algorithm consistently performs higher than the other algorithms according to the AUC, across the hyper-parameter range investigated. Furthermore, SMOTEing shows no improvement for any algorithm over the original data.

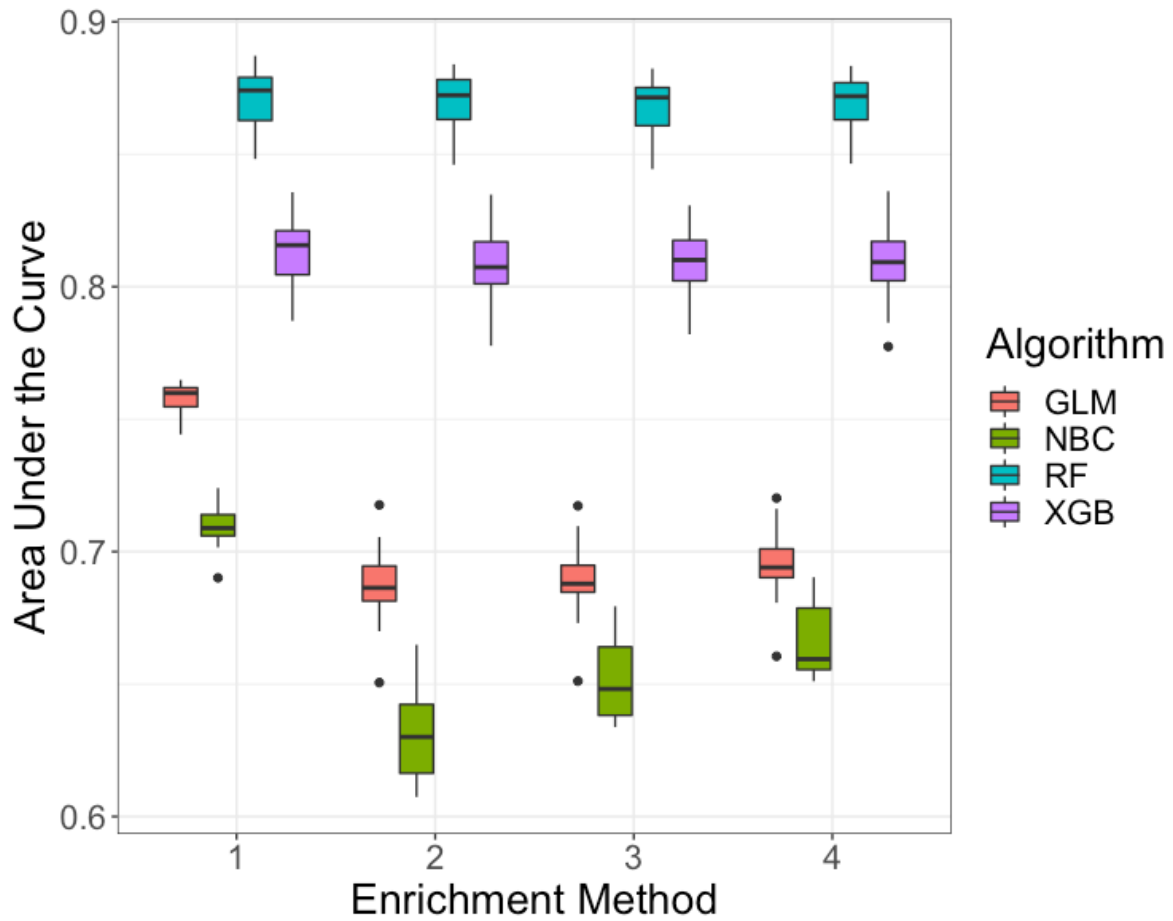


Figure 7.8: Boxplots of the area under the curve for each algorithm and enrichment method

Notes: Algorithms: GLM = Generalised Logistic Regression, NBC = Naïve Bayes Classification, RF = Random Forest, XGB = eXtreme Gradient Boosting.

Enrichment methods: (1) unenriched data, (2) high up-sampling SMOTE, (3) medium up-sampling SMOTE, (4) low up-sampling SMOTE.

As shown in Figure 7.9, an average balanced accuracy of greater than 70 is achieved by the GLM and XGBoost in the unenriched data, and by the RF algorithm for most enrichment methods. Optimising the classification threshold (based on the balanced accuracy in the training data) resulted in higher average performance in the validation data across all algorithms and enrichment methods.

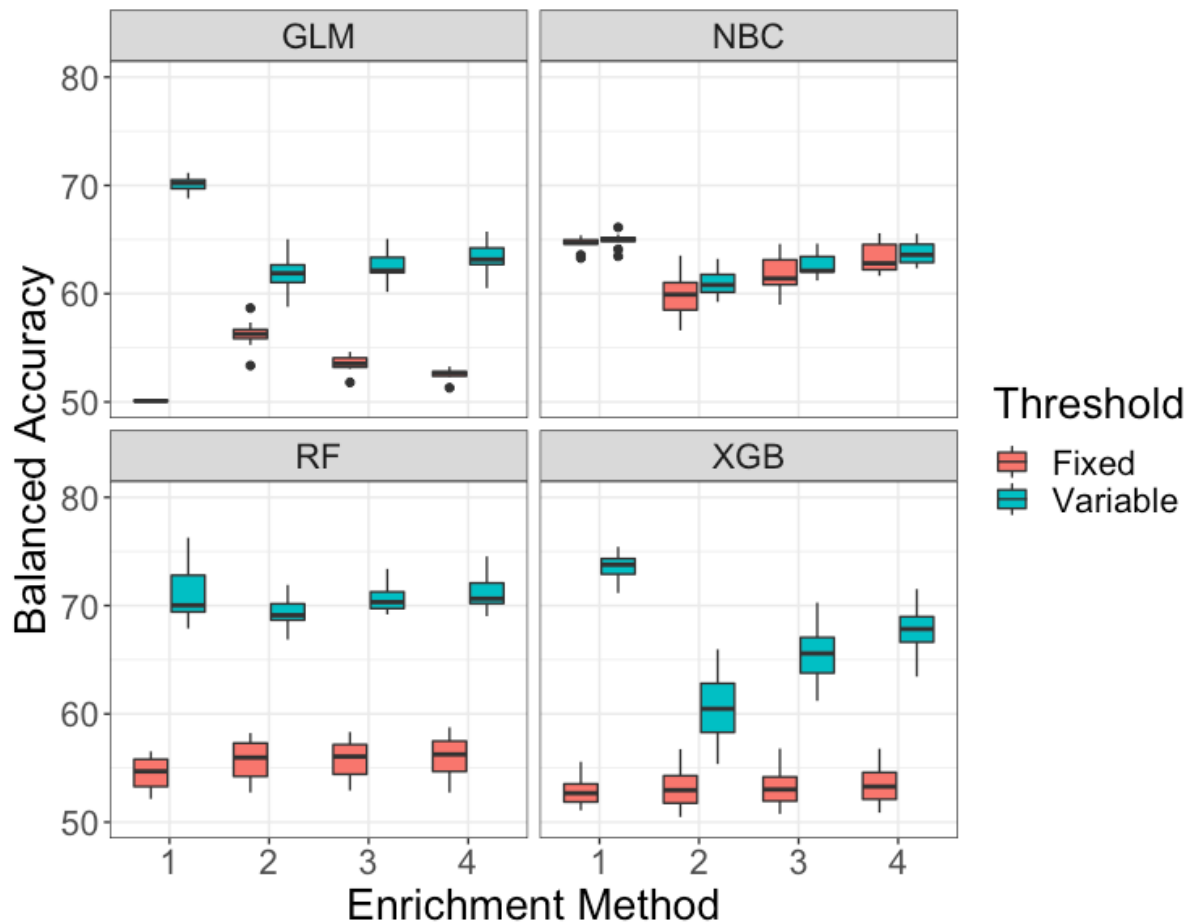


Figure 7.9: Boxplots of balanced accuracy for each algorithm, enrichment method, and classification threshold approach

Notes: Algorithms: GLM = Generalised Logistic Regression, NBC = Naïve Bayes Classification, RF = Random Forest, XGB = eXtreme Gradient Boosting.

Enrichment methods: (1) unenriched data, (2) high up-sampling SMOTE, (3) medium up-sampling SMOTE, (4) low up-sampling SMOTE.

Thresholds: Fixed = 0.5, Variable = balanced accuracy optimising threshold in training data.

As seen in Figure 7.10, all algorithms achieved greater than 70% specificity, with most achieving over 95% on average by enrichment method and thresholding approach. The specificity was much lower for the NBC than the other algorithms on average, and it was also the only algorithm for which the fixed classification threshold of 0.5 resulted in higher specificity than the balanced accuracy optimised threshold (specifically, in the unenriched data).

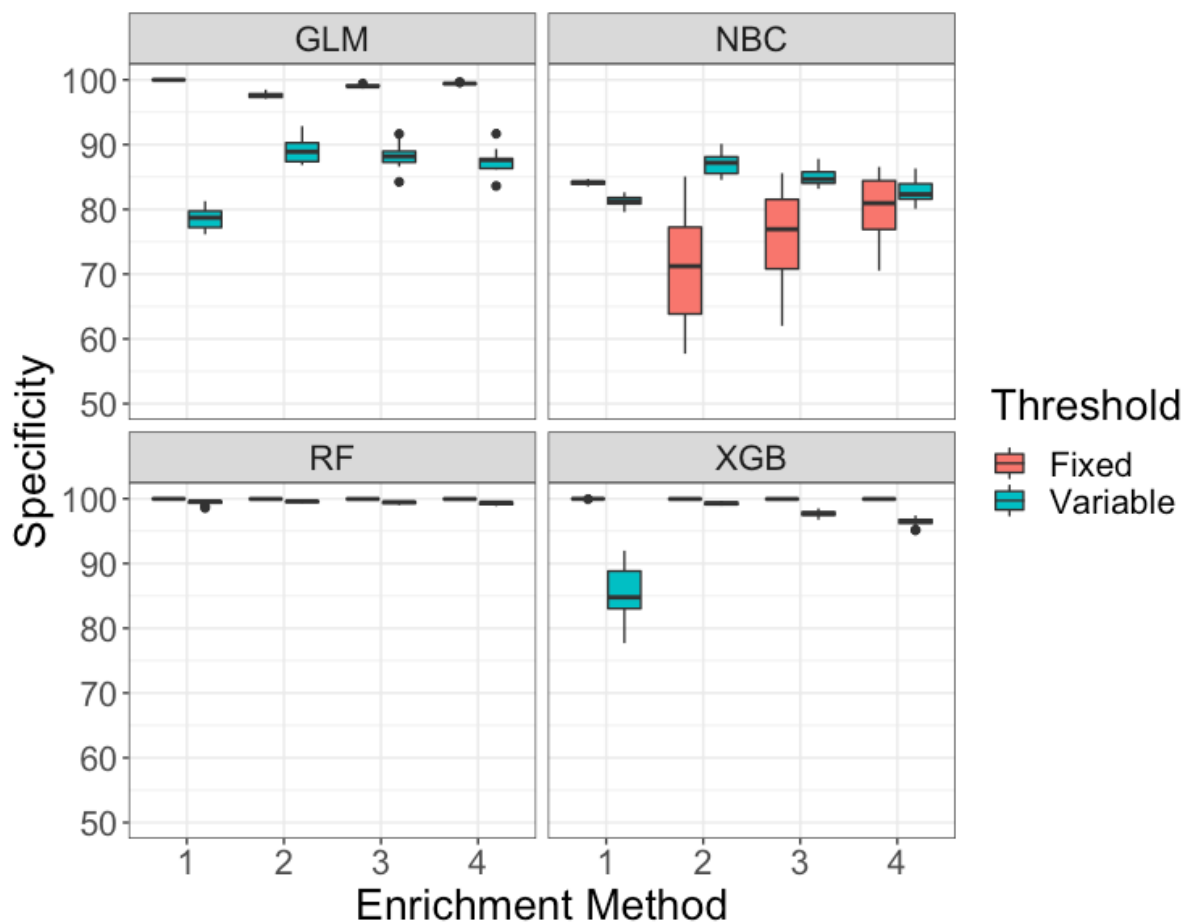


Figure 7.10: Boxplots of specificity for each algorithm, enrichment method, and classification threshold approach

Notes: Algorithms: GLM = Generalised Logistic Regression, NBC = Naïve Bayes Classification, RF = Random Forest, XGB = eXtreme Gradient Boosting.

Enrichment methods: (1) unenriched data, (2) high up-sampling SMOTE, (3) medium up-sampling SMOTE, (4) low up-sampling SMOTE.

Thresholds: Fixed = 0.5, Variable = balanced accuracy optimising threshold in training data.

In Figure 7.11, we see that the sensitivity follows a similar pattern to the AUC, with the unenriched, optimised threshold GLM and XGBoost achieving average sensitivity of over 60%. For the RF, the median was consistently around 40% for all enrichment methods, when the optimised threshold was used.

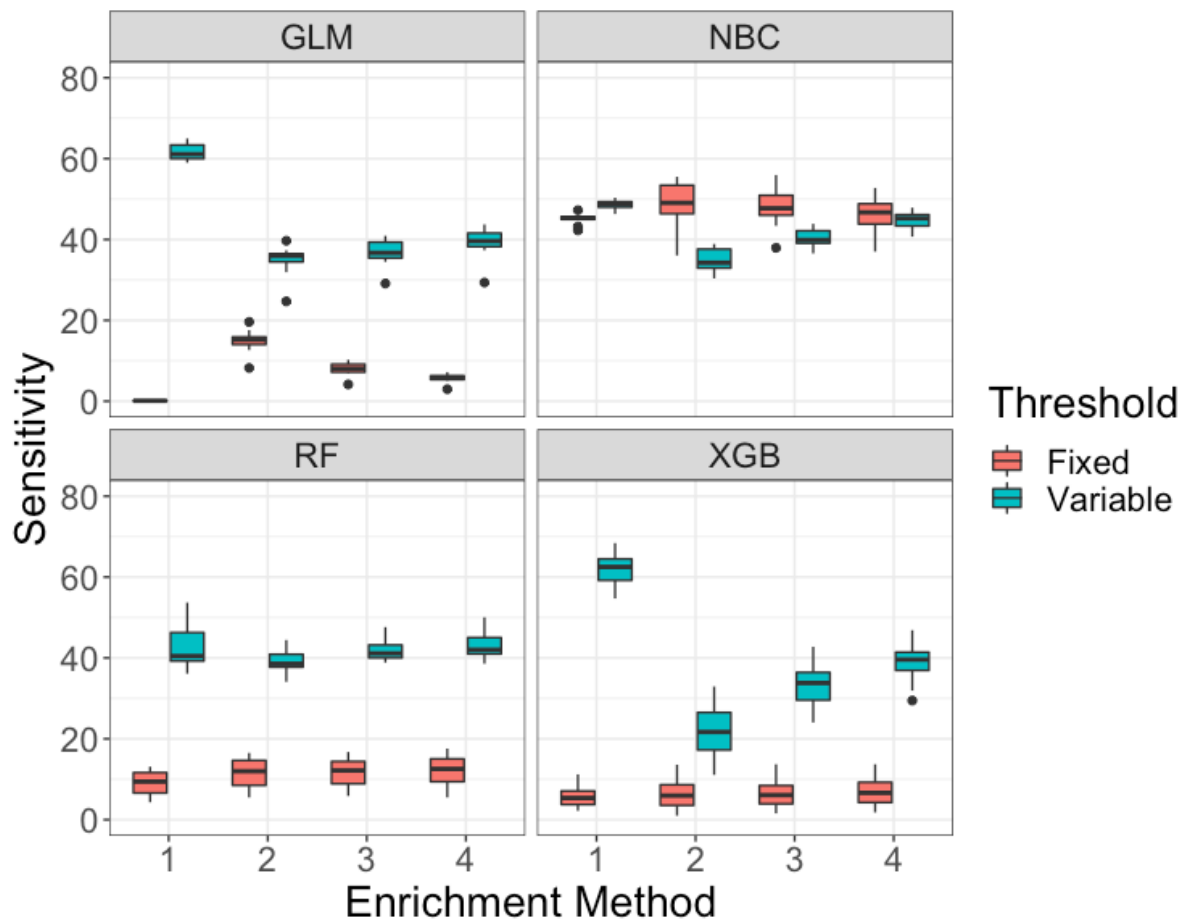


Figure 7.11: Boxplots of sensitivity for each algorithm, enrichment method, and classification threshold approach

Notes: Algorithms: GLM = Generalised Logistic Regression, NBC = Naïve Bayes Classification, RF = Random Forest, XGB = eXtreme Gradient Boosting.

Enrichment methods: (1) unenriched data, (2) high up-sampling SMOTE, (3) medium up-sampling SMOTE, (4) low up-sampling SMOTE.

Thresholds: Fixed = 0.5, Variable = balanced accuracy optimising threshold in training data.



As shown in Figure 7.12, however, the higher sensitivity for the GLM and XGBoost came at the expense of the PPV, which was below 25 for all (optimised threshold) algorithms except RF. Plots for additional performance measures accuracy, MCC, Brier Score, and NPV are presented in Appendix R.

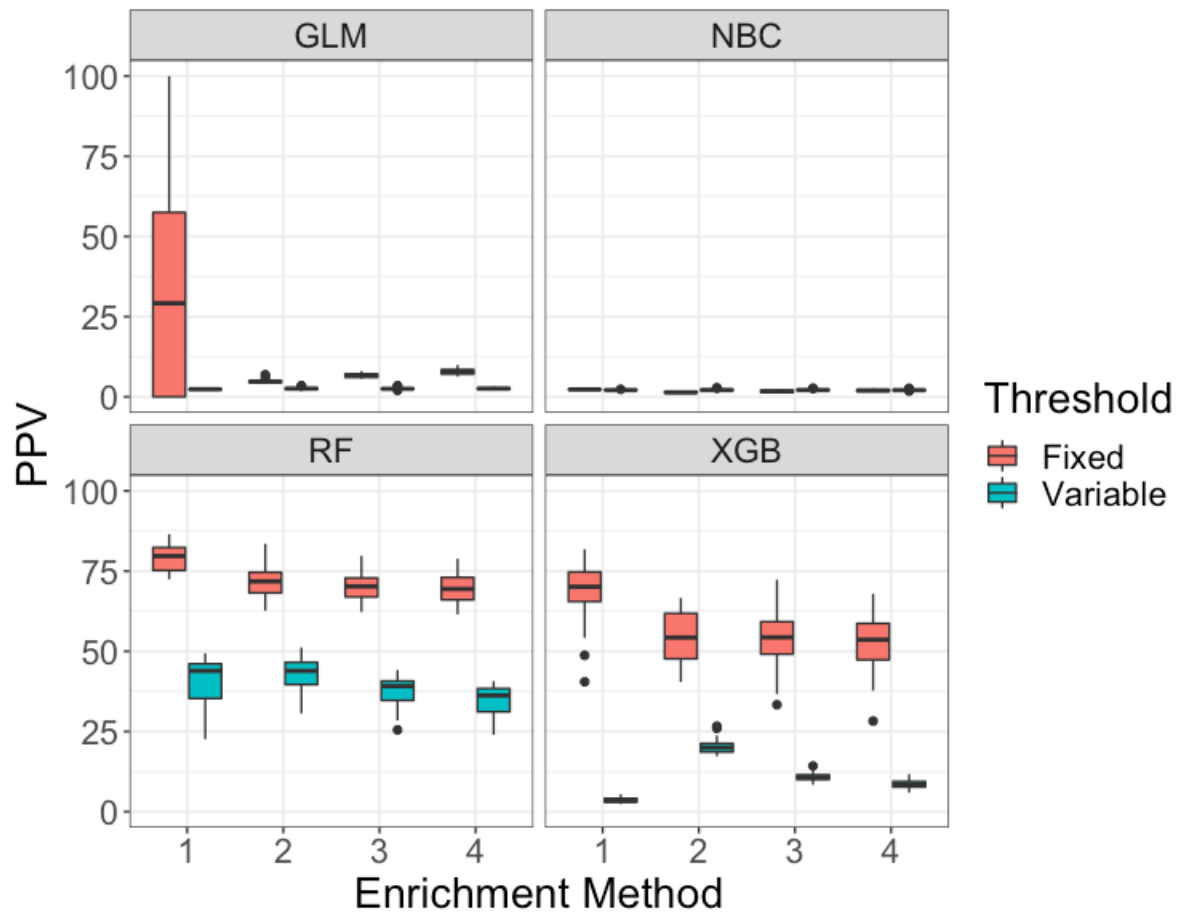


Figure 7.12: Boxplots of Positive Predictive Value (PPV) for each algorithm, enrichment method, and classification threshold approach

Notes: Algorithms: GLM = Generalised Logistic Regression, NBC = Naïve Bayes Classification, RF = Random Forest, XGB = eXtreme Gradient Boosting.

Enrichment methods: (1) unenriched data, (2) high up-sampling SMOTE, (3) medium up-sampling SMOTE, (4) low up-sampling SMOTE.

Thresholds: Fixed = 0.5, Variable = balanced accuracy optimising threshold in training data.

Overall, the RF appeared to be the best performing algorithm as it had the highest AUC, and the best balance of the sensitivity and PPV. The results were also very stable across iterations, with clear distinctions between the cases investigated. As such, I was confident in my selection of algorithm (RF) and enrichment method (none) and did not require further iterations to be added at this stage.

Consequently, the subsequent analyses focussed on identifying the optimal hyperparameters and classification threshold for the RF using the unenriched data, with balanced accuracy optimised classification threshold. Figure 7.13 shows the sensitivity, PPV, and MCC for the four RF models (varying values of the *mtry* parameter, which defines the number of features randomly sampled as candidates at each split).

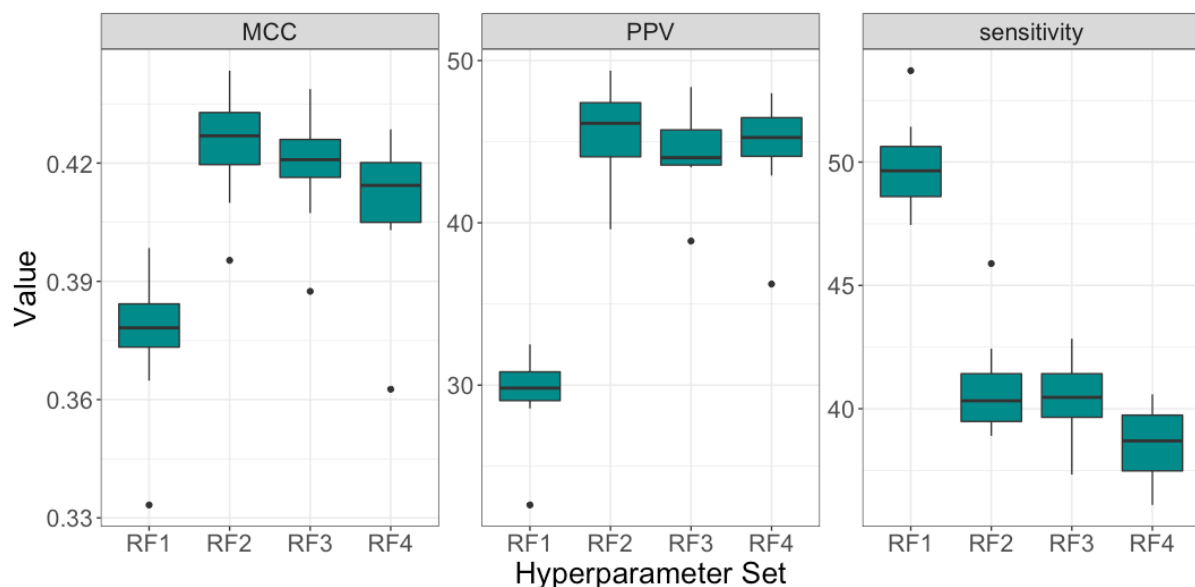


Figure 7.13: Sensitivity, Positive Predictive Value (PPV) and Matthews Correlation Coefficient (MCC) for the four RF models, using unenriched training data and optimised classification thresholds

Notes: Number of variables randomly sampled as candidates at each split (default square root of the number of predictors; k): RF1 =  $\text{floor}(\sqrt{k})$ , RF2 =  $\text{floor}(2*\sqrt{k})$ , RF3 =  $\text{floor}(4*\sqrt{k})$ , RF4 =  $\text{floor}(8*\sqrt{k})$

Generally, we see that the lowest *mtry* value results in the highest sensitivity but the lowest PPV. The MCC, which incorporates both of these values in its calculations, was used for the model selection. The second model, with  $\text{floor}(2 \cdot \sqrt{k})$  (equal to 22 of 124), variables randomly sampled as candidates at each split, was chosen. Once again, the boxplot demonstrates sufficient stability across iterations to make this selection with confidence.

#### 7.4.4 Model Performance

In Table 7.6, the summary statistics of a selection of model performance measures across the 100 iterations of derivation data partitioning are presented. There were no substantial differences to the results seen in the 10 iterations at the model selection phase.

The threshold that optimised the balanced accuracy in each training data partition ranged between 0.111 and 0.168, with a median of 0.150 (mean 0.148). In a linear regression, adjusted for the resultant AUC, lower threshold values were significantly associated with higher balanced accuracy in the validation partition ( $p < 0.001$ ), recalling that the threshold is optimised in the training data. The formula is shown below:

$$\text{balanced accuracy} = 42.7 + 40.43 * \text{AUC} - 50.97 * \text{threshold}$$

However, higher threshold values were associated with higher PPV (and lower threshold values with higher sensitivity). For model evaluation in the holdout partition, the median probability threshold of 0.149974 was chosen to classify samples as low or high predicted risk in the hold-out data. The confusion matrix in the holdout partition is presented in Table 7.7.

Table 7.6: Summary statistics of model performance measures from 100 data partition iterations, and the hold-out data partition

Performance Measure	ALHS Model Development Data						ALHS Hold-out Validation Data
	Minimum	25 <sup>th</sup> Percentile	Median	Mean	75 <sup>th</sup> Percentile	Maximum	
Sensitivity	37.84	40.74	41.71	41.72	42.56	46.83	47.70
Specificity	99.31	99.52	99.58	99.56	99.61	99.67	99.57
PPV	36.10	42.70	44.92	44.80	47.02	50.50	48.90
NPV	99.46	99.49	99.51	99.51	99.52	99.58	99.55
Accuracy	98.86	99.05	99.09	99.08	99.12	99.17	99.13
AUC	86.03	87.38	87.84	87.77	88.19	89.77	90.72
Balanced Accuracy	68.74	70.24	70.65	70.64	71.07	73.07	73.64
MCC	38.76	41.54	42.86	42.72	43.79	46.18	47.86

Note: Green cells indicate that this value was exceeded in the hold-out validation data partition.

As shown in the final column of Table 7.6, the performance in the holdout partition was consistently within, or above, the range observed in the derivation dataset. Performance in the holdout partition was within the top 25% of derivation iterations' performance for the PPV, NPV and the accuracy. In the sensitivity, AUC, balanced accuracy, and MCC, it was higher in the holdout partition (with a larger amount of training data, and a more robust classification threshold) than in any derivation iterations. This internal validation demonstrates the stability of the model performance within this data to perturbations of the sample set and confirms that the crossover in samples between the model selection and performance reporting subsets did not bias the results.

Table 7.7: Confusion matrix for model performance in holdout partition

		<b>Asthma Attack in the 4 weeks following an asthma consultation</b>	
		<b>Yes</b>	<b>No</b>
<b>Predicted Class</b>	<b>High Risk</b>	467	488
	<b>Low Risk</b>	512	113,937

The ROC curve is presented in Figure 7.14, with threshold values for each combination of sensitivity (true positive rate) and specificity (1- true negative rate) colour-coded. We can see that any threshold above 0.2 yields very poor sensitivity.

The density plot of the estimated probabilities by the observed outcome is shown in Figure 7.15. The median estimated attack probability in those who did have an attack was 13.7% (IQR = 2.2 – 31.6%) and was 0.2% in those who did not have an attack (IQR = <0.1 - 0.6%).

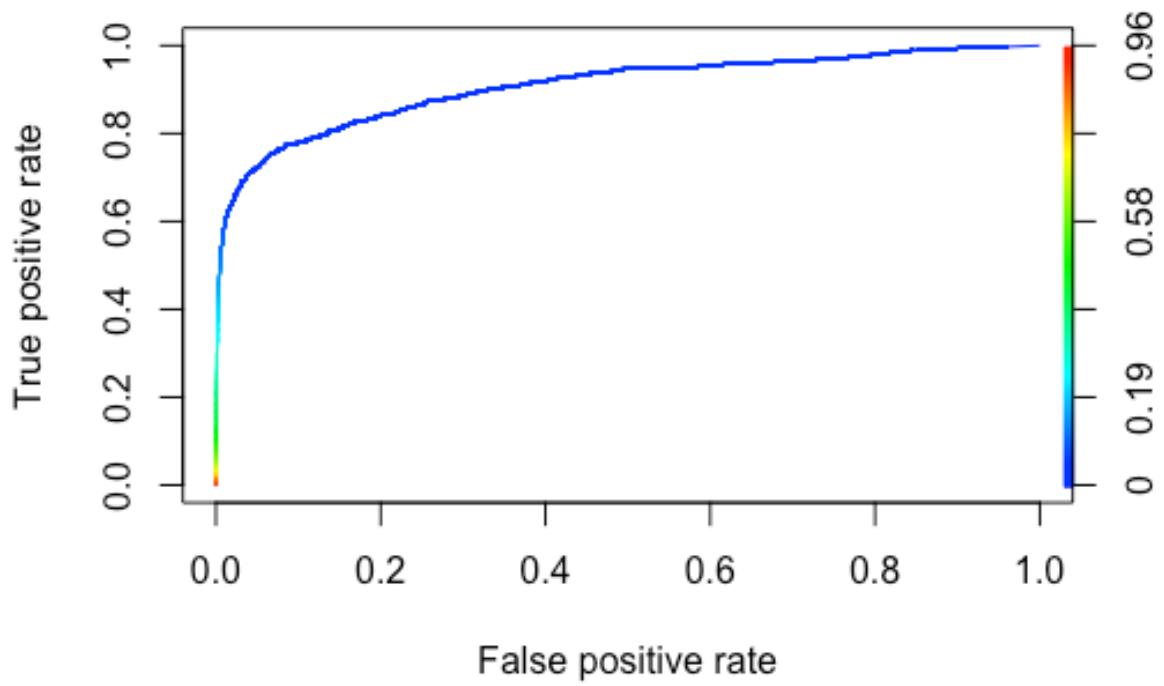


Figure 7.14: Receiver Operator Curve for model performance in holdout partition, with threshold values indicated by colour

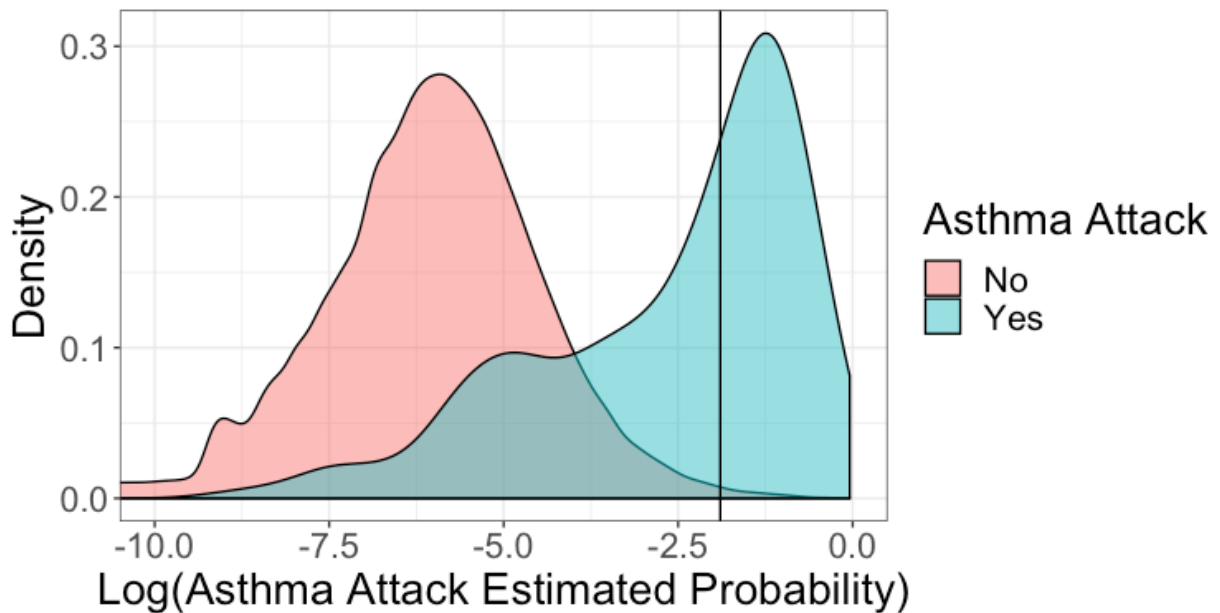


Figure 7.15: Density plot of logarithm of estimated probabilities by observed outcome in holdout partition

#### 7.4.5 Feature Importance

Appendix S lists the feature importance (mean decrease in impurity, defined in Section 5.6.1) of all features in the model, excluding the NUTS-3 area codes. The three most important features in the derivation dataset were CSA\_3, reliever medication use, and age (Figure 7.16). The least important features were AIDS diagnosis, hemiplegia diagnosis, and nasal polyps diagnosis more than five years ago. In fact, no features relating to nasal polyps (such as diagnosis in the last year) ranked in the top 100. The latter is surprising perhaps, given the high unadjusted odds ratio seen in Table 7.4.

Additionally, for tree-based methods, feature importance can be quantified by averaging the mean decrease in impurity (or the corresponding performance measure used to grow the tree) that would be achieved by using each feature as the splitting criterion for each parent node in each tree<sup>314</sup>. These values should be considered relative to each other, rather than in terms of the computed magnitude.

Nasal spray prescription, on the other hand, had a lower unadjusted odds ratio, but was the 19<sup>th</sup> most important feature (prescription in the last year), and the second most important comorbidity related feature behind chronic pulmonary disease (10<sup>th</sup> most important overall). As discussed in Section 7.4.1, this included explicit asthma diagnosis codes (rather than asthma management codes; everyone in the study had one or the other according to inclusion criteria), in addition to pulmonary fibrosis, asbestosis, and others. The importance of this feature highlights that medical coding is shrouded in nuance, and that the use of specific codes can hold more meaning than the description of the code itself. The use of certain codes changes over time, such as when new QOF guidelines are introduced, and so it may be important for further development of the model to include the year of data collection.

Despite such high levels of missingness (74%; Appendix O), blood eosinophil count under 400 cells per  $\mu\text{L}$  was the 34<sup>th</sup> most important feature (missing and over 400 cells per  $\mu\text{L}$  were 40<sup>th</sup> and 42<sup>nd</sup> most important features, respectively). Peak flow did not manage to remain important with 98% missingness, but interestingly missing and greater than 90% of baseline were 89<sup>th</sup> and 92<sup>nd</sup> most important features, with the former being 4.3 times more important than a peak flow measurement of under 70% of baseline.



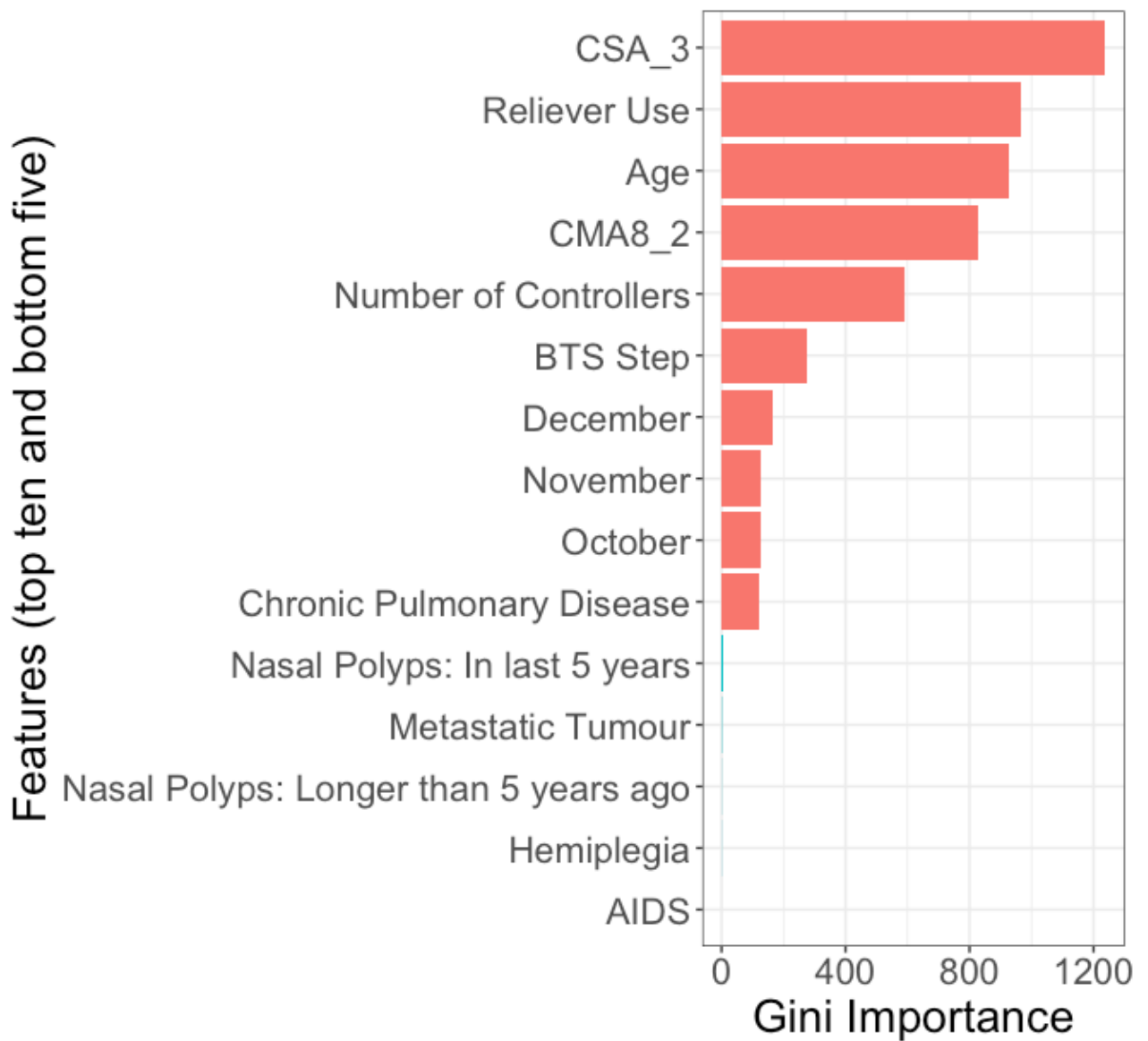


Figure 7.16: Top and bottom ranked features by Gini importance

Note: features relating to geographical areas (NUTS-3 codes) have been excluded from these rankings.

### 7.4.6 Model Calibration

Figure 7.17 shows the calibration between the estimated probability of an asthma attack and the observed rate of asthma attacks in the holdout partition. 98.7% of samples had estimated risk under 0.1, 0.7% had estimated risk between 0.1 and 0.2, and 0.6% had estimated risk over 0.2. The calibration-in-the-large (Section 5.4.4) was estimated as -4.77 and the calibration slope was estimated at 18.27. This demonstrates inappropriate scaling of estimated probabilities.

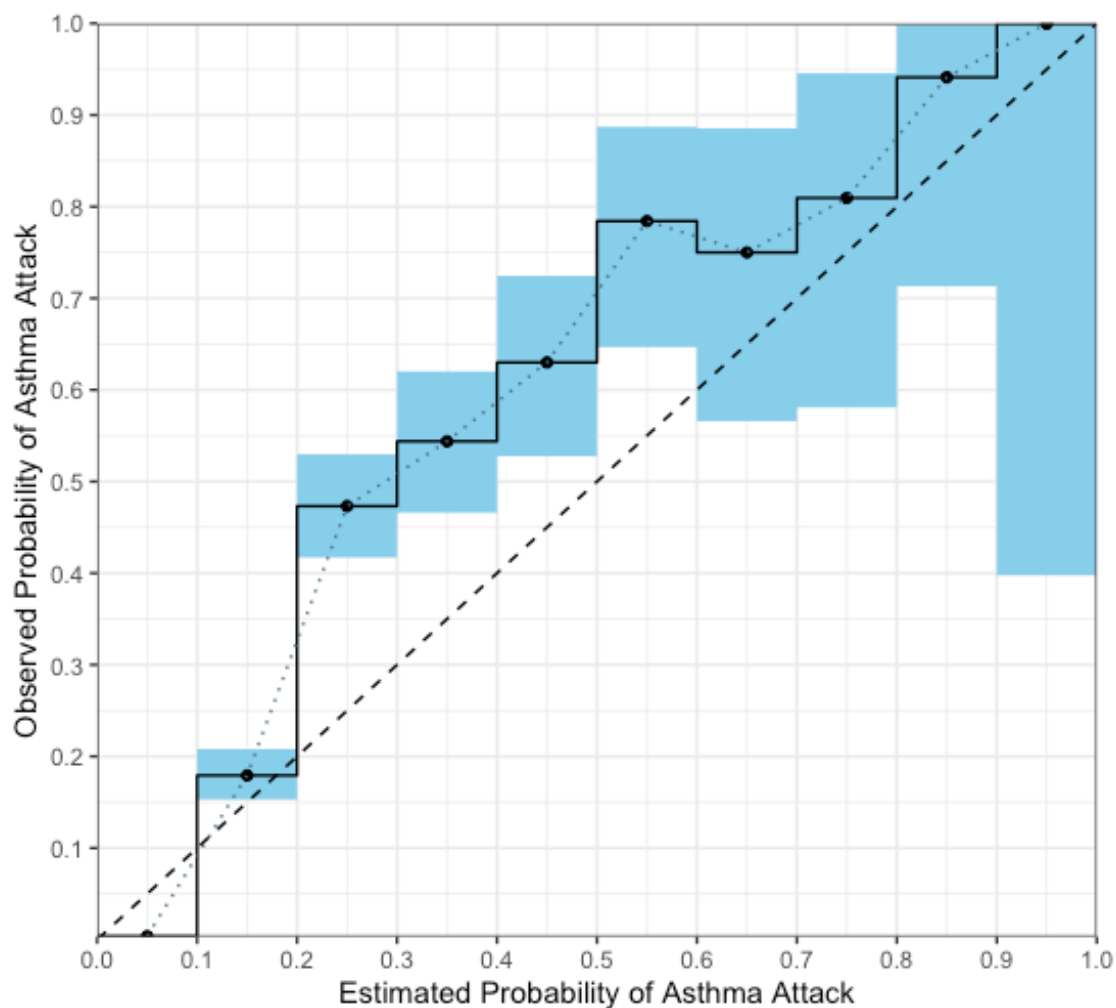


Figure 7.17: Model calibration by risk deciles in the holdout partition

Note: the bins used to approximate the observed probabilities are of width 0.1, and the confidence intervals were estimated using the exact binomial test.

### 7.4.7 Model Discrimination in Subgroups

The discrimination in selected population subgroups was evaluated by comparing the number of false negative and false positive predictions to the number of true positive predictions. As shown in Figure 7.18, those with previous asthma attacks have higher sensitivity (58.8%) than those without (38.9%), with modest differences in PPV (48.0% compared to 50.0%).

