



Swansea University
Prifysgol Abertawe



Swansea University E-Theses

Creating and utilising the Wales Asthma Observatory to support health policy, health service planning and clinical research

Al Sallakh, Mohammad A.

How to cite:

Al Sallakh, Mohammad A. (2018) *Creating and utilising the Wales Asthma Observatory to support health policy, health service planning and clinical research*. Doctoral thesis, Swansea University.
<http://cronfa.swan.ac.uk/Record/cronfa48569>

Use policy:

This item is brought to you by Swansea University. Any person downloading material is agreeing to abide by the terms of the repository licence: copies of full text items may be used or reproduced in any format or medium, without prior permission for personal research or study, educational or non-commercial purposes only. The copyright for any work remains with the original author unless otherwise specified. The full-text must not be sold in any format or medium without the formal permission of the copyright holder. Permission for multiple reproductions should be obtained from the original author.

Authors are personally responsible for adhering to copyright and publisher restrictions when uploading content to the repository.

Please link to the metadata record in the Swansea University repository, Cronfa (link given in the citation reference above.)

<http://www.swansea.ac.uk/library/researchsupport/ris-support/>

Creating and utilising the
Wales Asthma Observatory
to support health policy,
health service planning
and clinical research

Mohammad Anass Al Sallakh, MD, MSc

Submitted to Swansea University
in fulfilment of the requirements for the Degree of
Doctor of Philosophy in Medical and Health Care Studies

Swansea University

2018

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed (candidate)

Date

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s). Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed (candidate)

Date

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed (candidate)

Date

Acknowledgement

I thank my supervisors, Professors Gwyneth Davies, Sarah Rodgers, Ronan Lyons, and Aziz Sheikh for their precious support and advice during my candidature; Professor Damon Berridge for statistical assistance in Chapter 5; Stanley Musgrave, the collaboration with whom inspired me to develop a data extraction tool and make access to the Observatory easier for researchers; the SAIL Databank team and analysts for the fruitful discussions; and, not lastly, my wife, Dana, for her encouragement, support, and sacrifice.

This thesis was funded by a studentship from Health and Care Research Wales and Abertawe Bro Morgannwg University Health Board. The studentship underwent external peer review and was awarded by the Asthma UK Centre for Applied Research [AUK-AC-2012-01]. I acknowledge the training and support from the Asthma UK Centre for Applied Research and the Farr Institute of Health Informatics Research. The Farr Institute is supported by a 10-funder consortium: Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the Medical Research Council, the National Institute of Health Research, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates), the Wellcome Trust, (MRC Grant Nos: CIPHER MR/K006525/1, Scotland MR/K007017/1).



Dedicated to those who struggle to breathe

Summary

Asthma is a public health challenge in Wales. In order to understand and improve its outcomes and reduce its burden, reliable evidence on disease epidemiology is needed. In this thesis, I describe the development of a platform for asthma surveillance and research in Wales using routinely collected electronic health record (EHR) data in the Secure Anonymised Information Linkage (SAIL) Databank.

To inform the development of operational definitions for asthma and its outcomes, I examine the contemporary approaches to defining asthma and assessing its outcomes using EHR data, and describe significant variations and suboptimal reporting on these approaches. I highlight the need for valid, standardised methods to study asthma, and emphasise the increasing demand for improved reporting to support research transparency and reproducibility.

Acknowledging the infeasibility of reference standards to define asthma in SAIL, I describe the development of a latent class model to identify asthma patients in this databank. I assess the performance of this model in relation to other objective and self-reported measures of asthma.

I also describe other methodological aspects of the development of the Wales Asthma Observatory, including asthma data profiling and identification of important data gaps.

To demonstrate the Observatory's utility for health policy and service planning, I highlight the variations in asthma epidemiology in Wales across age groups, gender, and socioeconomic deprivation levels. I found that asthma patients living in the most deprived areas had higher healthcare utilisation for asthma, indicating worse disease control, than those in the least deprived areas.

Finally, I reflect on the experience of developing the Wales Asthma Observatory, recognising its strengths and limitations, and identify opportunities and challenges of maximising the use of routine data towards a learning health system for asthma in Wales.

Contents

List of Figures	ix
List of Tables	xii
Abbreviations	xv
1 Introduction	3
1.1 Overview of asthma	4
1.2 Asthma is a public health challenge	17
1.3 Asthma epidemiology in Wales	18
1.4 Asthma burden can be reduced	20
1.5 Routinely collected data: an overview	21
1.6 Routinely collected data can improve our understanding of asthma	22
1.7 Disease registries and observatories	23
1.8 An asthma observatory for Wales: opportunities, challenges, and solutions	25
1.9 Thesis aims, research questions, and objectives	26
1.10 Thesis structure	28
2 Defining and assessing asthma using EHR Data	33
2.1 Introduction	36
2.2 Methods: a systematic scoping review	38
2.3 Results	40
2.4 Discussion	45
2.5 Conclusion	50
3 Identifying asthma patients in Wales	55
3.1 Introduction	57
3.2 Objectives	71
3.3 Methods	72
3.4 Results	80
3.5 Discussion	97

3.6 Conclusion	103
4 Development of the Wales Asthma Observatory	107
4.1 Introduction	109
4.2 Purpose and context	109
4.3 Methods	110
4.4 Summary statistics	125
4.5 Quality of asthma-related events in the GP database	130
4.6 Discussion	135
4.7 Conclusion	142
5 Inequalities in asthma care and outcomes in Wales	145
5.1 Introduction	147
5.2 Aims and Objectives	154
5.3 Methods	154
5.4 Results	160
5.5 Discussion	177
5.6 Conclusion	188
6 General Discussion	191
6.1 Summary of findings	193
6.2 Original contributions	195
6.3 Strengths and limitations	196
6.4 Interpretation of findings in the light of related literature	199
6.5 Challenges	208
6.6 Implications and potential uses of the Observatory	212
6.7 Towards maximising population data benefits to improve asthma out-comes	214
6.8 Future work	218
6.9 Conclusions	222
References	225
Appendices	255
A Chapter 2 Appendix	255
A.1 Additional results for Chapter 2	255
A.2 Published paper related to Chapter 2	280
B Chapter 3 Appendix	295
B.1 Making sense of patient-reported currently treated asthma using rou-tinely collected data	295

B.2	Read Codes sets used in latent class analysis	298
B.3	Item-response probabilities for the competing latent class models . .	308
B.4	Study protocol: identifying patients with ACOS using LCA	321
C	Chapter 4 Appendix	333
C.1	SAIL Databank Datasets used in the Observatory	333
C.2	SAIL IGRP approval letter	364
C.3	A tool for automatic characterisation of cohorts using primary care data.	365
C.4	Read codes used in the assessment of data quality in Chapter 4 . . .	368
C.5	Density distributions of lung function test values	370
D	Chapter 5 Appendix	375
D.1	Meeting abstract	376
E	Clinical codes	379

List of Figures

1.1	Trends of self-reported and GP-reported treated asthma in Wales between 1995 and 2015.	19
2.1	Flowchart for study selection in this scoping review.	41
3.1	Visual representation of a latent class model.	64
3.2	The methodology followed in Chapter 3.	73
3.3	A histogram for the sample age at the beginning of year 2014.	80
3.4	Assignment of the sample individuals across the competing latent models.	83
3.5	Diagnostics for the competing latent class models.	84
3.6	Class prevalences and item-response probabilities of the eight-class model.	85
3.7	Visualisation of the eight-class latent class model using principal component analysis.	88
3.8	Visualisation of the 10-fold cross-validation results of the recursive partitioning.	90
3.9	A decision tree representation of the classification algorithm.	91
4.1	Compilation of the Wales Asthma Observatory.	111
4.2	Split-file approach to data anonymisation by a trusted third party.	118
4.3	Asthma case definitions as state variables in the Wales Asthma Observatory.	120
4.4	Incidence and prevalence of asthma in Wales between 2000 and 2017.	128
4.5	Asthma-related health care utilisation in patients with current asthma.	129
5.1	Map of Wales showing ranks of the 2011 Welsh Index of Multiple Deprivation for Lower-level Super Output Areas.	153
5.2	A flowchart of case selection.	161
5.3	Distribution of age across WIMD quintiles in the study cohorts.	165
5.4	Histograms of outcome event counts in the study cohorts.	167

5.5	Average counts of the outcome events between 2010 and 2014 per WIMD rank quintiles in Cohort 1 and Cohort 2.	168
5.6	Incidence rate ratios for the outcome variables in each of the deprivation quintiles.	170
5.7	Incidence rate ratios for the outcome variables across the age groups.	171
5.8	Quantile-quantile plot for the model residuals.	175
5.9	Rootograms illustrating the goodness of fit for the zero-inflated negative binomial models.	176
C.1.1	Frequency of events per year in the SAIL datasets.	363
C.2.1	SAIL IGRP approval letter.	364
C.3.1	A tool for automatic characterisation of cohorts using primary care data.	367
C.5.1	Beanplots showing density distributions for lung function event values.	370

List of Tables

1.1	Measures of performance of asthma tests.	9
1.2	Main clinical coding systems in the United Kingdom.	21
2.1	Practices of reporting or justifying the validity of algorithms to define and assess asthma using EHR-derived data.	45
3.1	Observed variables used in the latent class model.	75
3.2	Latent classes before and after merging.	89
3.3	Results of 10-fold cross-validation for the recursive partitioning model.	89
3.4	Confusion matrix and statistics for the cross-classification of the predicted classifications against the latent class analysis (LCA)-based labels.	93
3.5	Comparison between the classification algorithm and other commonly used case definitions.	95
4.1	Datasets used for the development of the Wales Asthma Observatory.	114
4.2	Case definitions used in the Wales Asthma Observatory.	119
4.3	Data tables in the asthma registry	121
4.4	All-time number of records and unique patients for each of the case definitions	126
4.5	Prevalence of asthma at national and health board levels in Wales in 2017.	127
4.6	Percentages of patients with at least one recording of key asthma-related GP events over specified periods from diagnosis.	132
4.7	Percentages of missing values for lung function event codes.	133
5.1	Characteristics of the source population.	160
5.2	Characteristics of the study cohorts.	162
5.3	Outputs of the zero-inflated negative binomial models.	172
A.1.1	Search query used in the systematic scoping review.	256

A.1.2	Charting table showing the data extracted from the reviewed articles.	257
A.1.3	Geographical distribution of the reviewed studies.	259
A.1.4	Study designs of the reviewed studies.	259
A.1.5	Types of EHR-derived data sources used in the reviewed articles.	259
A.1.6	Algorithms used to identify asthma patients.	260
A.1.7	Approaches used in identifying asthma patients.	263
A.1.8	Age restriction approaches used in asthma patient identification.	263
A.1.9	Co-morbidities and conditions based on which asthma patients were excluded.	264
A.1.10	Algorithms used to ascertain asthma severity using EHR data.	265
A.1.11	Algorithms used to ascertain asthma exacerbation using EHR data.	268
A.1.12	Algorithms used to assess asthma control using EHR data	274
B.2.1	Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3.	298
C.1.1	Data fields of SAIL datasets.	334
C.1.2	Frequency of events per year in the SAIL datasets.	362
C.4.1	Asthma diagnosis Read codes.	368
C.4.3	Asthma-related event groups chosen for the coding quality analysis.	369

Abbreviations

A&E	accident and emergency
ACBS	Asthma Call-back Survey
ACOS	asthma-COPD overlap syndrome
ACQ	Asthma Control Questionnaire
ACT	Asthma Control Test
ADDE	Annual District Death Extract
AHR	Airway hyperresponsiveness
AIC	Akaike information criterion
ALF	Anonymised Linking Field
AQLQ	Asthma Quality of Life Questionnaire
ARIMA	autoregressive integrated moving average
ARRISA-UK	“At-Risk Registers Integrated into primary care to Stop Asthma crises in the United Kingdom”
ATS	American Thoracic Society
AUC	area under the curve
AUKCAR	Asthma UK Centre for Applied Research
BIC	Bayesian information criterion
BRFSS	Behavioral Risk Factor Surveillance System
BSAR	Belgian Severe Asthma Registry
BTS	British Thoracic Society
CALIBER	Clinical research using LInked Bespoke studies and Electronic health Records
cAMP	cyclic adenosine monophosphate
CI	confidence interval
COPD	chronic obstructive pulmonary disease
CPRD	Clinical Practice Research Datalink
DALYs	disability-adjusted life years
ED	emergency department
EDDS	Emergency Department Data Set
EHR	electronic health record
EM	expectation-maximisation