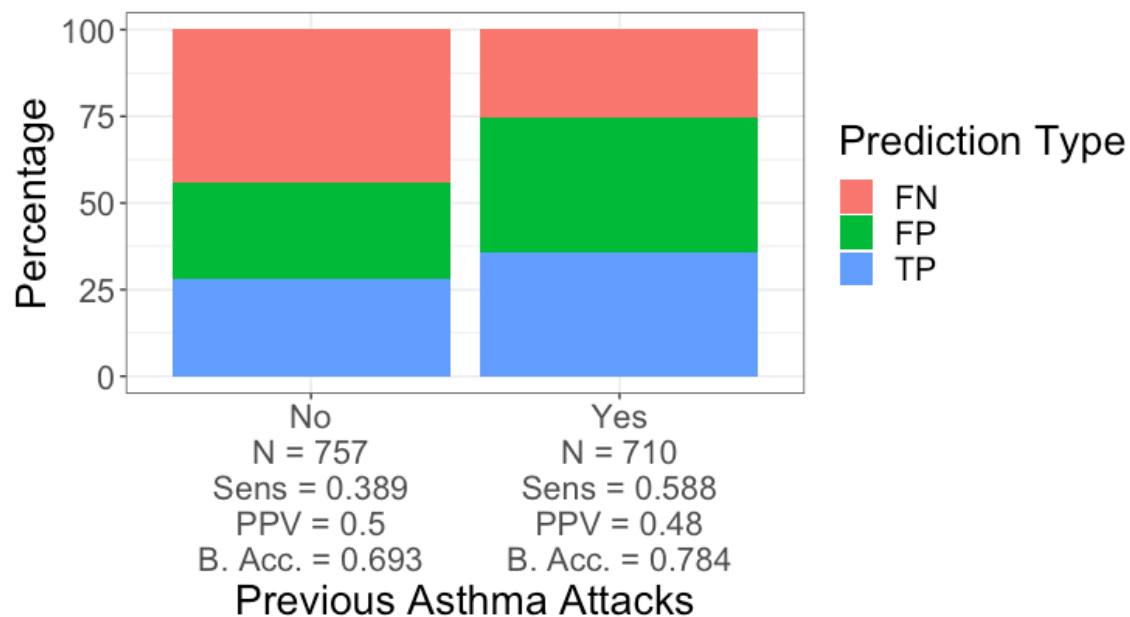


Figure 7.18: Discrimination by asthma attack history in the holdout partition

Notes: Sens = sensitivity, PPV = Positive Predictive Value, B. Acc. = Balanced Accuracy

Secondly, I wanted to know how well the model discriminated by asthma severity (as defined by BTS treatment step). Figure 7.19 shows that the sensitivity and PPV were both lower in those at BTS step 0 (no controller therapy; 29.5% and 48.6%, respectively) than at other steps (32.8% and 51.2%, for each BTS step, respectively).

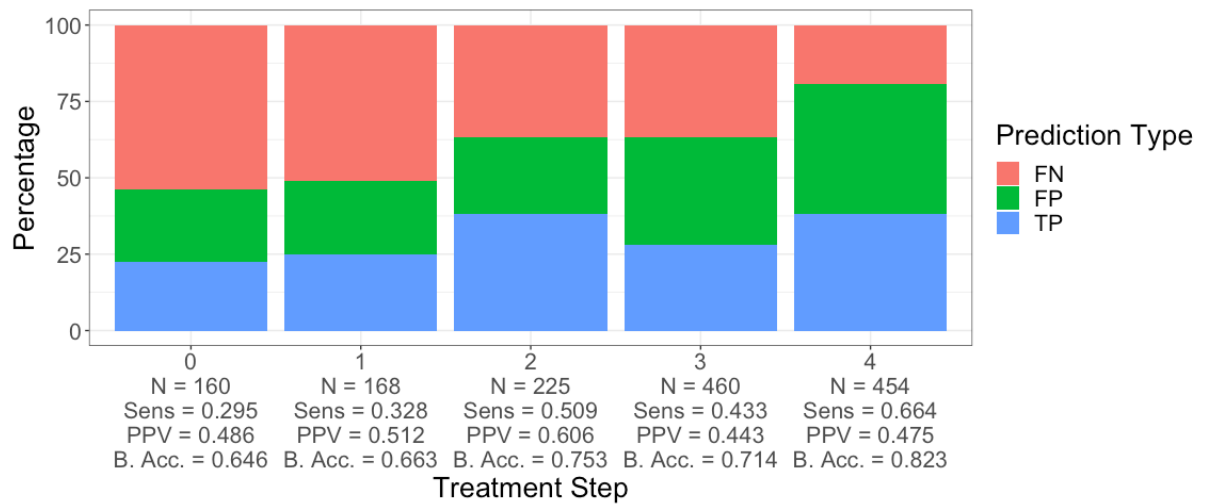


Figure 7.19: Discrimination by asthma severity, according to treatment step, in the holdout partition

Notes: Sens = sensitivity, PPV = Positive Predictive Value, B. Acc. = Balanced Accuracy

Next, I wanted to know whether the model was predicting severe attacks (A&E and hospitalisations) better or worse than attacks which were treated in primary care. The sensitivity for asthma attacks treated in primary care was 45.4%, and the PPV was 35.6% (Figure 7.20). For those treated in secondary care, the sensitivity was higher (51.3%) but the PPV was only 28.8%. For both of these analyses, the negative class was not having any attack, which is why both of the PPV values are lower than the overall average.

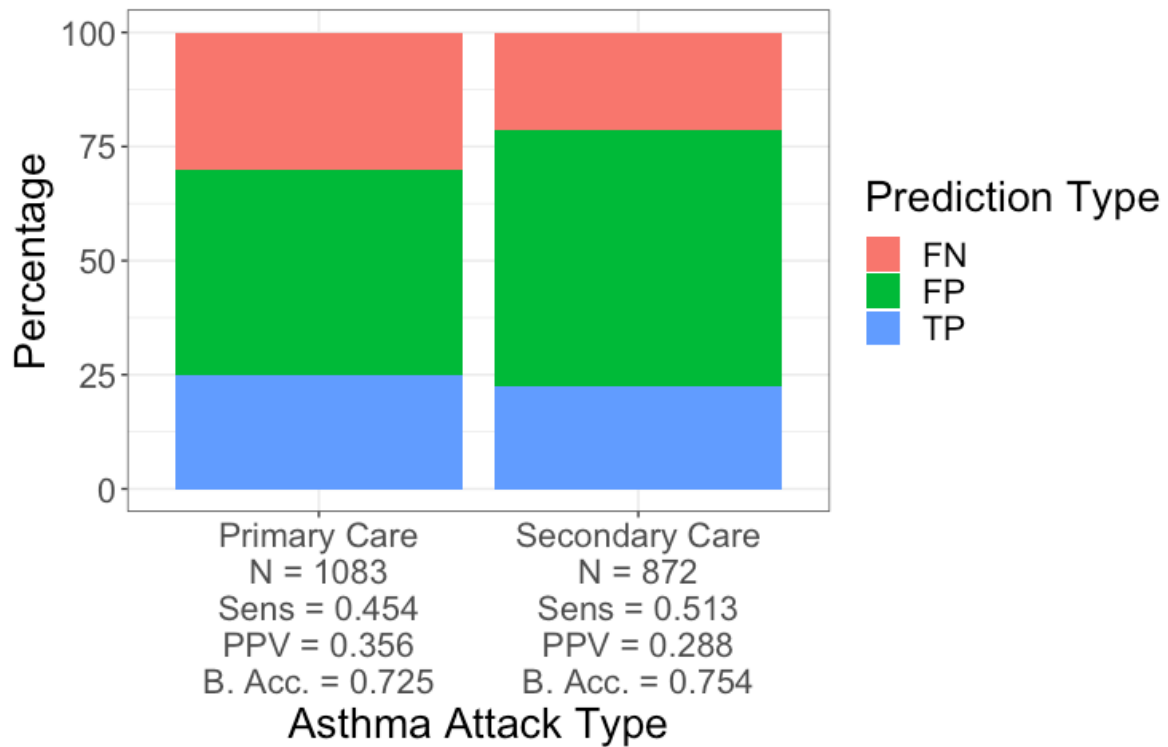


Figure 7.20: Discrimination by asthma attack severity in the holdout partition

Notes: Sens = sensitivity, PPV = Positive Predictive Value, B. Acc. = Balanced Accuracy

Finally, discrimination by smoking status was used to assess whether possibly undiagnosed COPD or ACOS might have affected model performance. The sensitivity was highest in former smokers (56.8%), and lowest in current smokers (41.8%). The PPV was similar between smoking statuses, but slightly higher in current smokers (50% compared to 48.9%; Figure 7.21)

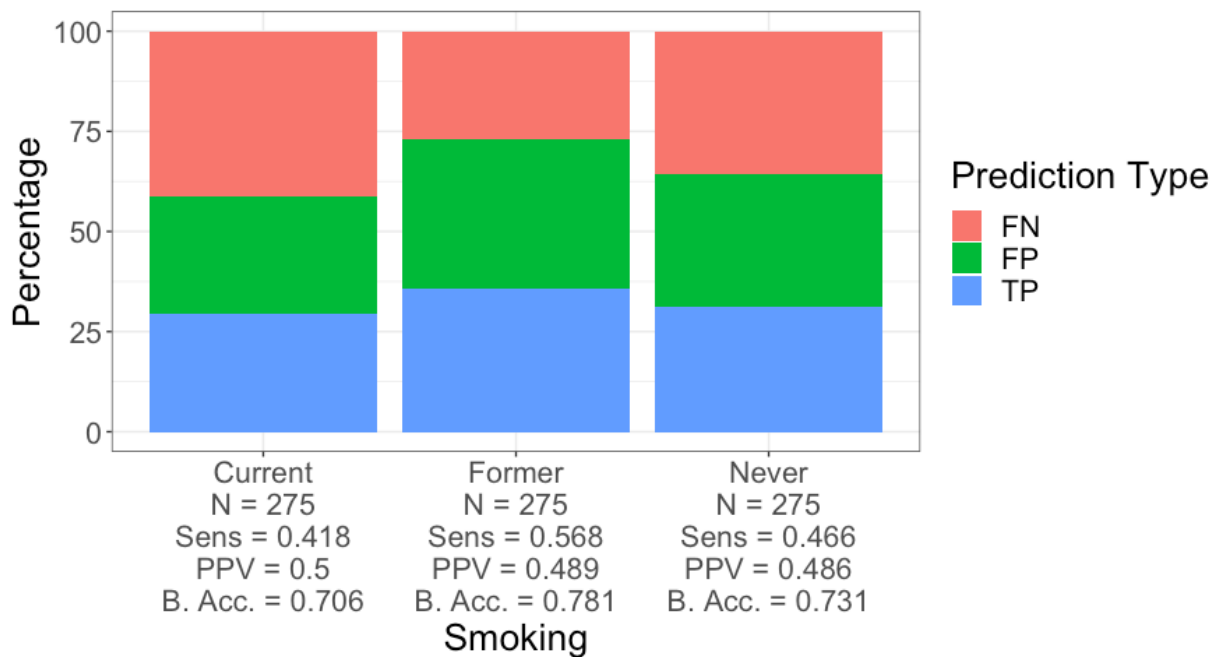


Figure 7.21: Discrimination by smoking status in the holdout partition

Notes: Sens = sensitivity, PPV = Positive Predictive Value, B. Acc. = Balanced Accuracy

### 7.4.8 Secondary Endpoints

For our primary endpoint, longer event horizons resulted in better performance according to the sensitivity, PPV, AUC, balanced accuracy, and MCC, with only modest reductions in specificity, NPV, and accuracy (Table 7.8). For the model predicting attacks within the next 12 weeks (and all subsequent event horizon models), the six benchmarks were clearly met. For example, the sensitivity was 83%, and the PPV was 72%. The same trend was seen in the models using only attacks presenting in secondary care, with comparable performance across outcome definitions at the same event horizon. This improvement in performance is likely a combination of better class balance, and more prominently that there are further important features for shorter-term prediction which are not available in EHRs, such as allergen exposure and weather.

Table 7.8: Performance measures for secondary endpoints

Performance Measure	All Attacks					Secondary Care Attacks				
	Event Horizon									
	1 Week	4 Weeks	12 Weeks	26 Weeks	52 Weeks	1 Week	4 Weeks	12 Weeks	26 Weeks	52 Weeks
Sensitivity	18.86	47.70	74.46	82.96	89.89	29.10	59.38	74.16	85.92	89.30
Specificity	99.84	99.57	98.84	98.72	98.49	99.89	99.70	99.55	99.25	99.28
PPV	23.83	48.90	57.94	71.94	80.15	24.22	39.79	57.02	61.20	73.80
NPV	99.79	99.55	99.45	99.32	99.31	99.92	99.86	99.79	99.80	99.76
Accuracy	99.64	99.13	98.33	98.13	97.94	99.81	99.57	99.34	99.06	99.06
AUC	79.49	90.72	96.25	97.96	98.69	82.90	93.59	97.38	98.48	98.98
Balanced Accuracy	59.35	73.64	86.65	90.84	94.19	64.50	79.54	86.86	92.58	94.29
MCC	21.02	47.86	64.86	76.30	83.80	26.46	48.40	64.71	72.08	80.72

While the top five features by importance were the same for all models, there were small differences in their relative (normalised by the maximum value for that model) importance. For example, the CSA\_3 (using the last three refills) adherence measure was more important than CMA8\_2 (using the last year's refills) for all event horizons except the very longest (attack in the next year) (Figure 7.22).

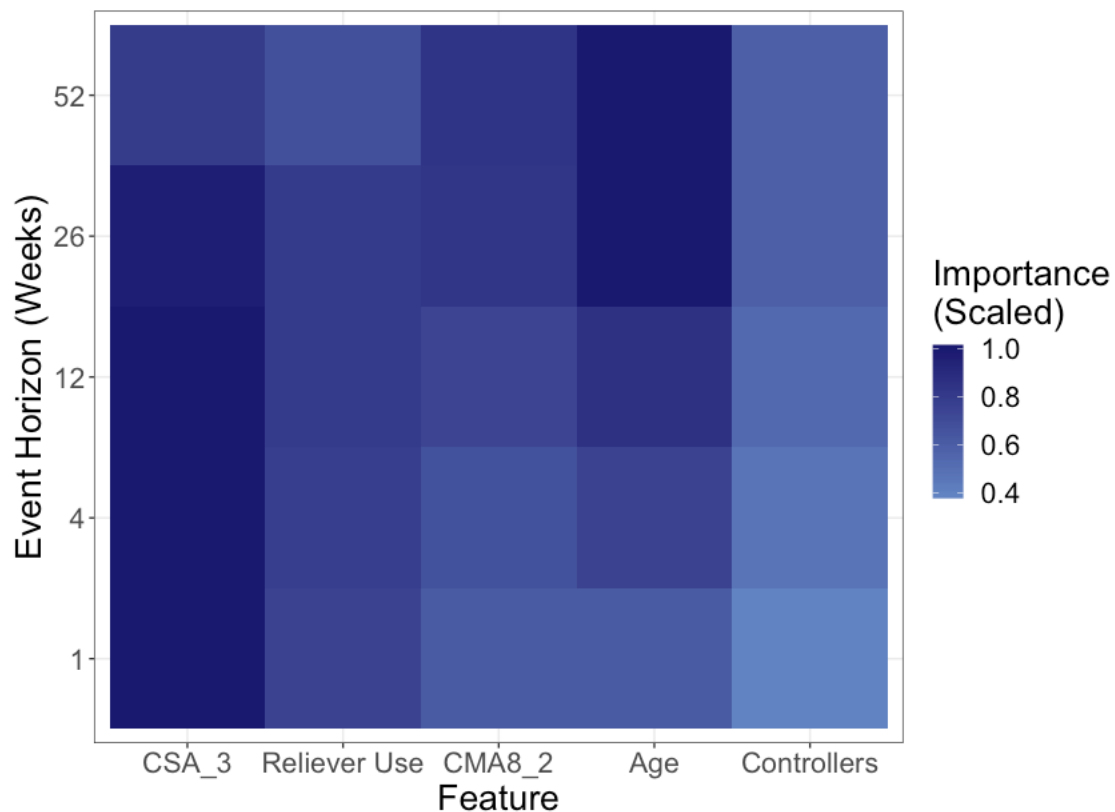


Figure 7.22: Adjusted feature importance of the top ten most important features in primary analysis, across secondary endpoints

Note: The importance of each feature within an endpoint is standardised by dividing the calculated feature importance by the maximum importance for that endpoint so that entries lie between 0 and 1.



Both age and controller use in the last year was also more important for longer event horizons, and reliever use (which is estimated using the SABA refill before last) was more important in near future prediction. The same results were observed in the models using only attacks presenting in secondary care (Figure 7.23).

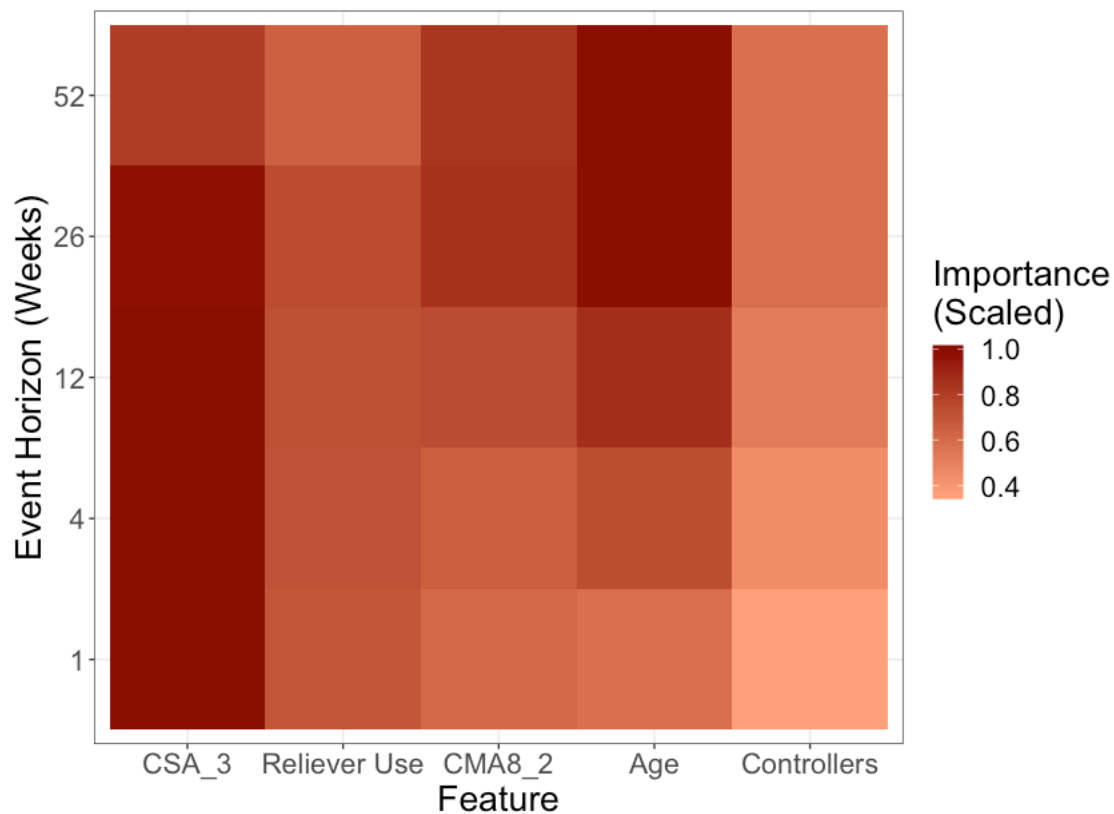


Figure 7.23: Adjusted feature importance of the top ten most important features in primary analysis, across secondary endpoints, for attacks presenting to secondary care only

Note: The importance of each feature within an endpoint is standardised by dividing the calculated feature importance by the maximum importance for that endpoint so that entries lie between 0 and 1.

I hypothesised that seasonal trends (measured by the month of the consultation) would be more influential in prediction models with shorter event horizons. The 52-week model, for example, would be positive if there was an asthma attack in any of the following seasons, whereas for the 1-week model the attack would be in the same

season as the consultation. To test this, I calculated the feature importance for each month in each model, and divided each month by the median feature importance for that model (recalling from Section 5.6.1 that feature importance should not be evaluated absolutely). Figure 7.24 shows this indeed to be true, but also suggests that consultation month was more important for predicting attacks in secondary care than in primary care.

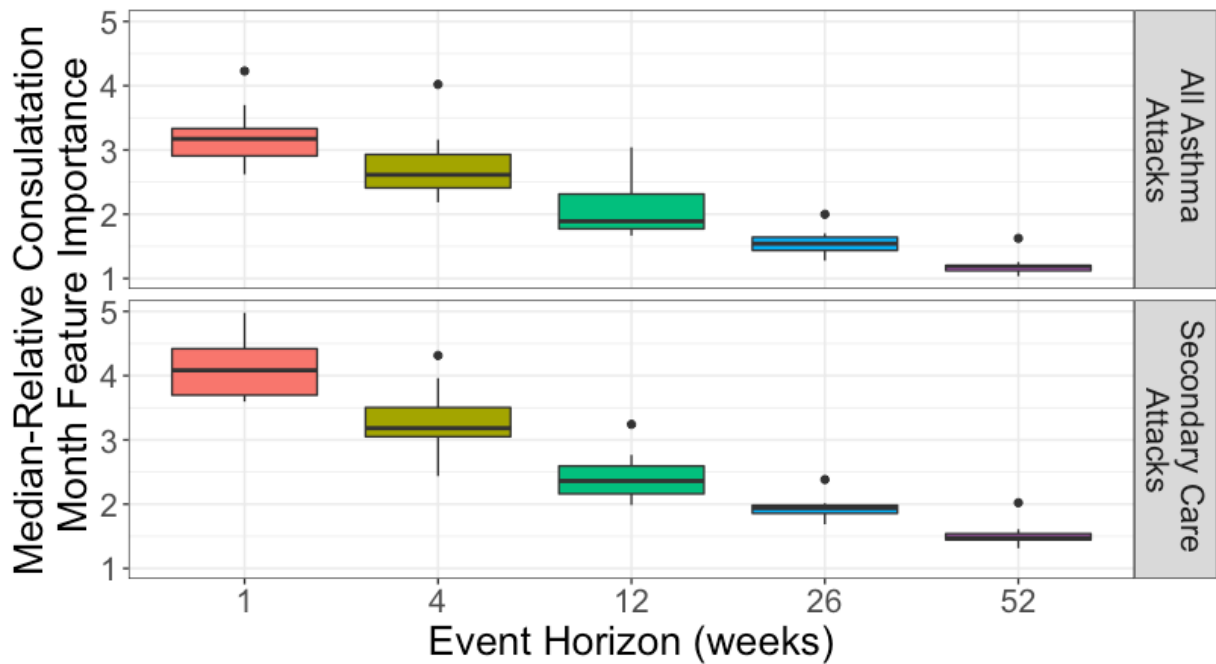


Figure 7.24: Median-relative feature importance for consultation month by health care setting and event horizon (weeks)

Note: The importance of each month feature was divided by the median feature importance for that model (event horizon and health care setting).

## 7.5 Conclusions

After evaluating multiple models, with different statistical learning algorithms, training data enrichment methods, hyper-parameters, and classification thresholds, I selected and trained a random forest model to predict asthma attacks in the next four weeks. In an unseen 'holdout' data partition, this model had a balanced accuracy of 73.6%. The sensitivity was 47.7%, and the specificity was 99.6%, largely due to the relatively small size of the positive class (low rate of asthma attacks in a four-week event horizon). The three most important features in the model were adherence averaged over the last three prescriptions (CSA\_3), the estimated daily reliever medication use, and age. The estimated probability of an asthma attack was poorly calibrated with the observed rate of attacks, with a high degree of variability seen especially in those with higher estimated risk. Using a longer prediction event horizon, such as 12 weeks rather than four weeks, substantially improved model performance.

## 8 Discussion

### 8.1 Key Findings

In this section I revisit the four major areas of my research, bringing together my key findings from studies that have: (1) compared adherence measures in EHRs, (2) compared predictive model performance measures in the context of imbalanced data, (3) quantified the incidence of asthma attacks in the general population, and (4) assessed the predictive performance of a short-term asthma attack prediction model using machine learning.

Adherence was repeatedly reported as a very important risk factor for asthma attack incidence, however the literature review of studies of adherence measured in EHRs highlighted the lack of clear guidance on best practices. Motivated by an aim to determine the most appropriate asthma adherence measures, I conducted a thorough appraisal of the literature, and a critical comparison of the methods in my data. I have cross compared findings with TB, and hence believe these findings may likely generalise in other settings where adherence measures are used. In particular, I found that rolling average windows of single-prescription measures, over a small number of prescriptions, were able to reduce the variance seen in single-prescription measures while reducing the risk of survivor bias observed in longer windows. This was particularly pertinent as 17% of all individuals with any asthma controller medications had only a single prescription over their median follow-up of 7.1 years. This also meant that many fixed-interval (such as calendar years) measures could not be calculated for a large proportion of the population. One fixed-interval measure, however, was able to account for this (CMA8). Only one fixed-interval measure (CMA1) had strong correlation ( $|IRI| > 0.7$ ) with any of the rolling average measures; a result of the different ways the calculations regarded good implementation during periods of persistence (before a discontinuation of treatment) within a fixed time period.

I conducted a theoretical and empirical investigation into the differences between various binary classification performance measures under certain conditions and investigated the three key factors which need to be considered in order to facilitate the identification of the most appropriate performance measure: the imbalance of the data, the application of the model, and the end user of the model. The results of these experiments were condensed into a simple decision graph to provide a generic roadmap to help researchers choose the likely most appropriate performance measure for the given application. In the context of my analysis, the central observations related to the impact of class imbalance; good discrimination in the risk prediction model could still be achieved by predicting all query samples to be in the major class. Consequently, the model calibration is more informative when ensuring that poor sensitivity is penalised; measures that prioritise prediction across an entire population, rather than for individuals, are more beneficial. The balanced accuracy was chosen accordingly as the most appropriate performance measure for this analysis and was used for my model selection process.

In patients with one or more asthma attack in the study follow-up (18%; median 7.8 years, interquartile range 6.3 to 8.1 years), 58% were never treated above BTS Step 1 (low-dose ICS without LABA or any add-on therapies). In unadjusted analyses, the four comorbidities with the strongest associations with asthma attacks were rheumatological diseases (OR = 4.36, 95% CI 3.07 – 6.19), nasal polyps (OR = 3.05, 95% CI 2.18 – 4.28), congestive heart disease (OR = 2.38, 95% CI 1.69 – 3.35), and anxiety/depression (OR = 2.39, 95% CI 2.21 – 2.58). The rate of asthma attacks in the analysis population was 687.7 per 10,000 person-years (95% CI = 675.5 - 699.9). Almost half of asthma attacks (47%) presented initially to secondary care, with no OCS courses prescribed in the two weeks prior. This clearly demonstrates the scope for improved early intervention in primary care. Primarily, OCS courses may be prescribed to diminish symptoms before the peak of the exacerbation, but also a timely consultation may facilitate discussion about triggers, home monitoring, and plans of action for if an attack does subsequently occur.

In an unseen data partition, the final model achieved 47.7% sensitivity and 99.6% specificity. The overall accuracy was 99.1%, and the balanced accuracy (which accounts for the relatively low incidence of the positive class) was 73.6%. The three most important features to the random forest were adherence averaged over the last three prescriptions (CSA\_3), the estimated daily reliever medication use, and age.

Using a longer prediction event horizon, such as 12 weeks rather than four weeks, improved model performance: up to 90% sensitivity and 98% specificity (80% PPV and 99% NPV) for prediction of attacks in the following year.

## 8.2 Strengths

In Chapter 2, I introduced the primary dataset for this thesis: the ALHS study data. As described therein, the study recruited 75 general practices in Scotland. These practices had over half a million patients registered, therefore covering about 9% of the population. In total, there were over 57 million entries for these patients in primary care (Read Codes and prescriptions), A&E presentations, hospital admissions, and deaths, between 2000 and 2017. The wide spread of this data resulted in a lower risk of selection bias than observational studies which individually recruited patients, as well as EHR based studies which were limited to a single geographic area or demographic (such as private US hospitals) but generalised to a much wider population.

In Section 7.4.1, I reported that 31,330 patients met the inclusion criteria of asthma consultation and treatment within the study period, without diagnosed COPD, before subsequent exclusions were made on the basis of age and missing demographic data. With a total of 682,396 individuals in the primary care registry, this equates to an estimated prevalence of 4.6% over the entire study duration. The demographics of the study population were compared to British studies in similar populations in the literature, and the (adult) age <sup>374</sup>, sex <sup>374–376</sup>, and obesity <sup>375,376</sup> distributions were comparable. Unlike some previous studies <sup>152</sup>, we found a fairly even distribution of patients amongst the socioeconomic deprivation scale, however we observed higher

rates of deprivation amongst those with attacks during the study. The most substantial difference compared to previous literature was baseline smoking status, in which a higher proportion of patients in our study were classed as non-smokers (never or former: 94%) in their baseline period than in previous <sup>374,375,377</sup> studies (79-85%). This was, however, solely due to how the baseline period was calculated for Table 7.3: the most recent smoking status before their first birthday within the study period, which will therefore be less than a year. When this period was extended to the most recent before their *second* birthday, the distribution was well aligned with the literature.

In Chapter 3, I provided a comprehensive review of the literature on asthma attack risk factors. The main strength of this chapter was my evaluation of the feasibility of extraction from EHRs, presentation of validated code lists, and critiques of algorithms and derivation methods employed in previous studies. As part of this, I conducted the most in-depth review of adherence measures in EHRs to date (presented in Chapter 4). As well as providing detailed guidance on the cleaning and processing of prescription records, I provided comparisons of several commonly used adherence measures, across multiple time-scales. This enables researchers to translate and replicate my methods for other datasets and medical conditions, and to make informed decisions about the most appropriate measure for their analysis.

Similarly, in Chapter 6 I conducted a thorough exploration of binary classification model performance measures in different settings with imbalanced data. As well as the breadth of performance measures covered in this analysis, a major strength was the resulting discussion and guidance on suggested use-cases, an analysis which informs problems across any binary classification problem in diverse applications. I was able to construct a tree diagram that a researcher could use to find the recommended performance measure for their analysis. In my original literature review, in which I sought guidance for the most appropriate way to handle the inevitable low incidence in my risk prediction model, I found that many studies provided findings from various experiments but did not provide any practical guidance <sup>353,355-</sup>

<sup>357</sup>.

Reporting guidelines relating to clinical risk prediction modelling and studies using EHRs were reviewed<sup>53,368–370</sup>, and relevant items on the checklists were combined in order to construct a thorough list of items to dictate how my analysis was conducted and my results were presented.

In Chapter 7, I constructed a list of composite benchmarks of model performance, based on a review of previously reported risk prediction models, which included a wide variety of model performance measurements. This multi-factor benchmark sets a threshold which any model exceeding can be considered state-of-the-art. My final model achieved four of the six benchmarking criteria, narrowly missing out on the final two criteria: 50% sensitivity and PPV (achieving 47.7% and 48.9%, respectively). As such, by my own benchmark, I cannot definitely class my 4-week event horizon model as the best-performing model ever. Despite this, I would still argue that it is a well-rounded model with good performance, and with a more clinically useful event horizon than many of the comparator models, listed in Table 7.1. Additionally, my model used a more heterogeneous population than the study by Zhang *et al.*<sup>346</sup> (diagnosed and treated asthma, and a recent history of asthma attacks), which was the only identified study with an event horizon of less than 6 months (3 days). As such, one might argue that my model's performance was both more clinically useful and harder to achieve.

The secondary analysis models using longer event horizons achieving significantly better performance than either the four-week or one-week horizon models. For example, the model predicting asthma attacks in the next 12 weeks achieved all benchmarks. The most clinically useful event horizon has yet to be determined and there is evidently further work to be done to find the optimal combination of model and intervention.

In my analyses, I included two different measures of adherence as features. One measure was a rolling average of the ratio of days' supply obtained to the length (in days) of the prescription refill interval, for the last three prescriptions. The second was a proportion of (eligible) days in the previous calendar year for which medication was available. These two adherence measures were both found to be important in the risk



prediction models across different event horizons, but their relative importance changed based on the event horizon in question. As well as allowing us to capture slightly different dimensions of medication adherence, using two adherence measures with different retrospective calculation durations affords us some estimate of the recent change in adherence. For example if the CMA8\_2 (last year) was high but the CSA\_3 (last three refills) was poor, we might expect that adherence was on a downward trajectory. The value of temporal trends, rather than single feature snapshots, has been discussed by other researchers in the past and is still being investigated<sup>378</sup>.

While both the data and the specifics of the implementation employed herein are both UK specific, the central processes can be applied to EHRs in any country. Additionally, it brings to focus key areas which need to be considered when assembling the infrastructure of a national digital health system, and areas to be reinforced in existing systems, which are discussed in Section 8.5.

### 8.3 Limitations

In this section, I will review the limitations of my thesis in the order of the research-based chapters.

First, the literature review conducted in Chapter 3 to identify candidate risk factors for asthma attacks was not a systematic review. By this, I mean that I did not construct a list of *composite search terms* and manually review every paper which matched these criteria, from an online academic repository such as Scopus or a similar established academic database. Composite search terms may use logical expressions such as A AND B, C OR D, or asterisk-denoted *wildcard* characters to allow for variations in words. For example, *asthma\** would include any word that began with *asthma*, including the word itself, *asthmatic*, and more.

The limitations of not conducting a systematic review are that it is possible that some important study might have been overlooked, which might have drastically improved

the performance of the model. For example, there might have been a less-commonly reported risk factor that I missed, a novel method of extracting a risk factor from the available data, or some important study on the effect of that risk factor which might have indicated a superior method of reporting that feature for prediction purposes. On the other hand, there were several practical benefits to my approach over a full systematic review. First, the review could be conducted much faster. The search criteria and study retrieval alone of a well-conducted systematic review often takes in excess of six hundred hours<sup>379,380</sup>. To include a systematic review of even a single of my research questions, let alone all of them, would have been so time-intensive it would have left me little time for any further research. Secondly, not conducting a full systematic review meant that it was not necessary for me to comprehensively report on the negative findings of my review, for example factors for which there was no evidence of increased asthma attack risk, or extraction methods which were demonstrated to be inferior to other methods in benchmarking studies. Nevertheless, I did include notes in Section 3.8 on a few risk factors which were not included in my final analysis but which I decided warranted a more detailed explanation for the reasons of their exclusion. This particularly pertained to risk factors for which there was substantial evidence of their predictive value, but which were not feasible to include in my analysis. Instead of a systematic review, however, I conducted a critical, systemised review<sup>381</sup> of the risk factors which I had identified from the non-systematic literature review, including evaluating the utility and feasibility of extracting each factor from EHRs.

Chapter 4 did not comprise an exhaustive list of all previously defined adherence measures in EHRs. Indeed, the seminal paper by Steiner and Prochazka described an adherence measure which they named the CMA<sup>189</sup>, but which was subtly different to any of the CMA measures defined by Vollmer *et al.*<sup>190</sup>. It was calculated similarly to Vollmer *et al.*'s CMA1, but instead of being calculated over a fixed interval, such as one year, it was a time-series which continuously updated at every prescription refill. This inspired my inclusion of the rolling-average CSA measures, but suffer from the same limitation as the CSA\_10 (10-interval rolling average) that recent, later intervals with poor adherence do not influence the overall output as much as we may desire.

Particularly over longer durations, Steiner and Prochazka's CMA has very clear drawbacks. One other inclusion in their paper, however, which I also did not include was the counter-measure to the CMA, the CSG. Indeed, no multiple interval measures of gaps were explicitly included in my analyses. However, as the CMA5 and CMA8 both use the proportion of days with supply, their natural complement is the proportion of days without supply, or the days with gaps. In this way, these two measures are conceptually distinct to the CMA1, which measures the total supply (not capped at 100%) and thus it cannot capture information relating to gaps. The justification for not including the other CMA measures as defined by Vollmer *et al.*<sup>190</sup> was listed in Section 4.2.1, however overall I believe I have captured an appropriate variety of measures which encompass the fundamental differences in adherence measurement.

Some of the limitations of the methods by which I processed the prescription data in Chapter 4 have already been described in Section 4.4.3. Briefly, my approach to handling the free-text fields was somewhat basic and allowed erroneous exclusions, such as excluding a prescription with the dose directions 'use inhaler after nasal spray' on the basis of the identification of the keyword 'nasal' and 'spray'. As previously stated, without the integration of complex natural language processing this problem is extremely difficult to circumvent with confidence. Such methods were beyond the scope of this thesis. Alternatively, however, a full manual review could have been conducted of all excluded records and exclusions could have been made on a case-by-case basis. There were 5269 records which were excluded on the basis of the presence of one or more exclusion keyword, but which did not have the formulation listed as a spray or drop. While this is a lot of records to manually review, it was also only a small percentage of the remaining ICS or ICS+LABA records (0.3%) and thus the potential effect was minimal.

Another point mentioned in Section 4.4.3 (and before, in Section 2.2.3.3) was the imperfect data linkage between prescribing and dispensing records in Scottish EHRs (conducted by National Services Scotland Information Services Division). Multiple medications prescribed simultaneously share a prescription event identifier and have no unique prescription item identifier. As such, differences in the ordering of items on

a single prescription may have resulted in incorrect dispensing data being assigned to a prescription. This may have been the cause of the 5269 records with exclusion keywords in their dose descriptions (from the dispensary) but not exclusion formulations listed on their prescription record. In order to evaluate the potential magnitude of this flawed linkage, I conducted a manual review of a random sample of 1000 asthma prescriptions (from the pool of 4,965,714 prescriptions) and identified fewer than 1% which either contained internal contradictions (either named a different medication or described a method of ingestion inherent to a different formulation, such as 'inject') or had empty data fields from the dispensing information. Overall, the integrity of the adherence measures, and indeed the other prescription-based model features (reliever inhaler use and BTS treatment step) is likely to have been compromised, however rare the occurrence.

Other points of concern are the high proportion of prescriptions for which the unit volume (doses per unit, such as inhaler) could not be extracted (84.8%) and were imputed. The unit volume was imputed based on the medication name and medication strength, the latter of which was also required to be imputed for the 8.4% of records without an extractable value. While review of the imputations demonstrated that there was a clearly prominent unit volume (over 80%) for 42/60 medication type (including brandname) and strength combinations, for 17.4% of imputed records the most prominent unit volume represented less than 60% of the reference prescriptions. Consequently, the imputation may not have been very precise. The most common medication for which confidence in the imputed dose was low (less than 60%) was both high dose and a combination ICS+LABA medication (Seretide 250mcg), and thus affected individuals were likely on a higher BTS step, there is a chance that adherence was systematically underestimated in the higher BTS steps.

Finally, although not pertinent to the results presented in Chapter 4, the review of primary asthma controller medications (ICS and combination ICS+LABA) had much more depth than the subsequent review of add-on therapies. The possible inclusion of add-on therapies which were being prescribed for indications other than asthma (and COPD, which was a participant exclusion criteria), such as theophylline for

apnoea or omalizumab for urticaria, may have biased the BTS step estimation to appear higher for those with higher numbers of comorbidities.

One limitation of Chapter 6 is that I did not include an exhaustive list of all performance measures. For example, I limited my focus on binary class classification settings, and hence have not investigated multi-class classification performance measures. Accuracy can naturally be applied to the multi-class setting, simply as the proportion of samples which were correctly classed. Others can easily be adapted, for example there is a multi-class adaptation of the MCC <sup>382</sup>. The class-specific performance measures such as the sensitivity/specificity and PPV/NPV easily generalise to the accuracy within any observed/predicted class, and they can also be averaged to compute a single summary performance measure. The *macro*-average of class-specific performance measures is the mean of the performance measure for each class, and thus gives equal weighting to the classes regardless of size imbalance. For example, the balanced accuracy (see Section 5.4.3.3) is the macro-average of the observed class accuracies: the sensitivity and specificity in the binary case. Other multi-class model performance measures include Cohen's Kappa and Multi-class Performance Score (MPS) <sup>331</sup>. Multi-class classification can be used in risk prediction to assign patients into ordinal risk categories (such as low, medium, and high risk). While interesting in the evaluation of model calibration, although less informative than simply using the estimated probabilities of the event themselves, asthma attacks are themselves a binary event and thus binary classification is the only way that discrimination can be evaluated.

In terms of the primary analyses presented in Chapter 7, one limitation is that short-term event prediction requires data to have been recorded in the recent past, and is thus reliant on frequent primary care consultations to accurately detect those at elevated risk. The model will inevitably provide less accurate predictions for those with infrequent primary care contact, or who have only recently joined a practice. One solution for this, however, would be to allow manual data entry to supplement, or overrule, the information that had been extracted automatically. Further discussion regarding the possible deployment of this model is presented in Section 8.4.

In section 7.3.1, I highlighted a study by Jentzsch *et al.* <sup>264</sup>, which found that, in children, adherence estimates derived from pharmacy dispensing data were often overestimated. One hypothetical reason for this, and the concern that lead me to focus my analysis on the adult population, was that the parent is typically coordinating medication refills, regardless of the child's medication taking. A limitation of this thesis is that adherence rates between children and adults were not compared in Chapter 4, and indeed children were excluded from the analyses in Chapter 7. Additionally, however, there is reason to hypothesise that differences might have been observed between the adult and elderly population (who might also have their refills coordinated by a third party, such as a carer) or for those with a high number of comorbidities (for whom asthma medications may well be refilled at the same time as other medications) <sup>383</sup>. These examples highlight a broader point, that further assessment of the calibration with regards to all risk factors is important in order to identify weaknesses in the model.

The criteria used to estimate whether steroid prescriptions were related to asthma symptoms are described in detail in Section 7.3.2, and while an obvious limitation is that this process might not have been perfectly accurate, a perhaps more interesting dilemma is that the steroids might not have been prescribed for immediate use <sup>376</sup>. If they were instead prescribed to be taken if needed, then the data sample would have been falsely labelled as an attack, and a potential future point in time might have been incorrectly labelled as 'not an attack'. Furthermore, the high prevalence of rheumatological diseases in people that were identified as having an asthma attack during the study (Table 7.3) may indicate that the identification of steroid prescriptions as asthma-related was imperfect, as steroids are commonly used to treat rheumatological flare-ups as well.

As discussed in Sections 3.4.4 and 3.8, three notable features which were not included in my risk prediction model were race/ethnicity, influenza vaccination, and synthetic hormone treatment (hormonal contraceptive or hormone replacement therapy). Race/ethnicity could not be linked from census data due to study ethics approval

limitations, and was not available in the patient registry data. Read Codes relating to hormonal therapies were not available for the ALHS study, and the identification of relevant medications from prescribing records was beyond the scope of this body of work. Additionally, hormonal contraceptives are often prescribed in secondary care and thus may not be reliably recorded in primary care data. Finally, influenza vaccination was not included as there was no supporting evidence for any mechanism of effect on asthma attack incidence aside from reduced risk of respiratory infections: a feature which had already been included in the analysis.

As discussed in Section 3.7.2, rhinitis was included as a predictor in the risk model as a marker of atopy. The Read Codes, listed in Appendix G, include “H18..” for *vasomotor* (non-allergic) rhinitis. Neither Price *et al.*<sup>117</sup> nor Blakey *et al.*<sup>75</sup> used any differentiation between allergic and vasomotor rhinitis in their risk prediction models. Luo *et al.*<sup>116</sup> used only allergic rhinitis ICD codes, and Engelkes *et al.*<sup>148</sup> also reported that they used allergic rhinitis specifically, although there was not sufficient detail in their methods to see how they specified this. While this demonstrates the evidence for the importance of both allergic and non-allergic rhinitis, the conflation of the two into a single risk factor may negatively impact the predictive ability of the model, especially in the tree-based models where compound effects with other markers of atopy may be attenuated. Further work is required to estimate whether the distinct mechanisms of the two types of rhinitis (atopic versus sinusal pathologies<sup>384</sup>) have an effect on the risk of asthma attacks, and should thus be considered separate risk factors. Additionally, the validity distinguishing between vasomotor and allergic rhinitis on the basis of Read Codes would need to be investigated.

There is some evidence that the sensitivity of anxiety and depression diagnostic Read coding was negatively affected by the Quality and Outcomes Framework (QOF). QOF is a points-based system, introduced in 2004, which financially rewards and remunerates GP surgeries for conducting certain practices envisioned to improve the quality of care provided. The first depression QOF points were initiated in 2006, for evaluating depression severity using specified symptom questionnaires up to 28 days after diagnosis with any depressive disorder<sup>385</sup>. A further criterion was added in 2009

awarding additional points for follow-up severity assessments 5-12 weeks after the initial diagnosis <sup>386</sup>. An unintended consequence of the QOF was that GPs became less likely to use diagnostic Read Codes (thus circumventing the QOF requirements) and instead use symptom-based coding <sup>387</sup> or prescribe antidepressants without any coding <sup>388,389</sup>. Doing so removed such patients from the denominator of the proportions with severity assessments, and thus enabled them to circumvent the time-consuming questionnaires without any loss to their overall percentage, and thus aid their awarding of the criteria points. Using prescriptions of anti-depressants in supplement to diagnostic coding may have improved the sensitivity of the feature, however only a limited selection of medications pertaining to the central nervous system were available in the ALHS dataset.

The final limitation to the methodology I will describe herein is that the range of algorithms that were investigated was relatively limited. I have already justified, in Sections 5.3.3 and 5.3.5, not including k-NNs or SVMs in my analyses. Briefly, k-NNs require the full training data to resolve query samples, and SVMs are very computationally intensive in large datasets, with complexity increasing exponentially with training sample size. One noteworthy algorithm that was not investigated is neural networks, described in detail in Hastie *et al.* (Book Chapter 11) <sup>290</sup>. Optimising the architecture of neural networks is considered an art by some and was outside of the scope of this thesis. However, their application to this problem may be considered in the future. Finally, there are other ensemble algorithms (see Section 5.3.6) that could be investigated, including bagging with base learners other than decision trees, and stacking.

Regarding the results themselves, in Section 7.4.6 the model was found to have poor calibration. The calibration slope (18.27) shows that the estimated probabilities were inappropriately scaled, with insufficient variation between the samples with low and high estimated risk. This was visualised in Figure 7.17, with significantly lower estimated risk than the observed rate for the higher-estimated risk samples. This could have been improved by optimising the model on calibration-based performance measures (Section 5.4.1) rather than the balanced accuracy, or by using recalibration



methods (Section 5.4.4). However, the focus of this thesis was on binary classification, and so these methods were not investigated. Primarily, this decision was made due to the known subjectivity of interpretation of probabilities<sup>390</sup>. Cognitive biases inherent to all humans may result in variation in test result interpretation and suboptimal management plans for patients<sup>391</sup>, which can contribute to increased health inequity.

## 8.4 Implementation

Despite the derivation of many asthma attack risk scores, none are currently endorsed for clinicians in guidelines by either the GINA (2020)<sup>392</sup> or BTS/SIGN (2019)<sup>135</sup>, and are not referenced at all in the 2020 National Institute of Clinical Excellence (NICE) guidelines<sup>393</sup>.

*“..., and the clinical usefulness of these, and other, classification and asthma prediction systems remain a subject of active investigation.”*

GINA Guidelines, 2020<sup>392</sup>

*“Clinical prediction models for quantifying risk need to be developed and prospectively validated in adults, children aged 5–12 and children under five years of age. Does risk assessment based on these factors improve outcomes when used prospectively in routine clinical practice?”*

BTS/SIGN Guidelines, 2019<sup>135</sup>

With so much effort and research into developing asthma risk prediction models and risk stratification tools, it is disheartening to see the minimal impact they have had on routine practice<sup>52</sup>. While insufficient predictive power of past models is undoubtedly a factor in this, it is not the only reason why the implementation of this work has hitherto been unsuccessful. The same phenomenon is seen in many medical fields, as discussed by Dekker *et al.* in their conversation piece entitled “Most clinical risk scores are useless”<sup>394</sup>. The authors argue that the most common causes are either too many (relative to the sample size) or too few predictor variables, poor reporting, and poor

methodology. Crucially, the authors note that “*the development of these scores is based on an underlying assumption that accurately predicted estimated probabilities improve a clinician’s decision-making or the patient’s quality of life.*” Similarly, Damen *et al.*<sup>395</sup> stated that “*most developed prediction models are insufficiently reported to allow external validation by others, let alone to become implemented in clinical guidelines or being used in practice*”. Having now developed a risk prediction model which is practical to implement in primary care, and with sufficiently good performance to justify it, further work is needed to evaluate the best way to implement such a model, and to measure the impact on clinical outcomes.

Typically, risk prediction models are implemented as Clinical Decision Support Systems (CDSSs), in which clinicians are provided with data-driven recommendations and statistics related to a patient (such as to inform their treatment recommendations) or process (such as prompting further investigations, or triaging)<sup>396</sup>. Data about an individual at the time of query is usually entered manually but may also be extracted automatically from the EHRs. As the cost of CDSS implementation and maintenance is not negligible, understanding what makes a system successful and cost-effectiveness is of great value. Perhaps the most comprehensive quantitative study was conducted by Roshanov *et al.* in 2013<sup>397</sup>; their meta-regression on 162 clinical trials of CDSSs evaluated factors associated with *successful* CDSS implementation: a significant change in provider activity or patient outcomes. For example, while multiple previous studies had reported that clinicians were often found to side with their own judgement when there was a substantial difference in the advice provided<sup>96,398</sup>, Roshanov *et al.* found that requiring a reason entered for over-riding the systems advice reduced this (unadjusted OR = 8.92, 95% CI = 2.01 to 39.61). They also found that presenting the output of the model as both advice to the practitioner and to the patient improved success rates (OR = 2.99, 95% CI = 1.20 to 7.42). Neither whether the advice was presented automatically in the workflow, nor whether the practitioner was required to enter the data manually, were associated with either success or failure, however advice being integrated into the electronic charting or order entry form was associated with reduced odds of success (OR = 0.53, 95% CI = 0.28 to 1.02).

Overall, Roshanov *et al.* found that 52-64% of the trials in their meta-regression showed significant improvements in processes of care, but only 15-31% of those evaluated for impact on patients' health showed positive impact<sup>397</sup>. Well implemented risk models have been used to triage chronic kidney disease patients, reducing wait time for high risk patients<sup>399</sup>, to reduce osteoporosis prescribing without any increase in fractures<sup>400</sup>, and to reduce inappropriate prescribing of antibiotics<sup>401</sup>. There have been modest successes in asthma management too: A 2012 RCT (known as ARRISA) also managed to demonstrate a slight reduction in asthma hospitalisations after introducing a very primitive (and unvalidated) high-risk patient flag, based on asthma treatment, attack history, and subjective psychosocial factors<sup>402</sup>. The number of asthma attacks did not decrease in the intervention practices, it was in actuality higher, however a lower number of the attacks were treated in secondary or out-of-hours care. As well as decreasing financial burden on the healthcare system, this is also less traumatic to the patient. The study team are currently conducting another RCT of a similar design, using a validated risk score in place of the previously used method, although the results have not yet been published<sup>403</sup>.

Not all CDSSs are based on data-driven risk prediction models, some are based on simple knowledge-based rulesets. The aforementioned original ARRISA patient flag, for example, was built on four simple criteria: age, asthma diagnosis, asthma severity, and evidence of some psychosocial problems. While such a crude model may not have the same predictive power as an algorithm with more features, interactions (or if-then statements), and weighting of features, they have the benefit of being incredibly transparent. The same cannot be said for all data-driven machine learning models, which can be incredibly difficult to interpret more generally than on the basis of individual predictions. In practice, many such models are what we call *black box models*: in which the input and output are recorded without any comprehensive understanding of the internal mechanism. For example, a simple decision tree like the one presented in Figure 5.2 is easy enough to follow, however if the depth were increased from two steps up to even 10 steps, the tree becomes much more complicated. Furthermore, understanding of a random forest of 500 or more distinct, deep decision trees is inevitably obfuscated.

Understanding how much individual features contribute to the model (through parametric model coefficients, or feature importance as discussed in Section 5.6.1) improves the interpretation of the model *globally*, but on a case-by-case basis understanding the model's *local* (such as for a single patient at a consultation) predictions may be more important. Methods such as Shapley values (Section 5.6.2) can be used to estimate the features which are most influencing the model's prediction for a specific patient, and by focussing on the modifiable risk factors this could make a powerful education tool.

Luo *et al.* conducted a secondary analysis of their asthma attack risk prediction model<sup>116</sup>, presented in Section 7.1, to evaluate a previously developed methodology<sup>404</sup> in this setting<sup>405</sup> (including an extension of the methodology for imbalanced data). In their methodology, the interpreter is used to generate a list of all possible decision rules (if *criteria* then *outcome*; generated using classification rule mining<sup>406</sup>) which are pruned according to certain global criteria (including use of *important* features, high confidence in rule, clinical guidance, interpretability, and generalisability). The set of rules matching both the predicted outcome (from the primary model) and the query sample (patient) characteristics are then presented. They were able to provide some explanation using this approach for 92% of the true positives in their previous model<sup>116</sup>, and 87% in an external dataset<sup>407</sup>. There was no way, however, to evaluate the interpreter fidelity, or indeed how well the presented rules aligned with the true drivers of the primary prediction model.

Black box prediction models are not likely to be a big concern to a user in practice when the model's prediction is well aligned with the user's instincts, or when the model is well known to outperform clinicians consistently. However, the hypothetical (and possibly real) concern about occasions where the model's predictions are very different to the clinician's view is more pertinent. In these cases, it is understandable that someone would be reluctant to put their patient's health (and indeed their legal liability) in the hands of a machine without clear and easily understandable rationale<sup>408</sup>. After all, the model cannot incorporate all knowledge that a patient-facing clinician

can detect. No prediction model or interface is perfect either, and inaccurate results and system malfunctions both decrease trust with users. A recent study of users of a CDSS at Brigham and Women's Hospital (US), used for drug interaction, allergy, test result and screening reminder alerts, found that two-thirds of those surveyed had experienced malfunctions at least annually <sup>409</sup>.

It is worth noting that while any risk prediction model is limited by the features provided in the training data, the additional information about a patient that can be observed from a face-to-face consultation (such as the severity of current seasonal allergies) does not necessarily translate to a better risk appraisal from a human than a machine. There are circumstances, for example, in which the user's instincts may be explicitly detrimental to the patient. Cognitive biases inherent to all humans may result in diagnostic inaccuracies and suboptimal treatment plans for patients <sup>391</sup>. This can contribute to increased health inequity. Theoretically, a CDSS could avert this problem by removing the personal judgement component. On the other hand, the system still needs data to inform the computer model which will not introduce bias of its own. The majority of routinely collected data are inherently populated by those with regular access to healthcare <sup>410</sup>, rather by those from vulnerable populations who may experience rates of healthcare events differently – such as the formerly incarcerated, refugee, or homeless populations. While the ALHS dataset is highly representative of the Scottish population it is inevitable that minority groups exist for which the prediction is substantially less reliable.

The final question is how the model is best presented to the user. As discussed in Section 8.3, there are limitations to presentation of risk prediction models as continuous values, which is why this thesis has focussed on binary classification. There are still more nuanced approaches to presenting this prediction than a simple 'yes' or 'no', however, such as presenting the uncertainty using the number needed to treat in order to prevent an attack. There also could be recommendations made based on the estimated risk, without presenting the risk itself. This would also be a way to discourage unnecessary treatment step-ups (which would inadvertently change the estimated risk, as treatment severity is a risk factor in the model) for people with

mild daily symptoms but prominent triggers, and to focus more on health education and monitoring.

Things to consider include whether patient data should be manually inserted (not pragmatic for a model of 45 features, but better for a simplified model using only the most important features, for example) or automatically extracted from EHRs (which will likely result in some extraction errors or missing data). Furthermore, there are potential unintended effects of manual data entry. The missing data mechanism primarily employed in this thesis is known as missing data indication: creating a category to flag that the value was missing (discussed in Section 3.2). The primary strength of this approach is that the fact that a feature was not noted, measured, or recorded, is acknowledged by the model. For example, a lack of peak flow measurement is likely to indicate that no clear reduction in lung function was evident. An important consideration for this approach, however, was noted in a recent paper by van Smeden *et al.*: if the clinician is required to manually input data into the model, the request for data which might otherwise be unrecorded might alter the meaning of the data capture (or lack thereof) and negatively influence model performance<sup>134</sup>. A similar effect might be observed with changes to clinical practice guidelines.

It might be preferred that model predictions were only presented when requested, or automatically during an asthma review, to prevent 'alert fatigue'. The model might be presented via an online webform, a mobile application, or an application integrated into the primary care EHR system. Some clinicians might prefer a binary suggestion, whereas others might prefer to see the estimated risk probability itself. The trajectory of expected change in risk caused by time-varying features (such as age and month) could be visualised<sup>411</sup>. Finally, the user would likely wish to see either some explanation of the results, using a points based score<sup>412</sup>, Shapley values (perhaps limited to the modifiable risk factors), the relevant path through a single-tree approximation, or a selection of rule-based statements like Luo *et al.*<sup>405</sup>, for example. Cai *et al.*'s recent survey of pathologist queries about a CDSS for prostate cancer diagnosis found that the most common theme was the capabilities and limitations of the system, including known population subgroups with lower accuracy and the

diversity of the training data <sup>413</sup>. Primarily, they wanted to know context that would help them decide whether to trust their own judgement over the model's estimates in the cases where they differed. Carroll *et al.* reported that many of the clinicians who tested their cardiovascular risk prediction model program particularly liked to be able to use the interface as a demonstration tool with patients, and so appreciated clear graphics and the option for a print-out to be generated <sup>411</sup>. In short, further consultation with patients and practitioners is essential in order to maximise the impact of any prediction model in clinical practice.

## 8.5 Future Work

This body of work highlights the value of enabling reliable and routine linkage between health data sources. While primary and secondary care records are simple enough to link for research purposes in the UK, using NHS or CHI personal identification numbers, linkage of prescribing and dispensing records, in which records must be linked not just by person but by event, is not so trivial. The system in place in Scotland facilitates basic linkage, although, as discussed in Section 2.2.3.3, it is not without limitations. In many countries, however, even this crude linkage is not conducted routinely. For example, in England, prescribing and dispensing of medications are recorded by separate processes. Since 2015, NHS Business Services Authority (NHSBSA) dispensing data have included a patient identifier (NHS number) <sup>414</sup>; this is, however, not routinely linked to primary care prescribing records held by Public Health England (PHE). The NHSBSA and PHE records also do not have a common unique prescribing event identifier. Therefore, even with a data sharing agreement in place, matching records (one-to-one) using common identifiers (known as *deterministic linkage*) is currently impossible. Therefore, it is necessary to link records *probabilistically*; estimating the likelihood that two records will match given the data they contain. As mentioned in Section 4.4.3, and described in Appendix J, I collaborated with the pharmaceutical company GlaxoSmithKline (GSK) and the Salford Lung Study (SLS) team) to design an algorithm for probabilistic linkage of prescribing and dispensing records for asthma controller medications <sup>265</sup>.

The high prevalence of primary non-adherence in asthma (reported incidence between 12-45%<sup>259,261,262,265,415,416</sup>) is an informative example of the utility of having linked prescribing and dispensing records. While further external validation of our promising findings<sup>265</sup> is needed on additional datasets, this work has gone a long way to improve the interoperability of my adherence research outside of Scotland. Making this data linkage routine is not only useful for research purposes but could also be integrated into primary care reporting systems to provide a preliminary assessment of whether patients are collecting their prescribed asthma medications, improve clinicians' understanding of patient adherence, and enable open discussions about barriers to adherence.

Recently, there has been increasing interest in the integration of patient data (including data from home monitoring devices, and PROMs) into EHRs<sup>417-419</sup>. Enabling patients to contribute data to their health records allows health professionals to see recent historical trends in time-varying data, such as symptoms and lung function<sup>420</sup>. It can also save time in consultations by encouraging patients to complete surveys and questionnaires in their own time. Patients being able to view their own EHRs might also enable errors to be spotted which would otherwise impact the quality of their care<sup>421</sup>. In medical research, this could improve the precision of time-varying features which are currently infrequently captured in primary care, and enable new features to be included in risk prediction models. When integrated EHRs are implemented, it will be a fantastic opportunity to examine how this richer data can improve prediction models.

In terms of more technical recommendations, Section 5.5 omits more recent developments to the field of training data enrichment than SMOTEing. For example, ADASYN (ADaptive SYNthetic sampling) is another synthetic data generation algorithm, which adds weight to the minority samples estimated to be the hardest to predict. It does so by generating synthetic samples which are equidistant (according to Euclidean distance in M-dimensional feature space) between a minority sample and another minority sample in its (K-nearest minority class) neighbourhood. The number of synthetic samples per minority sample is calculated by a function of the normalised



ratio of classes of their neighbours and the desired enriched data class (im)balance<sup>422</sup>. This algorithm seeks to overcome the biggest limitation of the SMOTE algorithm, which is that it does not take into account the nearby neighbours of a minority class sample which are in the majority class. This and further methods should be investigated to see if they are able to provide improvements on the performance achieved herein.

In Section 7.3.4, I described how the six continuous features were scaled using the min-max normalisation method. All the algorithms used herein are independent of the feature scale, however, so the (linear) scaling itself will have no effect on the models built on unenriched data. Algorithms which are not independent of feature scale are again those which are distance-based, such as k-NN (Section 5.3.3) and SVMs (Section 5.3.5). Additionally, algorithms which use *regularisation* (penalising coefficients for the purpose of feature selection) require all features to be scaled<sup>290</sup>. For more information, refer to Hastie *et al.* (Book section 3.4)<sup>290</sup>. Alternative feature scaling methods, such as Z-score normalisation (also known as standardisation), may be investigated in the future, if different algorithms or enrichment methods were used.

Another approach to the prediction of rare events is one-class classification (OCC; also known as *anomaly detection*<sup>423</sup>), in which models are trained on data sharing the primary characteristic (for example, emails which are not spam) and try to identify outliers which do not belong in that class. Typically, these OCC approaches attempt to find some minimum area in multi-dimensional space which encapsulate all the training samples, and then identify outliers simply as any query sample lying outside of this area<sup>424</sup>. OCC can be applied to imbalanced data prediction in the typically binary case and is best suited to cases in which the minor class lacks consistent characteristics, making it difficult for many binary classification to establish reliable decision rules, and often resulting in poor discrimination<sup>425</sup>. For example, asthma attacks resulting in death and those resulting in GP-prescribed oral steroids may be distinctive and trying to classify them both under the general class of ‘asthma attacks’ rather than anything other than stable asthma (one-class) leads to poor performance. Performance in the test data set can be evaluated in the same way as in a binary-

classification problem. The inclusion of this analytical approach was beyond the scope of this thesis, and while there were no identified studies in this area in the literature, one protocol paper which included OCC in its prospective methods was published in 2020 <sup>426</sup>. I eagerly anticipate the result and hope that it might provide further insights into areas of future investigation.

Five of the six most important features were derived from prescribing records (CSA\_3, reliever use, CMA8\_2, number of controllers in previous year, and treatment step). As such, it is likely that inaccurate data extraction in some individuals has contributed towards incorrect model predictions. Neither pharmacy nor primary care records are written with future linkage in mind, and as such they often require substantial pre-processing. Missing data are a common problem; in the aforementioned collaboration with GSK and SLS, we found that for the 17% of dispensing records for which a match could not be identified, both medication strength and dispensed quantity had approximately 60% missing data <sup>265</sup>. Furthermore, both primary and secondary care records often contain data in free text entry fields: areas that allow manual entry of information without finite options. These free text fields can contain information which is vital to correctly understand the record <sup>427,428</sup>. As an example, a 2012 study into health records in the UK General Practice Research Database found that cause of death was written in the free-text alone (and not in any coded or structured cells) in almost 20% of mortality records <sup>101</sup>. Yang *et al.* categorised quality-related events in free-text prescription fields, and found that the most common problem was missing dose quantity (e.g. 'two puffs'; 54% of quality flagged records) <sup>429</sup>.

Extracting free-text information requires intensive processing to ensure validity, often using methods known as Natural Language Processing (NLP) <sup>430,431</sup>. In contrast to simple rule-based methods, like searching for the word 'asthma' in inpatient admission records, NLP uses machine learning algorithms to distinguish nuanced segments, such as understanding that 'not asthma' is semantically different from 'asthma'. Although there have been many recent advances in free-text prescribing data extraction <sup>270,432,433</sup>, there is still the requirement for more research into the integration of specific medical domain knowledge <sup>434,435</sup>. The algorithm developed by McTaggart

*et al.*'s <sup>270</sup> is applied to all Scottish prescribing data, however as discussed in Section 4.2.3, my methods (which included asthma-specific domain knowledge) resulted in lower missingness, with above 99.6% agreement on non-missing extracted features.

Given the high rate of imputation of the doses per prescribed medication unit, and the importance of this information in adherence estimation, refining this process is empirical to improving the reliability of the estimates. The observed low confidence in the number of doses per unit for the 250mcg Seretide prescriptions can be used as a starting point for identifying areas for improvement. For example, further work could seek to identify distinguishing factors between the prescriptions of the 120-dose and 60-dose units in order to improve imputation across all prescriptions.

Finally, a very important future area for extending the work described in this thesis is extensive validation in the real-world setting. This validation is essential to confirm the generalisability of the model's performance to other datasets. Additionally, it is necessary to ensure that the model is suitable for deployment in the primary care setting, both in terms of user-interaction (as discussed in Section 8.4) and in impact on clinical outcomes. Given the exemplary performance at the 12-week event horizon, this might be the endpoint with the most potential to pursue further. Several models may be appraised for impact when paired with different interventions, in order to find the optimal implementation.

## 8.6 Conclusion

In conclusion, I have demonstrated that it is possible to predict asthma attacks in primary care with sufficiently high discrimination to guide clinical decision making, prompting further reviews, and initiating preventative treatments. Furthermore, I have demonstrated the importance of modifiable risk factors, including medication adherence and overuse of reliever medication. Crucially, the basis in electronic health records results in a prediction model which is feasible and clinically useful to implement in primary care, due to the use of routinely collected data, and near-future predictions.

The key to this achievement was a robust understanding of how adherence to medications can be estimated from electronic health records and a thorough extraction of risk factors from primary care data. Furthermore, I have built the tools and knowledge base to allow other researchers to more robustly build on my work – such as data linkage algorithms, and detailed investigations into prescribing records, adherence patterns, and measurement trends.

While the primary final model (prediction of ATS/ERS defined asthma attacks in the next four weeks) did not meet all components of my comprehensive, composite model benchmarking criteria, it succeeded in providing an implementable and well tested decision aid with good, balanced performance across a wide selection of quantitative measures. Furthermore, several of the secondary models, including predicting asthma attacks in the following 12 weeks, achieved state of the art performance and still had high potential clinical utility.

## References

1. World Health Organization. WHO | Asthma: Definition. *WHO* (2020).
2. American Academy of Allergy Asthma & Immunology. Asthma Defined. (2020). Available at: <https://www.aaaai.org/conditions-and-treatments/conditions-dictionary/asthma>. (Accessed: 25th November 2020)
3. Bush, A. & Griffiths, C. Improving treatment of asthma attacks in children. *BMJ* **359**, j5763 (2017).
4. World Health Organization. *Asthma Fact Sheet (2017)*. *World Health Organisation Fact Sheets* (World Health Organization, 2017).
5. Braman, S. S. The burden of asthma. *Chest* **130**, 4s-12s (2006).
6. Masoli, M., Fabian, D., Holt, S. & Beasley, R. The global burden of asthma: executive summary of the GINA Dissemination Committee Report. *Allergy* **59**, 469–478 (2004).
7. To, T. *et al.* Global asthma prevalence in adults: findings from the cross-sectional world health survey. *BMC Public Health* **12**, 204 (2012).
8. Global Asthma Network. *The Global Asthma Report 2018*. (2018). doi:10.1109/ICIP.2009.5414240
9. Stern, J., Pier, J. & Litonjua, A. A. Asthma epidemiology and risk factors. *Semin. Immunopathol.* **42**, 5–15 (2020).
10. Tran, P. & Tran, L. Comparisons between 2015 US asthma prevalence and two measures of asthma burden by racial/ ethnic group. *J. Asthma* **57**, 217–227 (2020).
11. World Health Organization. *Noncommunicable diseases Fact Sheet (2017)*. *World Health Organization Fact Sheets* (World Health Organization, 2017).
12. Mukherjee, M. *et al.* The epidemiology, healthcare and societal burden and costs of asthma in the UK and its member nations: analyses of standalone and linked national databases. *BMC Med.* **14**, 113 (2016).
13. Barnes, P. J. Inhaled Corticosteroids. *Pharmaceuticals* **3**, 514–540 (2010).
14. Barnes, P. J. Efficacy of inhaled corticosteroids in asthma. *J. Allergy Clin. Immunol.* **102**, 531–538 (1998).

15. Peters, S. P., Ferguson, G., Deniz, Y. & Reisner, C. Uncontrolled asthma: A review of the prevalence, disease burden and options for treatment. *Respir. Med.* **100**, 1139–1151 (2006).
16. Kim, S. Y. *et al.* Incidence and Risk Factors of Steroid-induced Diabetes in Patients with Respiratory Disease. *J Korean Med Sci* **26**, 264–267 (2011).
17. Suissa, S., Kezouh, A. & Ernst, P. Inhaled Corticosteroids and the Risks of Diabetes Onset and Progression. *Am. J. Med.* **123**, 1001–1006 (2010).
18. Blackburn, D., Hux, J. & Mamdani, M. Quantification of the risk of corticosteroid-induced diabetes mellitus among the elderly. *J. Gen. Intern. Med.* **17**, 717–720 (2002).
19. Price, D. B. *et al.* Adverse outcomes from initiation of systemic corticosteroids for asthma: long-term observational study. *J. Asthma Allergy* **11**, 193–204 (2018).
20. Adinoff, A. D. & Hollister, J. R. Steroid-Induced Fractures and Bone Loss in Patients with Asthma. *N. Engl. J. Med.* **309**, 265–268 (1983).
21. Van Staa, T. P., Leufkens, H. G. M., Abenhaim, L., Zhang, B. & Cooper, C. Use of oral corticosteroids and risk of fractures. *J Bone Min. Res* **15**, 993–1000 (2000).
22. Bloechliger, M. *et al.* Adverse events profile of oral corticosteroids among asthma patients in the UK: cohort study with a nested case-control analysis. *Respir. Res.* **19**, 75 (2018).
23. Dawson, K. L. & Carter, E. R. A steroid-induced acute psychosis in a child with asthma. *Pediatr. Pulmonol.* **26**, 362–364 (1998).
24. Kayani, S. & Shannon, D. C. Adverse behavioral effects of treatment for acute exacerbation of asthma in children: A comparison of two doses of oral steroids. *Chest* **122**, 624–628 (2002).
25. Sherwood Brown, E., Khan, D. A. & Nejtek, V. A. The psychiatric side effects of corticosteroids. *Ann. Allergy, Asthma Immunol.* **83**, 495–504 (1999).
26. Hui, R. W. H. Inhaled corticosteroid-phobia and childhood asthma: Current understanding and management implications. *Paediatr. Respir. Rev.* **33**, 62–66 (2020).
27. Boulet, L. P. Perception of the role and potential side effects of inhaled

- corticosteroids among asthmatic patients. *Chest* **113**, 587–592 (1998).
28. Zieck, S. E. *et al.* Asthma, bones and corticosteroids: Are inhaled corticosteroids associated with fractures in children with asthma? *J. Paediatr. Child Health* **53**, 771–777 (2017).
  29. Kouro, T. & Takatsu, K. IL-5-and eosinophil-mediated inflammation: from discovery to therapy. *Int. Immunol.* **21**, 1303–1309 (2009).
  30. Leckie, M. J. *et al.* Effects of an interleukin-5 blocking monoclonal antibody on eosinophils, airway hyper-responsiveness, and the late asthmatic response. *Lancet* **356**, 2144–2148 (2000).
  31. May, R. D. *et al.* Preclinical development of CAT-354, an IL-13 neutralizing antibody, for the treatment of severe uncontrolled asthma. *Br. J. Pharmacol.* **166**, 177–193 (2012).
  32. Kraft, M. Asthma Phenotypes and Interleukin-13: Moving Closer to Personalized Medicine. *N. Engl. J. Med.* **365**, 1141–1144 (2011).
  33. Gould, H. J. *et al.* The Biology of IgE and the Basis of Allergic Disease. *Annu. Rev. Immunol.* **21**, 579–628 (2003).
  34. Katsaounou, P. *et al.* Omalizumab as alternative to chronic use of oral corticosteroids in severe asthma. *Respir. Med.* **150**, 51–62 (2019).
  35. Global Initiative for Asthma. *Pocket Guide for Asthma Management and Prevention 2019.* (2019).
  36. Reddel, H. K. *et al.* GINA 2019: a fundamental change in asthma management: Treatment of asthma with short-acting bronchodilators alone is no longer recommended for adults and adolescents. *Eur. Respir. J.* **53**, 1901046 (2019).
  37. Asthma UK. *UK asthma death rates among worst in Europe.* (2017).
  38. Iacobucci, G. Asthma deaths rise 33% in past decade in England and Wales. *Br. Med. J.* **366**, l5108 (2019).
  39. Okpapi, A., Friend, A. J. & Turner, S. W. Asthma and other recurrent wheezing disorders in children (acute). *BMJ Clin. Evid.* **07**, 300 (2012).
  40. Rachelefsky, G. Treating Exacerbations of Asthma in Children: The Role of Systemic Corticosteroids. *Pediatrics* **112**, 382–397 (2003).
  41. British Thoracic Society, Research Unit of the Royal College of Physicians of

- London, King's Fund Centre & National Asthma Campaign. Guidelines For Management Of Asthma In Adults: I: Chronic Persistent Asthma. *Br. Med. J.* **301**, 651–653 (1990).
42. Rodrigo, G. Asthma in adults (acute). *BMJ Clin. Evid.* **04**, 1513 (2011).
  43. Martin, M. J., Beasley, R. & Harrison, T. W. Towards a personalised treatment approach for asthma attacks. *Thorax* **75**, 1119–1129 (2020).
  44. Rowe, B. H., Spooner, C., Ducharme, F., Bretzlaff, J. & Bota, G. Early emergency department treatment of acute asthma with systemic corticosteroids. *Cochrane Database Syst. Rev.* CD002178 (2001).  
doi:10.1002/14651858.cd002178
  45. Rowe, B. H., Spooner, C. H., Ducharme, F. M., Bretzlaff, J. A. & Bota, G. W. Corticosteroids for preventing relapse following acute exacerbations of asthma. *Cochrane Database Syst. Rev.* CD000195 (2007).  
doi:10.1002/14651858.CD000195.pub2
  46. Barr, R. G., Woodruff, P. G., Clark, S. & Camargo, C. A. Sudden-onset asthma exacerbations: Clinical features, response to therapy, and 2-week follow-up. *Eur. Respir. J.* **15**, 266–273 (2000).
  47. Woodruff, P. G., Emond, S. D., Singh, A. K. & Camargo, C. A. Sudden-onset severe acute asthma: Clinical features and response to therapy. *Acad. Emerg. Med.* **5**, 695–701 (1998).
  48. Turner, M. O. *et al.* Risk factors for near-fatal asthma: A case-control study in hospitalized patients with asthma. *Am. J. Respir. Crit. Care Med.* **157**, 1804–1809 (1998).
  49. Rodrigo, G. J. & Rodrigo, C. Rapid-onset asthma attack: A prospective cohort study about characteristics and response to emergency department treatment. *Chest* **118**, 1547–1552 (2000).
  50. Kolbe, J., Fergusson, W. & Garrett, J. Rapid onset asthma: A severe but uncommon manifestation. *Thorax* **53**, 241–247 (1998).
  51. Royal College of Physicians. *Why asthma still kills: The National Review of Asthma Deaths (NRAD)*. (2014).
  52. Triantafyllidis, A. K. & Tsanas, A. Applications of machine learning in real-life digital health interventions: Review of the literature. *J. Med. Internet Res.* **21**,



- e12286 (2019).
53. Luo, W. *et al.* Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J. Med. Internet Res.* **18**, e323 (2016).
  54. Loymans, R. J. B. *et al.* Exacerbations in Adults with Asthma: A Systematic Review and External Validation of Prediction Models. *J. Allergy Clin. Immunol. Pract.* **6**, 1942–1952 (2018).
  55. Bzdok, D., Altman, N. & Krzywinski, M. Statistics versus machine learning. *Nat. Methods* **15**, 233–234 (2018).
  56. Krumholz, H. M. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health Aff.* **33**, 1163–1170 (2014).
  57. Payne, T. H., Detmer, D. E., Wyatt, J. C. & Buchan, I. E. National-scale clinical information exchange in the United Kingdom: Lessons for the United States. *J. Am. Med. Informatics Assoc.* **18**, 91–98 (2011).
  58. Chaudhry, B. *et al.* Systematic Review: Impact of Health Information Technology on Quality, Efficiency, and Costs of Medical Care. *Ann. Intern. Med.* **144**, 742–752 (2006).
  59. DesRoches, C. M. *et al.* Electronic Health Records in Ambulatory Care — A National Survey of Physicians. *N. Engl. J. Med.* **359**, 50–60 (2008).
  60. Cebul, R. D., Love, T. E., Jain, A. K. & Hebert, C. J. Electronic Health Records and Quality of Diabetes Care. *N. Engl. J. Med.* **365**, 825–833 (2011).
  61. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *J. Am. Med. Informatics Assoc.* **20**, 117–121 (2013).
  62. ten Brinke, A. *et al.* Risk factors of frequent exacerbations in difficult-to-treat asthma. *Eur. Respir. J.* **26**, 812–818 (2005).
  63. Fernandes, A. G. O. *et al.* Risk factors for death in patients with severe asthma. *J Bras Pneumol* **40**, 364–372 (2014).
  64. Papi, A. *et al.* Relationship of Inhaled Corticosteroid Adherence to Asthma Exacerbations in Patients with Moderate-to-Severe Asthma. *J. Allergy Clin. Immunol. Pract.* **6**, 1989–1998 (2018).
  65. Eisner, M. D., Yegin, A. & Trzaskoma, B. Severity of asthma score predicts

- clinical outcomes in patients with moderate to severe persistent asthma. *Chest* **141**, 58–65 (2012).
66. Bateman, E. D. *et al.* Development and validation of a novel risk score for asthma exacerbations: The risk score for exacerbations. *J. Allergy Clin. Immunol.* **135**, 1457–1464 (2015).
  67. Loymans, R. J. B. *et al.* Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. *Thorax* **71**, 838–846 (2016).
  68. Turner, S. W., Murray, C., Thomas, M., Burden, A. & Price, D. B. Applying UK real-world primary care data to predict asthma attacks in 3776 well-characterised children: a retrospective cohort study. *npj Prim. Care Respir. Med.* **28**, 28 (2018).
  69. Van Vliet, D. *et al.* Prediction of asthma exacerbations in children by innovative exhaled inflammatory markers: Results of a longitudinal study. *PLoS One* **10**, e0119434 (2015).
  70. Robroeks, C. M. H. H. T. *et al.* Prediction of asthma exacerbations in children: Results of a one-year prospective study. *Clin. Exp. Allergy* **42**, 792–798 (2012).
  71. Haselkorn, T. *et al.* Recent asthma exacerbations predict future exacerbations in children with severe or difficult-to-treat asthma. *J. Allergy Clin. Immunol.* **124**, 921–927 (2009).
  72. Schatz, M., Cook, E. F., Joshua, A. & Petitti, D. Risk Factors for Asthma Hospitalizations in a Managed Care Organization: Development of a Clinical Prediction Rule. *Am. J. Manag. Care* **9**, 538–547 (2003).
  73. Lieu, T. A., Quesenberry, C. P., Sorel, M. E., Mendoza, G. R. & Leong, A. B. Computer-based models to identify high-risk children with asthma. *Am J Respir Crit Care Med* **157**, 1173–1180 (1998).
  74. Forno, E. *et al.* Risk factors and predictive clinical scores for asthma exacerbations in childhood. *Chest* **138**, 1156–1165 (2010).
  75. Blakey, J. D. *et al.* Identifying Risk of Future Asthma Attacks Using UK Medical Record Data: A Respiratory Effectiveness Group Initiative. *J. Allergy Clin. Immunol. Pract.* **5**, 1015–1024 (2017).

76. Williams, L. K. *et al.* Quantifying the proportion of severe asthma exacerbations attributable to inhaled corticosteroid non-adherence. *J. Allergy Clin. Immunol.* **128**, 1185–1191 (2011).
77. Grana, J., Preston, S., McDermott, P. D. & Hanchak, N. A. The Use of Administrative Data to Risk-Stratify Asthmatic Patients. *Am. J. Med. Qual.* **12**, 113–119 (1997).
78. Reddel, H. K. *et al.* An official American Thoracic Society/European Respiratory Society statement: Asthma control and exacerbations - Standardizing endpoints for clinical asthma trials and clinical practice. *Am. J. Respir. Crit. Care Med.* **180**, 59–99 (2009).
79. Buelo, A. *et al.* At-risk children with asthma (ARC): a systematic review. *Thorax* **73**, 813–824 (2018).
80. Altman, D. G. & Royston, P. What do we mean by validating a prognostic model? *Stat. Med.* **19**, 453–473 (2000).
81. Hazra, N. C., Rudisill, C. & Gulliford, M. C. Developing the role of electronic health records in economic evaluation. *Eur. J. Heal. Econ.* **20**, 1117–1121 (2019).
82. Schinasi, L. H., Auchincloss, A. H., Forrest, C. B. & Diez Roux, A. V. Using electronic health record data for environmental and place based population health research: a systematic review. *Ann. Epidemiol.* **28**, 493–502 (2018).
83. Casey, J. A., Schwartz, B. S., Stewart, W. F. & Adler, N. E. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu. Rev. Public Health* **37**, 61–81 (2016).
84. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. P. A. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Informatics Assoc.* **24**, 198–208 (2017).
85. Concato, J., Shah, N. & Horwitz, R. I. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. *N. Engl. J. Med.* **342**, 1887–1892 (2000).
86. Booth, C. M. & Tannock, I. F. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence.

- Br. J. Cancer* **110**, 551–555 (2014).
87. Pannucci, C. J. & Wilkins, E. G. Identifying and Avoiding Bias in Research. *Plast Reconstr Surg* **126**, 619–625 (2010).
  88. Kaptchuk, T. J. The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *J. Clin. Epidemiol.* **54**, 541–549 (2001).
  89. Kim, Y.-Y. *et al.* Level of Agreement and Factors Associated With Discrepancies Between Nationwide Medical History Questionnaires and Hospital Claims Data. *J. Prev. Med. Public Heal.* **50**, 294–302 (2017).
  90. Blake, T. L. *et al.* How does parent/self-reporting of common respiratory conditions compare with medical records among Aboriginal and Torres Strait Islander (Indigenous) children and young adults? *J. Paediatr. Child Health* **56**, 55–60 (2020).
  91. Fiederling, J., Shams, A. Z. & Haug, U. Validity of self-reported family history of cancer: A systematic literature review on selected cancers. *Int. J. Cancer* **139**, 1449–1460 (2016).
  92. Mongan, D., Curtis, E. & Drennan, D. Secondary Data Analysis. in *Quantitative Health Research: Issues and Methods* 373–382 (2013).
  93. Harpe, S. E. Using Secondary Data Sources for Pharmacoepidemiology and Outcomes Research. *Pharmacotherapy* **29**, 138–153 (2009).
  94. Brundin-Mather, R. *et al.* Secondary EMR data for quality improvement and research: A comparison of manual and electronic data collection from an integrated critical care electronic medical record system. *J. Crit. Care* **47**, 295–301 (2018).
  95. Wachter, R. M. *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age.* (McGraw-Hill Education, 2015).
  96. Hemingway, H. *et al.* Using nationwide 'big data' from linked electronic health records to help improve outcomes in cardiovascular diseases. *Program. Grants Appl. Res.* **5**, 1–330 (2017).
  97. Starfield, B. *et al.* Comorbidity: implications for the importance of primary care in 'case' management. *Ann. Fam. Med.* **1**, 8–14 (2003).
  98. Deeny, S. R. & Steventon, A. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. *BMJ*

- Qual. Saf.* **24**, 505–515 (2015).
99. McDonald, L., Lambrelli, D., Wasiak, R. & Ramagopalan, S. V. Real-world data in the United Kingdom: opportunities and challenges. *BMC Med.* **14**, 97 (2016).
  100. Jones, C. P., Papadopoulos, C. & Randhawa, G. Who's opting-in? A demographic analysis of the U.K. NHS Organ Donor Register. *PLoS One* **14**, e0209161 (2019).
  101. Shah, A. D., Martinez, C. & Hemingway, H. The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records. *BMC Med. Inform. Decis. Mak.* **12**, 88 (2012).
  102. Wang, Z. *et al.* Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One* **7**, e30412 (2012).
  103. Perera, G. *et al.* Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open* **6**, e008721 (2016).
  104. Soyiri, I. N. *et al.* Improving predictive asthma algorithms with modelled environment data for Scotland: an observational cohort study protocol. *BMJ Open* **8**, e23289 (2018).
  105. Kostkova, P. *et al.* Who Owns the Data? Open Data for Healthcare. *Front. Public Heal.* **4**, 7 (2016).
  106. Junghans, C., Feder, G., Hemingway, H., Timmis, A. & Jones, M. Recruiting patients to medical research: Double blind randomised trial of 'opt-in' versus 'opt-out' strategies. *Br. Med. J.* **331**, 940–942 (2005).
  107. Al-Shahi, R., Vousden, C. & Warlow, C. Bias from requiring explicit consent from all participants in observational research: Prospective, population based study. *Br. Med. J.* **331**, 942–945 (2005).
  108. Dunn, K. M., Jordan, K., Lacey, R. J., Shapley, M. & Jinks, C. Patterns of consent in epidemiologic research: Evidence from over 25,000 responders. *Am. J. Epidemiol.* **159**, 1087–1094 (2004).
  109. Woolf, S. H., Rothemich, S. F., Johnson, R. E. & Marsland, D. W. Selection bias from requiring patients to give consent to examine data for health services

- research. *Arch. Fam. Med.* **9**, 1111–1118 (2000).
110. Glass, D. C. *et al.* A telephone survey of factors affecting willingness to participate in health research surveys. *BMC Public Health* **15**, (2015).
  111. Spooner, R. & Towell, N. Fears that patients' personal medical information has been leaked in Medicare data breach. *The Canberra Times* (2016). Available at: <http://www.canberratimes.com.au/national/public-service/privacy-watchdog-called-after-health-department-data-breach-20160929-grr2m1.html>. (Accessed: 18th October 2017)
  112. Scottish Government National Statistics Publications. *Introducing The Scottish Index of Multiple Deprivation 2016*. (2016).
  113. Scottish Government. *Scottish Government Urban Rural Classification 2016*.
  114. Scottish Government. *Review of Nomenclature of Units for Territorial Statistics (NUTS) Boundaries*. (2016).
  115. Osborne, M. L. *et al.* Assessing future need for acute care in adult asthmatics: The profile of asthma risk study: A prospective health maintenance organization-based study. *Chest* **132**, 1151–1161 (2007).
  116. Luo, G., He, S., Stone, B. L., Nkoy, F. L. & Johnson, M. D. Developing a Model to Predict Hospital Encounters for Asthma in Asthmatic Patients: Secondary Analysis. *JMIR Med. Informatics* **8**, e16080 (2020).
  117. Price, D. *et al.* Predicting frequent asthma exacerbations using blood eosinophil count and other patient data routinely available in clinical practice. *J. Asthma Allergy* **9**, 1 (2016).
  118. Nissen, F. *et al.* Validation of asthma recording in the Clinical Practice Research Datalink (CPRD). *BMJ Open* **7**, e017474 (2017).
  119. Honkoop, P. J., Taylor, D. R., Smith, A. D., Snoeck-Stroband, J. B. & Sont, J. K. Early detection of asthma exacerbations by using action points in self-management plans. *Eur. Respir. J.* **41**, 53–59 (2013).
  120. Miller, M. K. *et al.* TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. *Eur. Respir. J.* **28**, 1145–1155 (2006).
  121. Bloom, C. I., Palmer, T., Feary, J., Quint, J. K. & Cullinan, P. Exacerbation patterns in adults with Asthma in England A population-based study. *Am. J. Respir. Crit. Care Med.* **199**, 446–453 (2019).

122. Hoskins, G. *et al.* Risk factors and costs associated with an asthma attack. *Thorax* **55**, 19–24 (2000).
123. Bloom, C. I. *et al.* Exacerbation risk and characterisation of the UK's asthma population from infants to old age. *Thorax* **73**, 313–320 (2018).
124. Sato, R. *et al.* The Strategy for Predicting Future Exacerbation of Asthma Using a Combination of the Asthma Control Test and Lung Function Test. *J. Asthma* **46**, 677–682 (2009).
125. Nissen, F. *et al.* Concomitant diagnosis of asthma and COPD: A quantitative study in UK primary care. *Br. J. Gen. Pract.* **68**, e775–e782 (2018).
126. Leung, J. M. & Sin, D. D. Asthma-COPD overlap syndrome: pathogenesis, clinical features, and therapeutic targets. *Br. Med. J.* **358**, j3772 (2017).
127. Hardin, M. *et al.* The clinical and genetic features of COPD-asthma overlap syndrome. *Eur. Respir. J.* **44**, 341–350 (2014).
128. Bourdin, A. *et al.* ERS/EAACI statement on severe exacerbations in asthma in adults: facts, priorities and key research questions. *Eur. Respir. J.* **54**, 1900900 (2019).
129. Al Sallakh, M. A. *et al.* Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. *Eur Respir J* **49**, 1700204 (2017).
130. Green, R. H., Brightling, C. E. & McKenna, S. Asthma exacerbations and eosinophil counts. A randomised controlled trial. *Lancet* **360**, 1715–21 (2002).
131. Bush, A. Pathophysiological mechanisms of asthma. *Front. Pediatr.* **7**, 68 (2019).
132. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* **339**, 157–160 (2009).
133. Pedersen, A. B. *et al.* Missing data and multiple imputation in clinical epidemiological research. *Clin. Epidemiol.* **9**, 157–166 (2017).
134. van Smeden, M., Groenwold, R. H. H. & Moons, K. G. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. *J. Clin. Epidemiol.* **125**, 188–190 (2020).
135. British Thoracic Society & SIGN. *British guideline on the management of asthma (2019 Edition)*. (2019).