FeNO	fractional exhaled nitric oxide
FEV1	forced expiratory volume in the first second
FVC	forced vital capacity
GP	general practitioner
HCRW	Health and Care Research Wales
HEDIS	Healthcare Effectiveness Data and Information Set
HRG	Healthcare Resource Group
ICD-10	the 10 th revision of the International Classification of Disease
ICS	inhaled corticosteroid
IgE	immunoglobulin E
IGRP	Information Governance Review Panel
IRR	incidence rate ratio
ISAAC	International Study of Asthma and Allergies in Childhood
ITP	immune thrombocytopenic purpura
JSON	JavaScript Object Notation
LABA	long-acting beta adrenoceptor agonist
LAMA	long-acting muscarinic antagonist
LCA	latent class analysis
LHS	learning health systems
LOS	length of stay
LR	likelihood ratio
LRTI	lower respiratory tract infections
LSOA	Lower Layer Super Output Area
LTA	latent transition analysis
LTRA	leukotriene receptor antagonist
MeSH	Medical Subject Headings
MIU	minor injury unit
MRC	Medical Research Council
NHS	National Health Services
NICE	National Institute for Health and Care Excellence
NIR	no information rate
NISCHR	National Institute for Social Care and Health Research
NLP	natural language processing
NPV	negative predictive value
NRAD	National Review of Asthma Deaths
NWIS	National Health Services Wales Informatics Service

OASIS	Ontario Asthma Surveillance Information System
OCS	oral corticosteroids
ONS	Office for National Statistics
OPD	Outpatient Dataset
PEAL	Population-Based Effectiveness in Asthma and Lung Diseases
PEDW	Patient Episode Database for Wales
PEFR	peak expiratory flow rate
PPV	positive predictive value
PROM	patient reported outcome measure
QOF	Quality of Outcomes Framework
RALF	Residential Anonymous Linking Fields
RCD	routinely collected data
RECORD	REporting of studies Conducted using Observational Routinely-collected health Data
ROC	receiver operating characteristic
SABA	short acting beta agonist
SAIL	Secure Anonymised Information Linkage
SAMA	short-acting muscarinic antagonist
SD	standard deviation
SIGN	Scottish Intercollegiate Guidelines Network
SNOMED-CT	Systematized Nomenclature of Medicine–Clinical Terms
SQL	Structured Query Language
STELAR	Study Team for Early Life Asthma Research
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
SWORD	Surveillance of Work-related and Occupational Respiratory Disease
UK	United Kingdom
UKAO	UK Asthma Observatory
US	United States
VC	vital capacity
WDS	Welsh Demographic Service
WHS	Welsh Health Survey
WIMD	Welsh Index of Multiple Deprivation
WLGP	Welsh Longitudinal General Practice
WRRS	Welsh Results Reports Service
WRS	Weekly Returns Service
ZINB	zero-inflated negative binomial

Chapter 1

Introduction

Opportunities for better understanding of asthma through routine health data

Asthma is a public health challenge in Wales. In order to reduce asthma burden and improve its outcomes, reliable population-based evidence on the disease epidemiology is needed. In this chapter, I present an overview of asthma, its public health burden, and endeavours to understand its epidemiology in Wales. Then, I present an overview of routinely collected health data, and highlight their unique potentials for better understanding of asthma. I then introduce my thesis aim of developing the Wales Asthma Observatory, based on routinely collected data, as a platform for asthma surveillance and research. I then describe the thesis objectives, including exploring the practices and challenges of studying asthma using routinely collected data, description of the Observatory's methodology, and demonstration of its utility for health policy using inequalities in asthma outcomes as an example. I conclude with describing the thesis structure.

Chapter Contents

1.1 Overview of asthma	4
1.1.1 Asthma subtypes	5
1.1.2 Pathophysiology	5
1.1.3 Diagnosis	7
1.1.3.1 Patient medical history and clinical examination	7
1.1.3.2 Diagnostic testing	8
1.1.4 Management	12
1.1.4.1 Pharmacological treatments	13
1.2 Asthma is a public health challenge	17
1.3 Asthma epidemiology in Wales	18
1.4 Asthma burden can be reduced	20
1.5 Routinely collected data: an overview	21
1.6 Routinely collected data can improve our understanding of asthma	22
1.7 Disease registries and observatories	23
1.7.1 Disease registries	23
1.7.2 Health and disease observatories	24
1.8 An asthma observatory for Wales: opportunities, challenges, and solutions	25
1.9 Thesis aims, research questions, and objectives	26
1.10 Thesis structure	28

1.1 Overview of asthma

Asthma is a chronic respiratory disorder typically characterised by cough, dyspnoea, chest tightness, and wheeze [1]. The clinical manifestations of asthma often exhibit a recurring and remitting pattern and a variable intensity over time [1]. The disease severity varies widely among patients, ranging from mild intermittent symptoms to a severe persistent disease. Patients with any level of asthma severity may develop *exacerbations*—temporary periods of acute or sub-acute deterioration in symptom control that can be life-threatening [2].

1.1.1 Asthma subtypes

Asthma is increasingly considered heterogeneous rather than a single disease. There is growing evidence that “asthma” should be regarded as an umbrella term comprising distinct phenotypes with different clinical presentations potentially explained by distinct endotypes with different pathophysiological mechanisms [3–5]. The aetiology of asthma is thought to vary between phenotypes, with a wide range of potential risk factors related to the host, genetics, and the environment [6]. Asthma can also be classified based on other criteria such as triggerability by allergens into *allergic* and *non-allergic*, and age of onset into *early-* and *late-onset* disease [7]. In addition, other recognised phenotypes include *infectious*, *aspirin-induced*, *occupational*, *exercise-induced*, and *obese asthma* [8].

The heterogeneous nature of asthma contributes to the challenges of diagnosis, treatment, and developing epidemiological definitions as discussed later in [Chapter 2](#) and [Chapter 3 \(Section 3.1.1\)](#).

1.1.2 Pathophysiology

Asthma signs and symptoms result from complex underlying pathophysiology in the airways that involves chronic inflammation, remodelling, hyperresponsiveness, and obstruction.

Airway inflammation

Airway inflammation results from an abnormal immune reaction to mostly exogenous stimuli (e.g., pollen, viruses, and bacteria). Type 2 T-helper cells are a major player in this immune reaction, which involves recruitment and activation of several other types of immune cells in the airway mucosa, including eosinophils, mast cells, basophils, neutrophils, monocytes, and macrophages [9].

Eosinophils have a unique role in asthma pathogenesis, and they are increased in number in the airway, sputum, and blood. They produce chemical mediators such as leukotrienes which cause airway smooth muscle contraction, recruitment of inflammatory cells, and an increase in mucus production [10].

Mast cells are increased in the airway mucosa and smooth muscle, and release histamine, leukotrienes, cytokines, chemokines, and growth factors, which target the smooth muscle, vessels, mucous glands, and sensory nerves [11].

Bradykinin, a natural vasoactive peptide and a by-product of the inflammatory process, causes bronchoconstriction and cough [12].

Airway remodelling

Chronic inflammation in the airways leads to several irreversible structural changes in the airways. These include airway wall thickening, hyperplasia and hypertrophy in the epithelium, mucous glands and smooth muscle cells, submucosal collagen deposition and basal membrane thickening, and neovascularisation and increased sizes of vessels [9, 13]. Airway remodelling leads to irreversible narrowing of the airways and contributes to the decline in the reversibility of airway obstruction [9, 11].

Airway hyperresponsiveness

Airway hyperresponsiveness (AHR) is a measure of variable airflow limitation. It represents the exaggerated response of airway smooth muscle to inhaled stimuli leading to bronchoconstriction, which causes a decline in airflow. This response may occur due to nerve stimulation, mast cell mediated events, or direct effects of stimuli on the airway smooth muscle [9, 13]. Several factors may contribute to development of AHR, including eosinophilic airway inflammation, airway epithelial damage, airway remodelling, and increased contractility of airway smooth muscle [14].

Airway obstruction

Airway obstruction in asthma results from a combination of several factors, such as increased airway thickness (due to smooth muscle cell hypertrophy and hyperplasia, and mucosal oedema), exaggerated bronchoconstriction, and excessive mucous secretion and plugs in the airway lumen [15]. Airway obstruction leads to dyspnoea, chest tightness, wheeze and a variable decline in lung function. Reversible airway obstruction is characteristic of asthma, but reversibility tends to

decline over time with the development of fixed obstruction due to airway remodelling. The variable nature of obstruction has been attributed to AHR [13].

1.1.3 Diagnosis

Asthma diagnosis is mainly clinical, although reasonably certain diagnosis often requires a combination of careful medical history taking, clinical examination, and objective tests [16]. The clinical definition of asthma by the British Thoracic Society (BTS)/Scottish Intercollegiate Guidelines Network (SIGN) guidelines requires at least two of the main symptoms (cough, dyspnoea, chest tightness, and wheeze) in addition to evidence of variable airflow obstruction [16]. With high clinical suspicion of asthma, trial treatment with bronchodilators can be initiated; a good response to this treatment, assessed with objective testing, allows confirmation of asthma diagnosis [16]. An intermediate probability of asthma, based on medical history and clinical examination, warrants testing for airway obstruction (variability and hyperresponsiveness) and airway inflammation [16]. Differential diagnoses include an extensive list of disorders that can masquerade as asthma such as chronic obstructive pulmonary disease (COPD), cystic fibrosis, vocal cord dysfunction, rhinitis, chronic cough syndrome, and gastro-oesophageal reflex [16–18].

1.1.3.1 Patient medical history and clinical examination

Medical history is useful to establish whether symptoms, their onset, patterns (e.g., episodic), and triggerability (e.g., association with known stimuli) are compatible with asthma. It may help to identify risk factors of asthma (e.g., atopy or a family history of asthma), assess disease severity, and to rule out alternative conditions.

On examination, the patient may show breathlessness (e.g., inability to complete sentences) and tachypnoea. Chest auscultation may reveal widespread expiratory wheeze, while chest percussion may reveal hyperresonance [13]. Non-pulmonary atopic findings such as atopic dermatitis and allergic rhinitis support the diagnosis of asthma. In moderate and severe disease, the use of accessory respiratory muscles, intercostal retractions, and pulsus paradoxus may be observed. Physical examination is useful to detect signs suggestive of alternative diagnoses (e.g., unilateral wheeze, focal lung abnormalities, and finger clubbing) [16].

1.1.3.2 Diagnostic testing

The diagnostic tests in asthma allow investigating airway inflammation and responsiveness, establishing variability and reversibility of airflow obstruction, and ruling out alternative diagnoses that mimic asthma. However, they can be negative during the asymptomatic periods of the disease and are therefore insufficient alone to establish or rule out asthma diagnosis [1, 16]. Table 1.1 shows measures of performance of tests used in asthma diagnosis.

Lung function tests

Lung function tests include spirometry, plethysmography, peak flow measurement, and diffusion capacity assessment.

A spirometer is a device that measures inspired and expired air volumes. The forced vital capacity (FVC) and the forced expiratory volume in the first second (FEV1) are central to the assessment of airflow limitation. Following a full inspiration, the FVC represents the maximum possible volume of exhaled air whereas FEV1 represents only the volume of exhaled air in the first second. Both measures decrease in airflow limitation in asthma, but FEV1 usually decreases to a larger extent than FVC. The ratio of the measured to predicted FEV1 often remains above 80% in mild disease and declines below 60% in untreated severe disease [19]. The lower limit of normal FEV1/FVC ratio ranges between 85% to 70% depending on age [16, 20]. Lower values for the respective age group are considered positive for airflow obstruction and are often present in moderate and severe asthma [16]. However, these airflow limitation criteria alone, are neither sensitive nor specific to asthma. More than half of patients with normal FEV1/FVC will have asthma [16, 21]. FEV1 is often normal in children with persistent disease and may fall in other respiratory diseases [22]. Therefore, a normal spirometry test performed when the patient is not symptomatic does not rule out asthma diagnosis. Furthermore, a single measurement of FEV1 correlates poorly with asthma severity as classified by symptoms and medications [23].

Reversibility to bronchodilators (also known as bronchodilator response) is the improvement in airflow following acute treatment with a beta-agonist bronchodilator. A positive test is defined as a 12% or more improvement, which should be at least 200 mL, in FEV1 over the baseline [24]. However, these criteria may be

Table 1.1: Measures of performance of asthma tests. This table is adapted from [16]. The reference tests were (spirometry and (bronchodilator reversibility or a challenge test)) with or without ‘typical history of attacks’, diurnal variation, physician diagnosis, documented history of wheeze on at least two occasions, and variability in FEV1 over time or during exercise testing.

Test	Description	Age group	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)
Spirometry	FEV1/FVC < 70%	adults	23–47	31–100	45–100	18–73
		children	52	73	75	49
Bronchodilator reversibility	Improvement in FEV1 of $\geq 12\%$ (and ≥ 200 ml in adults)	adults	17–69	55–81	53–82	22–68
		children	50	86		
Challenge tests	Methacholine PC(20) value of ≤ 8 mg/ml	adults	51–100	39–100	60–100	46–100
		children	47–86	36–97	20	94
	Mannitol: 635 mg cumulative dose causing a decrease in FEV1 of $\geq 15\%$	adults	65	75	80	49
		children	63	81		
	Exercise	adults	26–80	100	100	0
		children	69–72	69–72	90–99	5–73
Peak flow, mean variability over 2-4 weeks	$\geq 20\%$	adults	46	80	97	10
	$\geq 15\%$	adults	3-5	98-99	60-67	60
	$\geq 15\%$ (>3 days/week)	adults	20	97	82	64
	$\geq 12.3\%$	children	50	72	48	74
Fractional exhaled nitric oxide	≥ 40 parts per billion	adults	43–88	60–92	54–95	65–93
	≥ 35 parts per billion	school children	57	87	90	49
Blood eosinophils	$>4.15\%$	adults	15–36	39–100	39–100	27–65
	$\geq 4\%$	children	55–62	67–84	56–69	73
Immunoglobulin E	Any allergen-specific IgE > 0.35 kU/l	adults	54–93	67–73	5–14	95–99
	Total IgE > 100 kU/l	adults	57	78	5	99
Skin prick test	Wheal ≥ 3 mm	adults	61–62	63–69	14–81	39–96
		children	44–79	56–92	65–92	36–79

falsely negative in many asthma patients, especially in those on treatment [15]. In addition, reversibility to bronchodilators may occur in some patients with COPD [25].

Peak flow (also known as peak expiratory flow rate, PEF_R) is the maximum rate of expiratory flow during a short and maximally forceful expiratory effort following a full inspiration. It can be measured by the patient using a portable peak flow meter at home. Daily recording of peak flow is useful to demonstrate variability of airway obstruction. Ideally, the patient is asked to make at least two recordings (during the day and the night). A seven-day average of the diurnal differences of more than 20% is considered positive for variability in airflow obstruction [16]. However, while this criterion has high specificity and positive predictive value (PPV), it has low sensitivity and negative predictive value (NPV), which means it produces a high rate of false negative results [16]. Peak flow is also useful for asthma monitoring [16]. Limitations of peak flow measurement, compared to spirometry, include the dependency on the patient's effort and difficulty of controlling measurement quality [26].

Diffusion capacity of the lung (DL), the ability to transfer gas (such as carbon monoxide) from air to alveolar vessels, is usually normal or increased in asthma [11, 27].

Whole-body plethysmography may show an increase in airway resistance, total capacity of the lung, and residual volume; however, it is rarely needed in asthma [11].

Airway responsiveness testing

The exaggerated response of airways (i.e. AHR) in asthma, which is an indicator of airflow obstruction variability, can be investigated using bronchoconstrictor stimuli. These stimuli can be direct (such as methacholine or histamine) or indirect (e.g., mannitol or exercise) [13, 16]. A decline in FEV₁ of 20% or more is considered a positive result [16]. However, this test is not suitable for patients with significant decline in lung function [16].

Exhaled Nitric Oxide

Nitric oxide (NO) is produced by several types of cells in the airway, including eosinophils, during inflammation. The fractional exhaled nitric oxide (FeNO) can be therefore used as a biomarker of airway inflammation in asthma. It can be helpful in the diagnosis, disease categorisation (eosinophilic vs. non-eosinophilic),

and determining the responsiveness to, and adjustment of, corticosteroid treatment [28]. The diagnostic accuracy measures of FeNO range between 43-95%, depending on the studies [16].

Skin prick testing

Skin prick tests help determine whether the patient is allergic to a common allergen (e.g., house dust mite). The test involves introducing a small amount of an allergen into the superficial epidermis. A weal with diameter of 3 millimetres or more suggests the patient has specific immunoglobulin E (IgE) antibodies for the used allergen [29]. Skin prick tests can help differentiate between allergic and non-allergic asthma. However, due to its mediocre sensitivity and specificity to asthma diagnosis, it cannot be used to establish or rule out the disease on its own [16]. The current National Institute for Health and Care Excellence (NICE) guidelines on asthma (NG80) recommends that skin prick testing should not be offered for the purpose of establishing the diagnosis but rather to identify triggers after a formal asthma diagnosis is made [30].

Blood tests

A blood eosinophil count of more than 4% is suggestive to atopy. However, this threshold, alone, has a poor predictive value for asthma diagnosis [16]. It is worth noting that marked blood eosinophilia suggests alternative diagnoses (e.g., parasitic infestation or familial eosinophilia) [16].

Serum IgE can be helpful in asthma diagnosis. In adults, a serum level > 0.35 kU/l of specific IgE antibodies to seasonal and perennial allergens, or total serum IgE > 100 kU/I, indicates an atopic state. However, these thresholds have a very low PPV, i.e. positive results poorly predict asthma [30]. In contrast, normal serum level of IgE substantially reduces the probability of asthma in adults with an NPV of 95-99% [16].

The current NICE guidelines on asthma, however, recommends that blood eosinophil count and total and specific IgE tests should not be offered for the purpose of establishing asthma diagnosis [30].

Imaging

Chest X-ray is usually normal in asthma, but it may show lung hyperinflation in severe disease and pneumothorax in exacerbations and it is useful to exclude alternative diagnoses [11].

1.1.4 Management

Asthma cannot be medically cured. Therefore, asthma management aims to adequately control symptoms and maintain normal activity levels (i.e. minimise asthma impairment) and minimise the risk of asthma exacerbations with the least possible treatment adverse effects [1, 16].

Asthma management involves key clinical concepts [2, 19]. *Asthma control* is the extent to which symptoms are sufficiently eliminated or reduced through treatment to an acceptable target. Complete asthma control is defined as the absence of symptoms, activity limitation, need for rescue medication, and asthma attacks, with normal lung function and minimal side effects of treatment [16].

Responsiveness to treatment is the ease with which disease control is achieved. *Impairment* refers to symptom severity and the resultant functional limitation. *Risk* is the probability of future exacerbations, chronic morbidity, and adverse effects of medications. Lastly, *asthma severity* is a complex concept composed of the following components: *level of control* including level of impairment and exacerbations in the last 12 months, *level of current prescribed treatment*, *responsiveness to treatment*, and *risk*.

Management guidelines have been developed globally [1] and nationally [19, 31]. In the United Kingdom, the national guidelines recommend measures for primary and secondary prevention, pharmacological management, in addition to guided self-management [16].

Disease self-management is important in asthma and requires an adequate level of patient education. Asthma self-management includes effective trigger avoidance, adherence to treatment, appropriate inhaler technique, regular monitoring of peak flow, and following up the personalised action plan [13].

Most patients with asthma can be treated in primary care [32]. However, patients with more severe disease (e.g., requiring 'high-dose therapies' or oral steroids),

treatment-refractory disease, comorbidities, and/or risk of exacerbations need to be referred to specialist care to review the diagnosis and management plan [13, 16].

1.1.4.1 Pharmacological treatments

Asthma therapies can be categorised clinically into:

- Quick relief medications (rescuers), which include bronchodilators, namely short acting beta agonists, magnesium, and short-acting muscarinic receptor antagonists (in asthma attacks), and provide quick reversal of airway obstruction;
- Controller medications (preventers), which provide long-term symptom control. These include anti-inflammatory medications (inhaled and oral corticosteroids, leukotriene receptor antagonists, and anti-immunoglobulin E antibodies) and long-acting bronchodilators (long-acting beta adrenoceptor agonists, theophyllines, and long-acting muscarinic antagonists).

Pharmacological categories of therapies

Beta 2 adrenergic agonists

Beta 2 adrenergic agonists act by increasing intracellular cyclic adenosine monophosphate (cAMP) in airway smooth muscle cells. This inhibits contractility, decreases airway hyperresponsiveness, and improves lung function [11].

Inhaled short acting beta agonists (SABAs), such as salbutamol and terbutaline, have a quick onset on action (< 5 minutes) [33]. They are the main choice in relieving acute symptoms in asthma [16].

Long-acting beta adrenoceptor agonists (LABAs) have a slower onset of action (~5-30 minutes after inhalation) and their effects last longer (12 hours or more) [33]. They are usually used as additional controllers in combination with inhaled corticosteroids (ICSs), when the latter are not sufficient to control the symptoms and/or to reduce the side effects of ICS [16]. Examples of LABAs include salmeterol, formoterol, olodaterol, vilanterol, and indacaterol.

Side effects of β_2 adrenergic agonists may include muscle tremor, palpitations, and a mild decrease in serum potassium [33].

Anticholinergics

Muscarinic receptors antagonists inhibit cholinergic nerve-induced bronchoconstriction and mucus secretion [11]. Tiotropium, a long-acting muscarinic antagonist (LAMA), can be used as an additional controller if ICS and LABA combinations are not sufficient to control the symptoms [16]. Nebulised ipratropium, a short-acting muscarinic antagonist (SAMA), can be used along with a nebulised beta-2 agonist in severe acute asthma attacks to improve bronchodilation and accelerate recovery [16]. Side effects may include dry mouth, dizziness, cough, arrhythmias and, in the elderly, urinary retention and glaucoma [33].

Theophyllines

Theophyllines inhibit the metabolism of cAMP, which increase its activation of beta-adrenoreceptors, leading to relaxation of airway smooth muscles [11]. However, their narrow therapeutic window requires measuring their plasma concentration to adjust the dose. Side effects may include nausea, tachycardia, arrhythmias, and seizures [33].

Corticosteroids

Inhaled corticosteroids are the most effective controllers of asthma [16]. ICSs suppress airway inflammation mainly by suppressing the activation of the genes that produces the inflammatory mediators. They decrease the number and activity of inflammatory cells, particularly T lymphocytes, eosinophils, and mast cells, that are responsible for airway inflammation [11]. This improves the disease control by improving lung function, controlling airway hyperresponsiveness, and reducing asthma symptoms. Examples of ICSs include beclometasone, fluticasone, budesonide, mometasone, and ciclesonide [33]. Adverse effects may include oropharyngeal candidiasis, hoarseness, and, at high doses, adrenal suppression and increased bone turnover [33, 34].

Systemic corticosteroids, typically as oral prednisolone, are usually used as short courses to treat asthma exacerbations [16]. They can be also used as long-term treatment in patients with severe asthma that is uncontrolled with a high-dose ICS, a LABA, a leukotriene receptor antagonist, and theophylline [16, 33].

The dose of long-term corticosteroids in asthma should be carefully reviewed to ensure adequate disease control with the least possible side effects [16], which may include diabetes, cataracts, glaucoma, and osteoporosis.

Leukotriene receptor antagonists

Leukotriene receptor antagonists (LTRAs), or antileukotrienes, block either the synthesis of leukotrienes or their binding to leukotriene receptors, which suppresses bronchoconstriction, microvascular leakage, and eosinophilic inflammation [11, 13]. Examples of LTRAs include montelukast and zafirlukast [33]. LTRAs can cause modestly improve symptoms in exercise-induced asthma and in asthma patients with concomitant rhinitis [13, 15]. Common side effects include headache and gastrointestinal disturbances [33].

Biologic targeted therapy

Omalizumab is a monoclonal antibody that binds to circulating IgE and blocks its interaction with mast cells and basophils [11]. It is given as a subcutaneous injection [33]. Other biologic agents include mepolizumab and reslizumab, which are given by subcutaneous injection and intravenous infusion, respectively. These medications are very expensive and are reserved for patients with a distinct phenotype who fulfil NICE criteria; for omalizumab, patients with severe allergic asthma who have frequent exacerbations despite high doses of corticosteroids [16].

Side effects of omalizumab may include hypersensitivity reactions, leading to urticaria, hypotension, syncope, bronchospasm, and/or angioedema [33]. Rarely, eosinophilic granulomatosis with polyangiitis (EGPA, also known as Churg-Strauss syndrome) may occur due to corticosteroid withdrawal [35]. EGPA may manifest as hypereosinophilia, worsening pulmonary symptoms, vasculitic rash, eosinophilic myocarditis, and/or peripheral neuropathy [33].

Cromones

Cromones, such as cromolyn and nedocromil, inhibit mast cells and sensory nerve activation [11]. They have some benefits in adults and children aged > 5 years, and can be used in the control of exercised induced asthma [16]. They are listed

as alternate initial controller therapies for mild asthma in national and international guidelines, although ICSs are the preferred agents. Side effects are rare and include bronchospasm, cough, headache, eosinophilic pneumonia, rhinitis, and throat irritation [33].

Magnesium sulphate

Intravenous magnesium sulphate causes relaxation of airway smooth muscle and can be used in patients with acute severe asthma attacks who had no good initial response to inhaled bronchodilators [16].

Immunosuppressants

Immunosuppressants such as methotrexate may be initiated by specialists in patients with severe asthma to achieve disease control and reduce oral steroids, but there is no strong evidence base for their use [16, 33].

Treatment approach

The pharmacological management of asthma should follow a step-wise approach starting with the step most appropriate to the presenting disease severity. The 2016 BTS/SIGN guidelines on asthma management recommended the following treatment steps [16]:

For suspected asthma, monitored initiation of low-dose ICS treatment (very-low to low dose in children) is recommended.

For confirmed asthma:

- *Regular preventer*: Low-dose ICS (very low dose in children). In children < 5 years, LTRA inhalers can be used instead of ICS.
- *Initial add-on therapy*: Add LABA to ICS, normally as a combination inhaler. In children < 5 years, use LTRA with very low dose ICS.
- *Additional add-on therapy*: If no response to LABA, stop LABA, and increase ICS dose (to medium dose in adults, and low dose in children). If LABA addition was helpful but insufficient, continue LABA, but increase ICS (to medium dose in adults, and low dose in children) or try LTRA (in adults and children), theophylline or LAMA (in adults).

- *High-dose therapies*: Try increasing ICS (to high-dose in adults, and medium dose in children); adding a fourth drug (in adults: LTRA, theophylline, LAMA, or an oral beta-2 agonist; in children: theophylline).
- *Continuous or frequent use of oral corticosteroids*: Use daily corticosteroid tablet in the lowest dose that provides adequate disease control, maintain high-dose ICS (medium dose in children), and consider steroid-sparing therapies.