136. Schatz, M., Meckley, L. M., Kim, M., Stockwell, B. T. & Castro, M. Asthma Exacerbation Rates in Adults Are Unchanged Over a 5-Year Period Despite High-Intensity Therapy. *J. Allergy Clin. Immunol. Pract.* **2**, 570–574 (2014).
137. O'Connor, R. D. *et al.* Subacute lack of asthma control and acute asthma exacerbation history as predictors of subsequent acute asthma exacerbations: Evidence from managed care data. *J. Asthma* **47**, 422–428 (2010).
138. Osman, M., Hansell, A. L., Simpson, C. R., Hollowell, J. & Helms, P. J. Gender-specific presentations for asthma, allergic rhinitis and eczema in primary care. *Prim. Care Respir. J.* **16**, 28–35 (2007).
139. Baibergenova, A. *et al.* Sex differences in hospital admissions from emergency departments in asthmatic adults: A population-based study. *Ann. Allergy, Asthma Immunol.* **96**, 666–672 (2006).
140. Rosychuk, R. J. *et al.* Sex differences in outcomes after discharge from Alberta emergency departments for asthma: A large population-based study. *J. Asthma* **55**, 817–825 (2018).
141. Mazurek, J. M. & Syamlal, G. *Prevalence of Asthma, Asthma Attacks, and Emergency Department Visits for Asthma Among Working Adults - National Health Interview Survey, 2011-2016.* (2018).
142. Schatz, M., Clark, S. & Camargo, C. A. Sex differences in the presentation and course of asthma hospitalizations. *Chest* **129**, 50–55 (2006).
143. Goto, T., Tsugawa, Y., Camargo, C. A. J. & Hasegawa, K. Sex differences in the risk of hospitalization among patients presenting to US emergency departments with asthma exacerbation. *J. Allergy Clin. Immunol. Pract.* **4**, 149–151 (2016).
144. Singh, A. K., Cydulka, R. K., Stahmer, S. A., Woodruff, P. G. & Camargo, C. A. Sex differences among adults presenting to the emergency department with acute asthma. *Arch. Intern. Med.* **159**, 1237–1243 (1999).
145. Hill, J., Arrotta, N., Villa-Roel, C., Dennett, L. & Rowe, B. H. Factors associated with relapse in adult patients discharged from the emergency department following acute asthma: a systematic review. *BMJ Open Respir. Res.* **4**, e000169 (2017).
146. Baltrus, P. *et al.* Individual and county level predictors of asthma related



- emergency department visits among children on Medicaid: A multilevel approach. *J. Asthma* **54**, 53–61 (2017).
147. Chen, Y., Stewart, P., Johansen, H., McRae, L. & Taylor, G. Sex difference in hospitalization due to asthma in relation to age. *J. Clin. Epidemiol.* **56**, 180–187 (2003).
148. Engelkes, M. *et al.* Incidence, risk factors and re-exacerbation rate of severe asthma exacerbations in a multinational, multidatabase pediatric cohort study. *Pediatr. Allergy Immunol.* **31**, 496–505 (2020).
149. Balzano, G., Fuschillo, S., Melillo, G. & Bonini, S. Asthma and sex hormones. *Allergy* **56**, 13–20 (2001).
150. McCleary, N., Nwaru, B. I., Nurmatov, U. B., Critchley, H. & Sheikh, A. Endogenous and exogenous sex steroid hormones in asthma and allergy in females: A systematic review and meta-analysis. *J. Allergy Clin. Immunol.* **141**, 1510–1513 (2018).
151. Al-Sahab, B., Hamadeh, M. J., Ardern, C. I. & Tamim, H. Early menarche predicts incidence of asthma in early adulthood. *Am. J. Epidemiol.* **173**, 64–70 (2011).
152. Gupta, R. P., Mukherjee, M., Sheikh, A. & Strachan, D. P. Persistent variations in national asthma mortality, hospital admissions and prevalence by socioeconomic status and region in England. *Thorax* **73**, 706–712 (2018).
153. Alsallakh, M. A., Rodgers, S. E., Lyons, R. A., Sheikh, A. & Davies, G. A. Association of socioeconomic deprivation with asthma care, outcomes, and deaths in Wales: A 5-year national linked primary and secondary care cohort study. *PLoS Med.* **18**, e1003497 (2021).
154. Das, L. T. *et al.* Predicting frequent emergency department visits among children with asthma using EHR data. *Pediatr. Pulmonol.* **52**, 880–890 (2017).
155. Acosta, L. *et al.* Respiratory Emergency Department Visits More Common Among Native American Children than Non-Native American in South Dakota. in *Journal of Allergy and Clinical Immunology* **145**, AB112 (2020).
156. Stingone, J. A. & Claudio, L. Disparities in the use of urgent health care services among asthmatic children. *Ann. Allergy, Asthma Immunol.* **97**, 244–250 (2006).

157. Sheikh, A. *et al.* Ethnic variations in asthma hospital admission, readmission and death: a retrospective, national cohort study of 4.62 million people in Scotland. *BMC Med.* **14**, 3 (2016).
158. Backman, H. *et al.* A population-based cohort of adults with asthma: mortality and participation in a long-term follow-up. *Eur. Clin. Respir. J.* **4**, 1334508 (2017).
159. Douglas, A. *et al.* Pilot study linking primary care records to Census, cardiovascular hospitalization and mortality data in Scotland: Feasibility, utility and potential. *J. Public Heal. (United Kingdom)* **38**, 815–823 (2016).
160. Mukherjee, M. *et al.* Estimating the incidence, prevalence and true cost of asthma in the UK: Secondary analysis of national stand-alone and linked databases in England, Northern Ireland, Scotland and Wales-A study protocol. *BMJ Open* **4**, e006647 (2014).
161. Atkinson, M. D. *et al.* Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. *BMC Med. Inform. Decis. Mak.* **17**, 2 (2017).
162. Lewis, J. D. & Brensinger, C. Agreement between GPRD smoking data: A survey of general practitioners and a population-based survey. *Pharmacoepidemiol. Drug Saf.* **13**, 437–441 (2004).
163. Marston, L. *et al.* Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol. Drug Saf.* **19**, 618–626 (2010).
164. Quinto, K. B. *et al.* The association of obesity and asthma severity and control in children. *J. Allergy Clin. Immunol.* **128**, 964–969 (2011).
165. Schatz, M. *et al.* Overweight/obesity and risk of seasonal asthma exacerbations. *J. Allergy Clin. Immunol. Pract.* **1**, 618–622 (2013).
166. Mosen, D. M., Schatz, M., Magid, D. J. & Camargo, C. A. The relationship between obesity and asthma severity and control in adults. *J. Allergy Clin. Immunol.* **122**, 507–511 (2008).
167. Barros, R. *et al.* Obesity increases the prevalence and the incidence of asthma and worsens asthma severity. *Clin. Nutr.* **36**, 1068–1074 (2017).
168. Farmer, R. *et al.* Promises and pitfalls of electronic health record analysis.

- Diabetologia* **61**, 1241–1248 (2018).
169. Lieu, T. A., Capra, A. M., Quesenberry, C. P., Mendoza, G. R. & Mazar, M. Computer-based models to identify high-risk adults with asthma: Is the glass half empty or half full? *J. Asthma* **36**, 359–370 (1999).
  170. Peters, D., Chen, C., Markson, L. E., Allen-Ramey, F. C. & Vollmer, W. M. Using an asthma control questionnaire and administrative data to predict health-care utilization. *Chest* **129**, 918–924 (2006).
  171. Tanaka, A. *et al.* Predicting future risk of exacerbations in Japanese patients with adult asthma: A prospective 1-year follow up study. *Allergol. Int.* **66**, 568–573 (2017).
  172. Quezada, W. A. *et al.* Controlled Asthma Despite Inhaled Corticosteroid Treatment. *Ann Allergy Asthma Immunol* **116**, 112–117 (2017).
  173. Zeiger, R. S. *et al.* Evaluation of the National Heart, Lung, and Blood Institute guidelines impairment domain for classifying asthma control and predicting asthma exacerbations. *Ann. Allergy, Asthma Immunol.* **108**, 81–87 (2012).
  174. National Heart, Lung, and B. I. *National Heart, Lung, and Blood Institute National Asthma Education and Prevention Program Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma Full Report 2007.* (2007).
  175. Currie, G. P., Douglas, J. G. & Heaney, L. G. Difficult to treat asthma in adults. *BMJ* **338**, b494 (2009).
  176. Myatt, R. Measuring peak expiratory flow rate: what the nurse needs to know. *Nurs. Stand.* **31**, 40–44 (2017).
  177. Nunn, A. J. & Gregg, I. New regression equations for predicting peak expiratory flow in adults. *Br. Med. J.* **298**, 1068–1070 (1989).
  178. Reddel, H. K., Vincent, S. D. & Civitico, J. The need for standardisation of peak flow charts. *Thorax* **60**, 164–167 (2005).
  179. Ko, F. W. S. *et al.* Evaluation of the asthma control test: A reliable determinant of disease stability and a predictor of future exacerbations. *Respirology* **17**, 370–378 (2012).
  180. Lehtimäki, L. *et al.* Predictive value of exhaled nitric oxide in the management of asthma: A systematic review. *Eur. Respir. J.* **48**, 706–714 (2016).

181. Semprini, R. *et al.* Type 2 Biomarkers and Prediction of Future Exacerbations and Lung Function Decline in Adult Asthma. *J. Allergy Clin. Immunol. Pract.* **6**, 1982–1988 (2018).
182. Daugherty, J., Lin, X., Baxter, R., Suruki, R. & Bradford, E. The impact of long-term systemic glucocorticoid use in severe asthma: A UK retrospective cohort analysis. *J. Asthma* **55**, 651–658 (2018).
183. Samuels-Kalow, M. E. *et al.* A Predictive Model for Identification of Children at Risk of Subsequent High-Frequency Utilization of the Emergency Department for Asthma. *Pediatr. Emerg. Care* **36**, e85–e89 (2020).
184. World Health Organization. *Adherence to Long-Term Therapies: Evidence for action.* (2003).
185. Engelkes, M., Janssens, H. M., de Jongste, J. C., Sturkenboom, M. C. J. M. & Verhamme, K. M. C. Medication adherence and the risk of severe asthma exacerbations: a systematic review. *Eur Respir J* **45**, 396–407 (2015).
186. Stern, L. *et al.* Medication compliance and disease exacerbation in patients with asthma: A retrospective study of managed care data. *Ann. Allergy, Asthma Immunol.* **97**, 402–408 (2006).
187. Chongmelaxme, B., Chaiyakunapruk, N. & Dilokthornsakul, P. Association between adherence and severe asthma exacerbation: A systematic review and meta-analysis. *J. Am. Pharm. Assoc.* (2020). doi:10.1016/j.japh.2020.02.010
188. Vrijens, B. *et al.* What We Mean When We Talk About Adherence in Respiratory Medicine. *J. Allergy Clin. Immunol. Pract.* **4**, 802–812 (2016).
189. Steiner, J. F. & Prochazka, A. V. The assessment of refill compliance using pharmacy records: Methods, validity, and applications. *J. Clin. Epidemiol.* **50**, 105–116 (1997).
190. Vollmer, W. M. *et al.* Comparison of pharmacy-based measures of medication adherence. *BMC Health Serv. Res.* **12**, 155 (2012).
191. Hess, L. M., Raebel, M. A., Conner, D. A. & Malone, D. C. Measurement of Adherence in Pharmacy Administrative Databases: A Proposal for Standard Definitions and Preferred Measures. *Ann. Pharmacother.* **40**, 1280–1288 (2006).
192. Schatz, M. *et al.* Change in asthma control over time: Predictors and

- outcomes. *J. Allergy Clin. Immunol. Pract.* **2**, 59–64 (2014).
193. Pharmaceutical Press Joint Formulary Committee. Prednisolone. in *British National Formulary Version 80* (2019).
194. McGrath, K. W. *et al.* A large subgroup of mild-to-moderate asthma is persistently noneosinophilic. *Am. J. Respir. Crit. Care Med.* **185**, 612–619 (2012).
195. Brightling, C. E., Green, R. H. & Pavord, I. D. Biomarkers predicting response to corticosteroid therapy in asthma. *Treat. Respir. Med.* **4**, 309–316 (2005).
196. Gibson, P. G., Saltos, N. & Borgas, T. Airway mast cells and eosinophils correlate with clinical severity and airway hyperresponsiveness in corticosteroid-treated asthma. *J. Allergy Clin. Immunol.* **105**, 752–759 (2000).
197. American Academy of Allergy Asthma and Immunology. Atopy Defined. Available at: <https://www.aaaai.org/conditions-and-treatments/conditions-dictionary/atopy>. (Accessed: 19th April 2021)
198. Westerhof, G. A., Coumou, H., de Nijs, S. B., Weersink, E. J. & Bel, E. H. Clinical predictors of remission and persistence of adult-onset asthma. *J. Allergy Clin. Immunol.* **141**, 104-109.e3 (2018).
199. Liao, H. *et al.* Impact of viral infection on acute exacerbation of asthma in out-patient clinics: A prospective study. *J. Thorac. Dis.* **8**, 505–512 (2016).
200. Atmar, R. L. *et al.* Respiratory tract viral infections in inner-city asthmatic adults. *Arch. Intern. Med.* **158**, 2453–2459 (1998).
201. Wark, P. A. B. *et al.* Neutrophil degranulation and cell lysis is associated with clinical severity in virus-induced asthma. *Eur. Respir. J.* **19**, 68–75 (2002).
202. Nicholson, K. G., Kent, J. & Ireland, D. C. Respiratory viruses and exacerbations of asthma in adults. *Br. Med. J.* **307**, 982–986 (1993).
203. Saraya, T., Kimura, H., Kurai, D., Ishii, H. & Takizawa, H. The molecular epidemiology of respiratory viruses associated with asthma attacks. *Medicine* **96**, e8204 (2017).
204. Costa, L. D. C., Costa, P. S. & Camargos, P. A. M. Exacerbation of asthma and airway infection: Is the virus the villain? *J. Pediatr. (Rio. J.)* **90**, 542–555 (2014).
205. Papadopoulos, N. G. & Johnston, S. L. The role of viruses in the induction and

- progression of asthma. *Curr. Allergy Asthma Rep.* **1**, 144–152 (2001).
206. Tan, W. C. *et al.* Epidemiology of respiratory viruses in patients hospitalized with near-fatal asthma, acute exacerbations of asthma, or chronic obstructive pulmonary disease. *Am. J. Med.* **115**, 272–277 (2003).
207. Schroeder, A. R. & Mansbach, J. M. Recent evidence on the management of bronchiolitis. *Current Opinion in Pediatrics* **26**, 328–333 (2014).
208. Docherty, A. B. *et al.* Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *Br. Med. J.* **369**, m1985 (2020).
209. Morais-Almeida, M., Pité, H., Aguiar, R., Ansotegui, I. & Bousquet, J. Asthma and the Coronavirus Disease 2019 Pandemic: A Literature Review. *Int. Arch. Allergy Immunol.* **181**, 680–688 (2020).
210. Williamson, E. *et al.* *OpenSAFELY: factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients.* (Cold Spring Harbor Laboratory Press, 2020).  
doi:10.1101/2020.05.06.20092999
211. Hussein, M. H. *et al.* Asthma in COVID-19 patients: An extra chain fitting around the neck? *Respir. Med.* **175**, 106205 (2020).
212. Gershon, A. S., Wang, C., Guan, J. & To, T. Burden of comorbidity in individuals with asthma. *Thorax* **65**, 612–618 (2010).
213. Ledford, D. K. & Lockey, R. F. Asthma and comorbidities. *Curr. Opin. Allergy Clin. Immunol.* **13**, 78–86 (2013).
214. Cazzola, M., Segreti, A., Calzetta, L. & Rogliani, P. Comorbidities of asthma: Current knowledge and future research needs. *Curr. Opin. Pulm. Med.* **19**, 36–41 (2013).
215. Su, X. *et al.* Prevalence of comorbidities in asthma and nonasthma patients. *Medicine (Baltimore)*. **95**, e3459 (2016).
216. Irwin, M. R. & Miller, A. H. Depressive disorders and immunity: 20 years of progress and discovery. *Brain. Behav. Immun.* **21**, 374–383 (2007).
217. Deyo, R. A., Cherkin, D. C. & Ciol, M. A. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J. Clin. Epidemiol.* **45**, 613–619 (1992).

218. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic in longitudinal studies: development and validation. *Journal of Chronic Diseases* **40**, 373–383 (1987).
219. Nwaru, B. I. & Sheikh, A. Hormonal contraceptives and asthma in women of reproductive age: Analysis of data from serial national Scottish Health Surveys. *J. R. Soc. Med.* **108**, 358–371 (2015).
220. Nwaru, B. I. *et al.* Hormonal contraception and the risk of severe asthma exacerbation: 17-year population-based cohort study. *Thorax* **76**, 109–115 (2021).
221. Anenberg, S. C. *et al.* Estimates of the global burden of ambient PM<sub>2.5</sub>, ozone, and NO<sub>2</sub> on asthma incidence and emergency room visits. *Environ. Health Perspect.* **126**, 107004 (2018).
222. Alhanti, B. A. *et al.* Ambient air pollution and emergency department visits for asthma: A multi-city assessment of effect modification by age. *J. Expo. Sci. Environ. Epidemiol.* **26**, 180–188 (2016).
223. Gharibi, H. *et al.* Ozone pollution and asthma emergency department visits in the Central Valley, California, USA, during June to September of 2015: a time-stratified case-crossover analysis. *J. Asthma* **56**, 1037–1048 (2019).
224. Xie, S., Greenblatt, R., Levy, M. Z. & Himes, B. E. Enhancing Electronic Health Record Data with Geospatial Information. in *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* **2017**, 123–132 (2017).
225. Abrahamsen, R. *et al.* Association of respiratory symptoms and asthma with occupational exposures: Findings from a population-based cross-sectional survey in Telemark, Norway. *BMJ Open* **7**, e014018 (2017).
226. Laditka, J. N., Laditka, S. B., Arif, A. A. & Hoyle, J. N. Work-related asthma in the USA: Nationally representative estimates with extended follow-up. *Occup. Environ. Med.* **77**, 617–622 (2020).
227. Caridi, M. N. *et al.* Occupation and task as risk factors for asthma-related outcomes among healthcare workers in New York City. *Int. J. Hyg. Environ. Health* **222**, 211–220 (2019).
228. Lu, P. J., O'Halloran, A., Ding, H., Srivastav, A. & Williams, W. W. Uptake of

- Influenza Vaccination and Missed Opportunities among Adults with High-Risk Conditions, United States, 2013. *Am. J. Med.* **129**, 636 (2016).
229. Kassianos, G. *et al.* Influenza vaccination: Key facts for general practitioners in Europe-A synthesis by European experts based on national guidelines and best practices in the United Kingdom and the Netherlands. *Drugs Context* **5**, 212293 (2016).
230. Vasileiou, E. *et al.* Effectiveness of influenza vaccines in Asthma: A systematic review and meta-analysis. *Clin. Infect. Dis.* **65**, 1388–1395 (2017).
231. Bell, T. D., Chai, H., Berlow, B. & Daniels, G. Immunization with killed influenza virus in children with chronic asthma. *Chest* **73**, 140–145 (1978).
232. Murphy, A. C. *et al.* The relationship between clinical outcomes and medication adherence in difficult-to-control asthma. *Thorax* **67**, 751–753 (2012).
233. Friedman, C. *et al.* Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J. Am. Med. Informatics Assoc.* **22**, 43–50 (2015).
234. Granger, B. B. *et al.* Adherence to candesartan and placebo and outcomes in chronic heart failure in the CHARM programme: Double-blind, randomised, controlled clinical trial. *Lancet* **366**, 2005–11 (2005).
235. Osterberg, L. & Blaschke, T. Drug therapy: Adherence to medication. *N Engl J Med* **353**, 487–497 (2005).
236. Muzina, D. J. *et al.* Rate of non-adherence prior to upward dose titration in previously stable antidepressant users. *J. Affect. Disord.* **130**, 46–52 (2011).
237. Cutler, R. L., Fernandez-Llimos, F., Frommer, M., Benrimoj, C. & Garcia-Cardenas, V. Economic impact of medication non-adherence by disease groups: a systematic review. *BMJ Open* **8**, e016982 (2018).
238. Patel, A. R. *et al.* Burden of non-adherence to latent tuberculosis infection drug therapy and the potential cost-effectiveness of adherence interventions in Canada: A simulation study. *BMJ Open* **7**, e015108 (2017).
239. McKenzie, S. J. *et al.* The Burden of Non-Adherence to Cardiovascular Medications Among the Aging Population in Australia: A Meta-Analysis. *Drugs Aging* **32**, 217–225 (2015).



240. Sokol, M. C., McGuigan, K. A., Verbrugge, R. R. & Epstein, R. S. Impact of medication adherence on hospitalization risk and healthcare cost. *Med. Care* **43**, 521–530 (2005).
241. Chiatti, C. *et al.* The economic burden of inappropriate drug prescribing, lack of adherence and compliance, adverse drug events in older people a systematic review. *Drug Saf.* **35**, 73–87 (2012).
242. Cutrona, S. L. *et al.* Targeting cardiovascular medication adherence interventions. *J. Am. Pharm. Assoc.* **52**, 381–397 (2012).
243. Haberer, J. E. *et al.* Improving antiretroviral therapy adherence in resource-limited settings at scale: a discussion of interventions and recommendations. *J. Int. AIDS Soc.* **20**, 21371 (2017).
244. Normansell, R., Kew, K. M. & Mathioudakis, A. G. Interventions to improve inhaler technique for people with asthma. *Cochrane Database Syst. Rev.* **3**, CD012286 (2017).
245. Valgimigli, M. *et al.* Standardized classification and framework for reporting, interpreting, and analysing medication non-adherence in cardiovascular clinical trials: A consensus report from the Non-adherence Academic Research Consortium (NARC). *Eur. Heart J.* **40**, 2070–2085 (2019).
246. DeWorsop, D. *et al.* Feasibility and success of cell-phone assisted remote observation of medication adherence (CAROMA) in clinical trials. *Drug Alcohol Depend.* **163**, 24–30 (2016).
247. Chongmelaxme, B., Chaiyakunapruk, N. & Dilokthornsakul, P. Incorporating adherence in cost-effectiveness analyses of asthma: a systematic review. *J. Med. Econ.* **22**, 554–566 (2019).
248. Vrijens, B. *et al.* A new taxonomy for describing and defining adherence to medications. *Br. J. Clin. Pharmacol.* **73**, 691–705 (2012).
249. Lehmann, A. *et al.* Assessing medication adherence: options to consider. *Int. J. Clin. Pharm.* **36**, 55–69 (2014).
250. Williams, A. B., Amico, K. R., Bova, C. & Womack, J. A. A Proposal for Quality Standards for Measuring Medication Adherence in Research. *AIDS Behav.* **17**, 284–297 (2013).
251. Sajatovic, M., Velligan, D. I., Weiden, P. J., Valenstein, M. & Ogedegbe, G.

- Measurement of psychiatric treatment adherence. *J. Psychosom. Res.* **69**, 591–599 (2010).
252. Boissel, J.-P. & Nony, P. Using pharmacokinetic-pharmacodynamic relationships to predict the effect of poor compliance. *Clin. Pharmacokinet.* **41**, 1–6 (2002).
253. Williams, L. K. *et al.* Relationship between adherence to inhaled corticosteroids and poor outcomes among adults with asthma. *J. Allergy Clin. Immunol.* **114**, 1128–1293 (2004).
254. Ismaila, A. *et al.* Impact of adherence to treatment with fluticasone propionate/salmeterol in asthma patients. *Curr. Med. Res. Opin.* **30**, 1417–1425 (2014).
255. Tibble, H. *et al.* Measuring and reporting treatment adherence: What can we learn by comparing two respiratory conditions? *Br. J. Clin. Pharmacol.* **87**, 825–836 (2020).
256. Tibble, H. *et al.* Heterogeneity in asthma medication adherence measurement. in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)* (2019). doi:10.1109/BIBE.2019.00168
257. Tibble, H. *et al.* A data-driven typology of asthma medication adherence using cluster analysis. *Sci. Rep.* **10**, 14999 (2020).
258. Allemann, S. S., Dediu, D. & Dima, A. L. Beyond Adherence Thresholds: A Simulation Study of the Optimal Classification of Longitudinal Adherence Trajectories From Medication Refill Histories. *Front. Pharmacol.* **10**, 383 (2019).
259. Wu, A. C. *et al.* Primary Adherence to Controller Medications for Asthma Is Poor. *Ann. Am. Thorac. Soc.* **12**, 161–166 (2015).
260. Fischer, M. A. *et al.* Primary medication non-adherence: analysis of 195,930 electronic prescriptions. *J. Gen. Intern. Med.* **25**, 284–90 (2010).
261. Williams, L. K. *et al.* Patients with asthma who do not fill their inhaled corticosteroids: A study of primary nonadherence. *J. Allergy Clin. Immunol.* **120**, 1153–1159 (2007).
262. Liberman, J. N. *et al.* Determinants of primary nonadherence in asthma-controller and dyslipidemia pharmacotherapy. *Am. J. Pharm. Benefits* **2**, 111–

- 118 (2010).
263. Berger, Z. *et al.* Lower copay and oral administration: Predictors of first-fill adherence to new asthma prescriptions. *Am. Heal. Drug Benefits* **2**, 174–179 (2009).
264. Jentzsch, N. S., Camargos, P. A. M., Colosimo, E. A. & Bousquet, J. Monitoring adherence to beclomethasone in asthmatic children and adolescents through four different methods. *Allergy* **64**, 1458–1462 (2009).
265. Tibble, H. *et al.* Linkage of primary care prescribing records and pharmacy dispensing Records in the Salford Lung Study: application in asthma. *BMC Med. Res. Methodol.* **20**, 303 (2020).
266. Galozy, A., Nowaczyk, S., Sant’Anna, A., Ohlsson, M. & Lingman, M. Pitfalls of medication adherence approximation through EHR and pharmacy records: Definitions, data and computation. *Int. J. Med. Inform.* **136**, 104092 (2020).
267. Barnes, P. J. & Pedersen, S. Efficacy and Safety of Inhaled Corticosteroids in Asthma. *Am. Rev. Respir. Dis.* **148**, S1–S26 (1993).
268. Suissa, S., Ernst, P., Benayoun, S., Baltzan, M. & Cai, B. Low-Dose Inhaled Corticosteroids and the Prevention of Death from Asthma. *N. Engl. J. Med.* **343**, 332–336 (2000).
269. Pharmaceutical Press Joint Formulary Committee. *British National Formulary Version 80.* (2019).
270. McTaggart, S. *et al.* Use of text-mining methods to improve efficiency in the calculation of drug exposure to support pharmacoepidemiology studies. *Int. J. Epidemiol.* **47**, 617–624 (2018).
271. DeClercq, J. & Choi, L. Statistical considerations for medication adherence research. *Curr. Med. Res. Opin.* **36**, 1549–1557 (2020).
272. Bjarnadottir, M. V., Czerwinski, D. & Onukwugha, E. Sensitivity of the Medication Possession Ratio to Modelling Decisions in Large Claims Databases. *Pharmacoeconomics* **36**, 369–380 (2018).
273. Obeng-Kusi, M., Lubbe, M. S., Cockeran, M. & Burger, J. R. Comparison of adherence measures using claims data in the South African private health sector. *South African Med. J.* **110**, 932–936 (2020).
274. Buono, E. W. *et al.* Coming full circle in the measurement of medication

- adherence: opportunities and implications for health care. *Patient Prefer. Adherence* **11**, 1009–1017 (2017).
275. Ciechanowski, P. S., Katon, W. J. & Russo, J. E. Depression and diabetes: Impact of depressive symptoms on adherence, function, and costs. *Arch. Intern. Med.* **160**, 3278–3285 (2000).
  276. Carroll, C. L., Feldman, S. R., Camacho, F. T., Manuel, J. C. & Balkrishnan, R. Adherence to topical therapy decreases during the course of an 8-week psoriasis clinical trial: commonly used methods of measuring adherence to topical therapy overestimate actual use. *J. Am. Acad. Dermatol.* **51**, 212–216 (2004).
  277. Yeaw, J., Benner, J. S., Walt, J. G., Sian, S. & Smith, D. B. Comparing Adherence and Persistence Across 6 Chronic Medication Classes. *J. Manag. Care Pharm.* **15**, 728–740 (2009).
  278. Huiart, L. *et al.* Early discontinuation of tamoxifen intake in younger women with breast cancer: Is it time to rethink the way it is prescribed? *Eur. J. Cancer* **48**, 1939–1946 (2012).
  279. Rinfret, S. *et al.* Telephone contact to improve adherence to dual antiplatelet therapy after drug-eluting stent implantation. *Heart* **99**, 562–569 (2013).
  280. Rabe, K. F. *et al.* Worldwide severity and control of asthma in children and adults: The global Asthma Insights and Reality surveys. *J. Allergy Clin. Immunol.* **114**, 40–47 (2004).
  281. Izquierdo, J. L. *et al.* Misdiagnosis of patients receiving inhaled therapies in primary care. *Int. J. Chron. Obstruct. Pulmon. Dis.* **5**, 241–249 (2010).
  282. Rabe, K. F., Vermeire, P. A., Soriano, J. B. & Maier, W. C. Clinical management of asthma in 1999: the Asthma Insights and Reality in Europe (AIRE) study. *Eur. Respir. J.* **16**, 802–807 (2000).
  283. Anderson, H. R., Bland, J. M., Patel, S. & Peckham, C. The natural history of asthma in childhood. *J. Epidemiol. Community Health* **40**, 121–129 (1986).
  284. Bernstein, J. A. Occupational asthma. in *Allergy and Asthma: Practical Diagnosis and Management: Second Edition* 253–270 (Springer International Publishing, 2016). doi:10.1007/978-3-319-30835-7\_17
  285. Blaschke, T. F., Osterberg, L., Vrijens, B. & Urquhart, J. Adherence to

- Medications: Insights Arising from Studies on the Unreliable Link Between Prescribed and Actual Drug Dosing Histories. *Annu. Rev. Pharmacol. Toxicol* **52**, 275–301 (2012).
286. O’Byrne, P. M. *et al.* Inhaled Combined Budesonide–Formoterol as Needed in Mild Asthma. *N. Engl. J. Med.* **378**, 1865–1876 (2018).
287. Bateman, E. D. *et al.* As-Needed Budesonide–Formoterol versus Maintenance Budesonide in Mild Asthma. *N. Engl. J. Med.* **378**, 1877–1887 (2018).
288. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).
289. Vayena, E., Blasimme, A. & Cohen, I. G. Machine learning in medicine: Addressing ethical challenges. *PLoS Med.* **15**, e1002689. (2018).
290. Hastie, T., Tibshirani, R. & Friedman, J. *Elements of Statistical Learning (2nd edition)*. *Springer Series in Statistics* (2009).
291. Kuncheva, L. I. On the optimality of Naïve Bayes with dependent binary features. *Pattern Recognit. Lett.* **27**, 830–837 (2006).
292. Bouckaert, R. R. Naive Bayes classifiers that perform well with continuous variables. in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)* **3339**, 1089–1094 (Springer Verlag, 2004).
293. Majka, M. CRAN: Package ‘naivebayes’ (version 0.9.2). <https://cran.r-project.org/web/packages/naivebayes/naivebayes.pdf> (2018).
294. Chomboon, K., Chujai, P., Teerarassamdee, P., Kerdprasop, K. & Kerdprasop, N. An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm. in *Proceedings of the 3rd International Conference on Industrial Application Engineering 2015* 280–285 (2015). doi:10.12792/iciae2015.051
295. Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **27**, 857–871 (1971).
296. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and regression trees*. *Classification and Regression Trees* (Routledge, 1984). doi:10.1201/9781315139470
297. Raileanu, L. E. & Stoffel, K. Theoretical comparison between the Gini Index and Information Gain criteria. *Ann. Math. Artif. Intell.* **41**, 77–93 (2004).

298. Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188 (1936).
299. Chen, D.-R., Wu, Q., Ying, Y. & Zhou, D.-X. Support Vector Machine Soft Margin Classifiers: Error Analysis. *J. Mach. Learn. Res.* **5**, 1143–1175 (2004).
300. Platt, J. C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Adv. large margin Classif.* **10**, 61–74 (1999).
301. Erfani, S. M., Rajasegarar, S., Karunasekera, S. & Leckie, C. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognit.* **58**, 121–134 (2016).
302. Kuncheva, L. I., Hadjitodorov, S. T. & Todorova, L. P. Experimental comparison of cluster ensemble methods. in *2006 9th International Conference on Information Fusion, FUSION* (2006).  
doi:10.1109/ICIF.2006.301614
303. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* **3**, 21–45 (2006).
304. Chawla, N., Eschrich, S. & Hall, L. O. Creating Ensembles of Classifiers. in *Proceedings of the 2001 IEEE International Conference on Data Mining* (2001).
305. Brown, G. Ensemble Learning. in *Encyclopedia of Machine Learning and Data Mining* 393–402 (2017). doi:10.1007/978-1-4419-9326-7\_1
306. Ho, T. K. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 832–844 (1998).
307. Kuncheva, L. I. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Mach. Learn.* **51**, 181–207 (2003).
308. Giacinto, G. & Roli, F. Design of effective neural network ensembles for image classification purposes. *Image Vis. Comput.* **19**, 699–707 (2001).
309. Kohavi, R., Wolpert, D. H. & Others. Bias plus variance decomposition for zero-one loss functions. in *Machine Learning: Proc. 13th International Conference* 275–283 (1996). doi:10.1016/S1066-7938(00)80097-4
310. Huang, Y. S. & Suen, C. Y. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Trans. Pattern Anal.*

- Mach. Intell.* **17**, 90–94 (1995).
311. Cruz, R. M. O., Sabourin, R. & Cavalcanti, G. D. C. Dynamic classifier selection: Recent advances and perspectives. *Inf. Fusion* **41**, 195–216 (2018).
  312. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82 (1997).
  313. Caruana, R. & Niculescu-Mizil, A. An Empirical Comparison of Supervised Learning Algorithms. in *Proceedings of the 23rd International Conference on Machine Learning* (2006). doi:10.1145/1143844.1143865
  314. Breiman, L. Random Forest. *Mach. Learn.* **45**, 5–32 (2001).
  315. Dietterich, T. G. Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach. Learn.* **40**, 139–157 (2000).
  316. Freund, Y. & Schapire, R. E. Experiments with a New Boosting Algorithm. in *Proceedings of the 13th International Conference on Machine Learning* 148–156 (1996). doi:10.1.1.133.1040
  317. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016). doi:10.1145/2939672.2939785
  318. Shor, N. Z. The Subgradient Method. in *Minimization Methods for Non-Differentiable Functions* 22–47 (1985). doi:10.1007/978-3-642-82118-9\_3
  319. Ratliff, N. D., Andrew Bagnell, J. & Zinkevich, M. A. Maximum margin planning. in *ACM International Conference Proceeding Series* **148**, 729–736 (2006).
  320. Chen, T. *et al.* CRAN: package ‘xgboost’ (version 0.71.2). <https://cran.r-project.org/web/packages/xgboost/xgboost.pdf> (2018).
  321. Schmid, C. H. & Griffith, J. L. Multivariate Classification Rules: Calibration and Discrimination. in *Wiley StatsRef: Statistics Reference Online* (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118445112.stat05650
  322. Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E. & Vergouwe, Y. Updating methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61**, 76–86 (2008).
  323. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).

324. Lobo, J. M., Jiménez-valverde, A. & Real, R. AUC: A misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145–151 (2008).
325. Brier, G. W. Verification of Forecasts Expressed in Terms of Probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
326. Poses, R. M., Cebul, R. D., Collins, M. & Fager, S. S. Accuracy of Experienced Physicians' Probability Estimates for Patients with Sore Throats: Implications for Decision Making. *J. Am. Med. Assoc.* **254**, 925–929 (1985).
327. Bell, N. R. *et al.* Understanding and communicating risk. *Can. Fam. Physician* **64**, 181–185 (2018).
328. Wegwarth, O. & Gigerenzer, G. The barrier to informed choice in cancer screening: Statistical illiteracy in physicians and patients. in *Recent Results in Cancer Research* **210**, 207–221 (Springer New York LLC, 2018).
329. Paolo, W. F., Silaban, R., Nguyen, L., Wojcik, S. & Grant, W. Physicians' understanding of CT probabilities in ED patients with acute abdominal pain. *Am. J. Emerg. Med.* **36**, 1986–1992 (2018).
330. Jenny, M. A., Keller, N. & Gigerenzer, G. Assessing minimal medical statistical literacy using the Quick Risk Test: a prospective observational study in Germany. *BMJ Open* **8**, e20847 (2018).
331. Kautz, T., Eskofier, B. M. & Pasluosta, C. F. Generic performance measure for multiclass-classifiers. *Pattern Recognit.* **68**, 111–125 (2017).
332. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA)-Protein Struct.* **405**, 442–451 (1975).
333. Ranawana, R. & Palade, V. Optimized precision - A new measure for classifier performance evaluation. in *2006 IEEE Congress on Evolutionary Computation* 2254–2261 (IEEE, 2006). doi:10.1109/cec.2006.1688586
334. Steyerberg, E. W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* (Springer, 2009). doi:10.1007/978-3-030-16399-0
335. Crowson, C. S., Atkinson, E. J. & Therneau, T. T. Assessing Calibration of Prognostic Risk Scores. *Stat Methods Med Res.* **25**, 1692–1706 (2016).



336. Escandar, G. M., Damiani, P. C., Goicoechea, H. C. & Olivieri, A. C. A review of multivariate calibration methods applied to biomedical analysis. *Microchem. J.* **82**, 29–42 (2006).
337. Chen, W., Sahiner, B., Samuelson, F., Pezeshk, A. & Petrick, N. Calibration of medical diagnostic classifier scores to the probability of disease. *Stat. Methods Med. Res.* **27**, 1394–1409 (2018).
338. Zadrozny, B. & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 694–699 (2002). doi:10.1145/775047.775151
339. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).
340. Suruki, R. Y., Daugherty, J. B., Boudiaf, N. & Albers, F. C. The frequency of asthma exacerbations and healthcare utilization in patients with asthma from the UK and USA. *BMC Pulm. Med.* **17**, 74 (2017).
341. He, H. & Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowl. DATA Eng.* **21**, 1263–1284 (2009).
342. Rahman, M. M. & Davis, D. N. Addressing the Class Imbalance Problem in Medical Datasets. *Int. J. Mach. Learn. Comput.* **3**, 224–228 (2013).
343. Chawla, N. V. Data Mining for Imbalanced Datasets: An Overview. in *Data Mining and Knowledge Discovery Handbook* 853–867 (2005).
344. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
345. Finkelstein, J. & Jeong, I. cheol. Machine learning approaches to personalize early prediction of asthma exacerbations. *Ann. N. Y. Acad. Sci.* **1387**, 153–165 (2017).
346. Zhang, O., Minku, L. L. & Gonem, S. Detecting asthma exacerbations using daily home monitoring and machine learning. *J. Asthma* (2020). doi:10.1080/02770903.2020.1802746

347. Doshi-Velez, F. & Kim, B. *Towards A Rigorous Science of Interpretable Machine Learning*. (2017).
348. Guidotti, R. *et al.* A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**, 93 (2018).
349. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. Conditional variable importance for random forests. *BMC Bioinformatics* **9**, 307 (2008).
350. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Advances in Neural Informations Processing Systems* (2017).
351. Seliya, N., Khoshgoftaar, T. M. & Van Hulse, J. A study on the relationships of classifier performance metrics. in *IEEE International Conference on Tools with Artificial Intelligence* 59–66 (IEEE, 2009). doi:10.1109/ICTAI.2009.25
352. Brown, J. B. Classifiers and their Metrics Quantified. *Mol. Inform.* **37**, 1700127 (2018).
353. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
354. Luque, A., Carrasco, A., Martín, A. & de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **91**, 216–231 (2019).
355. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **45**, 427–437 (2009).
356. Alaiz-Rodríguez, R., Japkowicz, N. & Tischer, P. Visualizing Classifier Performance on Different Domains. in *2008 20th IEEE International Conference on Tools with Artificial Intelligence. Vol. 2.* (2008).
357. Kouznetsov, A. & Japkowicz, N. Using classifier performance visualization to improve collective ranking techniques for biomedical abstracts classification. in *Canadian Conference on Artificial Intelligence* 299–303 (Springer, Berlin, Heidelberg, 2010). doi:10.1007/978-3-642-13059-5\_33
358. UCI Machine Learning Repository. UCI Machine Learning Repository: default of credit card clients Data Set. Available at: <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>. (Accessed: 2nd August 2020)

359. Harris, D. M. & Harris, S. L. *Digital design and computer architecture, 2nd edition. Digital Design and Computer Architecture, 2nd Edition* (2012).  
doi:10.1016/C2011-0-04377-6
360. UCI Machine Learning Repository. UCI Machine Learning Repository: Poker Hand Data Set. Available at:  
<https://archive.ics.uci.edu/ml/datasets/Poker+Hand>. (Accessed: 2nd August 2020)
361. Pozzolo, A. D., Caelen, O. & Bontempi, G. When is undersampling effective in unbalanced classification tasks? in *Joint european conference on machine learning and knowledge discovery in databases*. 200–215 (2015).  
doi:10.1007/978-3-319-23528-8\_13
362. Briscoe, E. & Feldman, J. Conceptual complexity and the bias/variance tradeoff. *Cognition* **118**, 2–16 (2011).
363. Huang, Y., Li, W., Macheret, F., Gabriel, R. A. & Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Informatics Assoc.* **27**, 621–633 (2020).
364. Martin, A. *et al.* Development and validation of an asthma exacerbation prediction model using electronic health record (EHR) data. *J. Asthma* **57**, 1339–1346 (2020).
365. Xiang, Y. *et al.* Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: Retrospective cohort study. *J. Med. Internet Res.* **22**, e16981 (2020).
366. Yurk, R. A. *et al.* Predicting Patient-Reported Asthma Outcomes for Adults in Managed Care. *Am. J. Manag. Care* **10**, 321–328 (2004).
367. Zein, J. G., Wu, C.-P., Attaway, A. H., Zhang, P. & Nazha, A. Novel machine learning can predict acute asthma exacerbation. *Chest* (2021).  
doi:10.1016/j.chest.2020.12.051
368. Kerr, K. F., Meisner, A., Thiessen-Philbrook, H., Coca, S. G. & Parikh, C. R. RiGoR: Reporting guidelines to address common sources of bias in risk model development. *Biomark. Res.* **3**, 2 (2015).
369. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or

- diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* **13**, 1 (2015).
370. Benchimol, E. I. *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med.* **12**, e1001885 (2015).
371. Kiefer, J. Sequential minimax search for a maximum. in *Proceedings of the American Mathematical Society* (1953). doi:10.2307/2032161
372. Tibble, H. *et al.* Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model. *BMJ Open* **9**, e028375 (2019).
373. Torgo, L. CRAN: Package 'DMwR' (version 0.4.1). <https://cran.r-project.org/web/packages/DMwR/DMwR.pdf> (2015).
374. Hull, S. A. *et al.* Asthma prescribing, ethnicity and risk of hospital admission: An analysis of 35,864 linked primary and secondary care records in East London. *npj Prim. Care Respir. Med.* **26**, 16049 (2016).
375. Pavord, I. D. *et al.* The impact of poor asthma control among asthma patients treated with inhaled corticosteroids plus long-acting  $\beta$  2-agonists in the United Kingdom: A cross-sectional analysis. *npj Prim. Care Respir. Med.* **27**, 17 (2017).
376. Tran, T. N. *et al.* Oral corticosteroid prescription patterns for asthma in France, Germany, Italy and the UK. *Eur. Respir. J.* **55**, 1902363 (2020).
377. Price, C. *et al.* Large care gaps in primary care management of asthma: A longitudinal practice audit. *BMJ Open* **9**, e022506 (2019).
378. Luo, G. *et al.* Using Temporal Features to Provide Data-Driven Clinical Early Warnings for Chronic Obstructive Pulmonary Disease and Asthma Care Management: Protocol for a Secondary Analysis. *JMIR Research Protocols* **8**, (JMIR Research Protocols, 2019).
379. Allen, I. E. & Olkin, I. Estimating time to conduct a meta-analysis from number of citations retrieved. *J. Am. Med. Assoc.* **282**, 634–635 (1999).
380. Borah, R., Brown, A. W., Capers, P. L. & Kaiser, K. A. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* **7**, e012545 (2017).
381. Grant, M. J. & Booth, A. A typology of reviews: An analysis of 14 review types

- and associated methodologies. *Health Info. Libr. J.* **26**, 91–108 (2009).
382. Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **28**, 367–374 (2004).
383. Hansen, R. A. *et al.* Comparison of Methods to Assess Medication Adherence and Classify Nonadherence. *Ann. Pharmacother.* **43**, 413–422 (2009).
384. Romanet-Manent, S., Charpin, D., Magnan, A., Lanteaume, A. & Vervloet, D. Allergic vs nonallergic asthma: what makes the difference? *Allergy* **57**, 607–613 (2002).
385. Primary Care Strategy and NHS Contracts Group. *2006/07 General Medical Services (GMS) contract Quality and Outcomes Framework (QOF)*. (2006).
386. Primary Care Strategy and NHS Contracts Group. *2009/10 General Medical Services (GMS) contract Quality and Outcomes Framework (QOF)*. (2009).
387. Kendrick, T., Stuart, B., Newell, C., Geraghty, A. W. A. & Moore, M. Changes in rates of recorded depression in English primary care 2003-2013: Time trend analyses of effects of the economic recession, and the GP contract quality outcomes framework (QOF). *J. Affect. Disord.* **180**, 68–78 (2015).
388. Sheehan, R. *et al.* Mental illness, challenging behaviour, and psychotropic drug prescribing in people with intellectual disability: UK population based cohort study. *BMJ* **351**, h4326 (2015).
389. Mclintock, K. *et al.* The effects of financial incentives for case finding for depression in patients with diabetes and coronary heart disease: interrupted time series analysis. *BMJ Open* **4**, e005178 (2014).
390. Whiting, P. F. *et al.* How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open* **5**, e008155 (2015).
391. Saposnik, G., Redelmeier, D., Ruff, C. C. & Tobler, P. N. Cognitive biases associated with medical decisions: a systematic review. *BMC Med. Inform. Decis. Mak.* **16**, 138 (2016).
392. Global Initiative for Asthma. *Global Strategy for Asthma Management and Prevention (2020)*. *Global Strategy for Asthma Management and Prevention (2020)*. (2020).
393. National Institute of Health and Care Excellence. *Asthma: diagnosis, monitoring and chronic asthma management NICE guideline*. (2020).