Regardless of the treatment step, patients with symptomatic asthma should use a SABA inhaler as required.

The most recent BTS/SIGN guidelines emphasise the importance of supported self-management, with all those with asthma being offered education for self-management which includes a written personalised asthma action plan in addition to regular review by health care professionals [16].

Treatment should be reviewed every three months until adequate disease control is achieved. Stepping up the treatment is warranted if the disease is not satisfactorily controlled with the current step (e.g., when more than two SABA inhalers are needed per week).

However, stepping up the treatment may expose the patient to increased side effects. Therefore, it should only be done after ruling out suboptimal adherence with existing therapies, poor inhaler technique, and exposure to avoidable or modifiable triggering factors [16].

1.2 Asthma is a public health challenge

Asthma is a worldwide public health problem affecting more than 300 million people worldwide [36]. The International Study of Asthma and Allergies in Childhood (ISAAC) was an important project for estimating and comparing asthma prevalence among children in 98 countries around the world [37]. The prevalence of clinical asthma varies between countries, ranging from around 1% in Indonesia to more than 10% in North and South America, Australia, and the United Kingdom [38]. In many countries, asthma prevalence seems to be increasing, and the worldwide trends do not seem to be decreasing [39].

Although mortality from asthma is relatively low, it represents only the tip of an iceberg of a wide range of adverse outcomes and a significant public health burden of the disease. The disease is associated with expensive health care utilisation, increased morbidity, reduced quality of life, and wider societal impact such as school and work absenteeism [40–42]. The annual cost of asthma in the United Kingdom (UK) in 2011 has been estimated as £1.1 billion pounds (approximately US\$1.8 billion) [43]. Significant financial costs of asthma care are mainly driven by medication use, exacerbations, and hospitalisations [41, 43, 44]. The costs are particularly high for patients with severe refractory asthma [45]. Furthermore, 15 million disability-adjusted life years (DALYs) per year worldwide were attributed to asthma [38]. Due to its high burden, asthma has increasingly received considerable attention by health policy makers and researchers [46].

1.3 Asthma epidemiology in Wales

The United Kingdom has one of the highest asthma prevalences in the world across all age groups [38, 43, 47–49]. According to Asthma UK, there were estimated 5.4 million asthma patients in the UK in 2015, of whom 314,000 are in Wales [50]. The mortality rate from asthma in the UK is among the highest across Europe [51].

Asthma epidemiology in Wales has been studied over the recent decades, mainly by means of cross-sectional health surveys and, to a lesser extent, using prospective cohort studies and routinely collected health data.

Health surveys are descriptive investigations of systematically collected health determinants [52], and they are usually used for cross-sectional analyses. Health surveys have been important tools to investigate asthma epidemiology in Wales. The Welsh Health Survey (WHS) has been conducted in 1995, 1998, and annually since 2003 before it ceased in 2015 [53]. In 2014, the WHS estimated the prevalence of patient-reported currently treated asthma in children and adults as 9% and 10%, respectively [54, 55].

The WHS had limitations for studying asthma epidemiology. Self-report has been prone to recall bias. In addition, sampling might not sufficiently represent the entire population since it was limited to people living in private households and excluded the homeless, older people, some migrant workers, and special populations (e.g., armed forces and prisoners) as well as those living in care institutions

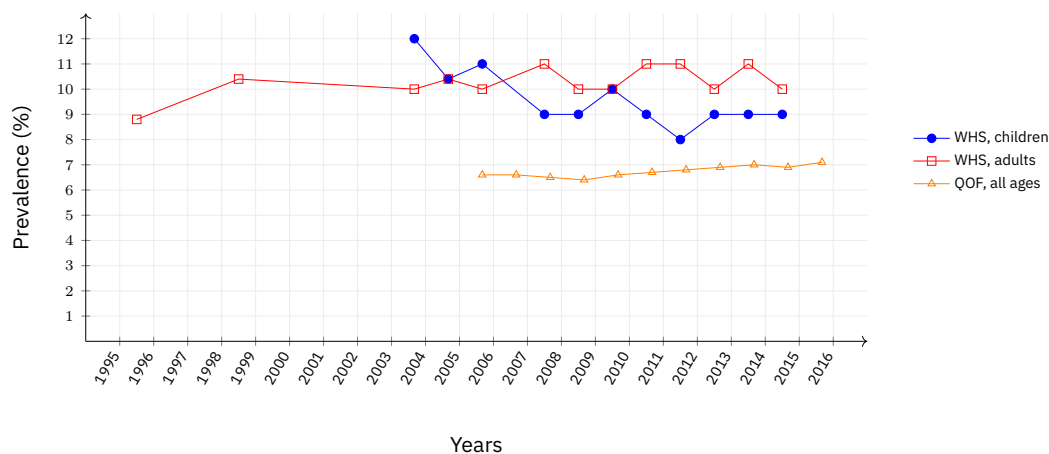


Figure 1.1: Trends of self-reported (in blue and red) and GP-reported (in orange) treated asthma in Wales between 1995 and 2015 according to the WHS [53] and the QOF [59], respectively.

and others due to language barriers [43, 56]. Being an annual study, the WHS had a limited value in providing timely disease insights. The differences in the asthma-related questions asked by the WHS and health surveys in the other UK countries hinder the comparability of self-reported asthma prevalence across the UK [43].

Electronic health record (EHR) data have been also used to estimate asthma epidemiology. Primary care data from a sample of Welsh general practitioner (GP) practices, covering around 1% of population in Wales, have been used by the Weekly Returns Service (WRS) of the Royal College of General Practitioners. However, the WRS reports did not produce separate asthma statistics for Wales [57]. Since April 2004, primary care data from more (and later all) GP practices in Wales have been used to produce clinical performance indicators as part of the Quality of Outcomes Framework (QOF) [58]. According to the QOF indicator of treated asthma (ASTHMA1 or AST001) the prevalence of patients with asthma diagnosis who received asthma treatment in the last 12 months, ranged from 6.4% in 2008–09 to 7.1% in 2015–16 [59]. These estimates were notably lower than the prevalence of patient-reported GP-diagnosed and treated asthma as estimated by the Welsh Health Surveys (see Figure 1.1).

A recent UK-wide analysis showed that Wales had a slightly higher asthma prevalence than the other member counties [43]. The annual prevalences of patient-reported doctor-diagnosed-and-treated and GP-reported-diagnosed-and-treated asthma in the fiscal year 2011–2012 were 9.8% and 6.9% in Wales compared to UK-wide estimates of 9.2% and 6.8%, respectively. That study demonstrated that asthma burden in Wales was high with estimated £74.7 million pounds being spent on

asthma care by the Welsh public sector in the fiscal year 2011-2012 [43]. This overall cost included £40.5 million on community prescribing, £9.7 million on GP and practice nurse consultations and out-of-hour calls, £3.3 million on accident and emergency (A&E) visits and ambulance trips, £8.4 million on hospitalisations, and £12.8 million on Disability Living Allowance.

1.4 Asthma burden can be reduced

Despite its high prevalence, asthma burden can be reduced by identifying modified risk factors of adverse outcomes, and improving allocation of health care resources [60, 61].

An investigation into asthma deaths in Wales between 1994-1996 [62] found that factors other than disease severity, such as inadequate treatment and patient factors, were identified in 70% of cases. The inquiry concluded that some of deaths that were attributed to asthma were preventable, and disease morbidity could be reduced.

The National Review of Asthma Deaths (NRAD), a UK-wide inquiry of the circumstances of deaths due to asthma, found that potentially modifiable risk factors played a significant role in the disease hospitalisation and mortality [51]. The NRAD report identified avoidable risk factor in two thirds of the reviewed asthma deaths. Adverse asthma outcomes can be avoided or ameliorated with early diagnosis, improved care, disease monitoring, patient education, personalised asthma action plan, and self-management [51, 63-65]. Exacerbations can be predicted using disease biomarkers and patient medical history [66-70].

Prevention of adverse outcomes can be boosted through better understanding of the disease epidemiology, trends, wider determinants, endotypes and phenotypes, and patterns of natural disease history. Systematic learning from health care data at the population level is crucial for improving asthma care and prevention of adverse outcomes. Effective surveillance for asthma requires a stream of real-time or near real-time data on asthma outcomes at the national level.

1.5 Routinely collected data: an overview

Documentation of health and care events is an integral part of health care. It serves a variety of purposes including supporting the delivery and continuity of care as well as administrative, financial, and legal purposes [71]. However, beyond these uses, large volumes of health data accumulate over time, and can be also useful for further purposes. Routinely collected data (RCD) are data that are regularly collected without *a priori* specific purpose into central repositories where they can be made available for secondary uses [72]. These data are usually collected in coded forms. Clinical coding systems helps standardising the documentation of clinical information and facilitates the identification of patients with a specific clinical profile. Table 1.2 shows examples of clinical coding systems used in the UK.

Common types of RCD include data on primary and secondary care (EHR data), laboratory tests, medication dispensing, medical insurance claims, vital events (e.g., births and deaths), social care, and education.

Examples of RCD uses include policy and service planning and evaluation [77], epidemiological studies, health surveillance, health technology assessment [78], comparative effectiveness research, and health economic analysis [43, 79].

Compared with traditional sources of health data, such as questionnaire surveys, medical record review, RCD have several advantages for research and surveillance. RCD are usually available in large volumes, and at relatively low costs [80].

Table 1.2: Main clinical coding systems in the United Kingdom.

Coding Scheme	Description
International Classification of Diseases, 10 th Revision (ICD-10) [73]	Produced and maintained by the World Health Organisation (WHO); aimed for international comparability of mortality and morbidity; used to classify health encounters based on diagnosis and health conditions for statistical and administrative purposes;
Read Codes [74]	The standard medical terminology for coding clinical information in the UK general practice.
Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [75]	A comprehensive medical terminology for documentation of variety of information types in clinical practice.
Office of Population Censuses and Surveys Classification of Interventions and Procedures (OPCS-4) [76]	Maintained by NHS Digital; used to code operations, procedures and interventions in UK secondary care.
British National Formulary (BNF) [33]	The reference book of coded medication and prescribing in the UK.

Many RCD sources cover entire populations, enabling generalisability of studies, reliable epidemiological estimates, and studying rare conditions [81, 82]. The contents of many RCD resources are rich and comprehensive enough to answer a wide range of research questions. Person-level record linkage of two or more RCD sources for the same population allows further research opportunities [83]. RCD are observational, objective data that are often recorded by qualified professionals rather than patients. They are therefore less prone to biases of self-report such as recall, learning, responses biases [84, 85]. Studies using RCD are not prone to experimenter bias since collection of data is independent from their secondary uses. The often-longitudinal nature of RCD allows conducting inexpensive, complex time series analyses. RCD reflect health, morbidity, and health care in the real world rather than in idealised or artificial conditions such as in clinical trials [81]. Therefore, they have been used in the various phases of clinical trials including patient follow-up and evaluation of real-world safety and effectiveness drugs [86, 87].

Nevertheless, being originally collected for other purposes, RCD usually have limitations for use in research. While clinical codes facilitate data standardisation, they do not cover all aspects of health and care. Important non-coded narrative data are usually missing from RCD. Incomplete, inaccurate, incorrect, and inconsistent capture and coding of data is a common concern about RCD [88–93]. Confounding is an important issue in RCD-based studies, and missing residual confounders limits causality inference [94, 95]. Users of RCD should consider their provenance and the circumstances under which they are collected. For instance, with incentivised documentation of care (e.g., QOF), some data items are better recorded than others [93].

With the increasing use and quality of RCD, they have been recognised as key sources of data in the strategic plans of several health research agencies, councils, institutes, and funders across the UK [96–98].

1.6 Routinely collected data can improve our understanding of asthma

Asthma is an exemplar of health problems in which RCD can be effectively used for surveillance and support health policy and research. Depending on their content,

RCD data can provide information on asthma symptoms, diagnosis, laboratory tests, disease severity, disease control, treatments, monitoring, exacerbations, healthcare utilisation, and care quality [99–105].

Asthma-related RCD have been increasingly used to study different clinical and epidemiological aspects of the disease. These include incidence, prevalence, and burden and trends of asthma and its outcomes [43, 83, 106–109], risk factors [110–114], and disease prediction [66, 68, 70, 115, 116].

RCD can be also useful inform management, prevention, and compare effectiveness of interventions [117–119].

RCD have been used in several countries to create disease registries and platforms for asthma surveillance and research [120–122]. In the next section, I present an overview of those projects.

1.7 Disease registries and observatories

1.7.1 Disease registries

A disease registry is a database that systematically tracks outcomes of interest for patients who have a particular health condition and live in a defined geographic area.

Disease registries are usually set up to support health policy, service planning and/or research [123]. They have been widely used in epidemiological studies and allowed assessment of disease outcomes and understanding the natural history of chronic diseases [124]. Some disease registries can be also used to support health care of individual patients [125].

In the UK, notable examples of asthma registries are the BTS Difficult Asthma Registry [126], its successor the UK Severe Asthma Registry,¹ and the UK Paediatric Difficult Asthma Network Registry.²

Traditionally, disease registries have relied on active, purpose-specific data collection. However, the growing number of EHR databases have provided inexpensive, rich, alternative sources of data. EHR-derived data are usually available in large volumes, and contain real-world data on patient care. These data have

¹<https://cl2.n3-dendrite.com/csp/asthma/frontpages/index.html>

²<http://rs2.e-dendrite.com/csp/paedasthma/frontpages/index.html>

been increasingly used in registries of long-term conditions such as chronic kidney disease [127], cancer [128], cardiovascular disease [129], and multiple sclerosis [130].

1.7.2 Health and disease observatories

Health observatories are projects often run by public health or academic institutions to monitor health status of the population.

There is no consensus on the definition of a health observatory [131, 132]. The term ‘observatory’ means that, unlike health authorities, health observatories observe and analyse health phenomena while ‘standing back’ from them [133]. This observational nature distinguishes health observatories from state-operated health surveillance systems [131–133]. They often combine academic rigour with the high quality of public health professional practice [132]. However, compared to academic research projects, observatories often seek to provide more timely answers to public health questions, supporting health service planners and policy makers’ continuous need for up-to-date evidence [132].

Common core functions of health observatories include:

- highlighting health problems requiring attention and measuring their prevalences and burdens,
- conducting health surveillance³ on those problems,
- producing actionable insights,
- evaluating service delivery, and
- forecasting the population health [132]

In undertaking these functions, health observatories need to identify the various sources of data that could be used to assess health problems, such as routinely collected data sets and disease registries [43]. They can also identify important gaps in data sources, and often seek to link different data sources for better understanding of health problems [43, 132, 134].

Health observatories often target several health problems [131, 132]. However, some are dedicated to specific public health problems, such as health inequality, or specific diseases, such as asthma.

³Health surveillance is the systematic, continuous analysis, interpretation, and feedback of data related to specific health problems in a population [52]

Several asthma-specific observatories and surveillance systems have been established around the world. In the UK, the Surveillance of Work-related and Occupational Respiratory Disease (SWORD) system, maintained data about workers with occupational asthma [135]. It used systematic, voluntary reporting from specialists of data about suspected new cases including demographics, occupation, and suspected causal agents. A similar system, the *Observatoire National de Asthmes Professionnels*, was established in France to monitor occupational asthma [136]. It was based on clinical, diagnostic and profession-related data about workers with occupational asthma, which were voluntarily reported by physicians using questionnaires. In Canada, the Ontario Asthma Surveillance Information System (OASIS) is based on a cumulative cohort of asthma patients in Ontario [134]. OASIS is based on health administrative data of physician billing, emergency department visits, and hospital admissions. It aims to monitor changes in asthma epidemiology, management, and health service use as well as variations in clinical practice and disease burden. In the United States (US), the Population-Based Effectiveness in Asthma and Lung Diseases (PEAL) Network is an asthma registry with wide research, surveillance, and public health applications [137]. The PEAL Network uses data of computerised medical billing and claims as well as pharmacy and laboratory data. Claim data for emergency department visits and hospitalisations have been also widely used for asthma surveillance elsewhere [105, 107, 120, 121, 138, 139].

1.8 An asthma observatory for Wales: opportunities, challenges, and solutions

Given the high asthma burden in Wales, the motivation of this doctoral project was the pressing need for a reliable tool to monitor and study the disease and its outcomes.

In the UK, including Wales, asthma is mainly managed in primary care. Primary and secondary care data have been routinely collected with high-to-complete geographical coverage across Wales. These data, as well as various other health data source, are maintained and linked in the Secure Anonymised Information Linkage (SAIL) databank at Swansea University [140, 141]. With the high asthma prevalence in Wales, large volumes of rich, representative, real-world, longitudi-

nal observational data on asthma patients are available in the SAIL Databank. This data-intensive environment provides a unique opportunity to develop an observatory for asthma in Wales.

Developing an observatory for asthma exploits the merits of RCD but is hindered by their limitations. Algorithms to identify patients with asthma and assess asthma outcomes using RCD data are essential in the observatory development. However, for a heterogeneous disease such as asthma, identifying cases and assessment of outcomes using RCD was a key challenge, and valid case definitions were needed. However, there were neither gold standards for the clinical definitions of asthma and its outcomes, nor consensus on their definitions using RCD.

With the above challenges, data-driven approaches can be employed to uncover patterns generated by the disease in the population. Populations health data ‘speak for themselves’. Their patterns, however, need to be carefully interpreted in the light of clinical and epidemiological knowledge and hypotheses, as well as data provenance and quality [142]. Using proper design and interpretation, data-driven approaches, such as mixture models (e.g., latent class analysis (LCA)), can be employed to identify asthma patients in a population.

1.9 Thesis aims, research questions, and objectives

The aims of this thesis are to describe the methodology and development of the Wales Asthma Observatory⁴ as a platform for asthma surveillance and research in Wales, and to demonstrate its utility for health policy. Throughout the thesis, I will refer to the Wales Asthma Observatory as *the Observatory*.

The thesis includes a systematic scoping review of approaches to define and assess asthma using RCD and their reporting quality. A particular focus of the thesis is the identification of asthma patients in Wales using RCD in the light of the absence of gold standard and inherent limitations of RCD.

Research questions

Throughout the thesis I will answer the following research questions:

⁴<http://www.wales-asthma-observatory.uk/>

1. What are the contemporary approaches and practices of identifying asthma patients and assessing asthma outcomes using RCD? How well have these approaches and their validity been described? ([Chapter 2](#))
2. Can a latent class model based on recorded primary care data identify clinically meaningful asthma-related groups? How does this model compare with commonly used doctor-reported and self-reported measures to identify asthma patients? ([Chapter 3](#))
3. How well have asthma-related primary care events been recorded in Wales? ([Chapter 4](#))
4. Do asthma patients across the socioeconomic scale in Wales have equal disease outcomes? ([Chapter 5](#))

Objectives

To address the research questions, I pursued the following objectives:

1. To review the current practices of studying asthma using routinely collected data ([Chapter 2](#)). This is performed through a systematic scoping review of the recent asthma literature with the following objectives:
 - a. To survey the algorithms used to define asthma and assess asthma outcomes using routinely collected EHR data;
 - b. To explore the practices of the reporting on the validity of those algorithms;
 - c. To assess the clarity of reporting on the implementation of these algorithms and other methodological aspects related to the use of RCD.
2. To discuss the challenges of identifying asthma patients using RCD ([Chapter 2](#) and [Chapter 3](#)).
3. To develop a data-driven reference identification for asthma in Wales based on routinely collected data ([Chapter 3](#)). This includes:
 - a. Development of a mixture model, namely latent class model, to identify patients with both ever and current asthma.
 - b. Evaluating the concordance of this model with other routine data-based case definitions as well as with self-reported measures.

4. To describe the purpose, context, and methodology used in the Observatory development, as well as data quality ([Chapter 4](#)). This will include the following sub-objectives:
 - a. To describe the purposes of the Observatory as a data-intensive platform for asthma surveillance and research.
 - b. To describe the logistics, technical infrastructure, and RCD databases based on which the Observatory was founded.
 - c. To define the Observatory's source population and the case definitions used to identifying asthma patients.
 - d. To describe the Observatory's data structure and the available variables.
 - e. To describe the approach used to improve efficiency and reproducibility of data interrogation.
 - f. To assess the quality of selected asthma-related primary care events in the Observatory.
5. To demonstrate the Observatory's utility for health policy by investigating inequalities in asthma outcomes ([Chapter 5](#)). This includes:
 - a. Investigate the variations in asthma-related outcomes across the deprivation groups in Wales, using a count regression model adjusted for age group and gender.
 - b. To interpret the model in the light of previous studies and strengths and limitations of routinely collected data used.
 - c. To reflect on the implications of the findings on health policy in Wales.
6. To reflect on the work presented in Chapters 2, 3, 4, and 5, identify main strengths and weaknesses of the Observatory, interpret the findings with related literature, discuss the opportunities and challenges to maximise the potentials of asthma data, and outline further research directions.

1.10 Thesis structure

The rest of the thesis is organised as follows:

In [Chapter 2](#), I present a systematic scoping review of the different approaches of defining asthma and assessing its key outcomes, and the clarity of reporting

on their implementation and validity. The review reveals wide variations in these approaches, and highlights the challenges of developing uniform RCD-based methods to study asthma. This systematic scoping review was published in a leading respiratory journal [102].

In [Chapter 3](#), I present an overview to approaches to develop valid methods to identify asthma patients using RCD. Recognising asthma heterogeneity, data limitations, and informed by the findings of the previous chapter, I justify using a data-driven approach to identify asthma patients. I describe a latent class model based on primary care data as a data-driven identification model for asthma in the Observatory. I also derived a reusable classification algorithm base on the latent class model. I compared the algorithm's performance with commonly used asthma case definitions based on objective and self-reported data.

In [Chapter 4](#), I describe the Observatory's context, purpose, data structure, case definitions, variables, information governance, and approaches to improve efficiency and reproducibility of the Observatory interrogation. I also examine the quality of asthma-related primary care data and discuss their implications on the Observatory utilisation.

In [Chapter 5](#), I demonstrate the Observatory's utility for health policy by examining the inequality gaps in asthma outcomes across the socioeconomic spectrum in Wales.

Finally, in [Chapter 6](#) I reflect on the work presented in this chapter, present with the thesis's original contributions, and discuss the Observatory's strength and weaknesses in the wider context of asthma and RCD research. I explore opportunities and challenges towards maximising the value of RCD to improve asthma outcomes. I then propose future research directions and developments to improve the Observatory's methodology and content.

Chapter 2

Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review

Same conditions, different definitions

This chapter is based on the following published paper:

Al Sallakh MA, Vasileiou E, Rodgers SE, Lyons RA, Sheikh A, and Davies GA. Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. *Eur Respir J* 49 (6 2017) 1 (see [Appendix A.2](#)).

In this doctoral project, I used routine data to create a registry and observatory for asthma in Wales. However, the heterogeneous nature of asthma and the limitations of routinely collected EHR data impose challenges on how these data could be used for this purpose. In addition, there is currently no consensus on approaches to defining asthma or assessing asthma outcomes using electronic health record (EHR)-derived data.

To inform the methodology of identifying of asthma patients and assessing the outcomes in this project, described in Chapter 3, in this chapter, I present a systematic scoping review of how routinely collected EHR data have been used in the recent international asthma literature.

I systematically searched for asthma-related articles published between 1-1-2014 and 31-12-2015. From the eligible studies, I extracted the algorithms used to identify asthma patients and assess severity, control and exacerbations from using different types of EHR-derived data sources. I also investigated how authors justified the validity of these algorithms, and how they reported on the aspects related to the use of EHR-derived data in their studies.

From 113 eligible articles, I found significant heterogeneity in the algorithms used to define asthma ($n = 66$ different algorithms), severity ($n = 18$), control ($n = 9$), and exacerbations ($n = 24$). For the majority of algorithms ($n = 106$), validity was not justified. In the remaining cases, approaches ranged from using algorithms validated in the same databases, to using non-validated algorithms that were based on clinical judgement or clinical guidelines. The implementation of these algorithms was sub-optimally described overall.

Although EHR-derived data are now widely used to study asthma, the approaches being used are significantly varied and are often underdescribed, rendering it difficult to assess the validity of studies and compare their findings. Given the substantial growth in this body of literature, it is crucial that scientific consensus is reached on the underlying definitions and algorithms.

Chapter Contents

2.1	Introduction	36
2.2	Methods: a systematic scoping review	38
2.2.1	Identifying the research questions	38
2.2.2	Identifying relevant studies: literature search strategy	38
2.2.3	Study selection	39
2.2.4	Data extraction, charting and synthesis	39
2.2.5	Collating, summarising and reporting the results	40
2.3	Results	40
2.3.1	Characteristics of studies and data sources	40
2.3.2	Defining asthma	41
2.3.2.1	Various diagnostic labels of asthma	41
2.3.2.2	Approaches to restrict study domain	42
2.3.3	Assessing asthma severity	43
2.3.4	Assessing asthma control	43
2.3.5	Defining exacerbations	43
2.3.6	Clarity of reporting	44
2.4	Discussion	45
2.4.1	Statement of main findings	45
2.4.2	Strengths and limitations	46
2.4.3	Interpretation in the light of previous studies	46
2.4.4	Implications for policy, practice and research	49
2.5	Conclusion	50

2.1 Introduction

Methods to define and assess asthma are key to the development of the Wales Asthma Observatory. To inform the development of Observatory, it was important to survey the contemporary methods to define and assess asthma using routinely collected electronic health record (EHR) data. This would provide insights into the different approaches and the challenges to define the disease from these data.

Asthma is in clinical practice a diagnosis based on the patient's history, examination and objective tests [143]. Typically, the diagnosis is made initially based on signs and symptoms such as wheezing, shortness of breath, tightness of chest, or cough that often follow characteristic fluctuating patterns and triggerability. Confirmation of diagnosis requires demonstrating variable expiratory obstruction of airflows over time [1]. However, asthma is not a single clinical entity. Instead, it is increasingly considered to represent a heterogeneous group of disorders with different phenotypes and endotypes [5]. In addition, several other diseases may masquerade as asthma leading to misdiagnosis [143]. Subsequently, the clinical definitions of asthma and its key outcomes, including asthma severity, control, and attacks/exacerbations have been the subject of vigorous debate with no consensus yet reached on a gold standard for diagnosis [2, 144–148].