394. Dekker, F. W., Ramspek, C. L. & Van Diepen, M. Most clinical risk scores are useless. *Nephrology Dialysis Transplantation* **32**, 752–755 (2017).
395. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **353**, i2416 (2016).
396. Sutton, R. T. *et al.* An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit. Med.* **3**, 17 (2020).
397. Roshanov, P. S. *et al.* Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. *BMJ* **346**, f657 (2013).
398. Moffat, J. & Green, T. *Clinical Decision Support Tool for Cancer (CDS) Project: Evaluation Report to the Department of Health.* (2014).
399. Hingwala, J. *et al.* Risk-based triage for nephrology referrals using the kidney failure risk equation. *Can. J. Kidney Heal. Dis.* **4**, (2017).
400. Leslie, W. D., Morin, S. & Lix, L. M. A Before-and-After Study of Fracture Risk Reporting and Osteoporosis. *Ann. Intern. Med.* **153**, 580–586 (2010).
401. Samore, M. H. *et al.* Clinical decision support and appropriateness of antimicrobial prescribing: A randomized trial. *J. Am. Med. Assoc.* **294**, 2305–2314 (2005).
402. Smith, J. R. *et al.* The at-risk registers in severe asthma (ARRISA) study: A cluster-randomised controlled trial examining effectiveness and costs in primary care. *Thorax* **67**, 1052–1060 (2012).
403. Smith, J. R. *et al.* At-risk registers integrated into primary care to stop asthma crises in the UK (ARRISA-UK): Study protocol for a pragmatic, cluster randomised trial with nested health economic and process evaluations. *Trials* **19**, 466 (2018).
404. Luo, G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. *Heal. Inf. Sci. Syst.* **4**, 2 (2016).
405. Luo, G., Johnson, M. D., Nkoy, F. L., He, S. & Stone, B. L. Automatically explaining machine learning prediction results on asthma hospital visits in patients with asthma: Secondary analysis. *JMIR Med. Informatics* **8**, e21965 (2020).

406. Thabtah, F. A review of associative classification mining. *Knowl. Eng. Rev.* **22**, 37–65 (2007).
407. Tong, Y., Messinger, A. I. & Luo, G. Testing the Generalizability of an Automated Method for Explaining Machine Learning Predictions on Asthma Patients' Asthma Hospital Visits to an Academic Healthcare System. *IEEE Access* **8**, 195971–195979 (2020).
408. Petkus, H., Hoogewerf, J. & Wyatt, J. C. What do senior physicians think about AI and clinical decision support systems: Quantitative and qualitative analysis of data from specialty societies. *Clin. Med.* **20**, 324–328 (2020).
409. Wright, A. *et al.* Analysis of clinical decision support system malfunctions: A case series and survey. *J. Am. Med. Informatics Assoc.* **23**, 1068–1076 (2016).
410. Greenhalgh, T., Snow, R., Ryan, S., Rees, S. & Salisbury, H. Six 'biases' against patients and carers in evidence-based medicine. *BMC Med.* **13**, 200 (2015).
411. Carroll, C. *et al.* Involving users in the design and usability evaluation of a clinical decision support system. *Comput. Methods Programs Biomed.* **69**, 123–135 (2002).
412. Bonnett, L. J., Snell, K. I. E., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* **365**, k1737 (2019).
413. Cai, C. J., Winter, S., Steiner, D., Wilcox, L. & Terry, M. 'Hello AI': Uncovering the Onboarding Needs of Medical Practitioners for Human–AI Collaborative Decision-Making. in *Proceedings of the ACM on Human-computer Interaction* 104 (2019). doi:10.1145/3359206
414. Henson, K. E. *et al.* Cohort profile: prescriptions dispensed in the community linked to the national cancer registry in England. *BMJ Open* **8**, e20980 (2018).
415. Blais, L. *et al.* Assessing adherence to inhaled corticosteroids in asthma patients using an integrated measure based on primary and secondary adherence. *Eur. J. Clin. Pharmacol.* **73**, 91–97 (2017).
416. Ducharme, F. M. *et al.* Clinical effectiveness of inhaled corticosteroids versus montelukast in children with asthma: prescription patterns and patient adherence as key factors. *Curr. Med. Res. Opin.* **28**, 111–119 (2012).

417. Basch, E. & Snyder, C. Overcoming barriers to integrating patient-reported outcomes in clinical practice and electronic health records. *Ann. Oncol.* **28**, 2332–2333 (2017).
418. Symons, J. D., Ashrafian, H., Dunscombe, R. & Darzi, A. From EHR to PHR: Let's get the record straight. *BMJ Open* **9**, 1–5 (2019).
419. Snyder, C. *et al.* *Users' Guide to Integrating Patient-Reported Outcomes in Electronic Health Records.* (2017).
420. Genes, N. *et al.* From smartphone to EHR: a case report on integrating patient-generated health data. *npj Digit. Med.* **1**, 1–6 (2018).
421. Bell, S. K. *et al.* Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. *JAMA Netw. open* **3**, e205867 (2020).
422. He, H., Bai, Y., Garcia, E. A. & Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. in *Proceedings of the International Joint Conference on Neural Networks* 1322–1328 (2008).  
doi:10.1109/IJCNN.2008.4633969
423. Bhuyan, M. H., Bhattacharyya, D. K. & Kalita, J. K. Network anomaly detection: Methods, systems and tools. *IEEE Commun. Surv. Tutorials* **16**, 303–336 (2014).
424. Crammer, K. & Chechik, G. A needle in a haystack: local one-class optimization. in *Twenty-first international conference on Machine learning - ICMfL '04* 26 (Association for Computing Machinery (ACM), 2004).  
doi:10.1145/1015330.1015399
425. Fernández, A. *et al.* *Learning from Imbalanced Data Sets.* (2018).  
doi:10.1007/978-3-319-98074-4
426. Hussain, Z., Shah, S. A., Mukherjee, M. & Sheikh, A. Predicting the risk of asthma attacks in children, adolescents and adults: protocol for a machine learning algorithm derived from a primary care-based retrospective cohort. *BMJ Open* **10**, e036099 (2020).
427. Hogan, W. R. & Wagner, M. M. Free-text fields change the meaning of coded data. in *Proceedings of the AMIA Annual Fall Symposium* 517–521 (1996).
428. Stein, H. D., Nadkarni, P., Erdos, J. & Miller, P. L. Exploring the degree of concordance of coded and textual data in answering clinical queries from a



- clinical data repository. *J. Am. Med. Informatics Assoc.* **7**, 42–54 (2000).
429. Yang, Y., Ward-Charlerie, S., Dhavle, A. A., Rupp, M. T. & Green, J. Quality and variability of patient directions in electronic prescriptions in the ambulatory care setting. *J. Manag. Care Spec. Pharm.* **24**, 691–699 (2018).
430. Chowdhary, K. R. Natural Language Processing. in *Fundamentals of Artificial Intelligence* 603–649 (Springer India, 2020). doi:10.1007/978-81-322-3972-7\_19
431. Spyns, P. Natural language processing in medicine: An overview. *Methods Inf. Med.* **35**, 285–301 (1996).
432. MacKinlay, A. & Verspoor, K. Extracting structured information from free-text medication prescriptions using dependencies. in *International Conference on Information and Knowledge Management* 35–39 (2012). doi:10.1145/2390068.2390076
433. Shah, A. D. & Martinez, C. An algorithm to derive a numerical daily dose from unstructured text dosage instructions. *Pharmacoepidemiol. Drug Saf.* **15**, 161–166 (2006).
434. Koleck, T. A., Dreisbach, C., Bourne, P. E. & Bakken, S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J. Am. Med. Informatics Assoc.* **26**, 364–379 (2019).
435. Sheikhalishahi, S. *et al.* Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med. Informatics* **7**, e12239 (2019).
436. Silverman, B. W. *Density estimation for statistics and data analysis.* (1986). doi:10.1201/9781315140919
437. Woodcock, A. *et al.* Effectiveness of fluticasone furoate plus vilanterol on asthma control in clinical practice: an open-label, parallel group, randomised controlled trial. *Lancet* **390**, 2247–2255 (2017).
438. Harron, K. L. *et al.* A guide to evaluating linkage quality for the analysis of linked data. *Int. J. Epidemiol.* **46**, 1699–1710 (2017).
439. Wright, M. N., Wager, S. & Probst, P. CRAN: Package ‘ranger’ (version 0.12.1). <https://cran.r-project.org/web/packages/ranger/ranger.pdf> (2020).

440. Chen, T. *et al.* CRAN: Package 'xgboost' (version 1.3.2.1). <https://cran.r-project.org/web/packages/xgboost/xgboost.pdf> (2021).

# Appendices

## Appendix A: Notable Events and Achievements

### Year 1

- 7th March 2018 – Matriculated into the University of Edinburgh
- 5<sup>th</sup> & 6<sup>th</sup> April 2018 – Outreach: *Edinburgh Science Festival* with AUKCAR
- 25th April 2018 – SURE safe researcher qualification awarded
- 5<sup>th</sup> June 2018 – Access approved for the LHS data in the Safe Haven
- 19<sup>th</sup>-22<sup>nd</sup> June 2018 – Hackathon: *EHMAthon* in Budapest. Result: Runner up
- 28<sup>th</sup>-29<sup>th</sup> June 2018 – Conference: *SICSA* in Aberdeen. Won a small hackathon about grant applications
- 29<sup>th</sup> June 2018 – Access approved for the asthma electronic monitoring dataset from UCL (Amy Chan)
- 4<sup>th</sup>-11<sup>th</sup> July 2018 – Hackathon: *EIT European Health Catapult* in Naples
- 21<sup>st</sup> August 2018 – AUKCAR affiliation approved
- 5<sup>th</sup> November 2018 – Presentation: *Centre for Medical Informatics Seminar*
- 29<sup>th</sup> November to 1<sup>st</sup> December 2018 – Conference: European Society for Patient Adherence, Compliance, and Persistence (*ESPACOMP*) in Dublin
- 8<sup>th</sup> February 2019 – Workshop: *Cluster Analysis*
- 21<sup>st</sup> – 24<sup>th</sup> February 2019 – Hackathon: *Product Forge*

### Year 2

- 7<sup>th</sup> March 2019 – SLS NDA Approved by all parties
- 8<sup>th</sup> March 2019 – Risk prediction model protocol Paper accepted to BMJ Open
- 11<sup>th</sup> March 2019 – First Year Review
- 12<sup>th</sup> March 2019 – Conference: *AUKCAR ASM* (poster – 2<sup>nd</sup> prize, Public and Patient Involvement (PPI) presentation, chairing a session)
- 6<sup>th</sup>-10<sup>th</sup> April 2019 – Outreach: *Edinburgh Science Festival Informatics Workshop*
- 15<sup>th</sup> June 2019 – Outreach: *Glasgow Science Festival* AUKCAR stand with paper maché airways
- 18<sup>th</sup> – 19<sup>th</sup> June 2019 – Conference: *SICSA* in Stirling (poster)

- 14<sup>th</sup> – 20<sup>th</sup> July 2019 – Summer School: *Public Health: From small island state to global population* with ACU in Mauritius
- 28<sup>th</sup> – 30<sup>th</sup> October 2019 – Conference: *IEEE BioInformatics and Biomedical Engineering (BIBE)* in Athens (oral presentation)
- 9<sup>th</sup> – 11<sup>th</sup> December 2019 – Conference: *Administrative Data Research* in Cardiff (poster presentation)
- 8<sup>th</sup> January 2020 – Accepted for poster presentation at *ISPE 2020* (subsequently cancelled due to COVID-19)
- 15<sup>th</sup> January 2020 – Conference: *Dealing with Data* in Edinburgh (oral presentation)
- 21<sup>st</sup> January 2020 – Second Year Review

### **Year 3**

- 11<sup>th</sup> February 2020 – Asthma phenotyping methods paper (First author Elsie Horne) accepted into JMIR
- 20<sup>th</sup> – 23<sup>rd</sup> February 2020 – Hackathon: *Product Forge* (Winner in track, emergency and unscheduled care, and winner overall)
- 27<sup>th</sup> February 2020 – Hormonal contraceptives and asthma onset paper (First Author Bright Nwaru) accepted into JACI
- 9<sup>th</sup> March 2020 – Started working from home, as per government guidelines. Currently, no access to ALHS dataset remotely.
- 23<sup>rd</sup> March 2020 – UK enters lockdown for COVID-19
- 26<sup>th</sup> March 2020 – Conference: Virtual AUKCAR ASM (three posters and an oral presentation)
- 30<sup>th</sup> June 2020 - Adherence in asthma and TB comparison paper accepted into BJCP
- 6<sup>th</sup> July 2020 – Remote access for ALHS dataset approved
- 20<sup>th</sup> August 2020 – Clustering of EMD adherence data in children paper accepted into Scientific Reports
- 10<sup>th</sup> September 2020 – 4-month funding extension confirmed
- 15<sup>th</sup> September 2020 – Conference: (Virtual) International Society for Pharmacoepidemiology Meeting (one poster)

- 24<sup>th</sup> September 2020 – Hormonal contraceptives and asthma outcomes paper (first Author Bright Nwaru) accepted into Thorax
- 25<sup>th</sup> November 2020 – Hormone replacement therapy and asthma onset paper (First Author Ahmar Shah) accepted into JACI
- 27<sup>th</sup> November 2020 - Linkage of primary care prescribing records and pharmacy dispensing records in the Salford Lung Study paper accepted into BMC Medical Research Methodology
- 1<sup>st</sup> December 2020 – Took over management of Master's course Medical Informatics
- 26<sup>th</sup> February 2021 - Hormone replacement therapy and asthma outcomes paper (First Author Ahmar Shah) accepted into JACI: In Practice

## Appendix B: Asthma Attack Risk Factor Read Codes (Version 2)

Terms have a maximum character length of 50, and as such may feature truncated expressions.

Smoking Status		
Read Code (V2)	Term	Class
1371.	Never smoked tobacco	Never Smoked
1377.	Ex-trivial smoker (<1/day)	Former Smoker
1378.	Ex-light smoker (1-9/day)	
1379.	Ex-moderate smoker (10-19/day)	
137A.	Ex-heavy smoker (20-39/day)	
137B.	Ex-very heavy smoker (40+/day)	
137F.	Ex-smoker - amount unknown	
137i.	Ex tobacco chewer	
137j.	Ex-cigarette smoker	
137K.	Stopped smoking	
137K0	Recently stopped smoking	
137L.	Current non-smoker	
137l.	Ex roll-up cigarette smoker	
137N.	Ex pipe smoker	
137S.	Ex smoker	
137T.	Date ceased smoking	
1372.	Trivial smoker - < 1 cig/day	
1373.	Light smoker - 1-9 cigs/day	
1374.	Moderate smoker - 10-19 cigs/d	
1375.	Heavy smoker - 20-39 cigs/day	
1376.	Very heavy smoker - 40+cigs/d	
137a.	Pipe tobacco consumption	
137b.	Ready to stop smoking	
137c.	Thinking about stopping smoking	
137C.	Keeps trying to stop smoking	
137d.	Not interested in stopping smoking	
137D.	Admitted tobacco consumption untrue	
137e.	Smoking restarted	
137f.	Reason for restarting smoking	
137G.	Trying to give up smoking	

<b>Smoking Status</b>		
<b>Read Code (V2)</b>	<b>Term</b>	<b>Class</b>
137h.	Minutes from waking to first tobacco consumption	Smoking Current
137H.	Pipe smoker	
137J.	Cigar smoker	
137M.	Rolls own cigarettes	
137M.	Rolls own cigarettes	
137m.	Failed attempt to stop smoking	
137P.	Cigarette smoker	
137Q.	Smoking started	
137R.	Current smoker	
137V.	Smoking reduced	
137..	Smoker - amount smoked	Current Smoker if >0
137E.	Tobacco consumption unknown	Current Smoker if non-missing
137g.	Cigarette pack-years	Current Smoker if >0
137X.	Cigarette consumption	Current Smoker if >0
137Y.	Cigar consumption	Current Smoker if >0
137Z.	Tobacco consumption NOS	Current Smoker if >0

<b>Obesity</b>		
<b>Read Code (V2)</b>	<b>Term</b>	<b>Code Type</b>
22K..	Body Mass Index	BMI – numerical value
22K3.	Body Mass Index low K/M2	BMI Low (Not Obese)
22K6.	Body mass index less than 20	
22K1.	Body Mass Index normal	BMI Normal (Not Obese)
22K8.	Body mass index 20-24 - normal	
22K2.	Body Mass Index high	BMI High (Not Obese)
22K4.	Body mass index index 25-29 - overweight	
22K5.	Body mass index 30+ - obesity	BMI Very High (Obese)
22K7.	Body mass index 40+ - severely obese	
22KC.	Obese Class I	
22KD.	Obese Class II	
22KE.	Obese Class III	



<b>Obesity</b>		
<b>Read Code (V2)</b>	<b>Term</b>	<b>Code Type</b>
229..	Height	Height – numerical value
22A..	Weight	Weight – numerical value

<b>Peak Flow</b>	
<b>Read Code (V2)</b>	<b>Term</b>
339A.	Peak flow rate before bronchodilation
339c.	Peak expiratory flow rate pre steroids

<b>Eosinophilia</b>	
<b>Read Code (V2)</b>	<b>Term</b>
42K..	Eosinophil count

## Appendix C: Illustration of Algorithm Used to Assign British Thoracic Society/Scottish Intercollegiate Guidelines Networks (2019) Treatment Steps

The 2019 BTS/SIGN Guidelines <sup>135</sup> present a single recommended medication dosage for each level of dosage: low, medium, or high. In practice, many regimens did not perfectly align with these guidelines.

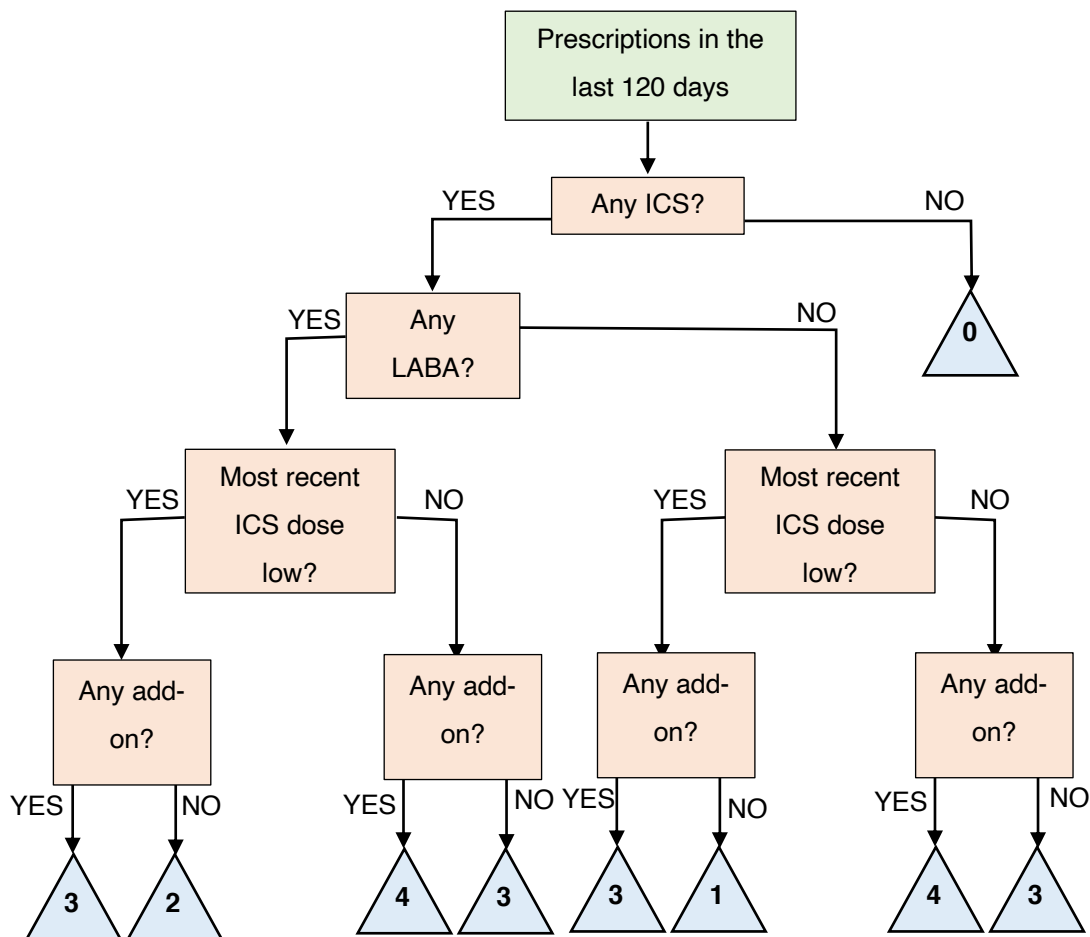
As such, conversion of the continuous ICS and ICS/LABA daily dose into the three levels (low, medium, and high) was based on ranges, accommodating all observed values, as listed in the table below. The range for the low-dose category was zero mcg/day up to the low-dose value in the guidelines. Medium-dose was assigned from one microgram higher than the low-dose value up to the medium-dose value, unless there was no recommended low-dose. In this case, half of the medium-dose value was used as the lower range limit. Similarly, the high-dose category was assigned from one microgram higher than the medium-dose value up to four times the medium-dose value, unless the medium-dose value was missing, in which case half of the high-dose value was used as the lower range limit and twice the high-dose value for the upper range limit. If the medication strength value recorded was above the upper limit of the high-dose category, then the medication strength category 'unknown' was assigned. Generic medications (or those with unlisted brand) will assume the category of the brand name group highlighted by asterisk.

Three Belcometasone inhaler brands (Becodisks, Asmabec, and Pulvinal) were no longer recommended in the 2019 BTS/SIGN guidelines, and so strength categories were taken from the most recent guidelines that they were respectively included in.

ICS Drug	Brand Names	Low Daily ICS Dose	Medium Daily ICS Dose	High Daily ICS Dose
Beclometasone	Clenil, Soprobeq, Becodisks *	0-400mcg	401-800mcg	801-3200mcg
	Qvar, Kelhale, Pulvinal	0-200mcg	201-400mcg	401-1600mcg
	Asmabec	0-200mcg	201-400mcg	N/A
Budesonide	Budelin	N/A	400-800mcg	801-3200mcg
	Pulmicort *	0-400mcg	401-800mcg	801-3200mcg
Fluticasone	Flixotide*	0-200mcg	201-500mcg	501-2000mcg
Mometasone	Asmanex Twisthaler *	0-400mcg	401-800mcg	N/A
Ciclesonide	Alvesco *	0-160mcg	161-320mcg	N/A
Beclometasone + Formoterol	Fostair *	0-200mcg	201-400mcg	401-1600mcg
Budesonide + Formoterol	Symbicort, DuoResp Spiromax *	0-400mcg	401-800mcg	801-3200mcg
	Fobumix	0-320mcg	321-640mcg	641-2560mcg
Fluticasone + Formoterol	Flutiform *	0-200mcg	201-500mcg	501-2000mcg
Fluticasone + Salmeterol	Seretide, Combisal *	0-200mcg	201-500mcg	501-2000mcg
	AirFluSal, Sirdupla, Sereflo, Aloflute, Fusacomb	N/A	250-500mcg	501-2000mcg
	Stalplex	N/A	N/A	500-2000mcg
Fluticasone + Vilanterol	Relvar Ellipta *	N/A	46-92mcg	93-368mcg

The 2019 BTS/SIGN guidelines recommend treating everyone with a minimum of as-needed low dose ICS (Step 1), however, prior to this, as-needed SABA only treatment was common in people with mild intermittent symptoms. As such, we have categorised this as a Step 0, as shown in the decision tree below. There were a negligible number of cases of adults being prescribed ICS solution monotherapy (to be used in nebulisers and similar devices, sometimes used as monotherapy in the paediatric setting), however these were also classed as Step 0. The second

BTS/SIGN treatment step is to add LABA to the ICS, using either a combination or a stand-alone inhaler. At Step 3, ICS may be increased to medium (or high) dose, and an LTRA may be added, with or without continuation of the LABA. I have included other add-on therapies with ICS (LTRA, theophylline, LAMA, MAb, and ICS solutions) into this step, however I have classed those with medium or high dose ICS with LABA and add-on therapies as a Step 4. BTS/SIGN classes Step 4 as the addition of specialist care. Finally, the BTS/SIGN Step 5 includes maintenance OCS treatment, however as it was not possible to identify the indication for OCS prescriptions, this treatment step was disregarded and Step 4 was considered the top level.



## Appendix D: UK Asthma Medication Brand and Generic Names, Formulations, and Medication Strength (Adults)

This table is an update of the classification used previously by Mukherjee et al. <sup>160</sup>. The updates include the addition of new brands added to the British National Formulary (brand highlighted in bold) and new therapies approved by NICE (also highlighted in bold, with NICE technical appraisal and evidence summary identifiers included). Additionally, bambuterol was corrected from being listed as SABA to LABA. The formulations and dosages approved for asthma treatment in adults were extracted from the British National Formulary on April 10<sup>th</sup>, 2020 or sourced from previous versions for medications which are no longer recommended.

Drug Type	Ingredients	Brand Names	Formulation	Medication Strength
SABA	Salbutamol / Albuterol	<i>Generic</i>	Tablet	2mg, 4mg
			Oral Solution	2mg/5ml
			Pressurised Inhaler	100mcg
			Inhalation Powder	100mcg, 200mcg
			Nebulising Solution	2.5mg/2.5ml, 5mg/2.5ml
		<b>Salamol</b>	Pressurised Inhaler	100mcg
			Nebulising Solution	5mg/2.5ml
		<b>Ventolin</b>	Infusion Ampoules	5mg/5ml
			Injection	500mcg/1ml
			Oral Solution	2mg/5ml
			Pressurised Inhaler	100mcg
			Inhalation Powder	200mcg
			Nebules	2.5mg, 5mg

Drug Type	Ingredients	Brand Names	Formulation	Medication Strength
SABA	Salbutamol / Albuterol	Ventolin	Nebulising Solution	5mg/1ml
		Airomir	Pressurised Inhaler	100mcg
		Salbulin	Inhalation Powder	100mcg
		AirSalb	Pressurised Inhaler	100mcg
		Ventmax	Capsule	4mg, 8mg
		Asmasal	Inhalation Powder	95mcg
		Pulvinal Salbutamol	Inhalation Powder	200mcg
LABA	Bambuterol	Bambec	Tablet	10mg
	Formoterol	<i>Generic</i>	Inhalation Powder	12mcg
		Atimos	Pressurised Inhaler	12mcg
		Foradil	Inhalation Powder	12mcg
		Oxis	Inhalation Powder	6mcg, 12mcg
	Salmeterol	Neovent	Pressurised Inhaler	25mcg
		Serevent	Pressurised Inhaler	25mcg
			Inhalation Powder	50mcg
	Terbutaline	Bricanyl	Tablet	5mg
			Injection	2.5mg/5ml, 500mcg/1ml
			Inhalation Powder	500mcg
			Nebulising Solution	5mg/2ml
	<b>Tiotropium [ESNM55]</b>	<b>Spiriva Respimat</b>	Pressurised Inhaler	2.5mg
LAMA	Ipratropium	<i>Generic</i>	Nebulising Solution	250mcg/1ml, 500mcg/2ml
		Atrovent	Pressurised Inhaler	20mcg

Drug Type	Ingredients	Brand Names	Formulation	Medication Strength
LAMA	Ipratropium	Atrovent	Nebulising Solution	250mcg/1ml, 500mcg/2ml
		<b>Inhalvent</b>	Pressurised Inhaler	20mcg
		<b>Ipravent</b>	Pressurised Inhaler	20mcg
		Respontin	Nebulising Solution	250mcg/1ml, 500mcg/2ml
LAMA + LABA	Ipratropium + Salbutamol	Ipramol	Nebulising Solution	(200mcg + 1mg) / 1ml
		Combivent	Nebulising Solution	(200mcg + 1mg) / 1ml
Theophylline	Theophylline	Uniphyllin	Tablet	200mcg, 300mcg, 400mcg
		Nuelin	Tablet	175 mg, 250 mg
		Slo-Phyllin	Tablet	60mg, 125mg, 250mg
	Aminophylline	Phyllocontin	Tablet	225mg, 350mg
			Injection	250mg/10ml
ICS	Beclometasone / Beclomethasone	<i>Generic</i>	Inhalation Powder	200mcg
		Clenil	Pressurised Inhaler	50mcg, 100mcg, 200mcg, 250mcg
		Qvar	Pressurised Inhaler	50mcg, 100mcg
		<b>Kelhale</b>	Pressurised Inhaler	50mcg, 100mcg
		<b>Soprobe</b>	Pressurised Inhaler	50mcg, 100mcg, 200mcg, 250mcg
		Becodisks	Inhalation Powder	100mcg, 200mcg, 400mcg

Notes: Dosage for combination LAMA+LABA medications are listed in the same order as the ingredients

Drug Type	Ingredients	Brand Names	Formulation	Medication Strength
ICS	Beclometasone / Beclomethasone	Pulvinal	Inhalation Powder	100mcg, 200mcg, 400mcg
		Asmabec	Pressurised Inhaler	100mcg, 250mcg
	Budesonide	<i>Generic</i>	Nebulising Solution	250mcg/2ml, 500mcg/2ml, 1mg/2ml
			Inhalation Powder	100mcg, 200mcg, 400mcg
		Budelin	Inhalation Powder	200mcg
		Pulmicort	Inhalation Powder	100mcg, 200mcg, 400mcg
			Respules	0.5mg, 1mg
	Fluticasone	Flixotide	Pressurised Inhaler	50mcg, 125mcg, 250mcg
			Inhalation Powder	50mcg, 100mcg, 250mcg, 500mcg
			Nebules	0.5mg/2ml, 2mg/2ml
	Mometasone	Asmanex / Twisthaler	Inhalation Powder	200mcg, 400mcg
	Ciclesonide	Alvesco	Pressurised Inhaler	80mcg, 160mcg
ICS + LABA	Beclometasone + Formoterol	Fostair	Pressurised Inhaler	100mcg+6mcg, 200mcg+6mcg
			Inhalation Powder	100mcg+6mcg, 200mcg+6mcg

Notes: Dosage for combination ICS+LABA medications are listed in the same order as the ingredients



Drug Type	Ingredients	Brand Names	Formulation	Medication Strength
ICS + LABA	Budesonide + Formoterol	Symbicort	Pressurised Inhaler	200mcg+6mcg
			Inhalation Powder	100mcg+6mcg, 200mcg+6mcg, 400mcg+12mcg
		<b>DuoResp Spiromax</b>	Inhalation Powder	160mcg+4.5mcg, 320mcg+12mcg
		<b>Fobumix</b>	Inhalation Powder	50mcg+4.5mcg, 160mcg+4.5mcg 320mcg+9mcg
	Fluticasone + Formoterol	Flutiform	Pressurised Inhaler	50mcg+5mcg, 125mcg+5mcg, 250mcg+10mcg
			Pressurised Inhaler	50mcg+25mcg, 125mcg+25mcg, 250mcg+25mcg
	Fluticasone + Salmeterol	Seretide	Inhalation Powder	100mcg+50mcg, 250mcg+50mcg, 500mcg+50mcg
			Pressurised Inhaler	125mcg+25mcg, 250mcg+25mcg
		Airflusal	Pressurised Inhaler	125mcg+25mcg, 250mcg+25mcg
			Inhalation Powder	500mcg+50mcg
<b>Sirdupla</b>		Pressurised Inhaler	250mcg+25mcg	

Notes: Dosage for combination ICS+LABA medications are listed in the same order as the ingredients

Drug Type	Ingredients	Brand Names	Formulation	Medication Strength
ICS + LABA	Fluticasone + Salmeterol	<b>Sereflo</b>	Pressurised Inhaler	125mcg+25mcg, 250mcg+25mcg
		<b>Aloflute</b>	Pressurised Inhaler	125mcg+25mcg, 250mcg+25mcg
		<b>Combisal</b>	Pressurised Inhaler	50mcg+25mcg, 125mcg+25mcg, 250mcg+25mcg
		<b>Fusacomb</b>	Inhalation Powder	100mcg+50mcg, 500mcg+50mcg
		<b>Stalpex</b>	Inhalation Powder	500mcg+50mcg
	Fluticasone + Vilanterol	Relvar Ellipta	Inhalation Powder	92mcg+22mcg, 184mcg+22mcg
LTRA	Montelukast	<i>Generic</i>	Tablet	4mg, 5mg, 10mg
			Sachet for Solution	4mg
		Singulair	Tablet	5mg, 10mg
			Sachet for Solution	4mg
	Zafirlukast	<i>Generic</i>	Tablet	10mg, 20mg
		Accolate	Tablet	10mg, 20mg
	Cromolyn / Sodium Cromoglicate	Intal	Pressurised Inhaler	5mg
	Nedocromil	Tilade	Pressurised Inhaler	2mg
Steroid	Prednisolone	<i>Generic</i>	Tablet	1mg, 2.5mg, 5mg, 10mg, 20mg, 25mg, 30mg

Notes: Dosage for combination ICS+LABA medications are listed in the same order as the ingredients

LTRA category includes cromoglicates and related therapies

<b>Drug Type</b>	<b>Ingredients</b>	<b>Brand Names</b>	<b>Formulation</b>	<b>Medication Strength</b>
Steroid	Prednisolone	<i>Generic</i>	Oral Solution	5mg/5ml, 10mg/1ml
		Deltacortril	Tablet	2.5mg, 5mg
		<b>Dilacort</b>	Tablet	2.5mg, 5mg
		Deltastab	Injection	25mg/1ml
		<b>Pevanti</b>	Tablet	2.5mg, 5mg, 10mg, 20mg, 25mg
MAb	Omalizumab	Xolair	Injection	75mg/0.5ml, 150mg/1ml
	<b>Mepolizumab [TA431]</b>	<b>Nucala</b>	Injection	100mg/1ml
	<b>Benralizumab [TA565]</b>	<b>Fasenra</b>	Injection	30mg/1ml
	<b>Reslizumab [TA479]</b>	<b>Cinqaero</b>	Infusion Solution	25mg/2.5ml, 100mg/10ml

## Appendix E: Asthma Diagnosis and Management Read Codes (Version 2)

Terms have a maximum character length of 50, and as such may feature truncated expressions.

Read Code (V2)	Term
173A.	Exercise induced asthma
H3120	Chronic asthmatic bronchitis
H33..	Asthma
H330.	Extrinsic (atopic) asthma
H3300	Extrinsic asthma without status asthmaticus
H3301	Extrinsic asthma with status asthmaticus
H330z	Extrinsic asthma NOS
H331.	Intrinsic asthma
H3310	Intrinsic asthma without status asthmaticus
H3311	Intrinsic asthma with status asthmaticus
H331z	Intrinsic asthma NOS
H332.	Mixed asthma
H334.	Brittle asthma
H335.	Chronic asthma with fixed airflow obstruction
H33z.	Asthma unspecified
H33z0	Status asthmaticus NOS
H33z1	Asthma attack
H33z2	Late-onset asthma
H33zz	Asthma NOS
H3B..	Asthma-chronic obstructive pulmonary disease overlap syndrome
663..	Respiratory disease monitoring
6632.	Follow-up respiratory assessment
6636.	Inhaler technique shown
6637.	Inhaler technique observed
663a.	Oral steroids used since last appointment
663B.	Resp. treatment changed
663d.	Emergency asthma admission since last appointment
663e.	Asthma restricts exercise
663e0	Asthma sometimes restricts exercise
663e1	Asthma severely restricts exercise
663F.	Oral steroids started

<b>Read Code (V2)</b>	<b>Term</b>
663f.	Asthma never restricts exercise
663G.	Oral steroids stopped
663g.	Inhaled steroids use
663g0	Not using inhaled steroids
663g1	Using inhaled steroids - normal dose
663g2	Using inhaled steroids - high dose
663g3	Increases inhaled steroids appropriately
663H.	Inhaler technique - good
663h.	Asthma - currently dormant
663I.	Inhaler technique - poor
663J.	Airways obstruction reversible
663j.	Asthma - currently active
663L.	Bronchodilators used more than once daily
663M.	Bronchodilators used a maximum of once daily
663m.	Asthma accident and emergency attendance since last visit
663N.	Asthma disturbing sleep
663n.	Asthma treatment compliance satisfactory
663N0	Asthma causing night waking
663N1	Asthma disturbs sleep weekly
663N2	Asthma disturbs sleep frequently
663O.	Asthma not disturbing sleep
663O0	Asthma never disturbs sleep
663P.	Asthma limiting activities
663p.	Asthma treatment compliance unsatisfactory
663Q.	Asthma not limiting activities
663q.	Asthma daytime symptoms
663R.	Service of nebuliser
663r.	Asthma causes night symptoms 1 to 2 times per month
663S.	Peak flow meter at home
663s.	Asthma never causes daytime symptoms
663T.	No peak flow meter at home
663t.	Asthma causes daytime symptoms 1 to 2 times per month
663U.	Asthma management plan given
663u.	Asthma causes daytime symptoms 1 to 2 times per week
663V.	Asthma severity
663v.	Asthma causes daytime symptoms most days
663V0	Occasional asthma

<b>Read Code (V2)</b>	<b>Term</b>
663V1	Mild asthma
663V2	Moderate asthma
663V3	Severe asthma
663W.	Asthma prophylactic medication used
663w.	Asthma limits walking up hills or stairs
663X.	Irritable airways
663x.	Asthma limits walking on the flat
663Y.	Steroid dose inhaled daily
663y.	Number of asthma exacerbations in past year
663Z.	Resp. disease monitoring NOS
663z.	Number of times bronchodilator used in one week

## Appendix F: Asthma Primary Care Encounter Read Codes (Version 2)

Primary care encounters relating to asthma were identified by the presence of any Read Codes used for asthma diagnosis or management (Appendix E), and any of the following additional asthma encounter codes. These codes were not seemed sufficient in isolation to indicate a diagnosis, but assuming that a diagnosis had been confirmed could be assumed to be relating to asthma management. Terms have a maximum character length of 50, and as such may feature truncated expressions.

Read Code (V2)	Term
173c.	Occupational asthma
173d.	Work aggravated asthma
178..	Asthma trigger
1O2..	Asthma confirmed
388t.	Royal College of Physicians asthma assessment
66Y0.	Number of times bronchodilator used in 24 hours
66Y1.	Peak expiratory flow rate - technique poor
66Y2.	Peak expiratory flow rate - technique moderate
66Y3.	Peak expiratory flow rate - technique good
66Y4.	Inhaler technique - moderate
66Y5.	Change in asthma management plan
66Y6.	Peak expiratory flow rate - compliance good
66Y7.	Peak expiratory flow rate - compliance moderate
66Y8.	Peak expiratory flow rate - compliance poor
66Y9.	Step up change in asthma management plan
66YA.	Step down change in asthma management plan
66Ya.	Reversibility trial by bronchodilator
66Yb.	Reversibility trial by anticholinergic
66YC.	Absent from work or school due to asthma
66Yc.	Number of cons days less than 80% peak expiratory flow rate
66YE.	Asthma monitoring due
66YF.	Nebulizer technique good
66YG.	Nebulizer technique poor
66YJ.	Asthma annual review
66YK.	Asthma follow-up
66Ym.	Inhaler device in use
66YN.	Peak expiratory flow rate compliance

<b>Read Code (V2)</b>	<b>Term</b>
66YO.	Peak expiratory flow rate technique
66YP.	Asthma night-time symptoms
66YQ.	Asthma monitoring by nurse
66YR.	Asthma monitoring by doctor
66YV.	Does not use spacer device
66YW.	No nebulisation since last appointment
66YX.	Peak expiratory flow rate monitoring
66YY.	Peak expiratory flow rate monitoring using diary
66YZ.	Does not have asthma management plan
679J.	Health education - asthma
8B3j.	Asthma medication review
8CE2.	Asthma leaflet given
8CR0.	Asthma clinical management plan
8HTT.	Referral to asthma clinic
9N1d.	Seen in asthma clinic
9N18.	Asthma outreach clinic
9OJ..	Asthma monitoring admin.
9OJ1.	Attends asthma monitoring
9OJA.	Asthma monitoring check done



## Appendix G: Comorbidity Read Codes (Version 2)

Terms have a maximum character length of 50, and as such may feature truncated expressions.

<b>Nasal Polyps</b>	
<b>Read Code (V2)</b>	<b>Term</b>
H11..	Nasal Polyps

<b>Anaphylaxis</b>	
<b>Read Code (V2)</b>	<b>Term</b>
SN50.	Anaphylactic shock
SN500	Anaphylactic shock due to adverse food reaction
SN501	Anaphylactic shock due to adverse effect of correct drug or medicament properly administered
SN59.	Allergic reaction to venom
SP34.	Anaphylactic shock due to serum

<b>Rhinitis</b>	
<b>Read Code (V2)</b>	<b>Term</b>
H17..	Allergic rhinitis
H170.	Allergic rhinitis due to pollens
H171.	Allergic rhinitis due to other allergens
H1710	Allergy to animal
H172.	Allergic rhinitis due to unspecified allergen
H17z.	Allergic rhinitis NOS
H18..	Vasomotor rhinitis
Hyu21	Other allergic rhinitis

<b>Eczema</b>	
<b>Read Code (V2)</b>	<b>Term</b>
M11..	Atopic dermatitis and related conditions
M111.	Atopic dermatitis/eczema
M112.	Infantile eczema
M113.	Flexural eczema
M114.	Allergic (intrinsic) eczema
M11z.	Atopic dermatitis NOS
M12z0	Dermatitis NOS
M12z1	Eczema NOS

<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
8G94.	Anxiety Management Training
E2...	Neurotic; Personality And Other Nonpsychotic Disorders
E20..	Neurotic Disorders
E200.	Anxiety States
E2000	Anxiety State Unspecified
E2001	Panic Disorder
E2002	Generalised Anxiety Disorder
E2003	Anxiety With Depression
E2004	Chronic Anxiety
E2005	Recurrent Anxiety
E200z	Anxiety State NOS
E201.	Hysteria
E2010	Hysteria Unspecified
E2011	Hysterical Blindness
E2012	Hysterical Deafness
E2013	Hysterical Tremor
E2014	Hysterical Paralysis
E2015	Hysterical Seizures
E2016	Other Conversion Disorder
E2017	Hysterical Amnesia
E2018	Hysterical Fugue
E2019	Multiple Personality
E201A	Dissociative Reaction Unspecified
E201B	Compensation Neurosis
E201C	Phantom Pregnancy
E201z	Hysteria Nos
E202.	Phobic Disorders
E2020	Phobia Unspecified
E2021	Agoraphobia With Panic Attacks
E2022	Agoraphobia Without Mention Of Panic Attacks
E2023	Social Phobia, Fear Of Eating In Public
E2024	Social Phobia, Fear Of Public Speaking
E2025	Social Phobia, Fear Of Public Washing
E2026	Acrophobia
E2027	Animal Phobia
E2028	Claustrophobia
E2029	Fear Of Crowds
E202A	Fear Of Flying

<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
E202B	Cancer Phobia
E202C	Dental Phobia
E202D	Fear Of Death
E202E	Fear Of Pregnancy
E202z	Phobic Disorder NOS
E203.	Obsessive-Compulsive Disorders
E2030	Compulsive Neurosis
E2031	Obsessional Neurosis
E203z	Obsessive-Compulsive Disorder Nos
E205.	Neurasthenia - Nervous Debility
E206.	Depersonalisation Syndrome
E207.	Hypochondriasis
E20y.	Other Neurotic Disorders
E20y0	Somatization Disorder
E20y1	Writer's Cramp Neurosis
E20y2	Other Occupational Neurosis
E20y3	Psychasthenic Neurosis
E20yz	Other Neurotic Disorder NOS
E20z.	Neurotic Disorder Nos
E21..	Personality Disorders
E210.	Paranoid Personality Disorder
E211.	Affective Personality Disorder
E2110	Unspecified Affective Personality Disorder
E2111	Hypomanic Personality Disorder
E2112	Depressive Personality Disorder
E2113	Cyclothymic Personality Disorder
E211z	Affective Personality Disorder NOS
E26..	Physiological Malfunction Arising From Mental Factors
E260.	Psychogenic Musculoskeletal Symptoms
E2600	Psychogenic Paralysis
E2601	Psychogenic Torticollis
E260z	Psychogenic Musculoskeletal Symptoms Nos
E261.	Psychogenic Respiratory Symptoms
E2610	Psychogenic Air Hunger
E2611	Psychogenic Cough
E2612	Psychogenic Hiccough
E2613	Psychogenic Hyperventilation
E2614	Psychogenic Yawning

<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
E2615	Psychogenic Aphonia
E261z	Psychogenic Respiratory Symptom NOS
E262.	Psychogenic Cardiovascular Symptoms
E2620	Cardiac Neurosis
E2621	Cardiovascular Neurosis
E2622	Neurocirculatory Asthenia
E2623	Psychogenic Cardiovascular Disorder
E262z	Psychogenic Cardiovascular Symptom Nos
E263.	Psychogenic Skin Symptoms
E2630	Psychogenic Pruritus
E263z	Psychogenic Skin Symptoms Nos
E264.	Psychogenic Gastrointestinal Tract Symptoms
E2640	Psychogenic Aerophagy
E2642	Cyclical Vomiting - Psychogenic
E2643	Psychogenic Diarrhoea
E2644	Psychogenic Dyspepsia
E2645	Psychogenic Constipation
E264z	Psychogenic Gastrointestinal Tract Symptom Nos
E265.	Psychogenic Genitourinary Tract Symptoms
E2650	Psychogenic Genitourinary Tract Malfunction Unspecified
E2651	Psychogenic Vaginismus
E2652	Psychogenic Dysmenorrhea
E2653	Psychogenic Dysuria
E265z	Psychogenic Genitourinary Tract Symptom NOS
E266.	Psychogenic Endocrine Malfunction
E267.	Psychogenic Symptom Of Special Sense Organ
E26y.	Other Psychogenic Malfunction
E26y0	Bruxism (Teeth Grinding)
E26yz	Other Psychogenic Malfunction NOS
E26z.	Psychosomatic Disorder
E278.	Psychalgia
E2780	Psychogenic Pain Unspecified
E2781	Tension Headache
E2782	Psychogenic Backache
E278z	Psychalgia Nos
E28..	Acute Reaction To Stress
E280.	Acute Panic State Due To Acute Stress Reaction
E281.	Acute Fugue State Due To Acute <i>Stress</i> Reaction

<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
E283.	Other Acute <i>Stress</i> Reactions
E2830	Acute Situational Disturbance
E2831	Acute Posttrauma <i>Stress</i> State
E283z	Other Acute Stress Reaction NOS
E284.	Stress Reaction Causing Mixed Disturbance Of Emotion/Conduct
E28z.	Acute <i>Stress</i> Reaction NOS
E29..	Adjustment Reaction
E2900	Bereavement Reaction
E292.	Adjustment Reaction, Predominant Disturbance Other Emotions
E2920	Separation Anxiety Disorder
E2921	Adolescent Emancipation Disorder
E2922	Early Adult Emancipation Disorder
E2923	Specific Academic Or Work Inhibition
E2924	Adjustment Reaction With Anxious Mood
E2925	Culture Shock
E292y	Adjustment Reaction With Mixed Disturbance Of Emotion
E292z	Adjustment Reaction With Disturbance Of Other Emotion NOS
E293.	Adjustment Reaction With Predominant Disturbance Of Conduct
E2930	Adjustment Reaction With Aggression
E2931	Adjustment Reaction With Antisocial Behaviour
E2932	Adjustment Reaction With Destructiveness
E293z	Adjustment Reaction With Predominant Disturbance Conduct NOS
E294.	Adjustment Reaction With Disturbance Emotion And Conduct
E29y.	Other Adjustment Reactions
E29y0	Concentration Camp Syndrome
E29y1	Other Post-Traumatic Stress Disorder
E29y2	Adjustment Reaction With Physical Symptoms
E29y3	Elective Mutism Due To An Adjustment Reaction
E29y4	Adjustment Reaction Due To Hospitalisation
E29y5	Other Adjustment Reaction With Withdrawal
E29yz	Other Adjustment Reactions Nos
E29z.	Adjustment Reaction Nos
Eu4..	[X]Neurotic; Stress - Related And Somofrom Disorders

<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
Eu40.	[X]Phobic Anxiety Disorders
Eu400	[X]Agoraphobia
Eu401	[X]Social Phobias
Eu402	[X]Specific (Isolated) Phobias
Eu403	[X]Needle Phobia
Eu40y	[X]Other Phobic Anxiety Disorders
Eu40z	[X]Phobic Anxiety Disorder, Unspecified
Eu41.	[X]Other Anxiety Disorders
Eu410	[X]Panic Disorder [Episodic Paroxysmal Anxiety]
Eu411	[X]Generalized Anxiety Disorder
Eu412	[X]Mixed Anxiety And Depressive Disorder
Eu413	[X]Other Mixed Anxiety Disorders
Eu41y	[X]Anxiety Hysteria
Eu41z	[X]Anxiety Nos
Eu42.	[X]Obsessive - Compulsive Disorder
Eu420	[X]Predominantly Obsessional Thoughts Or Ruminations
Eu421	[X]Predominantly Compulsive Acts [Obsessional Rituals]
Eu422	[X]Mixed Obsessional Thoughts And Acts
Eu42y	[X]Other Obsessive-Compulsive Disorders
Eu42z	[X]Obsessive-Compulsive Disorder; Unspecified
Eu43.	[X]Reaction To Severe Stress; And Adjustment Disorders
Eu430	[X]Acute Stress Reaction
Eu431	[X]Post - Traumatic Stress Disorder
Eu432	[X]Adjustment Disorders
Eu43y	[X]Other Reactions To Severe Stress
Eu43z	[X]Reaction To Severe Stress; Unspecified
Eu44.	[X]Dissociative [Conversion] Disorders
Eu440	[X]Dissociative Amnesia
Eu441	[X]Dissociative Fugue
Eu442	[X]Dissociative Stupor
Eu443	[X]Trance And Possession Disorders
Eu444	[X]Dissociative Motor Disorders
Eu445	[X]Dissociative Convulsions
Eu446	[X]Dissociative Anaesthesia And Sensory Loss
Eu447	[X]Mixed Dissociative [Conversion] Disorders
Eu44y	[X]Other Dissociative [Conversion] Disorders
Eu44z	[X]Dissociative [Conversion] Disorder; Unspecified
Eu45.	[X]Somatoform Disorders

<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
Eu450	[X]Somatization Disorder
Eu451	[X]Undifferentiated Somatoform Disorder
Eu452	[X]Hypochondriacal Disorder
Eu453	[X]Somatoform Autonomic Dysfunction
Eu454	[X]Persistent Somatoform Pain Disorder
Eu455	[X]Globus Pharyngeus
Eu45y	[X]Other Somatoform Disorders
Eu45z	[X]Somatoform Disorder; Unspecified
Eu46.	[X]Other Neurotic Disorders
Eu460	[X]Neurasthenia
Eu461	[X]Depersonalization - Derealization Syndrome
Eu46y	[X]Other Specified Neurotic Disorders
Eu46z	[X]Neurotic Disorder; Unspecified
ZN114	Anxiety Management
ZS7C7	Post-Traumatic Mutism.
1B17.	Depressed
62T1.	Puerperal Depression
6G00.	Postnatal Depression Counselling
8CAa.	Patient Given Advice About Management Of Depression
9H90.	Depression Annual Review
9H91.	Depression Medication Review
9H92.	Depression Interim Review
E03y2	Organic Affective Syndrome
E03y3	Unspecified Puerperal Psychosis
E11..	Depressive Psychoses
E112.	Single Major Depressive Episode
E1120	Single Major Depressive Episode, Unspecified
E1121	Single Major Depressive Episode, Mild
E1122	Single Major Depressive Episode, Moderate
E1123	Single Major Depressive Episode, Severe, Without Psychosis
E1124	Single Major Depressive Episode, Severe, With Psychosis
E1125	Single Major Depressive Episode, Partial Or Unspec Remission
E1126	Single Major Depressive Episode, In Full Remission
E112z	Single Major Depressive Episode NOS
E113.	Recurrent Major Depressive Episode
E1130	Recurrent Major Depressive Episodes, Unspecified

<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
E1131	Recurrent Major Depressive Episodes, Mild
E1132	Recurrent Major Depressive Episodes, Moderate
E1133	Recurrent Major Depressive Episodes, Severe, No Psychosis
E1134	Recurrent Major Depressive Episodes, Severe, With Psychosis
E1135	Recurrent Major Depressive Episodes, Partial/Unspec Remission
E1136	Recurrent Major Depressive Episodes, In Full Remission
E1137	Recurrent Depression
E113z	Recurrent Major Depressive Episode NOS
E118.	Seasonal Affective Disorder
E11y2	Atypical Depressive Disorder
E11y3	Other Mixed Manic- <i>Depressive</i> Psychoses
E11yz	Other And Unspecified Manic- <i>Depressive</i> Psychoses NOS
E11z.	Other And Unspecified Affective Psychoses
E11z0	Unspecified Affective Psychoses NOS
E11z1	Rebound Mood Swings
E11z2	Masked Depression
E11zz	Other Affective Psychosis NOS
E135.	Agitated Depression
E204.	Neurotic Depression Reactive Type
E290.	Brief Depressive Reaction
E290z	Brief Depressive Reaction NOS
E291.	Prolonged Depressive Reaction
E2B..	Depressive Disorder NEC
E2B0.	Postviral Depression
E2B1.	Chronic Depression
Eu3..	[X]Mood - Affective Disorders
Eu32.	[X]Depressive Episode
Eu320	[X]Mild Depressive Episode
Eu321	[X]Moderate Depressive Episode
Eu322	[X]Severe Depressive Episode Without Psychotic Symptoms
Eu324	[X]Mild Depression
Eu32y	[X]Other Depressive Episodes
Eu32z	[X]Depressive Episode, Unspecified
Eu33.	[X]Recurrent Depressive Disorder
Eu330	[X]Recurrent Depressive Disorder, Current Episode Mild



<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
Eu450	[X]Somatization Disorder
Eu451	[X]Undifferentiated Somatoform Disorder
Eu452	[X]Hypochondriacal Disorder
Eu453	[X]Somatoform Autonomic Dysfunction
Eu454	[X]Persistent Somatoform Pain Disorder
Eu455	[X]Globus Pharyngeus
Eu45y	[X]Other Somatoform Disorders
Eu45z	[X]Somatoform Disorder; Unspecified
Eu46.	[X]Other Neurotic Disorders
Eu460	[X]Neurasthenia
Eu461	[X]Depersonalization - Derealization Syndrome
Eu46y	[X]Other Specified Neurotic Disorders
Eu46z	[X]Neurotic Disorder; Unspecified
ZN114	Anxiety Management
ZS7C7	Post-Traumatic Mutism.
1B17.	Depressed
62T1.	Puerperal Depression
6G00.	Postnatal Depression Counselling
8CAa.	Patient Given Advice About Management Of Depression
9H90.	Depression Annual Review
9H91.	Depression Medication Review
9H92.	Depression Interim Review
E03y2	Organic Affective Syndrome
E03y3	Unspecified Puerperal Psychosis
E11..	Depressive Psychoses
E112.	Single Major Depressive Episode
E1120	Single Major Depressive Episode, Unspecified
E1121	Single Major Depressive Episode, Mild
E1122	Single Major Depressive Episode, Moderate
E1123	Single Major Depressive Episode, Severe, Without Psychosis
E1124	Single Major Depressive Episode, Severe, With Psychosis
E1125	Single Major Depressive Episode, Partial Or Unspec Remission
E1126	Single Major Depressive Episode, In Full Remission
E112z	Single Major Depressive Episode NOS
E113.	Recurrent Major Depressive Episode
E1130	Recurrent Major Depressive Episodes, Unspecified

<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
E1131	Recurrent Major Depressive Episodes, Mild
E1132	Recurrent Major Depressive Episodes, Moderate
E1133	Recurrent Major Depressive Episodes, Severe, No Psychosis
E1134	Recurrent Major Depressive Episodes, Severe, With Psychosis
E1135	Recurrent Major Depressive Episodes, Partial/Unspec Remission
E1136	Recurrent Major Depressive Episodes, In Full Remission
E1137	Recurrent Depression
E113z	Recurrent Major Depressive Episode NOS
E118.	Seasonal Affective Disorder
E11y2	Atypical Depressive Disorder
E11y3	Other Mixed Manic- <i>Depressive</i> Psychoses
E11yz	Other And Unspecified Manic- <i>Depressive</i> Psychoses NOS
E11z.	Other And Unspecified Affective Psychoses
E11z0	Unspecified Affective Psychoses NOS
E11z1	Rebound Mood Swings
E11z2	Masked Depression
E11zz	Other Affective Psychosis NOS
E135.	Agitated Depression
E204.	Neurotic Depression Reactive Type
E290.	Brief Depressive Reaction
E290z	Brief Depressive Reaction NOS
E291.	Prolonged Depressive Reaction
E2B..	Depressive Disorder NEC
E2B0.	Postviral Depression
E2B1.	Chronic Depression
Eu3..	[X]Mood - Affective Disorders
Eu32.	[X]Depressive Episode
Eu320	[X]Mild Depressive Episode
Eu321	[X]Moderate Depressive Episode
Eu322	[X]Severe Depressive Episode Without Psychotic Symptoms
Eu324	[X]Mild Depression
Eu32y	[X]Other Depressive Episodes
Eu32z	[X]Depressive Episode, Unspecified
Eu33.	[X]Recurrent Depressive Disorder
Eu330	[X]Recurrent Depressive Disorder, Current Episode Mild

<b>Anxiety or Depression</b>	
<b>Read Code (V2)</b>	<b>Term</b>
Eu331	[X]Recurrent Depressive Disorder, Current Episode Moderate
Eu332	[X]Recurr Depress Disorder Cur Epi Severe Without Psyc Symp
Eu334	[X]Recurrent Depressive Disorder, Currently In Remission
Eu33y	[X]Other Recurrent Depressive Disorders
Eu33z	[X]Recurrent Depressive Disorder, Unspecified
Eu34.	[X]Persistent Mood Affective Disorders
Eu340	[X]Cyclothymia
Eu341	[X]Dysthymia
Eu34y	[X]Other Persistent Mood Affective Disorders
Eu34z	[X]Persistent Mood Affective Disorder, Unspecified
Eu3y.	[X]Other Mood Affective Disorders
Eu3y0	[X]Other Single Mood Affective Disorders
Eu3y1	[X]Recurrent Brief Depressive Episodes
Eu3yy	[X]Other Specified Mood Affective Disorders
Eu3z.	[X]Unspecified Mood Affective Disorder

<b>GERD</b>	
<b>Read Code (V2)</b>	<b>Term</b>
J101.	Oesophagitis
J10y4	Oesophageal reflux without mention of oesophagitis
J10y6	Barrett's oesophagus
J1011	Reflux oesophagitis
J1016	Barratt's oesophagitis
J101z	Oesophagitis NOS
J1025	Barrett's ulcer of oesophagus
J1020	Peptic ulcer of oesophagus
1957.	Gastric reflux

<b>LRTI</b>	
<b>Read Code (V2)</b>	<b>Term</b>
H2...	Pneumonia and influenza
H20..	Viral pneumonia
H201.	Pneumonia due to respiratory syncitial virus
H20y.	Viral pneumonia NEC
H20z.	Viral pneumonia NOS

<b>LRTI</b>	
<b>Read Code (V2)</b>	<b>Term</b>
H20y0	Severe acute respiratory syndrome
H21..	Lobar (pneumococcal pneumonia)
H22..	Other bacterial pneumonia
H220.	Pneumonia due to klebsiella pneumoniae
H222.	Pneumonia due to haemophilus influenzae
H223.	Pneumonia due to streptococcus
H224.	Pneumonia due to staphylococcus
H22y.	Pneumonia – other specific bacteria
H22y2	Pneumonia-legionella
H22yz	Pneumonia due to bacteria NOS
H22z.	Bacterial pneumonia NOS
H23..	Pneumonia due to other specified organisms
H231.	Pneumonia due to mycoplasma pneumoniae
H23z.	Pneumonia due to specified organism NOS
H24..	Pneumonia with infectious diseases EC
H24y.	Pneumonia with other infectious diseases EC
H24yz	Pneumonia with other infectious diseases EC NOS
H24y2	Pneumonia with pneumocystis carinii
H24z	Pneumonia with infectious diseases EC NOS
H25..	Bronchopneumonia due to unspecified organism
H26..	Pneumonia due to unspecified organism
H260.	Lobar pneumonia due to unspecified organism
H2600	Lung consolidation
H261.	Basal pneumonia due to unspecified organism
H262.	Postoperative pneumonia
H263.	Pneumonitis, unspecified
H27..	Influenza
H270.	Influenza with pneumonia
H2700	Influenza with bronchopneumonia
H2701	Influenza with pneumonia, influenza virus identified
H271.	Influenza with other respiratory manifestation
H2710	Influenza with laryngitis
H27z.	Influenza NOS
H28..	Atypical pneumonia
H2A..	Influenza due to Influenza A virus subtype H1N1
H2B..	Community acquired pneumonia
H2C..	Hospital acquired pneumonia
H2y..	Other specified pneumonia or influenza

<b>LRTI</b>	
<b>Read Code (V2)</b>	<b>Term</b>
H2z..	Pneumonia or Influenza NOS
H5400	Hypostatic pneumonia
H5401	Hypostatic bronchopneumonia
Hyu08	Other viral pneumonia
Hyu0A	Other bacterial pneumonia
Hyu0B	Pneumonia due to other specified infectious organisms
Hyu0H	Other pneumonia, organism unspecified
G5203	Acute myocarditis – influenzal

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
A788%	Acquired immune deficiency syndrome	AIDS
A789%	Human immunodef virus resulting in other disease	AIDS
AyuC%	[X]Human immunodeficiency virus disease	AIDS
B....	Cancers	Cancer
B0%	Carcinoma of lip, oral cavity and pharynx	Cancer
B1%	Carcinoma of digestive organs and peritoneum	Cancer
B2%	Carcinoma of respiratory tract and intrathoracic organs	Cancer
B3%	Carcinoma of bone, connective tissue, skin and breast	Cancer
B4%	Malignant neoplasm of genitourinary organ	Cancer
B5...	Malignant neoplasm of other and unspecified sites	Cancer
B50%	Malignant neoplasm of eye	Cancer
B51%	Malignant neoplasm of brain	Cancer
B52%	Malig neop of other and unspecified parts of nervous system	Cancer
B53..	Malignant neoplasm of thyroid gland	Cancer

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
B54%	Malig neop of other endocrine glands and related structures	Cancer
B55%	Malignant neoplasm of other and ill-defined sites	Cancer
B6%	Malignant neoplasm of histiocytic tissue	Cancer
By%	Neoplasms otherwise specified	Cancer
Bz...	Neoplasms NOS	Cancer
ZV10%	[V]Personal history of malignant neoplasm	Cancer
1477.	H/O: cerebrovascular disease	Cerebrovascular disease
70043	Evacuation of intracerebral haematoma NEC	Cerebrovascular disease
14A7.	H/O: stroke	Cerebrovascular disease
662M.	Stroke monitoring	Cerebrovascular disease
F11x2	Cerebral degeneration due to cerebrovascular disease	Cerebrovascular disease
G6%	Cerebrovascular disease	Cerebrovascular disease
Gyu6%	[X]Cerebrovascular diseases	Cerebrovascular disease
S62%	Subarachnoid haemorrhage following injury	Cerebrovascular disease
H30%	Bronchitis unspecified	Chronic pulmonary disease
H31%	Chronic bronchitis	Chronic pulmonary disease
H325	Emphysema	Chronic pulmonary disease
H33%	Asthma	Chronic pulmonary disease
H34..	Bronchiectasis	Chronic pulmonary disease
H35..	Extrinsic allergic alveolitis	Chronic pulmonary disease
H3z%	Chronic obstructive pulmonary disease NOS	Chronic pulmonary disease
H40%	Coal workers' pneumoconiosis	Chronic pulmonary disease
H415	Asbestosis	Chronic pulmonary disease

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
H42%	Silica and silicate pneumoconiosis	Chronic pulmonary disease
H43%	Pneumoconiosis due to other inorganic dust	Chronic pulmonary disease
H440.	Byssinosis	Chronic pulmonary disease
H441.	Cannabinosis	Chronic pulmonary disease
H442.	Flax-dressers' disease	Chronic pulmonary disease
H45%	Pneumoconiosis NOS	Chronic pulmonary disease
H4605	Bronchitis and pneumonitis due to chemical fumes	Chronic pulmonary disease
1761.	C/O bronchial catarrh	Chronic pulmonary disease
1780.	Aspirin induced asthma	Chronic pulmonary disease
14B4.	H/O: asthma	Chronic pulmonary disease
173A.	Exercise induced asthma	Chronic pulmonary disease
173c.	Occupational asthma	Chronic pulmonary disease
1O2..	Asthma confirmed	Chronic pulmonary disease
663%	Asthma monitoring	Chronic pulmonary disease
66YC.	Absent from work or school due to asthma	Chronic pulmonary disease
8H2P.	Emergency admission, asthma	Chronic pulmonary disease
66YP.	Asthma night-time symptoms	Chronic pulmonary disease
8H2P.	Emergency admission, asthma	Chronic pulmonary disease
9OJ1.	Attends asthma monitoring	Chronic pulmonary disease

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
9OJA.	Asthma monitored	Chronic pulmonary disease
H47y0	Detergent asthma	Chronic pulmonary disease
H4y10	Chronic pulmonary fibrosis following radiation	Chronic pulmonary disease
H4z..	Lung disease due to external agents NOS	Chronic pulmonary disease
H57y.	Lung disease with diseases EC	Chronic pulmonary disease
H57yz	Lung disease with diseases EC NOS	Chronic pulmonary disease
H581.	Interstitial emphysema	Chronic pulmonary disease
H582.	Compensatory emphysema	Chronic pulmonary disease
Hyu30	[X]Other emphysema	Chronic pulmonary disease
Hyu40	[X]Pneumoconiosis due to other dust containing silica	Chronic pulmonary disease
Hyu41	[X]Pneumoconiosis due to other specified inorganic dusts	Chronic pulmonary disease
Hyu43	[X]Hypersensitivity pneumonitis due to other organic dusts	Chronic pulmonary disease
SK07.	Subcutaneous emphysema	Chronic pulmonary disease
66YP.	Asthma night-time symptoms	Chronic pulmonary disease
8H2P.	Emergency admission, asthma	Chronic pulmonary disease
9OJ1.	Attends asthma monitoring	Chronic pulmonary disease
9OJA.	Asthma monitored	Chronic pulmonary disease
H47y0	Detergent asthma	Chronic pulmonary disease
H4y10	Chronic pulmonary fibrosis following radiation	Chronic pulmonary disease



<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
H4z..	Lung disease due to external agents NOS	Chronic pulmonary disease
H57y.	Lung disease with diseases EC	Chronic pulmonary disease
H57yz	Lung disease with diseases EC NOS	Chronic pulmonary disease
H581.	Interstitial emphysema	Chronic pulmonary disease
H582.	Compensatory emphysema	Chronic pulmonary disease
Hyu30	[X]Other emphysema	Chronic pulmonary disease
Hyu40	[X]Pneumoconiosis due to other dust containing silica	Chronic pulmonary disease
Hyu41	[X]Pneumoconiosis due to other specified inorganic dusts	Chronic pulmonary disease
Hyu43	[X]Hypersensitivity pneumonitis due to other organic dusts	Chronic pulmonary disease
SK07.	Subcutaneous emphysema	Chronic pulmonary disease
14A6.	H/O: heart failure	Congestive heart disease
14AM.	H/O: Heart failure in last year	Congestive heart disease
1O1..	Heart failure confirmed	Congestive heart disease
662W.	Heart failure annual review	Congestive heart disease
8B29.	Cardiac failure therapy	Congestive heart disease
8CL3.	Heart failure care plan discussed with patient	Congestive heart disease
8H2S.	Admit heart failure emergency	Congestive heart disease
G232.	Hypertensive heart&renal dis wth (congestive) heart failure	Congestive heart disease
G5540	Congestive cardiomyopathy	Congestive heart disease
G58%	Heart failure	Congestive heart disease
SP111	Heart failure as a complication of care	Congestive heart disease
1461.	H/O: dementia	Dementia
E00%	Senile/presenile dementia	Dementia
E041%	Dementia in conditions EC	Dementia

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
Eu00%	[X]Dementia in Alzheimer's disease	Dementia
Eu01%	[X]Vascular dementia	Dementia
Eu02%	[X]Dementia in other diseases classified elsewhere	Dementia
1434.	H/O: diabetes mellitus	Dementia
66A%	Diabetic monitoring	Diabetes
8A13.	Diabetic stabilisation	Diabetes
8BL2.	Patient on maximal tolerated therapy for diabetes	Diabetes
8H2J.	Admit diabetic emergency	Diabetes
C10..	Diabetes mellitus	Diabetes
Cyu2.	[X]Diabetes mellitus	Diabetes
G73y0	Diabetic peripheral angiopathy	Diabetes
L1805	Pre-existing diabetes mellitus, insulin-dependent	Diabetes
L1806	Pre-existing diabetes mellitus, non-insulin-dependent	Diabetes
L180X	Pre-existing diabetes mellitus, unspecified	Diabetes
2BB%	O/E - diabetic retinopathy	Diabetes with complications
C104%	Diabetic nephropathy	Diabetes with complications
C105%	Diabetes mellitus with ophthalmic manifestation	Diabetes with complications
C106%	Diabetic amyotrophy	Diabetes with complications
C108%	Type 1 diabetes mellitus with complications	Diabetes with complications
C109%	Type 2 diabetes mellitus with complications	Diabetes with complications
C10E%	Type 1 diabetes mellitus with complications	Diabetes with complications
C10F%	Type 2 diabetes mellitus with complications	Diabetes with complications
F372.	Diabetic polyneuropathy	Diabetes with complications

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
F374z	Polyneuropathy in disease NOS	Diabetes with complications
F3813	Diabetic amyotrophy	Diabetes with complications
F3y0.	Diabetic mononeuropathy	Diabetes with complications
F420%	Diabetic retinopathy	Diabetes with complications
F4640	Diabetic cataract	Diabetes with complications
K01x1	Kimmelstiel - Wilson disease	Diabetes with complications
2833.	O/E - hemiplegia	Hemiplegia
2835.	O/E - paraplegia	Hemiplegia
F141.	Hereditary spastic paraplegia	Hemiplegia
F22%	Hemiplegia	Hemiplegia
F230%	Paraplegia - congenital	Hemiplegia
F241%	Paraplegia	Hemiplegia
B153.	Secondary malignant neoplasm of liver	Metastatic tumour
B56%	Lymph node metastases	Metastatic tumour
B57%	Metastases of respiratory and/or digestive systems	Metastatic tumour
B58%	Secondary carcinoma of other specified sites	Metastatic tumour
B59zX	Malignant neoplasm of unspecified site	Metastatic tumour
B5y..	Malignant neoplasm of other and unspecified site OS	Metastatic tumour
B5z..	Malignant neoplasm of other and unspecified site NOS	Metastatic tumour
ByuC%	[X]Malignant neoplasm of ill-defined, secondary and unspeci	Metastatic tumour
C3104	Glycogenosis with hepatic cirrhosis	Mild liver disease
C3500	Pigmentary cirrhosis of liver	Mild liver disease
J6002	Acute yellow atrophy	Mild liver disease
J6012	Subacute yellow atrophy	Mild liver disease

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
J61%	Cirrhosis and chronic liver disease	Mild liver disease
J633.	Hepatitis unspecified	Mild liver disease
J6356	Toxic liver disease with fibrosis and cirrhosis of liver	Mild liver disease
Jyu71	[X]Other and unspecified cirrhosis of liver	Mild liver disease
760F3	Rigid oesophagoscopy injection sclerotherapy oesoph varices	Mod liver disease
A704z	Other specified viral hepatitis with hepatic coma NOS	Mod liver disease
G85%	Oesophageal varices	Mod liver disease
Gyu94	[X]Oesophageal varices in diseases classified elsewhere	Mod liver disease
J622.	Hepatic coma	Mod liver disease
J623.	Portal hypertension	Mod liver disease
J624.	Hepatorenal syndrome	Mod liver disease
J62y.	Other sequelae of chronic liver disease	Mod liver disease
J62z.	Liver abscess and chronic liver disease causing sequelae NOS	Mod liver disease
14AH.	H/O: Myocardial infarction in last year	Myocardial infarction
G30%	Heart attack	Myocardial infarction
G32..	Personal history of myocardial infarction	Myocardial infarction
1956.	Peptic ulcer symptoms	Peptic ulcer disease
7627.	Operations on duodenal ulcer	Peptic ulcer disease
76121	Balfour excision of gastric ulcer	Peptic ulcer disease
76125	Resection of gastric ulcer by cautery	Peptic ulcer disease
76270	Closure of perforated duodenal ulcer	Peptic ulcer disease
761D6	Endoscopic injection haemostasis of gastric ulcer	Peptic ulcer disease
761J.	Stomach ulcer operations	Peptic ulcer disease
761J.	Operations on gastric ulcer	Peptic ulcer disease

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
761J0	Closure of perforated gastric ulcer	Peptic ulcer disease
761J1	Closure of gastric ulcer NEC	Peptic ulcer disease
761J1	Suture of ulcer of stomach NEC	Peptic ulcer disease
761Jy	Other specified operation on gastric ulcer	Peptic ulcer disease
761Jz	Operation on gastric ulcer NOS	Peptic ulcer disease
J1020	Peptic ulcer of oesophagus	Peptic ulcer disease
J11%	Gastric ulcer - (GU)	Peptic ulcer disease
J12%	Duodenal ulcer - (DU)	Peptic ulcer disease
J13%	Peptic ulcer	Peptic ulcer disease
J14%	Stomal ulcer	Peptic ulcer disease
ZV127	[V]Personal history of peptic ulcer	Peptic ulcer disease
ZV12C	[V] Personal history of gastric ulcer	Peptic ulcer disease
14AE.	H/O: aortic aneurysm	Peripheral vascular disease
14NB.	H/O: Peripheral vascular disease procedure	Peripheral vascular disease
2I16.	O/E - gangrene	Peripheral vascular disease
7A112	Y graft of abdominal Aortic aneurysm (emergency)	Peripheral vascular disease
7A113	Y graft abdominal Aortic aneurysm	Peripheral vascular disease
7A13.	Emergency repair of aortic aneurysm	Peripheral vascular disease
7A134	Tube graft abdominal Aortic aneurysm (emergency)	Peripheral vascular disease
7A14.	Aortic aneurysm repair	Peripheral vascular disease
7A144	Tube graft of Abdominal aortic aneurysm	Peripheral vascular disease
C107.	Diabetes with gangrene	Peripheral vascular disease

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
G71%	Aortic aneurysm	Peripheral vascular disease
G73%	Peripheral ischaemic vascular disease	Peripheral vascular disease
Gyu71	[X]Aortic aneurysm of unspecified site, ruptured	Peripheral vascular disease
Gyu72	[X]Aortic aneurysm of unspecified site, nonruptured	Peripheral vascular disease
Gyu74	[X]Other specified peripheral vascular diseases	Peripheral vascular disease
R054%	[D]Gangrene	Peripheral vascular disease
14D1.	H/O: nephritis	Renal disease
1Z10.	Chronic kidney disease stage 1	Renal disease
1Z11.	Chronic kidney disease stage 2	Renal disease
1Z12.	Chronic kidney disease stage 3	Renal disease
1Z13.	Chronic kidney disease stage 4	Renal disease
1Z14.	Chronic kidney disease stage 5	Renal disease
K0%	Nephritis, nephrosis and nephrotic syndrome	Renal disease
K1000	Chronic pyelonephritis without medullary necrosis	Renal disease
K1001	Chronic pyelonephritis with medullary necrosis	Renal disease
K1010	Acute pyelonephritis without medullary necrosis	Renal disease
K1011	Acute pyelonephritis with medullary necrosis	Renal disease
Kyu2%	[X]Renal failure	Renal disease
F3712	Polyneuropathy in rheumatoid arthritis	Rheumatological disease
F3961	Myopathy due to disseminated lupus erythematosus	Rheumatological disease

<b>Charlson Comorbidity Index Categories</b>		
<b>Read Code (V2) *</b>	<b>Parents Term Header</b>	<b>Charlson Comorbidity</b>
F3964	Myopathy due to rheumatoid arthritis	Rheumatological disease
F3966	Myopathy due to scleroderma	Rheumatological disease
G5yA.	Rheumatoid carditis	Rheumatological disease
H570.	Rheumatoid lung	Rheumatological disease
H572.	Lung disease with systemic sclerosis	Rheumatological disease
H57y1	Lung disease with polymyositis	Rheumatological disease
H57y4	Lung disease with systemic lupus erythematosus	Rheumatological disease
K01x4	Nephrotic syndrome in systemic lupus erythematosus	Rheumatological disease
N000%	Systemic lupus erythematosus	Rheumatological disease
N001%	Scleroderma	Rheumatological disease
N004.	Polymyositis	Rheumatological disease
N04%	Rheumatoid arthritis and other inflammatory polyarthropathy	Rheumatological disease
N060.	Endemic polyarthritis	Rheumatological disease
N20%	Polymyalgia rheumatica	Rheumatological disease
N2314	Polymyositis ossificans	Rheumatological disease
N240%	Rheumatism and fibrositis unspecified	Rheumatological disease
N2y..	Other specified nonarticular rheumatism	Rheumatological disease
N2z..	Nonarticular rheumatism NOS	Rheumatological disease
Nyu10	[X]Rheumatoid arthritis organs or systems	Rheumatological disease
Nyu11	[X]Other seropositive rheumatoid arthritis	Rheumatological disease
Nyu12	[X]Other specified rheumatoid arthritis	Rheumatological disease
Nyu1G	[X]Seropositive rheumatoid arthritis, unspecified	Rheumatological disease
Nyu43	[X]Other forms of systemic lupus erythematosus	Rheumatological disease
Nyu45	[X]Other forms of systemic sclerosis	Rheumatological disease

## Appendix H: Visualisation of CMA Adherence Measures

KEY	
	Days Excluded from Analysis Window
	Non-Excluded Day Prior to First Refill
	Day in Interval of First Refill
	Day in Interval of Second Refill
	Day in Interval of Third Refill
{x}	x days of supply obtained on this date
[[x]]	x days of supply remaining on this date

CMA1 & CMA3						
M	T	W	T	F	S	S
		{28}				
{28}						
			{28}			

Window starts on the day of the first dispensing in observation period (day 10)

Window ends on the day prior to the last dispensing in observation period (day 66)

Duration of analysis window = 57 days (refill 1= 26 days, refill 2 = 31 days)

Supply dispensed =  $28 \times 2 = 56$

$CMA1 = 56/57 = 0.98$   
 $CMA3 = \min(CMA1, 1) = 0.98$

CMA2 & CMA4						
M	T	W	T	F	S	S
		{28}				
{28}						
			{28}			

Window starts on the day of the first dispensing in observation period (day 10)

Window ends on the last day in observation period (day 70)

Duration of analysis window = 61 days (refill 1= 26 days, refill 2 = 31 days, refill 3 = 4 days)

Supply dispensed =  $28 \times 3 = 84$

$CMA2 = 84/61 = 1.37$   
 $CMA4 = \min(CMA2, 1) = 1$



CMA5						
M	T	W	T	F	S	S
		{28}				
{28}						
			{28}			

Window starts on the day of the first dispensing in observation period (day 10)

Window ends on the day prior to the last dispensing in observation period (day 66)

Duration of analysis window = 57 days  
(refill 1 = 33 days, refill 2 = 24 days)

Days with medication available in window = 52 days  
(refill 1 = 28/33 days, refill 2 = 24/24 days)

CMA5 =  $52/57 = 0.91$

CMA6						
M	T	W	T	F	S	S
		{28}				
{28}						
			{28}			

Window starts on the day of the first dispensing in observation period (day 10)

Window ends on the last day in observation period (day 70)

Duration of analysis window = 61 days  
(refill 1 = 26 days, refill 2 = 31 days, refill 3 = 4 days)

Days with medication available in window = 56 days  
(refill 1 = 28/33 days, refill 2 = 24/24 days, refill 3 = 4/4 days)

CMA6 =  $56/61 = 0.92$

CMA7						
M	T	W	T	F	S	S
[[3]]						
		{28}				
{28}						
			{28}			

Window starts on first day in observation period (day 1)

Window ends on the last day in observation period (day 70)

Duration of analysis window = 70 days

Days with medication available in window = 59 days  
(before first refill = 3/9 days, refill 1 = 28/33 days, refill 2 = 24/24 days, refill3 = 4/4 days)

CMA7 = 59/70 = 0.84

CMA8						
M	T	W	T	F	S	S
[[3]]						
		{28}				
{28}						
			{28}			

Window starts on the day when the supply remaining at the start of the observation period is exhausted (day 4)

Window ends on the last day in observation period (day 70)

Duration of analysis window = 67 days

Days with medication available in window = 56 days  
(before first refill = 0/6 days, refill 1 = 28/33 days, refill 2 = 24/24 days, refill3 = 4/4 days)

CMA7 = 56/67 = 0.84

## Appendix I: Density Plots of Adherence Measures

The following plots show the Kernel Density Estimates (KDEs) calculated with Gaussian kernels for the adherence measures described in this analysis. Kernel density estimation aims to provide a smooth density estimate. Computationally, it works by fitting a kernel (a weighting function; denoted  $K_\lambda(x)$ ) over each sample (each observed value of an adherence measure, in our case;  $x_i$ ), using the samples in its neighbourhood and the number of query samples,  $n$ . The width of the kernel (the confidence around the sample) imposed over each sample is defined by the parameter  $\lambda > 0$ , known as the bandwidth. In the base R KDE implementation, the bandwidth,  $\lambda$ , is selected using the Silverman's Rule of Thumb method <sup>436</sup>, according to the standard deviation ( $\widehat{\sigma}$ ) and the interquartile range ( $I$ ) of the samples:

$$\lambda = 0.9 * \min \left( \widehat{\sigma}, \frac{I}{1.34} \right) n^{-\frac{1}{5}}$$

The KDE of  $x$  (the adherence value in this instance) is then defined as follows:

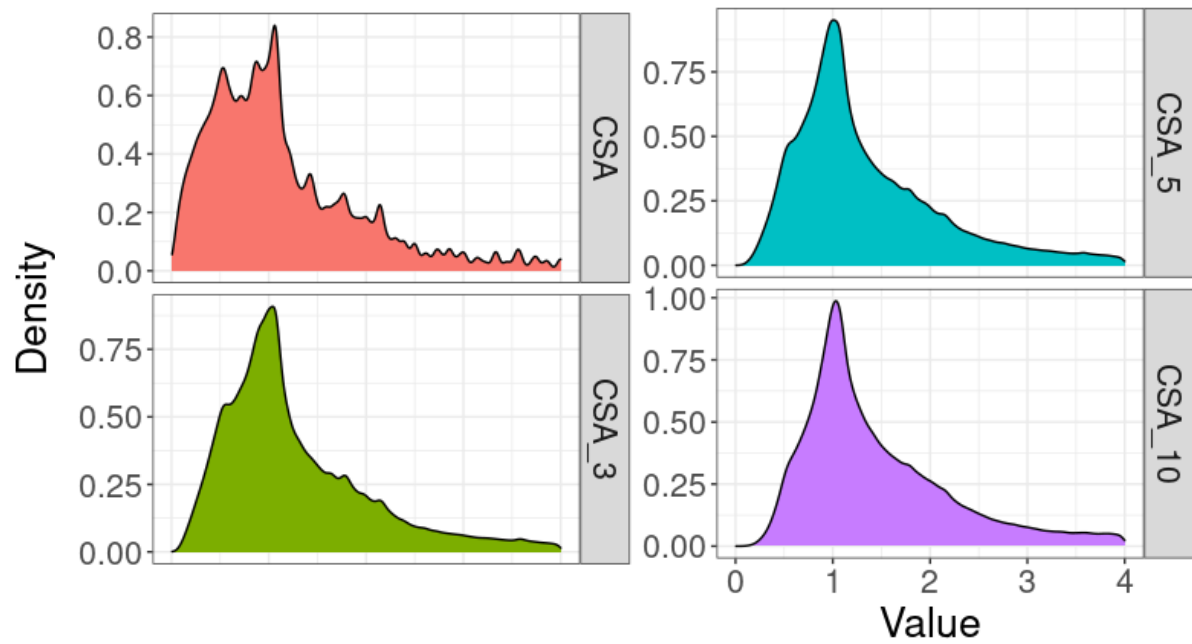
$$\hat{f}_\lambda(x) = \frac{1}{n} \sum_{i=1}^n K_\lambda(x - x_i)$$

The Gaussian kernel function used herein,  $G_\lambda(x)$ , is defined as follows:

$$G_\lambda(x) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-x^2}{2\lambda^2}}$$

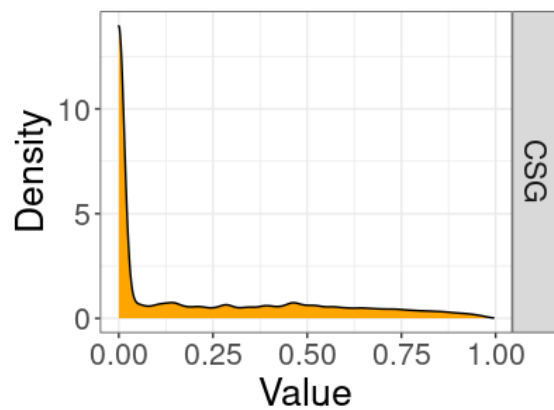
For more information, please refer to Chapter 6 of Hastie et al. <sup>290</sup>.