The uncertainty in the clinical definition of asthma has significant implications for research. Particular challenges arise in the context of epidemiological studies where groups of populations rather than individual patients are compared, and in which validated, standardised, operational case definitions are needed [149, 150]. Various types of health data have been used in these studies. Traditionally, these data were often self-reported or collected by investigators specifically for the purpose of research. In addition, these studies have been increasingly conducted using data derived from EHR such as health administrative data, health insurance data, primary care data, dispensing data, and disease registries. These data usually have the advantages of being inexpensive, objective, read-world data, and are usually available in large volumes. However, they have inherent limitations such as incorrect, inconsistent, or missing recording of health care events, resulting in many key clinical variables being missing or of low quality [82, 151] (see Section 1.5).

The limitations of routinely collected EHR data add a further layer of complexity and challenges to the use of these data in asthma research. Investigators often resort to indirectly assess low-quality variables using surrogate variables of higher quality or algorithms based on those surrogate variables. Although the face validity is important, formal validation of these algorithms is critical to ensure reproducibility of findings of studies conducted using routinely collected EHR data [152]. With the increasing and widespread use of EHR data in asthma literature, it is however unclear to which extent algorithms to define and assess asthma using these data have been supported by sufficient evidence of validity.

Clear reporting of methodology in studies conducted using routinely collected data (RCD) is critical not only for scientific transparency and understandability by their consumers (e.g., researchers, service providers, policy makers), but also for reproducibility and comparability of research findings as well as for evidence synthesis and meta-analysis. Guidelines to improve the reporting of observational studies have been available for several years [153]. However, adherence to these guidelines has been suboptimal [154, 155]. Furthermore, a recent assessment found that the reporting of studies conducted using routinely collected health data was often inadequate [156]. Until recently there were no clear guidelines on how to report specific methodological aspects related to the use of EHR-derived data in research. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement [157], published in September 2015, was an extension to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement which argues authors to completely and clearly communicate, in addition to the main methodology, important aspects related to the use of routinely collected health data in their study. To my knowledge, there was no existing assessment of the clarity of reporting on the use of EHR data in the recent asthma literature.

To assess the approaches and practices of using routinely collected EHR data in asthma research, I systematically interrogated the recent EHR-based asthma literature with the following objectives:

- To describe the different methods of defining asthma and assessing disease severity, control and exacerbations in EHR-based studies;
- To assess the clarity of reporting on the implementation and validity of these methods.

2.2 Methods: a systematic scoping review

I conducted a systematic scoping review based on Arksey and O'Malley's five-stage framework [158]. This methodology included identifying the research question, identifying relevant studies through literature search, selection of eligible studies, data charting and collating, and summarising and reporting the results. The research questions were:

- How were asthma and its key outcomes defined using EHR data in the recent literature?
- How did authors report on the validity of their EHR-based algorithms?
- How clearly was the reporting on the implementation of these algorithms and other EHR-related methods?

2.2.1 Identifying the research questions

Since this framework is intended for exploratory review, research questions could be iteratively developed during the review process. The primary research question behind this scoping review is:

- How has asthma and its key outcomes been defined using RCD in the recent literature?

In addition, while reviewing the literature, two other related research questions emerged to address validation of case definitions and the clarity of reporting of methodological aspects related to routinely collected EHR data:

- How did authors report on the validity of their RCD-based case definitions?
- More generally, how clear were the RCD-related methods reported?

2.2.2 Identifying relevant studies: literature search strategy

I searched PubMed using a broad query (Table A.1.1) to retrieve asthma studies that used EHR-derived data and were published between January 1, 2014 and December 31, 2015. The search query was iteratively improved by adding many variations and equivalents of the keywords "EHR" and "routinely collected data" as well as named data sources found in the literature. Only articles written in English were included.

2.2.3 Study selection

I excluded non-relevant articles by reviewing titles and abstracts, referring to the full-text when needed. I included only articles where asthma was a main finding. For the purpose of this review, I limited the concept of EHR-derived data to coded, objective, individual-level data that were generated as a by-product of routine health care. Therefore, I excluded studies in which only text-based medical records were used, the coded nature of EHR-derived data was unclear, the used data were aggregated, or no asthma-specific variables were measured from EHR-derived data. For example, if the only variables measured from routine data for asthma patients were related to co-morbidities or birth weight the study was excluded.

2.2.4 Data extraction, charting and synthesis

From each of the eligible articles, I extracted and summarised information from the full text and online supplements, including basic bibliography, setting (country) and design; names and types of EHR-derived data sources used; algorithms to identify asthma patients, assess disease severity, control, exacerbation; and how authors reported on algorithm validity. In this context, I referred to ‘validation’ as any attempt to assess the algorithm’s concurrent¹ or construct validity.² I used the RECORD Statement’s 13-items checklist [157] to assess the completeness and clarity of reporting of methodological aspects related the use of EHR-derived data in the study. I investigated whether in each study authors provided detailed information on how they identified asthma populations and assessed asthma outcomes. Ideally, the checklist requires authors to provide detailed description of the algorithms used, including complete lists of clinical codes and any validation performed previously or in the same study. In addition, information about the data sources used should be provided, including their content and validity, their catchment areas, level of access to them by authors (i.e. whether they had access to the whole or part of the dataset), explanation of any record-linkage performed, and in which date range the data used in the study were originally recorded. Authors should also explain how they prepared and cleaned the data for the purpose

¹*Concurrent (criterion) validity* is the extent to which the algorithm agrees with a concurrent measure, the validity of which to establish the diagnosis was previously assessed.

²*Construct validity* is the extent to which the algorithm accurately measures the real disease state.

of their study (e.g., how they processed inconsistent and invalid values). Also, they should ideally publish the programming source codes used for data extraction, preparation, and analysis. To enable the readers critically evaluate the validity of studies, authors should adequately communicate any implications of using EHR-derived data sources in their studies to assess a complex condition such as asthma. [Table A.1.2](#) describes the data extraction and charting tool used in this review. Article screening and data extraction were performed independently by two researchers (myself and Eleftheria Vasileiou³) with my first supervisor, Gwyneth Davies, arbitrating.

2.2.5 Collating, summarising and reporting the results

I summarised the general characteristics of the reviewed articles including the country to which the study source population belonged, study design, and types of routine data sources used in the study (e.g., health insurance claims, primary or secondary care, or pharmacy dispensing data). I highlighted the clinical labels used in asthma algorithms as they appeared in the studies. I also identified the approaches used in these algorithms which aimed at improving their accuracy. I also summarised the practices of justifying the validity of algorithms (e.g., by citing a previous validation of the same algorithm in the same population) as well as aspects related to the use of EHR data.

2.3 Results

2.3.1 Characteristics of studies and data sources

I included 113 articles in the review. [Figure 2.1](#) shows the study selection process. Most studies were conducted in the United States (US), Taiwan, and Canada ([Table A.1.3](#)), and employed longitudinal designs ([Table A.1.4](#)). The most commonly used data types were health insurance claims followed by medical record repositories and dispensing databases ([Table A.1.5](#)).

³A PhD student at Asthma UK Centre for Applied Research, University of Edinburgh.

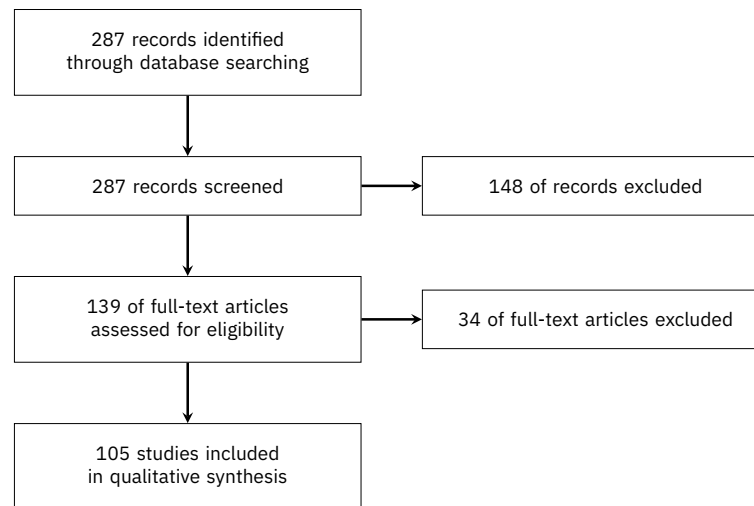


Figure 2.1: Flowchart for study selection in this scoping review.

2.3.2 Defining asthma

2.3.2.1 Various diagnostic labels of asthma

I identified 66 different algorithms to define asthma under seven diagnostic labels (Table A.1.6).

'Persistent asthma' was defined over 12 and 24 months using the US Healthcare Effectiveness Data and Information Set (HEDIS) criteria [159], which involved assessing for any of the following asthma-related events: (1) emergency department (ED) visit, (2) hospitalisation, (3) outpatient visit and two asthma prescriptions, or (4) four asthma prescriptions [160–163]; by HEDIS criteria except “four asthma prescriptions” [164]; and by any asthma encounter (hospitalisation or ED visit) or using oral corticosteroids (OCS) for three or more days [165].

'Current asthma' was defined by any asthma encounter in the last three years [166].

'Current general practitioner (GP)-reported and diagnosed asthma' was defined as any asthma encounter in the last 12 months, and *'current GP-reported, diagnosed and treated asthma'* as the same plus any asthma prescription in the same period [167].

Patients with treated asthma were otherwise required to have at least three dispensing events of asthma treatments in three different quarters of the year [168].

'Acute asthma' was defined using any asthma diagnosis codes in ED or inpatient data [169].

In the remaining studies, the label 'asthma' was defined using various algorithms, some of which were similar to those of the aforementioned more specific labels.

The intervals over which asthma diagnostic/management and prescription codes were queried were specified in 31 and 8 studies, respectively. The positions of diagnostic codes in the encounter (i.e. primary or secondary) were specified in 37 studies.

I identified five approaches in these algorithms: requiring diagnostic/management events, prescription events, or both (Table A.1.7).

2.3.2.2 Approaches to restrict study domain

To reduce the risk of misclassification, some studies applied additional non-asthma selection criteria which were meant to exclude individuals who are unlikely to have asthma. These criteria were based on restricting the study population based on age and co-morbidities.

Age restriction

In 12 studies Table A.1.8, age restriction was applied to asthma definitions as an indirect way of excluding those with co-morbidities that are common at age extremes and to acknowledge the uncertainty of asthma diagnosis in these ages. The minimum age limits were 2 (n = 4), 3 (n = 2), 5 (n = 1) and 12 (n = 1), while the maximum age limits were 40, 55 and 60 (n = 1 for each).

Excluding patients with specific co-morbidities

Eighteen studies (Table A.1.9) applied additional criteria to exclude asthma patients who also had other conditions, most commonly cystic fibrosis and chronic obstructive pulmonary disease (COPD). The list also includes "smoker over the age of 60" as a proxy for COPD diagnosis. The complete list is shown in Table A.1.9.

2.3.3 Assessing asthma severity

Eighteen studies used 20 different algorithms to assess asthma severity (Table A.1.10), as binary (i.e. severe vs. non-severe asthma) [160, 168, 170–183] or ordinal variables (mild, moderate, and severe asthma [184]; or low, moderate, and high-risk asthma [185]). The algorithms were based on one or more of the following asthma-related variables: number and/or dosage of prescriptions—namely short acting beta agonist (SABA), inhaled corticosteroid (ICS), OCS, and leukotriene receptor antagonist (LTRA)—and number of hospitalisations, ED and outpatient visits. Almost all algorithms (17) used prescriptions (either alone or with other variables), while one algorithm was based only on hospitalisations and ED visits [181]. The intervals over which asthma severity was assessed were three [174], six [183], 12 [160, 168, 173, 175–177, 179, 181, 182, 184, 185], 24 months [178, 180], or unclear [171, 172].

2.3.4 Assessing asthma control

Nine studies assessed asthma control using 11 algorithms, in 9 of which the interval was 12 months, in one 1-3 months, and in the remaining study this was unclear (Table A.1.12). Uncontrolled asthma was defined by a minimum number/dose of SABA prescriptions [175, 176, 184, 186, 187]; any or short-course OCS prescriptions [175, 176, 186–189]; any hospitalisation or ED visit with either diagnosis of asthma [172, 175, 176, 186–188, 190] or — in already diagnosed asthma patients — diagnosis of status asthmaticus, pneumonia, dyspnoea, or respiratory insufficiency [175]; unscheduled outpatient visits for asthma or lower respiratory tract infections (LRTI) [176]; and GP consultations for LRTI requiring antibiotics in asthma patients [176]. Asthma impairment was defined based on the required SABA use, namely an average of more than two salbutamol puffs per day [176]. One study assessed asthma control based on number of OCS and SABA prescriptions per year (without giving any further details about the actual algorithm) [186].