## CSA Measures

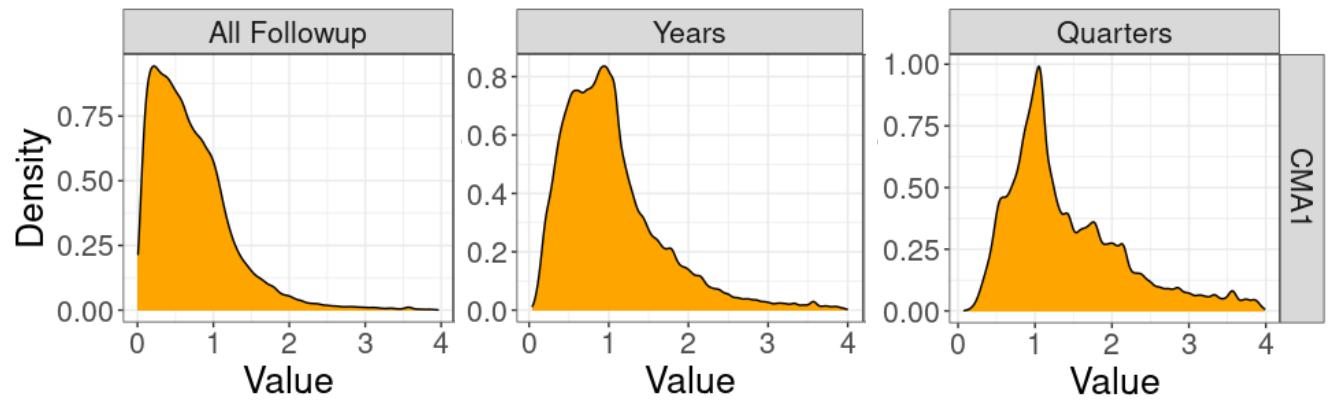


Note: the upper range of x-axis has been cropped at the 4, as there is a long tail (very low-density area).

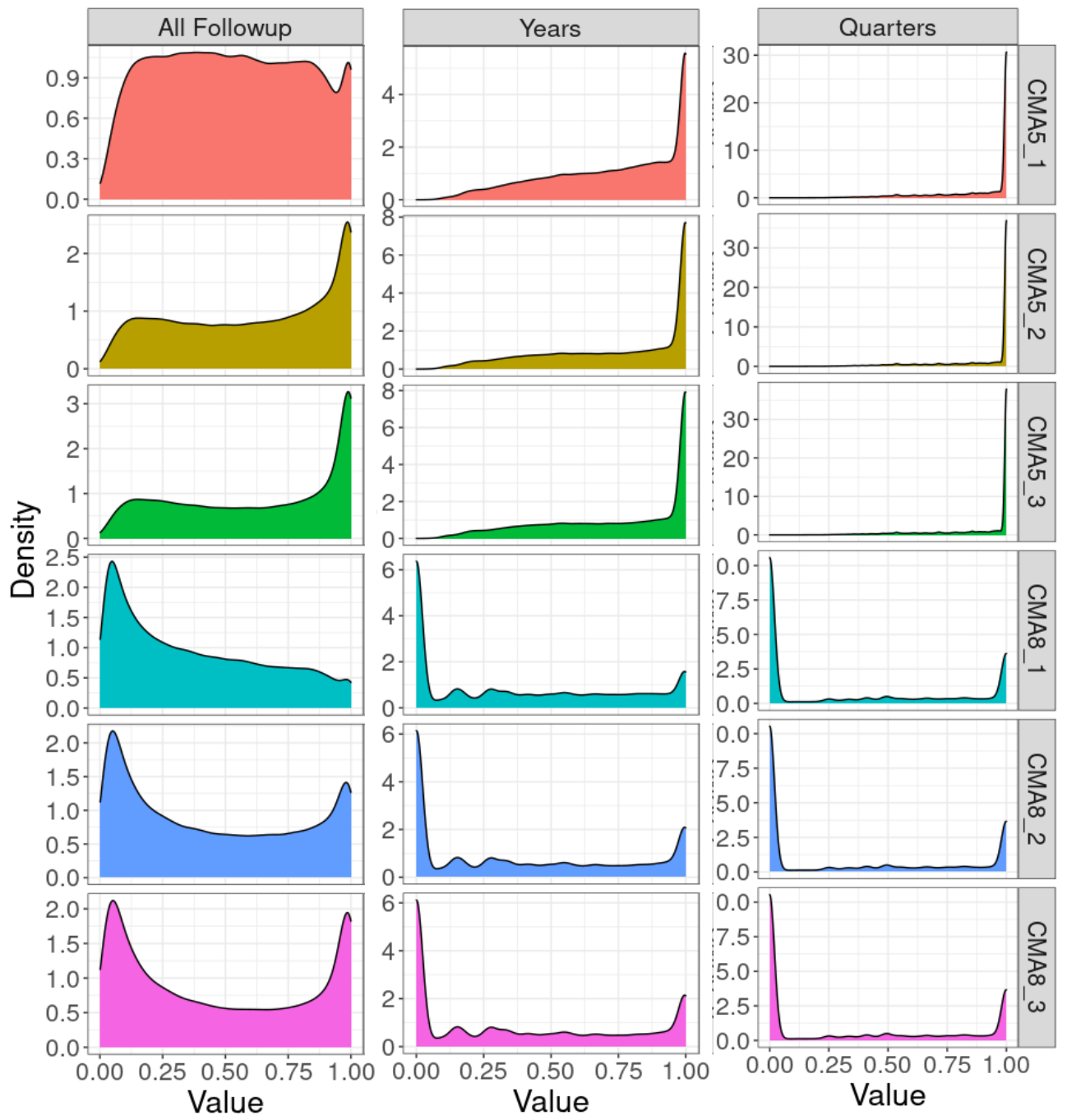
## CSG



## CMA1 Measures



Note: the upper range of x-axis has been cropped at the 4, as there is a long tail (very low-density area).



## Appendix J: Linkage of Primary Care Prescribing Records and Pharmacy Dispensing Records in the Salford Lung Study: Application in Asthma

I published the following in BMC Medical Research Methodology in 2020, and it is currently available online in its full form <sup>265</sup>. The following is a short summary, aiming to highlight the rationale for such an algorithm, and provide a brief overview of the methodology and key results.

### Background

Failure to collect the initial asthma prescription (*primary non-adherence*) has wide-varying reported incidence in studies of linked (or integrated) prescribing and dispensing records of between 12-45% <sup>259,261,262,415,416</sup>, with high variance due to differences in the right censoring point. Various components of the text processing that were necessary for my analysis (in the estimation of BTS treatment step and medication adherence) leant themselves naturally to the derivation of a linkage algorithm, which I hoped would facilitate the replication and validation of my work outside of Scotland.

As described in Section 8.4, prescribing and dispensing of medications are recorded by separate processes, and the data held by separate bodies, in England. These data also do not have a common unique prescribing event identifier, and as such matching records (one-to-one) using common identifiers (known as *deterministic linkage*) is currently impossible. Therefore, it is necessary to link records *probabilistically*; estimating the likelihood that two records will match given the data they contain.

The linkage of prescribing and dispensing records can enable the extraction of information about adherence to prescribed medications, including the identification of uncollected medications. In this study, we sought to develop a novel methodology linking primary care prescribing and dispensing records without a common identifier, using heuristics and features extracted from free-text fields.

## Methods

The Salford Lung Study (SLS) was a prospective, 12-month, open-label, parallel group, RCT conducted in 74 general practice clinics in Salford and South Manchester, UK<sup>437</sup>. A total of 4,233 participants with asthma were recruited in primary care settings by the healthcare professionals who provided their normal everyday care, and randomly allocated to either initiate a combination fluticasone furoate/vilanterol treatment or to continue their maintenance therapy (“usual care”). The dispensing data contained 225,235 records, for 4,197 unique participants, between 27<sup>th</sup> November 2012 and 9<sup>th</sup> December 2016. The prescribing dataset contained 339,792 records for 4,233 unique participants between 22<sup>nd</sup> November 2012 and 17<sup>th</sup> January 2017, however records outside of the dispensing data period were excluded.

Asthma controller medications were identified by the predecessor of the refined process described in Section 4.2.2. Key differences include the refinement of Appendix D and Table 4.2 (process described in full in publication).

The datasets of prescribing and dispensing records were merged such that a record (a *candidate link*) was generated for each eligible (common patient identifier and medication class) pair of records for matching. We note that the medication class keyword, composed of the active ingredients identified, was used in the place of a brand name such that generic substitutions would be identified as appropriate candidates for matching records. Pairs of records were eligible if the suggested dispensing date occurred after the prescription was written, but no more than six months *after* the prescription was written, at which point the prescription became invalid.

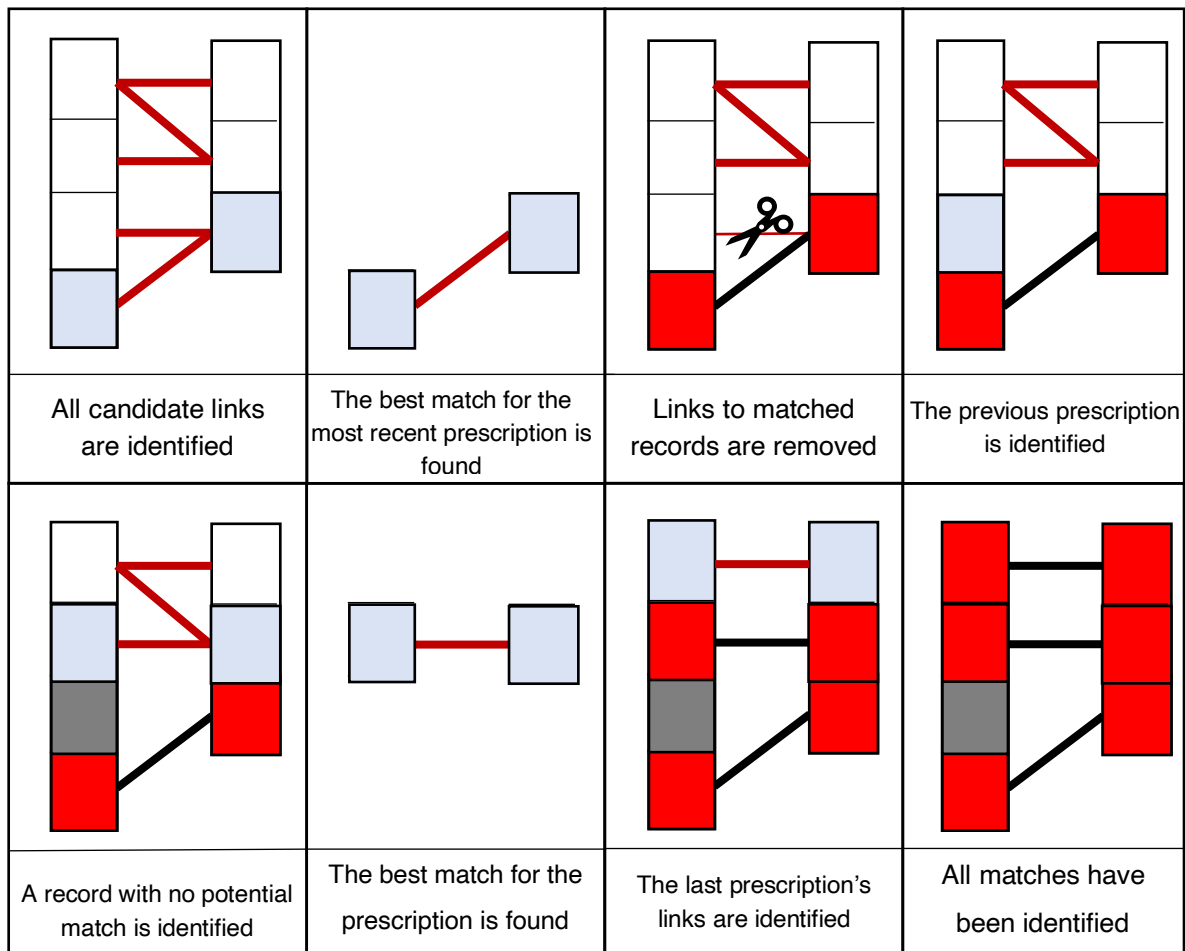
Probabilistic linkage, which aims to match records based on multiple non-unique features, utilizes *weights* to determine the strength of a link. These weights are numerical values representing the similarity of two records, derived using domain knowledge about the prevalence of dissimilarities between features in true matches.



In this linkage, a rule-based approach, based on a simplified posterior multivariate distribution of clerically reviewed data and previous literature, was used to weight candidate links for estimated likelihood of being a true match. Candidate links could then be ranked, and those with a linkage weight lower than 70% excluded. Each set of remaining dispensing records for each person-medication combination were looped through from the last to first through, as follows:

1. Identified the candidate in which the dispensing record occurs most recently after the prescription was written (record with highest match weight chosen if two candidate links on the same day were identified); this is a match between records,
2. Removed all other candidate links which contain the dispensing record or the prescribing records relating to this match,
3. Progressed to the previous dispensing for this person-medication.

The process is illustrated in the following figure:



The most recent prescribing record before the dispensing was prioritised over more distant records with a higher match weight, as we considered it more likely that prescription records for the same person within such a short time window were for the same medication, recorded differently, rather than a new treatment. Prescriptions that did not match any dispensing record were marked as unclaimed. We also noted dispensing records that were not matched (implying no corresponding prescription event) to assess linkage quality.

## Results

202,659 candidate links of identified asthma medications were processed, and 53,289 candidate links were confirmed as matches: 69.5% of prescribing records (n=76,680), and 83.2% of dispensing records (n=64,065). The median percentage of prescriptions claimed by an individual was 79%, with an interquartile range of 50-92% (range 0-100%). 23% of individuals claimed fewer than 50% of their prescriptions.

We inspected 23,391 prescribing records (31%) and 10,776 dispensing records (17%) for which a match could not be made (including those with candidate links which were not matched by the matching algorithm). In the non-matched prescriptions, 9% (n=2,109/23,391) had missing medication dosage, and <1% (n=87/23,391) had missing data on quantity (both missing in less than <0.1%). In the non-matched *dispensing* records, however, it was 62% (n=6,639/10,776) and 58% (n=6,222/10,776), respectively (both missing in 55%).

## Discussion

Our finding that 30% of prescriptions were labelled as uncollected, known as primary non-adherence, was a substantially higher proportion than the 8-20% found in previous asthma studies in US administrative health data studies<sup>259–263</sup>. One might assume that subsidised prescriptions, as we have in England, would result in higher primary adherence rates, as a barrier to adherence has been removed. On the contrary, a recent study in Canada, where prescriptions are subsidised and thus

considerably more affordable than in the USA, found that the fill rate for new asthma prescriptions was only 69% in adults <sup>415</sup>.

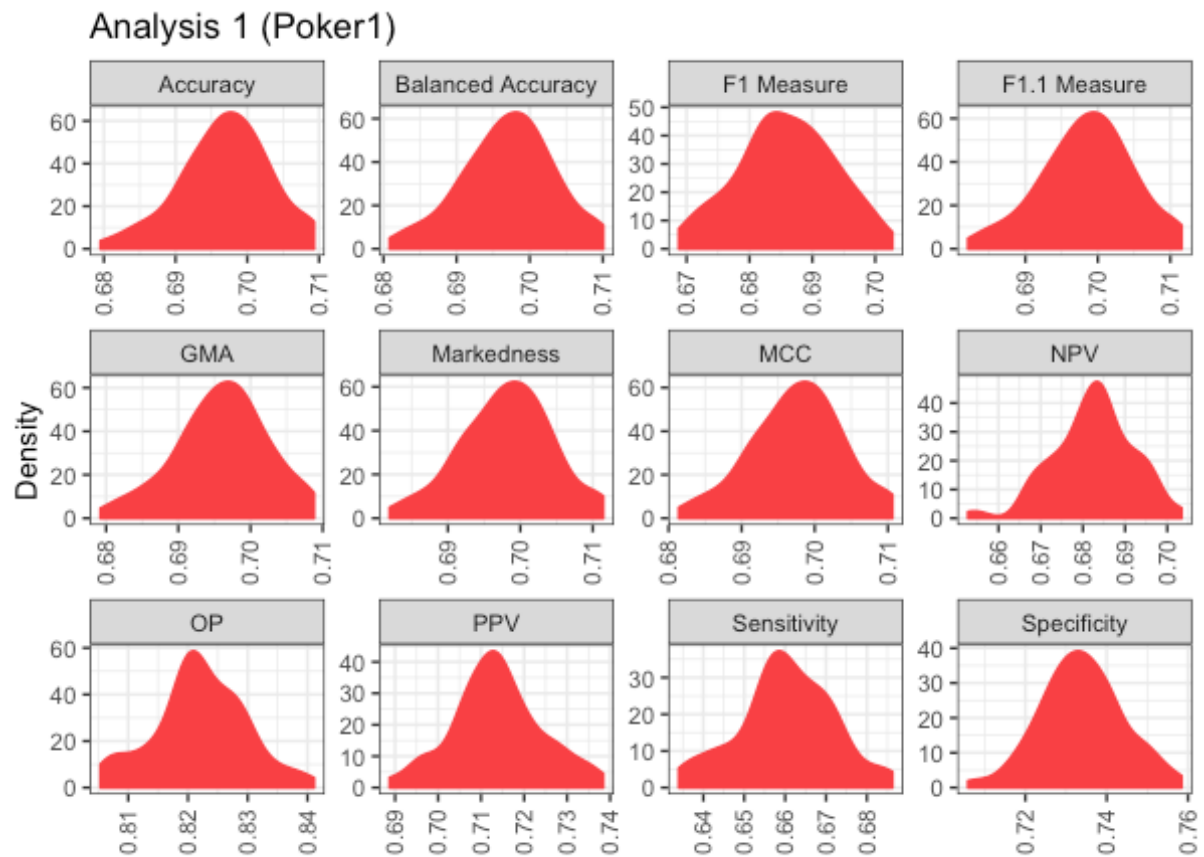
In lieu of a ground truth for comparison of our matches, we conducted quality assurance comparing features of the matched and unmatched records, as recommended by Harron *et al.*'s guidelines <sup>438</sup>. We observed that prescriptions (for which the status of being non-matched might imply either medication non-initiation, or not being correctly matched using the proposed algorithm) had missed medication strength in fewer than 10% of records, and missing quantity in fewer than 1%. In the non-matched dispensing records (which should occur only in rare emergency prescriptions and indicate shortcomings in matching prescription and dispensing records), 62% had missing medication strength and 58% had missing quantity. This indicates that one of the biggest barriers to successful record linkage was poor medication dispensing record quality.

The frequency of non-matched dispensing records was our best indicator as to the quality of our linkage, however we found that 95% of these records that were missing quantity (58%) were also missing medication strength. As such, reducing the weight threshold from 70% to 50%, would have had a substantial effect on the pool of candidate links allowed to be used in the matching algorithm. With so much missing data, however, the veracity of these matches would be hard to ascertain. In its current state, the algorithm will not match records with high amounts of missing data even if no other match is identified.

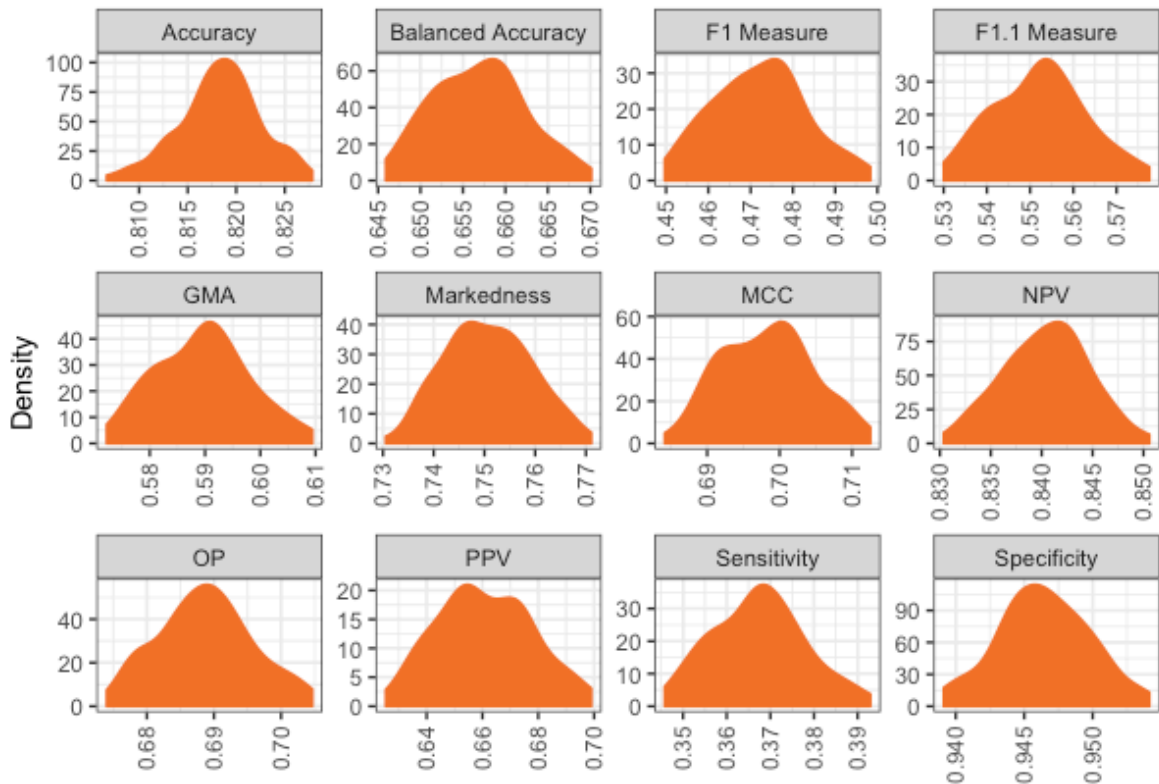
The presented methodology towards probabilistic record linkage enables preliminary assessment of whether patients are collecting their prescribed asthma medications and can improve clinicians' understanding of patient adherence. Further external validation of these promising findings on additional datasets is needed given the uncertainty around linkage quality.

## Appendix K: Density Plots of Performance Measures in Iterations of Empirical Data Analyses

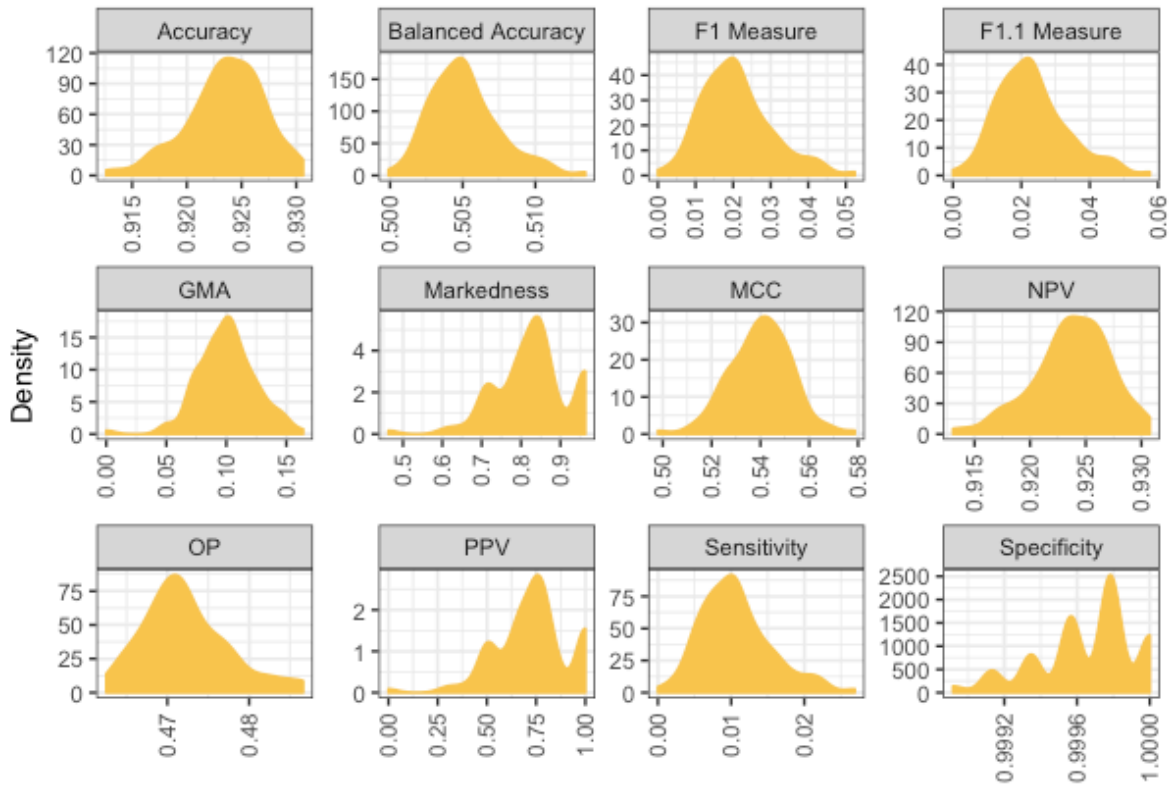
For details on the derivation of the kernel density estimators plotted in this appendix, please see Appendix I.



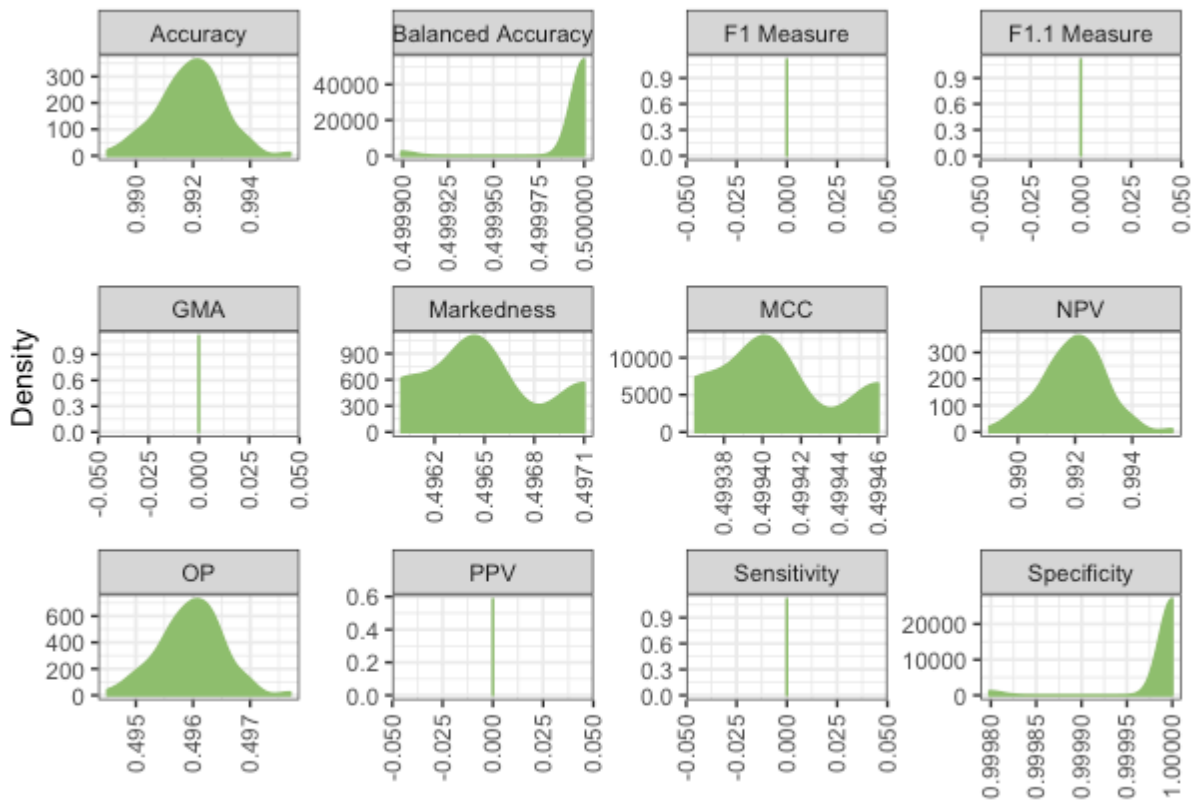
### Analysis 2 (Default)



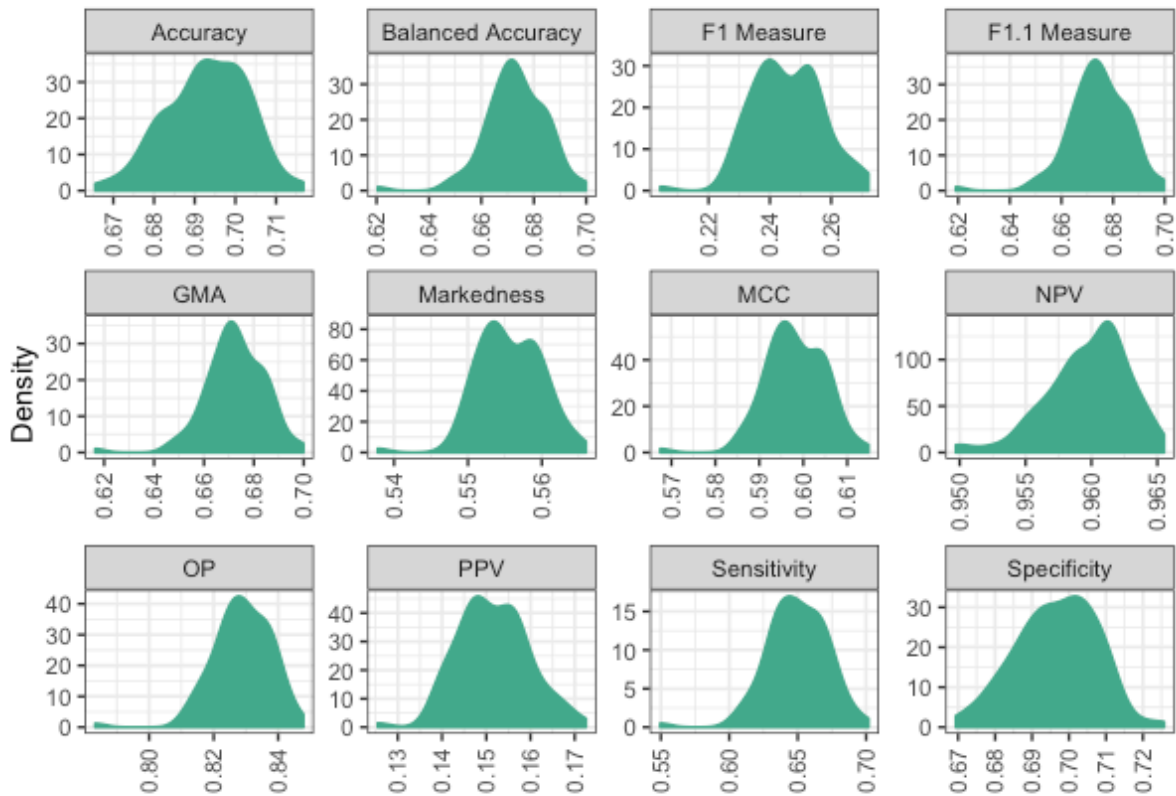
### Analysis 3 (Poker2)



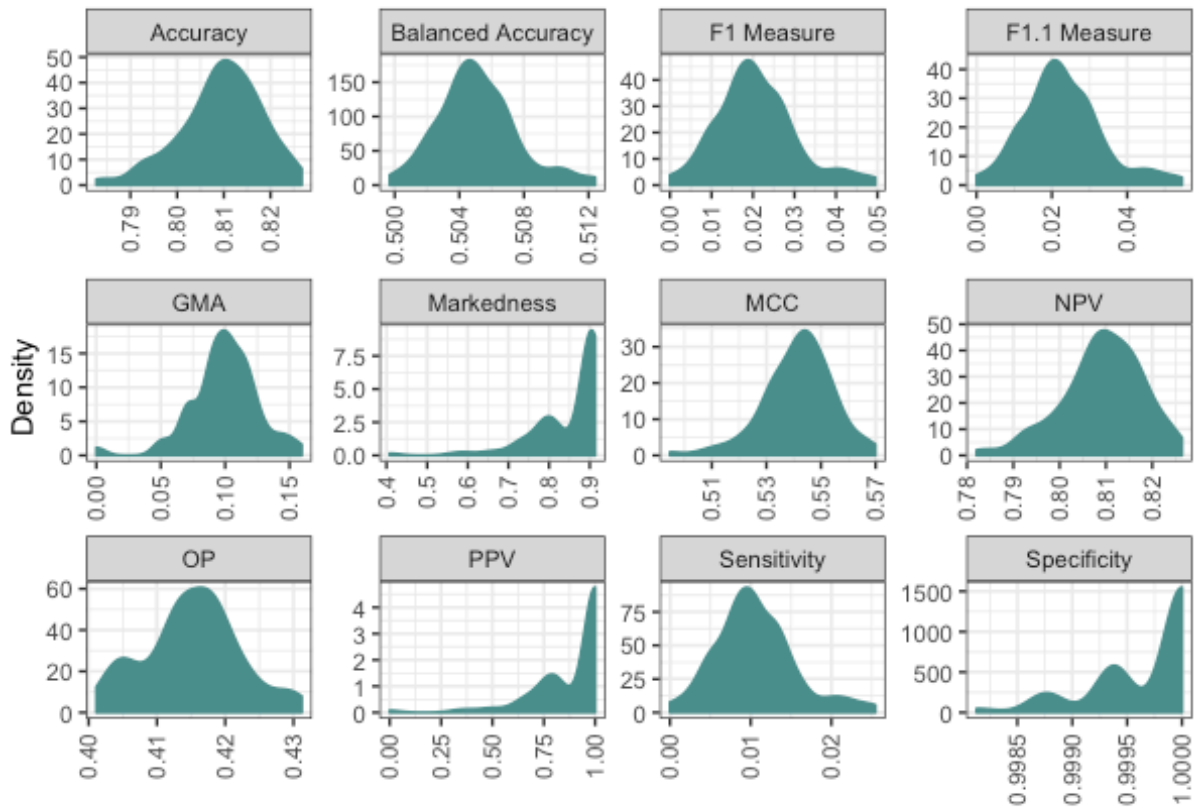
### Analysis 4 (Poker3)



### Analysis 5 (Enriched Poker2)



### Analysis 6 (Poker2 with More Balanced Testing)



## Appendix L: Application of Sokolova and Lapalme's Performance Measure Invariance Properties to Further Performance Measures

Invariance Property	[Accuracy]	Balanced Accuracy	[F1 Measure]	F1.1 Measure	GMA	MCC	Markedness	OP
I1	+	+	-	-	+	+	+	+
I2	-	-	+	-	-	-	-	-
I3	-	-	-	-	-	-	-	-
I4	-	-	-	-	-	-	-	-
I5	-	-	-	-	-	-	-	-
I6	+	+	+	+	+	+	+	+
I7	-	-	-	-	-	-	+	-
I8	-	+	-	+	+	-	-	-

Notes: + means positive for invariance, and - means negative for invariance (measure is variant)

The accuracy and F1 Measures were both included in the original analysis by Sokolova and Lapalme: the other measures have been included here for thoroughness.

Summary of invariance properties: (1) what is considered the positive and negative classes are switched, (2) The number of true negatives is changed, (3) The number of true positives is changed, (4) The number of false negatives is changed, (5) The number of false positives is changed, (6) All cell counts are changed by a consistent factor, (7) the observed positive and negative columns are changed by two distinct factors, (8) the predicted positive and negative rows are changed by two distinct factors.



## Appendix M: Relevant Risk Prediction Model Guidelines Items and Location within Thesis

Section and Topic	Item	Thesis Section
Rationale	Review the current practice and the rationale for the investigation being reported	1.1.3
	State specific objectives, including any prespecified hypotheses	1.3
	Review the state-of-the-art in predictive accuracy	7.1
	Explain the practical costs of misclassification errors	1.1.2, 1.1.3
Methods	State the ethics approval number for data access	2.2.2
	Describe the extent to which the investigators had access to the database population	2.2.2
	Describe the population selection criteria	7.3.1
	List the codes or algorithms used to identify the study population	Appendix E, 4.2.2
	Present a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage	7.4.1
	Report on the beginning and end dates of the study period	2.2.1
	Describe how the outcome was defined	7.3.2
	Document the model performance measures, and the methods to quantify uncertainty	7.3.4
	Describe the class imbalance	7.4.2
	Describe the feature selection process	3.9
	Describe the feature pre-processing performed, including how missing data were handled	7.3.3
	Provide a complete list of codes used to classify exposures	Appendix B, Appendix G
	Report any categorical features which predominantly (more than 95% samples) take the same value	Appendix O
	Document the algorithm used to develop the model	5.3

<b>Section and Topic</b>	<b>Item</b>	<b>Thesis Section</b>
Methods	Document the methodology used to avoid model-selection bias	7.3.4
	Document the methodology used to avoid resubstitution bias	7.3.4
	Describe the methodology for assessing internal validation and calibration	7.3.4
Results	Report the clinical and demographic characteristics of the study population	7.4.1
	Report on the final risk model	7.4.3
	Report the estimates of model performance with measures of uncertainty	7.4.4
	Report evidence of model calibration	7.4.6
	Quantify predictive value of features	7.4.5
External Validation	Describe the methodology for assessing external validation	N/A
	Report on the differences between the development and external validation dataset study populations	
	Present the results of the external validation	
Discussion	Discuss the prospects of the final model for satisfying the research goal, including the clinical implications	7.4.4, 7.4.5, 8.5
	Discuss known and possible limitations to generalizability or applicability of the model	8.3
	Report on differences between the final methodology and the published study protocol, where appropriate	Appendix Q

## Appendix N: Chronic Obstructive Pulmonary Disease Diagnosis Read Codes (Version 2)

Terms have a maximum character length of 50, and as such may feature truncated expressions.

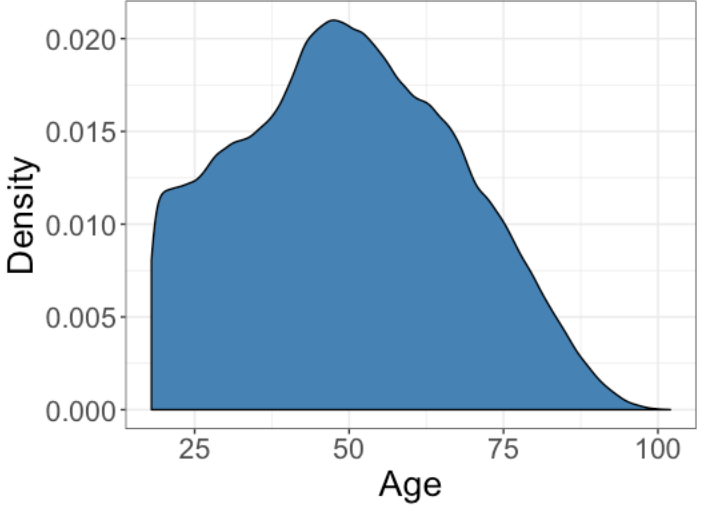
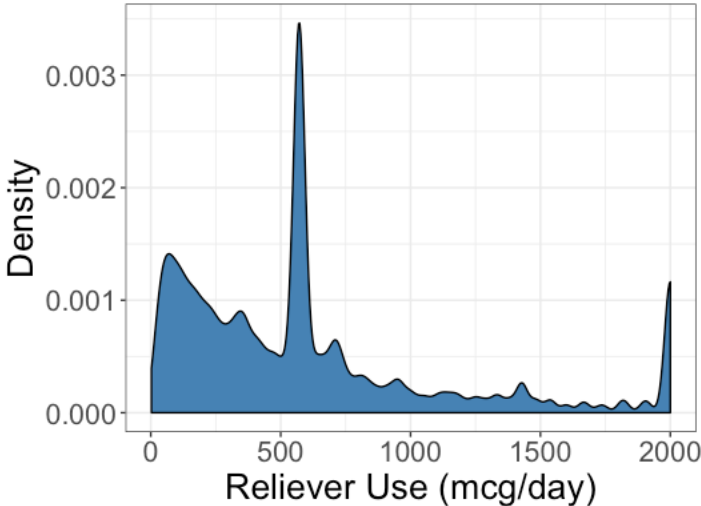
Read Code (V2)	Term
173A.	Exercise induced asthma
H3120	Chronic asthmatic bronchitis
H33..	Asthma
H330.	Extrinsic (atopic) asthma
H3300	Extrinsic asthma without status asthmaticus
H3301	Extrinsic asthma with status asthmaticus
H330z	Extrinsic asthma NOS
H331.	Intrinsic asthma
H3310	Intrinsic asthma without status asthmaticus
H3311	Intrinsic asthma with status asthmaticus
H331z	Intrinsic asthma NOS
H332.	Mixed asthma
H334.	Brittle asthma
H335.	Chronic asthma with fixed airflow obstruction
H33z.	Asthma unspecified
H33z0	Status asthmaticus NOS
H33z1	Asthma attack
H33z2	Late-onset asthma
H33zz	Asthma NOS
H3B..	Asthma-chronic obstructive pulmonary disease overlap syndrome
663..	Respiratory disease monitoring
6632.	Follow-up respiratory assessment
6636.	Inhaler technique shown
6637.	Inhaler technique observed
663a.	Oral steroids used since last appointment
663B.	Resp. treatment changed
663d.	Emergency asthma admission since last appointment
663e.	Asthma restricts exercise
663e0	Asthma sometimes restricts exercise
663e1	Asthma severely restricts exercise
663F.	Oral steroids started

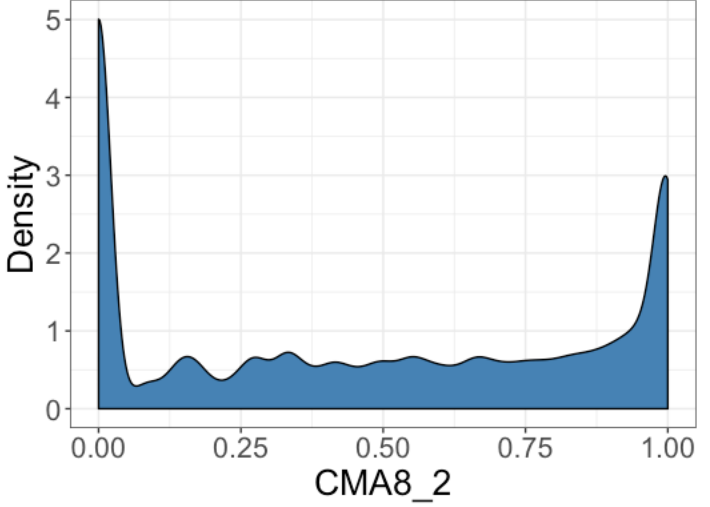
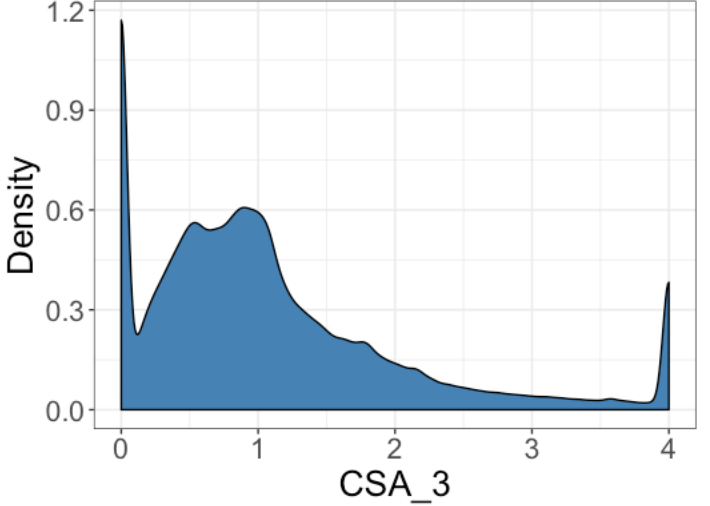
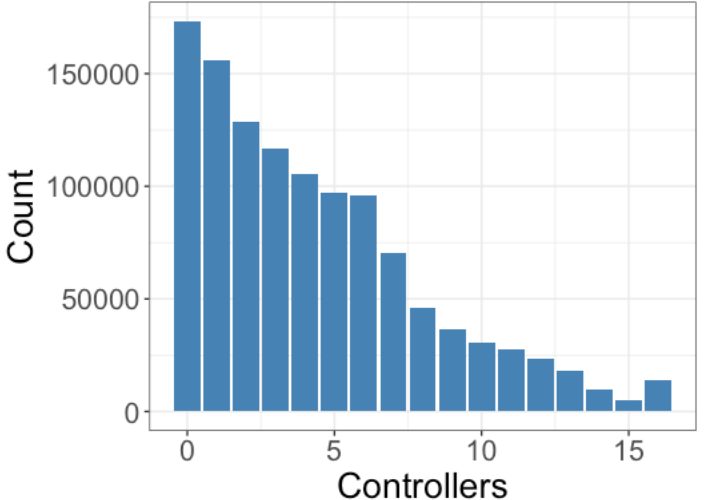
<b>Read Code (V2)</b>	<b>Term</b>
663f.	Asthma never restricts exercise
663G.	Oral steroids stopped
663g.	Inhaled steroids use
663g0	Not using inhaled steroids
663g1	Using inhaled steroids - normal dose
663g2	Using inhaled steroids - high dose
663g3	Increases inhaled steroids appropriately
663H.	Inhaler technique - good
663h.	Asthma - currently dormant
663l.	Inhaler technique - poor
663J.	Airways obstruction reversible
663j.	Asthma - currently active
663L.	Bronchodilators used more than once daily
663M.	Bronchodilators used a maximum of once daily
663m.	Asthma accident and emergency attendance since last visit
663N.	Asthma disturbing sleep
663n.	Asthma treatment compliance satisfactory
663N0	Asthma causing night waking
663N1	Asthma disturbs sleep weekly
663N2	Asthma disturbs sleep frequently
663O.	Asthma not disturbing sleep
663O0	Asthma never disturbs sleep
663P.	Asthma limiting activities
663p.	Asthma treatment compliance unsatisfactory
663Q.	Asthma not limiting activities
663q.	Asthma daytime symptoms
663R.	Service of nebuliser
663r.	Asthma causes night symptoms 1 to 2 times per month
663S.	Peak flow meter at home
663s.	Asthma never causes daytime symptoms
663T.	No peak flow meter at home
663t.	Asthma causes daytime symptoms 1 to 2 times per month
663U.	Asthma management plan given
663u.	Asthma causes daytime symptoms 1 to 2 times per week
663V.	Asthma severity
663v.	Asthma causes daytime symptoms most days
663V0	Occasional asthma
663V1	Mild asthma

<b>Read Code (V2)</b>	<b>Term</b>
663V2	Moderate asthma
663V3	Severe asthma
663W.	Asthma prophylactic medication used
663w.	Asthma limits walking up hills or stairs
663X.	Irritable airways
663x.	Asthma limits walking on the flat
663Y.	Steroid dose inhaled daily
663y.	Number of asthma exacerbations in past year
663Z.	Resp. disease monitoring NOS
663z.	Number of times bronchodilator used in one week

## Appendix O: Summary of Features in Risk Prediction Model

### Continuous (non-ordinal) features (n=5)

Feature	Summary Statistics	Density Plot or Bar Chart (after capping, when appropriate)
Age	Minimum = 18 LQL = 36 Median = 50 Mean = 50.1 UQL = 63 Max = 102	 <p>A density plot for the 'Age' feature. The x-axis is labeled 'Age' and ranges from 25 to 100 with major ticks at 25, 50, 75, and 100. The y-axis is labeled 'Density' and ranges from 0.000 to 0.020 with major ticks at 0.000, 0.005, 0.010, 0.015, and 0.020. The plot shows a blue-filled area under a curve that starts at approximately 0.011 at age 25, rises to a peak of about 0.021 at age 50, and then gradually declines to 0.000 at age 100.</p>
Reliever Medication Usage (mcg/day)	Minimum = 2.35 LQL = 215.1 Median = 571.4 Mean = 821.4 UQL = 740.7 Max = 80000  5.76% over 4*median to 1 significant figure (2000; upper cap)	 <p>A density plot for the 'Reliever Medication Usage (mcg/day)' feature. The x-axis is labeled 'Reliever Use (mcg/day)' and ranges from 0 to 2000 with major ticks at 0, 500, 1000, 1500, and 2000. The y-axis is labeled 'Density' and ranges from 0.000 to 0.003 with major ticks at 0.000, 0.001, 0.002, and 0.003. The plot shows a blue-filled area under a curve with a very sharp peak at approximately 600 mcg/day, reaching a density of about 0.0035. There is also a smaller peak at 0 mcg/day with a density of about 0.0014, and a very thin spike at 2000 mcg/day with a density of about 0.0012.</p>

Feature	Summary Statistics	Density Plot or Bar Chart (after capping, when appropriate)
CMA8_2	Minimum = 0.000 LQL = 0.000 Median = 0.485 Mean = 0.479 UQL = 0.878 Max = 1.000	 <p>A kernel density estimate plot for the feature CMA8_2. The x-axis is labeled 'CMA8_2' and ranges from 0.00 to 1.00. The y-axis is labeled 'Density' and ranges from 0 to 5. The plot shows a bimodal distribution with a sharp peak at 0.00 (density ~5) and another sharp peak at 1.00 (density ~3). There is a low density plateau between these two peaks.</p>
CSA_3	Minimum = 0.000 LQL = 0.432 Median = 0.881 Mean = 1.319 UQL = 1.459 Max = 162.041  3.97% over 4*median to 1 significant figure (4; upper cap)	 <p>A kernel density estimate plot for the feature CSA_3. The x-axis is labeled 'CSA_3' and ranges from 0 to 4. The y-axis is labeled 'Density' and ranges from 0.0 to 1.2. The plot shows a distribution with a sharp peak at 0 (density ~1.15) and a broader peak around 1 (density ~0.6). The distribution has a long right tail extending to 4.</p>
Number of controller medications	Minimum = 0 LQL = 1 Median = 4 Mean = 4.4 UQL = 6 Max = 67  0.91% over 4*median (16; upper cap)	 <p>A bar chart showing the count of controller medications. The x-axis is labeled 'Controllers' and ranges from 0 to 15. The y-axis is labeled 'Count' and ranges from 0 to 150,000. The distribution is right-skewed, with the highest count at 0 (over 150,000) and counts decreasing as the number of controllers increases.</p>

Note: For details on the derivation of the kernel density estimators plotted in this appendix, please see Appendix I, LQL = Lower Interquartile Limit, UQL = Upper Interquartile Limit

## Ordinal features (n=1)

Feature	Value	Proportion
BTS Step	0	13.75%
	1	25.40%
	2	20.09%
	3	31.87%
	4	8.89%

## Binary features (n=22)

Feature	Proportion Positive
Recent LRTI	1.19%
Recent Asthma Encounters	13.04%
Recent Steroid Prescriptions	1.29%
Nebulised SABA	1.76%
Obesity	28.14%
AIDS	<0.01%
Cancer	1.32%
Cerebrovascular disease	1.07%
Chronic pulmonary disease	21.97%
Congestive heart disease	0.40%
Dementia	0.28%
Diabetes (without complications)	2.02%
Diabetes with complications	0.99%
Hemiplegia	0.03%
Metastatic tumour	0.05%
Mild liver disease	0.19%
Moderate liver disease	0.11%
Myocardial infarction	0.50%
Peptic ulcer disease	0.25%
Peripheral vascular disease	0.36%
Renal disease	1.24%
Rheumatological disease	0.05%



## Categorical features (n=17)

Feature	Value	Proportion
Sex	Female	60.83%
	Male	39.17%
Socioeconomic Status (SIMD)	Quintile 1 (Most Deprived)	22.72%
	Quintile 2	20.93%
	Quintile 3	16.97%
	Quintile 4	21.24%
	Quintile 5 (Least Deprived)	15.84%
	Missing	2.31%
Local Area Code	Not listed: most common 13.8%, least common <0.01%	
Rurality (UR6)	Level 1 (Large Urban Areas)	31.78%
	Level 2 (Other Urban Areas)	37.03%
	Level 3 (Small Towns)	8.60%
	Level 4 (Rural Areas)	3.48%
	Level 5 (Accessible)	10.63%
	Level 6 (Remote)	5.56%
	Missing	2.92%
Smoking Status	Current	13.87%
	Former	15.57%
	Non-Smoker	70.56%
Peak Expiratory Flow	>90%	1.60%
	80-90%	0.24%
	70-80%	0.10%
	Less than 70%	0.05%
	Missing	98.00%
Time Since Last Asthma Attack recorded in Primary Care	One to two years	3.73%
	Six months up to one year	3.19%
	Three up to six months	1.92%
	One up to three months	1.56%
	In the last month	1.46%
	None in the last two years	88.14%
Eosinophilia	≥400 cells per $\mu$ L	6.27%
	<400 cells per $\mu$ L	19.73%
	Missing	74.00%

<b>Feature</b>	<b>Value</b>	<b>Proportion</b>
Month	January	7.86%
	February	7.40%
	March	7.91%
	April	6.88%
	May	7.04%
	June	7.15%
	July	7.12%
	August	7.24%
	September	7.23%
	October	8.13%
	November	8.75%
	December	18.29%
Rhinitis Diagnosis	Never	97.38%
	In the past year	0.80%
	One up to 5 years ago	1.57%
	Longer than five years ago	0.24%
Eczema Diagnosis	Never	96.51%
	In the past year	1.39%
	One up to 5 years ago	1.93%
	Longer than five years ago	0.18%
Anxiety/Depression Diagnosis	Never	88.72%
	In the past year	4.67%
	One up to 5 years ago	5.74%
	Longer than five years ago	0.87%
Nasal Polyps Diagnosis	Never	99.46%
	In the past year	0.19%
	One up to 5 years ago	0.31%
	Longer than five years ago	0.05%
Anaphylaxis Diagnosis	Never	99.92%
	In the past year	0.02%
	One up to 5 years ago	0.05%
	Longer than five years ago	0.01%
GERD Diagnosis	Never	96.9%
	In the past year	1.03%
	One up to 5 years ago	1.85%
	Longer than five years ago	0.23%

<b>Feature</b>	<b>Value</b>	<b>Proportion</b>
Corticosteroid Nasal Sprays	Never	68.89%
	In the past year	19.72%
	One up to 5 years ago	9.95%
	Longer than five years ago	1.24%
Time since last LRTI	In the past two weeks	1.04%
	Between two weeks and up to two months ago	0.88%
	Between two months and up to six months ago	1.80%
	Between six months and up to twelve months ago	2.20%
	Between one year and up to two years ago	3.04%
	None in the last two years	91.04%

## Appendix P: Machine Learning Classification Algorithms: Functions for Implementation in R, and Hyper-parameter Ranges

### Logistic Regression

Implemented using the base R function *glm*.

No hyper-parameters.

### Naïve Bayes Classifier

Implemented using the R function *naivebayes*, from the package of the same name

<sup>293</sup>.

No hyper-parameters.

### Random Forests

Implemented using the R function *ranger*, from the package of the same name <sup>439</sup>.

- *MTRY* = Number of features randomly sampled as candidates at each split (default is the rounded down integer of the square root of the number of features;  $k$ ):  $\text{floor}(\sqrt{k})$ ,  $\text{floor}(2*\sqrt{k})$ ,  $\text{floor}(4*\sqrt{k})$ ,  $\text{floor}(8*\sqrt{k})$  – in which floor represents the rounded-down integer value.

All other hyper-parameters take implementation default values.

Note that the ‘floor’ function denotes the rounded down integer of a value.

### Extreme Gradient Boosting

Implemented using the R function *xgboost*, from the package of the same name <sup>440</sup>.

- *ETA* = Step size shrinkage: 0.1, 0.25, 0.5
- *NROUNDS* = the number of decision trees in the final model: 100, 200

All other hyper-parameters take implementation default values.

## Appendix Q: Deviations between the Final Analysis and the Published Protocol Paper Analysis Plan

In this appendix, I describe the deviations between my analysis and the original protocol which I had published before commencing the analysis, published in *BMJ Open*<sup>372</sup>.

Topic	Original plan: quote from protocol paper	Revised action
Record Right-Censoring	“All records [will be] right-censored at March 2017, in order to align with the mortality, primary care, and inpatient hospital admission records”	Records should be right-censored at the earliest of death, asthma resolution Read code, or the end of the study period (the end date of the data from the dataset in ALHS which ends first)
External Validation	“In order to verify that the prediction model performance is not limited to the development dataset and that it generalizes well in new, unseen data ... we will evaluate its performance using an external cohort study dataset, the second Seasonal Influenza Vaccination Effectiveness (SIVE II) cohort study ...”	Access to the SIVE II dataset was not available for the duration required for analysis to be conducted, due to GDPR requirements for data deletion after the conclusion of the original study. Thus, unfortunately this external validation was not possible to conduct.
Model Features	“Active diagnoses of rhinitis, eczema, gastroesophageal reflux disease (GERD), nasal polyps, and anaphylaxis will be recorded”	I created a feature for the time since the last diagnosis code was recorded, allowing both recent and past diagnoses to be included, categorised as {‘Never’, ‘In the past year’, ‘One up to five years ago’, ‘Longer than five years ago’}
	N/A	In line with the findings of Price <i>et al.</i> <sup>117</sup> , anxiety and depression were also included (as a single feature, recorded categorically by time since last diagnostic code, as above.)

Topic	Original plan: quote from protocol paper	Revised action
Model Features	N/A	<p>Given the strength of the evidence for association between both nasal polyps and rhinitis with asthma exacerbation risk (Section 3.7.2), and the relative ease of identifying corticosteroid nasal sprays in the prescribing data (by virtue of needing to exclude them from the pool of asthma prescribing records), time since the most recent prescription of nasal spray corticosteroids was included as a risk factor (coded categorically as above).</p>
	N/A	<p>LRTIs (including pneumonia and influenza) were added on the basis of the evidence presented in Section 3.7.3. They were measured by two distinct features: a binary flag for whether more than one had been recorded in the last year (a proxy for susceptibility) and the time since the last recorded infection (to flag periods of recovery). This feature was categorised as: {'In the past two weeks', 'Between two weeks and up to two months ago', 'Between two months and up to six months ago', 'Between six months and up to twelve months ago', 'Between one year and up to two years ago', 'None in the last two years'}</p>

Topic	Original plan: quote from protocol paper	Revised action
Model Features	“the number of primary care asthma encounters (days on which at least one asthma related code was recorded) in the previous year will be derived”	I created a binary flag for whether or not there were more than one in the previous year. The decision to binarize the data was due to the wide range observed in the counts, which resulted in the differences between lower counts being quashed by the normalisation process. The decision boundary was based on the observed low median number of past encounters observed across the whole analysis population.
	“the prior number of attacks ... will be considered time-dependent and accurate at the weekly level.”	As above, this feature was amended to a binary indicator of whether there was more than one asthma attack either in the previous calendar year, or in the current year to date.
	“The mean Short-Acting Beta-2 Agonist (SABA) dose per day will be estimated retroactively by examining the dates between prescriptions”	The mean SABA dose was refined to only include <i>inhaled</i> SABA medications, however an additional feature was added to indicate that a nebulised SABA medication had been prescribed in the last 90 days.
	“Adherence to preventer therapy will be approximated using the medication possession ratio, calculated from primary care prescribing records.”	As per the analyses described in Chapter 4, two measures of adherence were used as risk factors in my prediction model: CSA_3 and CMA8_2. The Medication Possession Ratio (MPR; equivalent to the CMA1) was previously selected based on its use by Blakey <i>et al.</i> <sup>75</sup> , however upon further investigation the requirement for at least two prescriptions to calculate excluded too many people (16.8% in Chapter 4 analyses).

Topic	Original plan: quote from protocol paper	Revised action
Models tested	<p>“[We will] employ more advanced state of the art principled supervised learning algorithmic tools such as support vector machines...”</p>	<p>Upon further investigation of the SVM algorithm, I decided it was no longer feasible to include: it was likely to be very computationally intensive and it was not a scenario in which SVMs typically excelled over other methods (see Section 5.3.5).</p>
	<p><b>Random Forest classifier hyper-parameters:</b></p> <p>“</p> <ul style="list-style-type: none"> <li>- NTREE = Number of trees to grow (default 500): 500, 750, 1000</li> <li>- MTRY = Number of variables randomly sampled as candidates at each split (default square root of the number of predictors; k): <math>\text{floor}(0.5 * \sqrt{k})</math>, <math>\text{floor}(\sqrt{k})</math>, <math>\text{floor}(2 * \sqrt{k})</math> – in which floor represents the rounded-down integer value.</li> </ul> <p>”</p>	<p>For RFs, higher values of mtry (candidate features at each split) were tested (<math>\text{floor}(\sqrt{k})</math>, <math>\text{floor}(2 * \sqrt{k})</math>, <math>\text{floor}(4 * \sqrt{k})</math>, and <math>\text{floor}(8 * \sqrt{k})</math>), but the models with higher numbers of trees (<i>ntrees</i>) were removed. More trees generally result in a better variance-bias trade-off, and thus a lower risk of overfitting, but the improvement is not always efficient relative to the increased training time.</p>
	<p>“Implemented using the r function randomForest, from the package of the same name”</p>	<p>The R implementation was changed to the faster <i>ranger</i> package.</p>
	<p><b>Extreme Gradient Boosting</b></p> <p>“Implemented using the r package <i>xgboost</i>, with 10-fold cross validation, repeated 3 times.”</p>	<p>For XGBoost, repeated cross-validation was not used, and the hyper-parameters were instead evaluated in the same way as the RFs, for consistency.</p>



Topic	Original plan: quote from protocol paper	Revised action
Models tested	<p><b>XGBoost classifier hyper-parameters:</b></p> <p>“</p> <ul style="list-style-type: none"> <li>- NROUNDS = maximum number of iterations (default 100): 50,100</li> <li>- MAXDEPTH = Maximum depth of each tree (default = 6): (1:5)^2</li> <li>- ETA = step size of each boosting step (default = 0.3): 0.25, 0.5, 1</li> </ul> <p>”</p>	<p>For XGBoost, lower values of the learning rate (eta) were used (0.1, 0.25, and 0.5, instead of 0.25, 0.5 and 1). Although this increased the computation time, it vastly increased the stability of the model’s performance across iterations, which was important to ensure that the first 10 iterations were sufficient to evaluate the model performance compared to the other algorithms. In line with the lowered learning rate, higher numbers of boosting rounds were tested (100 and 200, rather than 50 and 100). To reduce the number of models being tested, only the default maximum tree depth (6 branches deep) was used.</p>
	<p><b>“Ensemble: Stacking</b></p> <p>Combining models from different classifiers, with an over-arching supervisor model which determines the best way to use all sources of information for prediction. The base set of weak learners will comprise all aforementioned model and hyper-parameter combinations, and the meta-learner (random forest with 500 trees and mtry = <math>\text{floor}(0.5 * \sqrt{k})</math>) will use all weak learners with a validation set performance in the top 50%. “</p>	<p>There was insufficient memory, particularly once the parallel computing was implemented, for model stacking to be tested, as it requires all of the trained models to be kept in the memory simultaneously.</p>

Topic	Original plan: quote from protocol paper	Revised action
Analysis Plan	<p>“We will run 100 iterations [of each model] for statistical confidence, each time randomly permuting samples prior to determining the three subsets” (training, testing and validation).</p>	<p>The data partitioning procedure was altered such that instead of running 100 iterations of every model, the model selection process was only based on the first ten iterations. As such, to ensure that there was no overlap between the validation and model selection partitions, a 10% hold-out set was used, and the partitioning in the remaining 90% was changed to 90% training and 10% testing.</p>
	<p>“we will identify the highest performing model as that with the highest mean MCC”</p>	<p>As described in Chapter 6, the balanced accuracy replaced the MCC as the primary performance measure, used to optimise the classification threshold.</p>
	<p>“Performance in the testing datasets will be assessed using ... the Bayesian Information Criterion (BIC) to obtain a trade-off between model complexity and accuracy.”</p>	<p>The BIC was no longer reported, as it is not appropriate for tree-based algorithms.</p>

Topic	Original plan: quote from protocol paper	Revised action
Analysis Plan	<p>“A selection of training enrichment methods will be trialled, in order to assess how to best overcome poor performance as a result of low outcome prevalence. Typically, modelling rare events results in reduced sensitivity (the proportion of those who had attacks that were detected), so those predicted to be low-risk will have a high rate of asthma attacks. As such, this start of this process (the first 20 iterations of training each model) will be repeated five times, using:</p> <ol style="list-style-type: none"> <li>1. the original analysis dataset,</li> <li>2. original data with additional duplicates of the positive outcome records (a method known as over-sampling),</li> <li>3. original data, with a selection of the negative outcome records removed (under-sampling),</li> <li>4. original data with additional slightly modified duplicates of the positive outcome records, with a selection of the negative outcome records removed (Synthetic minority over-sampling; SMOTE)</li> <li>5. original data, using the outcome classification threshold to maximise the primary metric”</li> </ol>	<p>Due to the extreme class imbalance, the pure under-sampling and over-sampling approaches would have resulted in either a very low sample size, or a dataset with almost 50% replicated samples, respectively. As such, three SMOTE tests were conducting, using different balanced of the under and over-sampling parameters, as described in Sections 5.5 and 7.3.5.</p>

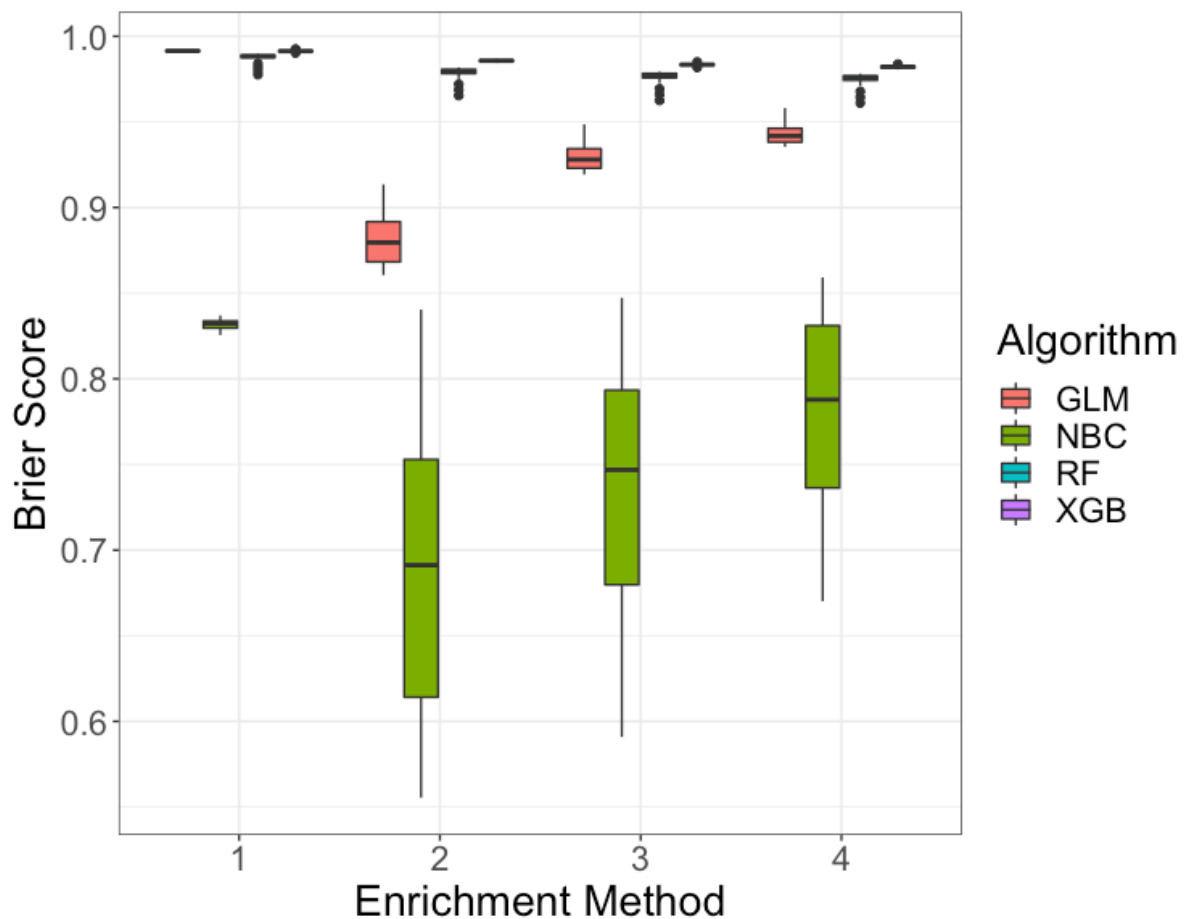
Topic	Original plan: quote from protocol paper	Revised action
	<p>“We will re-train the model using the hyper-parameter specifications from the best performing model, with a modified version of the derivation dataset which incorporates data extracted from secondary care records (such as A&amp;E presentations for asthma attack not captured in primary care records) in the determination of the risk factors. This allows us to evaluate the added value of secondary care data linkage in the prediction of impending asthma attacks, and will be determined by the same metrics used for the primary model evaluation”</p>	<p>I was not able to conduct the planned analyses of the increased predictive accuracy when features extracted from secondary care data sources (such as the accurate number of previous A&amp;E presentations) were used, due to time constraints.</p>

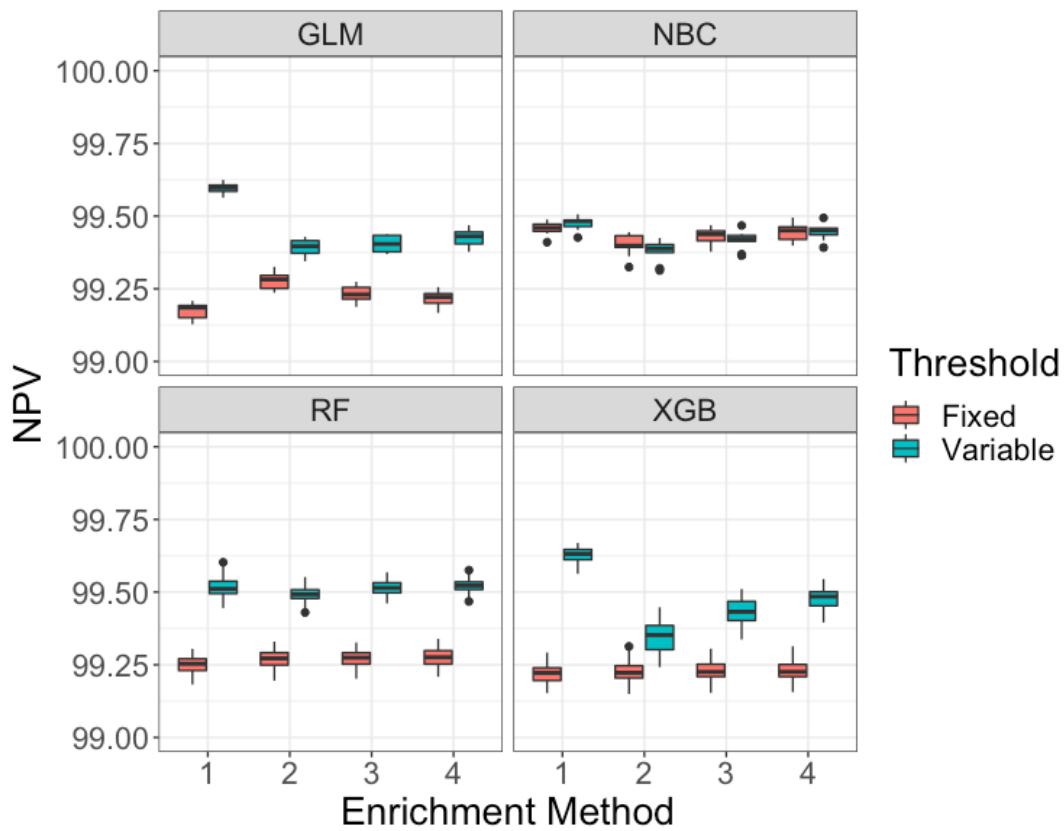
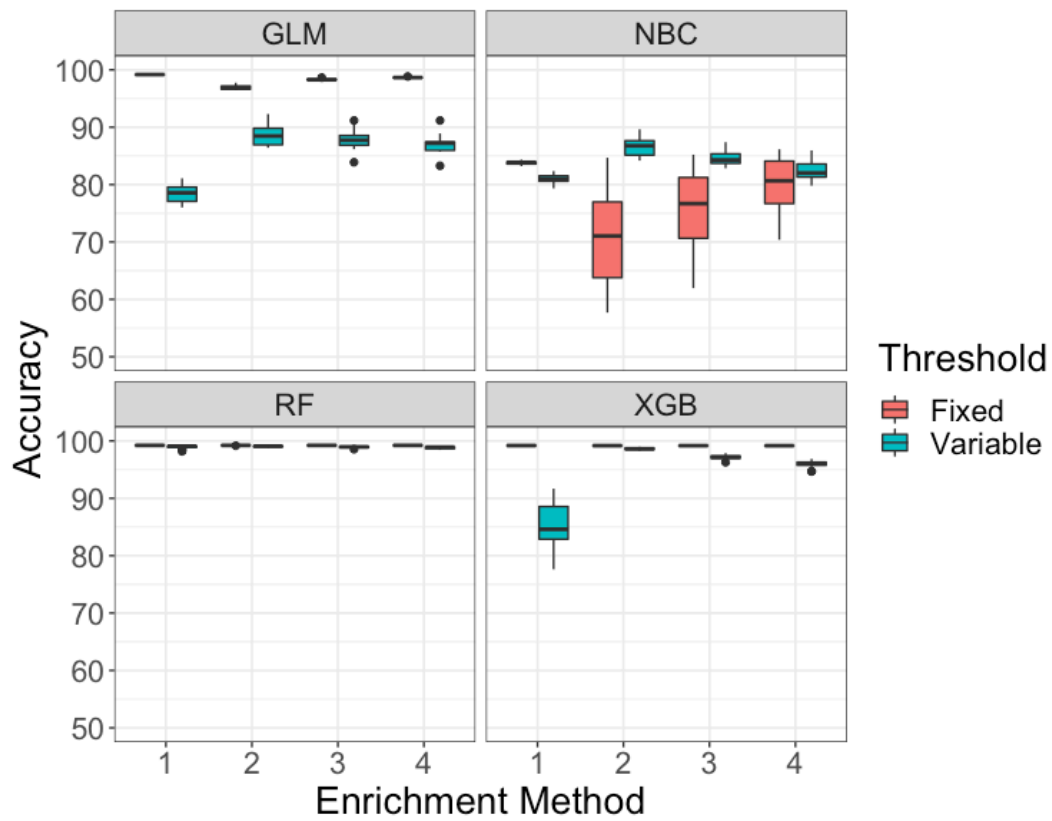
## Appendix R: Algorithm and Enrichment Selection: Additional Performance Measure Boxplots

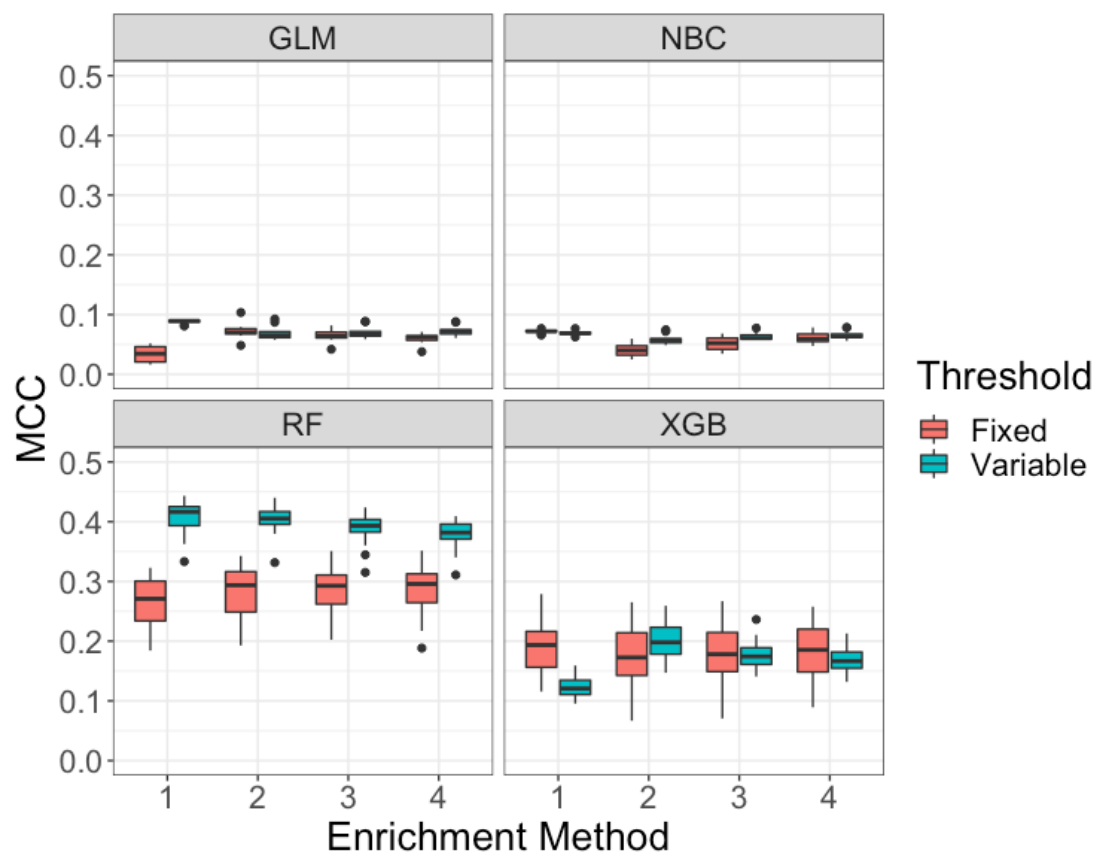
Notes: Algorithms: GLM = Generalised Logistic Regression, NBC = Naïve Bayes Classification, RF = Random Forest, XGB = eXtreme Gradient Boosting.

Enrichment methods: (1) unenriched data, (2) high up-sampling SMOTE, (3) medium up-sampling SMOTE, (4) low up-sampling SMOTE.

Thresholds: Fixed = 0.5, Variable = balanced accuracy optimising threshold in training data.







## Appendix S: Feature Importance

NUTS-3 Area code features have been omitted from the table, with feature importance ranks preserved.

Ranking	Feature	Importance
1	CSA_3	1238.73256
2	Reliever medication usage	966.64807
3	Age	926.46366
4	CMA8_2	825.19142
5	Number of asthma controller medications	589.03353
6	BTS Step	273.78548
7	December	164.15359
8	November	129.10426
9	October	124.52131
10	Chronic pulmonary disease	122.74018
11	January	117.98583
12	September	117.86288
13	Recent steroid prescriptions	116.18237
14	August	109.61743
15	SIMD Quintile 1 (Most Deprived)	104.14281
16	February	103.75893
17	March	103.06641
18	April	99.16979
19	Nasal spray: in last year	97.32321
20	SIMD Quintile 2	96.5793
21	Nasal spray: never	96.37738
22	July	95.96794
23	SIMD Quintile 3	93.21225
24	June	93.13822
25	Obese	89.75654
26	May	89.25433
27	Nasal spray: most recent prescription between 1 and 5 years ago	86.63695
28	Recent asthma encounters	85.14908
29	SIMD Quintile 4	82.48751
30	Last asthma attack recorded in primary care: more than 2 years ago, or never	77.7579
31	UR6 Level 2 (Other Urban Areas)	75.46633
32	SIMD Quintile 5 (Least Deprived)	74.38821
33	Smoking: former	74.02668
34	Blood eosinophil count: <400 cells per $\mu$ L	71.97436



Ranking	Feature	Importance
37	Nebulised SABA	70.6153
38	Smoking: never	70.07
39	Smoking: current	68.90283
40	Blood eosinophil count: missing	67.40511
41	UR6 Level 5 (Accessible)	66.0845
42	Blood eosinophil count: $\geq 400$ cells per $\mu\text{L}$	66.01362
45	UR6 Level 1 (Large Urban Areas)	63.32584
47	Anxiety or depression: diagnosis in the last year	60.42213
48	Last asthma attack recorded in primary care: in the last month	58.7913
49	UR6 Level 3 (Small Towns)	58.21762
50	Sex: male	57.37407
51	Anxiety or depression: never diagnosed	57.13349
52	Sex: female	56.91534
55	Anxiety or depression: diagnosis more than five years ago	53.13463
56	Last asthma attack recorded in primary care: between 1-3 months ago	51.32111
58	Last asthma attack recorded in primary care: between 6-12 months ago	49.97327
59	Last asthma attack recorded in primary care: between 3-6 months ago	46.64843
60	Most recent ARI: none in the last 2 years	46.00958
61	Most recent ARI: between 1-2 years ago	41.73519
62	Most recent ARI: between 6-12 months ago	41.44628
64	UR6 Level 6 (Remote)	39.98814
65	Most recent ARI: less than 2 weeks ago	38.15304
66	Nasal spray: most recent prescription longer than 5 years ago	38.01244
67	Recent ARI	37.93222
69	Last asthma attack recorded in primary care: between 1-2 year ago	36.95823
70	Most recent ARI: between 2-6 months ago	36.88576
72	Most recent ARI: between 2 weeks and 2 months ago	32.7854
73	Eczema: never diagnosed	32.52421
74	GERD: diagnosis between 1 and 5 years ago	31.71125
75	Eczema: diagnosis in the last year	31.12032
76	Eczema: diagnosis between 1 and 5 years ago	30.28
77	Diabetes	30.17845
78	Renal disease	30.08415
79	GERD: never diagnosed	29.88308
81	UR6 Level 4 (Rural Areas)	26.98316

Ranking	Feature	Importance
82	Rhinitis: never diagnosed	26.06488
83	GERD: diagnosis in the last year	24.20393
84	Cancer	24.16607
85	Cerebrovascular disease	22.64548
86	Rhinitis: diagnosis between 1 and 5 years ago	21.8377
87	Anxiety and/or depression: diagnosis longer than 5 years ago	21.74429
88	Rhinitis: diagnosis in the last year	21.40076
89	Peak flow: missing	20.44507
90	UR6: Missing	20.04361
92	Peak flow: Over 90% of previous best	18.47437
94	Diabetes with complications	17.48473
95	Peptic ulcer disease	16.07865
96	Anaphylaxis: never diagnosed	15.82673
98	Congestive heart disease	15.28804
99	Rheumatological disease	13.47202
100	SIMD: missing	13.38391
102	Myocardial infarction	12.51991
103	Peak flow: between 70-80% of previous best	12.37725
104	Peak flow: between 80-90% of previous best	8.99995
105	Peripheral vascular disease	8.07812
106	GERD: diagnosis longer than 5 years ago	7.6681
107	Nasal polyps: diagnosed in the last year	6.77403
108	Nasal polyps: never diagnosed	6.12656
109	Moderate liver disease	5.88974
110	Eczema: diagnosis longer than 5 years ago	5.85268
111	Anaphylaxis: diagnosis longer than 5 years ago	5.38636
112	Anaphylaxis: diagnosis between 1 and 5 years ago	5.27932
113	Anaphylaxis: diagnosis in the last year	5.14537
114	Rhinitis: diagnosis longer than 5 years ago	4.95995
115	Peak flow: less than 70% of previous best	4.73282
116	Dementia	4.30758
117	Mild liver disease	3.88159
118	Nasal polyps: diagnosis between 1 and 5 years ago	3.80023
120	Metastatic tumour	1.97517
121	Nasal polyps: diagnosis longer than 5 years ago	0.86597
122	Hemiplegia	0.82026
123	AIDS	0.00062

# Glossary

Phrase	Definition
Asthma Attack	A sudden increase of constriction to the airways, leading to a drastic worsening of symptoms
Asthma Exacerbation	Another term for an asthma attack, discouraged by patient advocates due to perceptions about lack of severity
Interleukins	Immune response signalling molecules
Parameters	Finite unknown values
Parametric Algorithms	Algorithms which find estimates of parameters using training data
Non-Parametric Algorithms	Algorithms not requiring the specification of parameters, and whose functional form is instead inferred as part of the statistical learning process
Primary Care	Care provided at a GP surgery
Secondary Care	Care provided at a hospital or other community health service
Electronic Health Records	Digitised medical records, including primary care and secondary care
Cross-sectional studies	Studies using a single time-point per person
Longitudinal studies	Studies which follow people over a duration of time, and have multiple time-points per person
Sample	A single data point, or observation
Sample Size	The number of samples
Analysis Population Size	The number of individuals in the study data (equal to the sample size for cross-sectional studies)
Feature	A measurable property or characteristic of a sample, either comprised of raw data values, or some function of the raw data
Characteristic	The value of a feature, such as 'green' for the feature 'eye colour'
Outcome	The response, or label; that which we are attempting to estimate or predict

<b>Phrase</b>	<b>Definition</b>
Labelled Data	Data which includes a corresponding outcome for each sample
Algorithm	A mathematical process which specifies the steps for solving a problem. In machine learning, these steps tend to be iterative and run until a specific criterion is met
Supervised Learning	Determining a functional form associating a set of features with outcomes
Training Data	Data used to build a statistical model
Query Sample	A sample which was not part of the training data, which is presented in the statistical learning model to estimate the outcome
Test Data	Labelled data which are used to test the performance of a constructed statistical model by the comparison of the predicted and observed outcome
Class	A categorical outcome
Classification	A form of supervised learning that assigns query samples on a finite number of classes, as observed in the training data
Regression	A form of supervised learning that estimates a continuous or ordered outcome
Model	The product of applying a machine learning algorithm to training data, allowing estimation or prediction of outcomes for unseen test data
Over-fitting	Learning very well the training data but failing to generalize in new, unseen data
Under-fitting	Failing to capture the trends observed in the training data
Validation	The process of establishing the reliability of the model's performance in unseen data
Selection Bias	The deviation from a true estimate resulting from samples which are not representative of the population under analysis
Read Codes	5-byte (or 4-byte prior to 2010) hierarchical, case-sensitive, and ordered character strings, describing some factor of medical care, such as a diagnosis, a test, a survey, or a measurement
Metadata	Information about the specifications of the data, including size and format

Phrase	Definition
Ground Truth	The observed outcome (class or value)
Modifiable Risk Factor	A characteristic of a sample which can be purposefully changed in some way, such as smoking status or weight
Spirometry	A pulmonary function test conducted in specialist care
Atopy	The tendency to develop allergic conditions
Adherence	“The extent to which a person’s behaviour taking medication, following a diet, and/or executing lifestyle changes, corresponds with agreed recommendations from a health care provider” <sup>184</sup>
Initiation (Adherence)	Taking the first dose of a prescribed medication
Primary Adherence	Relating to the collection of an initial prescription after it is written
Discontinuation	The act of ceasing to take a prescribed medication
Implementation	The execution of the recommended treatment plan, during the period between initiation and discontinuation.
Persistence	The continuity of adherence, including both the duration of the time between treatment initiation and discontinuation
Treatment Intermission	A period of <i>non-persistence</i> , in which medication is not taken continuously for a duration
Re-initiation	The act of restarting a treatment after an intermission of treatment
(Medical) Electronic monitoring devices	Devices which enable the real-time tracking of medication-related device use, such as Bluetooth enabled inhalers which record the date and time of each dose actuation
Ensemble Learning	The method of combining multiple <i>base models</i> (also known as <i>weak learners</i> ), either in parallel or in sequence, in order to improve out-of-sample performance
Confusion matrix	A 2*2 table (in binary classification, or more generally m*m for multi-class classification problems) of the true and predicted classes
Discrimination	The assessment of how well the predicted risk allows us to distinguish between positive and negative samples
Calibration	The assessment of how well the predicted risk of an outcome corresponds to the observed outcome

<b>Phrase</b>	<b>Definition</b>
Imbalanced Data	One class is substantially larger than another
Training Data Enrichment	The artificial modification of data used for model training in order to improve estimation in unseen data
Over-sampling	A training data enrichment method which increases the size of the minor class, either by replication or generation of new synthetic samples
Under-sampling	A training data enrichment method which reduces the size of the major class
Synthetic Minority Over-Sampling (SMOTE)	A training data enrichment method which combines over-sampling in the minor class and under-sampling in the major class
Model Interpretability	The characteristic of a model such that the reasons why a prediction was made are intuitively understandable to a human
Feature Importance	The predictive value of each feature to a statistical model
Parallel Programming	The mode of executing sections of a program simultaneously and in parallel, rather than sequentially