2.3.5 Defining exacerbations

Twenty-four studies defined exacerbations using EHR-derived data (Table A.1.11), as a dichotomous variable (absent vs. present) [161, 162, 168, 172, 175–177, 180, 182–184, 187–189, 191–199], or stratified into absent, moderate and severe

[200]. Oral corticosteroid prescriptions were used as a marker for exacerbations in 17 studies, either alone [168, 175, 176, 180, 184, 187, 192, 193, 198] or with a concurrent asthma encounter (e.g., a GP, outpatient, or ED visit, or hospitalisation within five or seven days) [161, 162, 177, 182, 183, 191, 197, 199]. In one study, exacerbations were defined by a minimum of six SABA prescriptions per year [192]. Other definitions included an outpatient code of ‘asthma exacerbation’ [197], asthma hospitalisation [168, 175, 177, 180, 182, 184, 188, 189, 191, 193, 195, 196, 198–200], asthma ED visit [161, 175–177, 180, 182, 183, 188, 189, 191, 193, 196–199], or hospitalisation with diagnosis of status asthmaticus, or — in already diagnosed asthma patients — diagnosis of pneumonia, dyspnoea, or respiratory insufficiency [175].

2.3.6 Clarity of reporting

Overall, the reporting of methodological aspects of using EHR-derived data was suboptimal. The majority of studies presented no information on the algorithms’ validity. Among studies that reported on the validity, I identified 10 practices of reporting or on or justifying the validity of algorithms (Table 2.1): (1) performing validation or concordance analysis in the same study against other measures based on different data sources (e.g., medical record review or patient-reported measures); (2) referring to previous validation of similar algorithms in the same or (3) different databases; (4) referring to previous validation of similar algorithms for different diseases in the same or (5) different database (6); using algorithms ‘consistent’ with previous studies in the same or (7) different databases; (8) using nationally developed algorithms; (9) using algorithms based on clinical guidelines; (10) and relying on previous validation of the database content. Some studies did not provide clear algorithms for asthma severity or control, but only referred to their components [168, 180, 182, 183, 186].

Of the 113 reviewed studies, 40 studies used record-linkage, of which 17 mentioned it in the abstract, and 28 provided at least some explanation in the full text. The geographical region, time frame of data, and types or names of the data sources were mentioned in 83, 91, and 104 abstracts, respectively. Eighty-three studies reported their extent of access to the data sources. The intervals over which the algorithms were applied were often not reported. One hundred and eleven studies touched on the implications of using EHR data to study asthma. Of

Table 2.1: Practices of reporting or justifying the validity of algorithms to define and assess asthma using EHR-derived data.

Algorithm validity was justified by	Number of algorithms				Total per category
	Identifying asthma patients	Assessing severity	Assessing control	Defining exacerbation	
Validation of the same algorithm in the same database	14	1	1	1	17
Validation of the same algorithm in different database(s)	2	6	3	2	13
Validation of other diseases' algorithms in the same database	2	0	0	0	2
Validation of other diseases' algorithms in different database(s)	1	0	0	0	1
Being consistent with similar studies in the same database	1	0	1	0	2
Being consistent with similar studies in different database(s)	1	0	0	1	2
Validation or concordance analysis in the same study	4	0	0	0	4
Being based on nationally developed algorithms	3	0	0	2	5
Relying on the validity of database coding	5	0	0	0	5
Being based on clinical guidelines	0	3	0	0	3
Not justified	76	8	4	18	106

these, 64 and 63 studies discussed the risk of misclassification bias and unmeasured confounding, respectively. Six studies acknowledged the possible changes over time in data quality and coding practices and the entailing changes in case definition eligibility and accuracy. Five studies explained their data cleansing procedures. Finally, no study shared the programming codes of data preparation and analysis.

2.4 Discussion

2.4.1 Statement of main findings

There is a considerable international activity in using EHR-derived data to study a variety of asthma populations and outcomes. This systematic analysis of the contemporaneous asthma literature provides a high-level view on how asthma and its main outcomes have been defined using routinely collected EHR data. Importantly, I found wide variations in the approaches used with limited attention being paid to the validity of the underlying algorithms used and suboptimal reporting on the

methodology. This poses a major challenge to the interpretation and reproducibility of this important, emerging body of research inquiry.

2.4.2 Strengths and limitations

To my knowledge, this is the first systematic exercise to investigate the quality of reporting on EHR-based studies, especially the validity of measures, in the context of asthma. In undertaking this work, I used robust approaches which involved two people independently selecting studies and undertaking data extraction. The findings of wide variations and suboptimal reporting of methods and their validity may also apply to other chronic diseases. This review had no geographic limits, but it was confined to assessing the recent literature. Examining the most recent asthma literature is most likely to provide meaningful insights on current practices. Finding studies conducted using routinely collected EHR data in the literature was challenging as there was no standardised method to do this. A dedicated Medical Subject Headings (MeSH) term for “routinely collected health data” in MEDLINE database was previously suggested [201]. However, I believe my broad search query, including a long list of synonyms routinely collected health data and clinical codes for asthma diagnosis reasonably increased the search recall. In few studies, it was a challenge to separate criteria used to define asthma from the study-specific inclusion and exclusion criteria. A limitation was that I did not systematically check whether the references provided by a study to support the claimed validity of algorithms in question actually provided sufficient evidence of validity. For example, slight differences might exist between the algorithms used in a given study and those validated by previous studies.

2.4.3 Interpretation in the light of previous studies

Although EHR-derived data are convenient resources for research, they are originally collected for other purposes, and usually suffer from missing or incorrect data and potential biases [82, 151, 202]. Asthma-related EHR data potentially suffer from significant levels of uncertainty due to a set of factors. Some of these factors are inherent to asthma as a heterogeneous disease with fluctuating and variable natural history. Other factors include the wide variability in health care provision, and in the practices of documentation and coding of clinical data. In addition, EHR systems usually fail to capture complete and accurate clinical in-

formation at the point of care due to design limitations and inefficient use of these systems by clinicians to document clinical data [203, 204]. Altogether, these factors create high levels of variability and inconsistency in the information recorded on asthma patients, even those with similar clinical profiles. Furthermore, many EHR-derived databases often lack important variables, such as lung function, indication of dispensed medications, adherence to treatment, and lifestyle, which are vital for identifying and assessing asthma patients.

The aforementioned issues impose challenges on the interpretation of asthma-related health events in EHR-derived data and their use to identify and assess asthma patients. These issues are further discussed in [Section 3.1.1](#). In this review, asthma diagnosis codes were commonly used solely for identification of asthma patient. However, these codes may be recorded after a trial or wrong diagnosis, and do not capture undiagnosed patients [205]. In addition, although an asthma diagnosis code may be recorded during any asthma-related health encounter, it does not necessarily imply an active or treated disease. More complex approaches for asthma patient identification included the requirement of medication codes in addition to an asthma code either to identify treated patients or to increase the specificity of “any asthma” case definition. Patients with “any asthma” were also ascertained by only medication codes where only dispensing data were available. However, this approach could exclude patients with active disease who did not receive prescriptions. In addition, it could incorrectly include non-asthma patients since some asthma medications, such as SABA and oral corticosteroids, could be prescribed in other conditions.

These challenges are however not insurmountable. In this analysis, I found several approaches, in addition to asthma-related criteria, which were intended to improve the specificity of algorithms such as age limitation, exclusion by comorbidities, and diagnosis position restriction. I was able to distinguish these approaches from study-specific patient selection criteria. Age restriction was driven by the uncertainty of asthma diagnosis in age extremes. In early childhood the clinical diagnosis of asthma is difficult to establish, while in the elderly COPD may be misdiagnosed as asthma [206]. Excluding patients, who already satisfied asthma criteria, but who also had other specific co-morbidities was a common practice. In the absence of data on respiratory symptoms, lung function, and laboratory tests, excluding adult “asthma” patients who also had COPD diagnosis and/or smoking history reflects the assumption that they were unlikely to have

asthma, and was assumed to increase the specificity of asthma case definitions. However, although misdiagnosis between the asthma and COPD is common, they may coexist in what is known as asthma-COPD overlap syndrome (ACOS) [207], although whether this should be considered a distinct third entity is under active debate [208–211]. Many other respiratory conditions can mimic asthma. Unless excluding asthma cases with co-morbidities is required by the study's scope, it may introduce risk of *diagnostic purity bias* [212] which compromises the study's external validity.

Ultimately, validity of the EHR-based algorithms should be assessed. Ideally, these algorithms should be validated in the databases in which they are intended to be used. However, this was often not the case. Instead, using algorithms with only reasonable face validity based on clinical guidelines or clinical judgement is a very common practice in EHR-based studies [213, 214]. These approaches implicitly assume that clinical codes in the database accurately represent the patient's actual health care events [213], which is a questionable assumption. It is worth noting that the validity of algorithms is not necessarily portable across datasets or populations [215]. This means that a case-finding algorithm that performs well in a dataset may be inaccurate if used in another dataset. Populations may differ in demographics, health parameters, and health care. EHR-derived datasets, even those with the same type of healthcare data, may differ in their content and quality. In this review, however, I found that it is common for algorithms of asthma and its outcomes to be re-used in other populations without re-validation in the target datasets. I believe this problem is underappreciated and deserves more attention.

Reproducibility and replicability are crucial issues in medical research and require complete, clear, and transparent reporting of methods. Under-reporting on the implementation details and the validity of methods compromises transparency and reproducibility. It has been previously found that in EHR-based studies, full lists of clinical codes were often not reported [216, 217]. A recent, large-scale reproducibility exercise identified similar challenges due to suboptimal reporting of EHR-based studies, particularly sharing code lists and algorithms [218]. Under-reporting of a study methodology means that the time and resources invested in conducting that study is wasted [219].

For complex clinical variables such as asthma and its outcomes, the lack of standardisation of the clinical definitions and the wide variability in the EHR-based

algorithms undermine the validity and replicability of studies [152]. The significant methodological heterogeneity I found in the EHR-based asthma assessment algorithms reflects, in addition to the content differences between the databases used, the lack of consensus on the clinical definitions in the first place despite continuous standardisation efforts [2, 146, 220, 221]. The focus of this work was to examine asthma algorithms and their validity specifically in the context of EHR-derived data, but this highlights the fundamental need to reach consensus on clinical asthma definitions and the appropriate validation of asthma diagnosis. For example, there is still an active debate on whether lung function is essential to establish asthma diagnosis [147, 148]. A recent study also found significant variation in algorithms to assess asthma severity from health insurance data [222]. Unjustified inter-study variation in the operational definitions of the same clinical concepts creates challenges for comparability [121], meta-analysis and evidence synthesis. These issues have been raised for asthma [223] and other allergic conditions such as peanut allergy [224, 225] and anaphylaxis [226], where wide variations in findings were potentially attributed to inconsistent case definitions. The findings in this chapter are likely to be applicable to a wide range of chronic conditions when defined and assessed using EHRs.

2.4.4 Implications for policy, practice and research

This review sheds light on the opportunities offered by the increasingly ubiquitous EHRs, but also highlights considerable heterogeneity and suboptimal reporting of EHR-based asthma assessment algorithms and the implications of these practices on comparability and reproducibility of studies.

Developing reliable algorithms to identify asthma patients and assess asthma outcomes using EHR data is a non-trivial challenge. Standardising asthma algorithms used in research, where possible, is a crucial need. However, this may be impractical since databases usually differ in their content, validity may not hold across different populations [215], and no best practice currently exists [222]. Similar challenges arise when comparing asthma epidemiology between multiple populations [227], as the availability and quality of data may differ across those populations and a single case definition may not work for all of them. These methodological issues, in addition to suboptimal reporting, should be considered when interpreting and synthesising evidence from geographically dispersed studies.

With the accelerating availability of EHR-derived data and their rapidly growing use to study asthma, I believe the global asthma research community needs to pay more attention to the methodological issues related to the use of these data in asthma research. I believe that consideration needs to be given to convening an international task force to work on the harmonisation of those algorithms under uniform and consistent clinical labels, while considering the differences between populations and databases. In addition, validation of these algorithms in the respective populations should be given a high priority [215]. Furthermore, to allow more accurate assessment of asthma from EHR-derived data, efforts are needed to improve the capture and coding of asthma-related data at the point of care [228] which requires more efficient EHR systems [203, 204]. In addition, emerging data sources such as patient-generated data and wearables need to be harnessed [229]. Finally, to improve the clarity of reporting on EHR-related methodological aspects, I strongly advocate the adoption of the RECORD Statement as an extension of the STROBE Statement by authors, journal editors, and peer reviewers [156, 157]. Optimal reporting should include complete code lists, detailed algorithms and validity assessment. Implications of using EHR-derived data to study a complex condition such as asthma should be clearly communicated to enable judgement of internal and external validity.

2.5 Conclusion

This systematic scoping review showed considerable international interest in exploiting EHR-derived data to study asthma. Asthma diagnosis, exacerbation, severity and control, have been assessed from different types of EHR-derived data using various approaches. However, there were considerable variations and inconsistency in these approaches. These variations were compounded by sub-optimal reporting of methods, their validity, and other aspects concerning the use of EHR-derived data for research. Reusing algorithms of asthma outcomes in new populations without re-validation, and relying only on clinical judgement and face validity were common practices. These issues make it difficult to assess the reproducibility of research and perform evidence synthesis and meta-analyses. Given the substantial investments taking place in EHRs globally, the number of EHR-based asthma studies is likely to grow substantially in the coming years. Unless addressed, these issues will aggravate the reproducibility problem and increase

the avoidable waste in this important body of research. It is therefore important that the asthma-interested research community works to place it on a solid footing in order to ensure the quality and reproducibility of this work. Improving the reporting of these aspects using standardised guidelines such as the RECORD Statement would improve the rigour, transparency and reproducibility of asthma research.

The findings in this chapter will inform the discussion on the challenges of defining asthma using routine collected EHR data in [Chapter 3](#). In that chapter, I will explore and demonstrate the usefulness of using data-driven methods, instead of manually developed algorithms, to identify asthma patients.

Chapter 3

Identifying asthma patients in Wales

Defining a complex condition using real-world data

In this chapter, I discuss the challenges of developing accurate case definitions for asthma. I then highlight the common methods used in validating asthma definitions from a variety of data sources. In the absence of a gold standard for asthma definition, latent class analysis (LCA) can be used to identify hidden clusters in a population using the observed data. I describe the development of an LCA model to identify patients with asthma, particularly those with currently treated asthma, using routinely collected primary care data in the Secure Anonymised Information Linkage Databank in Wales. Based on this model, I trained a classification algorithm to identify asthma patients that can be used in this Databank and similar data settings.

Chapter Contents

3.1	Introduction	57
3.1.1	Challenges of developing accurate case definitions	57
3.1.1.1	Asthma is a heterogeneous condition	57
3.1.1.2	Variations and changes in clinical coding practice	58
3.1.1.3	Limitations of EHR-derived data	58
3.1.2	Validity assessment of asthma case definitions is needed	59
3.1.2.1	Approaches for validation of routine data-based case definitions	59
3.1.3	Latent class analysis: An overview	62
3.1.3.1	Concept and assumptions	63
3.1.3.2	Model specification	64
3.1.3.3	Parameter estimation	65
3.1.3.4	Membership probabilities	66
3.1.3.5	Model homogeneity and separation	66
3.1.3.6	Model assessment and selection	67
3.1.3.7	Model interpretation	69
3.1.4	Using latent class analysis to identify asthma patients in the Wales Asthma Observatory	70
3.2	Objectives	71
3.3	Methods	72
3.3.1	Data sources	72
3.3.2	Patient population	73
3.3.3	Latent class modelling	74
3.3.3.1	Observed variables	74
3.3.3.2	Number of classes and model selection	75
3.3.3.3	Statistical tool	76
3.3.4	Derivation of a classification algorithm	76
3.3.5	Comparison of the classification algorithm with other case definitions	78
3.3.5.1	Case definition used for comparison	78
3.3.5.2	Statistical analysis	79
3.4	Results	80
3.4.1	Sampling and sample characteristics	80
3.4.2	Summary of the competing latent class models	81
3.4.3	Model selection	83
3.4.4	Model interpretation	84
3.4.5	Class merging	88
3.4.6	Derivation of classification algorithm	89
3.4.7	Comparing the classification algorithm with other case identification methods	94

3.5 Discussion	97
3.5.1 Summary and interpretation of the findings	97
3.5.2 Strengths and limitations	98
3.5.2.1 Strengths	98
3.5.2.2 Limitations	99
3.5.3 Comparison with related works	101
3.5.4 Future directions	102
3.6 Conclusion	103

3.1 Introduction

In the light of the literature review findings presented in [Chapter 2](#), I further discuss the challenges of defining and assessing asthma using routinely collected data (RCD) considering asthma heterogeneity, data limitations, and the absence of a gold standard to define asthma. I then justify the use of data-driven approaches to identify people with ever and current asthma in the Wales Asthma Observatory.

3.1.1 Challenges of developing accurate case definitions

Developing case definitions is an essential step in the development of disease observatories and registries [137] and is crucial to the interpretation of epidemiological estimates and research findings [213]. However, particular challenges exist when identifying patients with a complex disease such as asthma. In addition, many disease registries increasingly use electronic health record (EHR)-derived data, which adds particular challenges related to the limitations of these data [102, 151, 202]. I summarise these challenges below:

3.1.1.1 Asthma is a heterogeneous condition

Asthma is not a single disease entity; there is increasing recognition that the disease is an umbrella of heterogeneous sub-entities at the molecular, pathological and clinical levels (i.e. endotypes and phenotypes) [4, 5, 230]. This means that patients who are diagnosed with ‘asthma’ do not all have the same underlying disease process. Furthermore, the natural course of asthma exhibits variability within and between patients, who also differ in their response to treatment. There-

fore, it is practically difficult to find a single, precise clinical definition of asthma that everyone agrees upon [2, 144-148].

3.1.1.2 Variations and changes in clinical coding practice

EHR-based case definitions for asthma are often based on recorded events related to patients' health, such as diagnosis, physician's visits, and prescriptions. However, physicians differ in their diagnostic skills and prescribing behaviour [231]. In addition, their practices are subject to changes by the continuous evolution of clinical guidelines, which can potentially affect clinical coding practices [109], as well as by incentives and resources. Variations in clinical practice mean that there is variation in the recorded data, even for clinically similar asthma patients. Further variations in the recorded data may result from differences in the documentation and coding of clinical data between physicians.

3.1.1.3 Limitations of EHR-derived data

Along the pathway of clinical data, from the point of care to central data repositories, several factors introduce error, uncertainty, and information loss into these data [102, 151, 202, 232]. EHRs usually capture both unstructured and structured data that are collected during clinical encounters. At the data entry stage, however, incorrect, invalid, or inaccurate data may be recorded due to both human and computer related factors. Due to design limitations of EHR systems [203, 204], not all clinical data are recorded and/or coded by physicians [228, 233]. In addition, clinical coding aims to reduce the detailed data captured during the encounter into few clinical codes on diagnosis, clinical findings, disease management events, and prescriptions. However, the commonly used clinical coding systems usually have limited granularity compared to the captured narrative data. These coding systems, while facilitating standardised coding [233], fail to codify all the details that are in the patient record. These non-codified data are important for secondary uses. Furthermore, in practice, clinical coding is a barrier to data recording [233], and physicians use only few of the available codes and under-utilise the granularity of coding schemes [228]. Finally, there is an increasing trend to link routinely collected health data from different sources, thereby introducing a further threat to data quality. Record-linkage can potentially introduce errors and biases to data analyses [234].

3.1.2 Validity assessment of asthma case definitions is needed

Due to the aforementioned challenges to define asthma, a single epidemiological case definition for asthma using routinely collected EHR data does not exist. At the same time, these challenges highlight the need to assess the validity of case definitions and clinical variables measured using these data [213, 235, 236]. Validity assessment has a particular importance in the development of disease registries and observatories as well as research databases since the validity and interpretation of epidemiological estimates and research findings depend on accurate patient identification and characterisation [236]. Reporting guidelines promote complete and transparent reporting on the methods and their validity. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement [157], which is an extension of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement [153], argues that authors of studies conducted using routine data should clarify whether the variable definitions used in their study have been validated. A comprehensive check-list of diagnostic values [201] has been proposed to be used in the reporting of validation studies including cross-tabulation, sensitivity, specificity, accuracy, positive and negative predictive values (PPV and NPV), likelihood ratios (LRs), Cohen's kappa, receiver operating characteristic (ROC) curve including the area under the curve (AUC), and case prevalence along with 95% confidence intervals when applicable. Several approaches to validity assessment are discussed in this section.

3.1.2.1 Approaches for validation of routine data-based case definitions

In epidemiological studies, validity of case definitions can be assessed using a variety of methods. A recent systematic review identified various methods to validate asthma diagnosis in EHRs [237]. These included manual review of the narrative medical records, medication data, and questionnaire data. Since a gold standard is usually unavailable or inaccessible, acceptable reference standards are used. A reference standard can be simple or composed of multiple reference measures, in which case it is called a *composite reference standard* [238]. When used, a composite reference standard is thought to have a discriminatory property that is greater than those of each of its components alone [238].

When no accepted reference standard can be available, the researcher can perform concordance analysis or latent variable modelling. Concordance analysis evaluates the agreement between a case definition and independent measures, usually derived from other data sources [217, 236, 238]. In this method, neither the case definition under assessment, nor the independent measure, are assumed to represent the ground truth (i.e. the real disease status as confirmed by clinical diagnosis and/or laboratory tests). Instead, the level of agreement between the two identification methods can provide insights into the reliability and meaning of each method. Alternatively, labelled classes in a well-specified latent variable model can be used to train a case identification algorithm which can be used as an 'internal reference' within a particular dataset [238].

In other approaches, aggregate data are used to compare rates or distributions of two case definitions [217, 236]. These approaches can be used when the researchers have no access to individual level data necessary to test one of the compared case definitions, but instead have access to aggregate results.

Practically, the choice of appropriate validation methods therefore depends on the available data sources, and the availability of a gold or reference standard, as well as the research questions and design.

Manual review of medical records

While routinely collected EHR data are usually coded, the source medical records, whether paper-based or computerised, usually contain more detailed data, including narrative clinical notes, about patient care. Point-of-care data are a good reference for validating case definitions because they contain more detailed data about patient care before being coded. For this purpose, medical records for a sample of individuals in a specified cohort are reviewed to confirm whether they had a confirmed asthma diagnosis at a certain point of time. Often, researchers send questionnaires to physicians [239, 240], nurses [241], or an expert panel to review patient records [242]. Clinical examination and/or laboratory tests such as lung function tests on the validation sample could be sometimes repeated to confirm or rule out the diagnosis. Ideally, the reviewer should be blind to the clinical codes to avoid confirmation bias [239, 241].

Analysis of concordance with dispensing data

Individual-level pharmacy dispensing datasets usually contain valuable information on asthma treatment. Analysis of concordance can be performed between case definitions based on dispensing data and other case definitions based on different data sources such as hospitalisation records, questionnaires filled by general practitioners (GPs), and self-administered questionnaires [237, 240, 243, 244].

A limitation of dispensing data is that they often lack information on diagnosis. Not all asthma medications are specific to asthma, and therefore it is difficult to ascertain whether an 'asthma medication' was actually prescribed to treat asthma or another co-morbidity. For example, inhaled corticosteroid (ICS) (in combination with bronchodilators) can be also prescribed for chronic obstructive pulmonary disease (COPD), and oral corticosteroids are indicated to a wide range of non-respiratory conditions. It is therefore difficult to accurately find asthma patients solely from dispensing data [245, 246].

Self-reported asthma measures

Self-reported data about asthma are commonly used in research. Questions about having asthma symptoms, severity of asthma symptoms, 'ever asthma', 'current treated asthma', and/or using specific asthma medications are commonly included in national health surveys and asthma questionnaires [54, 247, 248]. Self-reported asthma measures have been used as a reference to assess the validity of asthma codes in primary care records [249] and in analysis of agreement with billing and health insurance data [250, 251].

Self-reported data have unique advantages and disadvantages. They represent patients' experience and understanding of their own health conditions which are not always appreciated by physicians [252]. However, some patients do not have clear understanding of their health status. For example, some asthma patients may believe they have asthma while in fact they have been diagnosed with COPD or hay fever. Self-perception of health status can be influenced by variety of other factors such as educational attainment and employment [253]. Self-reported data are usually prone to recall bias, i.e. patients with more health problems are often more likely to report previous exposures or health events. I found that, using

the Welsh Health Survey (WHS), patient-reported currently treated asthma had suboptimal concordance with the ‘ever-diagnosed currently treated asthma’ ascertained from general practice data [254].¹ For these reasons, patient-reported doctor-diagnosed asthma should not be considered as a ‘gold standard’ case definition for asthma [255].

Where no accepted reference standard exists

The above approaches to obtain reference standards for validation of routine data-based case definitions can be expensive, time-consuming, and/or labour-intensive. Where accepted reference standards based on independent data sources are unavailable or unfeasible, other approaches can be followed. For example, case definitions can be developed based on medical knowledge, clinical guidelines, and/or relevant literature, as well as knowledge of data recording and coding practices (see [Section 2.3](#) for examples from the systemic scoping review in [Chapter 2](#)). These approaches often assume the completeness and accuracy of data recordings and may rely on heuristic techniques assumed to improve accuracy. For example, a researcher may decide to exclude from a case definition people in whom the diagnosis may not be ‘certain’; e.g., those with differential diagnoses or particular morbidities (see [Chapter 2](#)).

A different approach to follow in the absence of accepted reference standards is to use data-driven methods, such as latent class analysis (LCA). Such methods use computational techniques to understand the hidden population structure in relation to specific observed characteristics [256]. In order to ensure meaningful findings, these methods should be performed in the light of the established knowledge about the disease pathophysiology, clinical course, and epidemiology as well as about data provenance and quality [142].

3.1.3 Latent class analysis: An overview

In the absence of accepted referenced standards for asthma case definitions, it is possible to identify asthma patients from RCD by examining the recorded events related to asthma. Examples of such variables include the following dichotomous variables:

¹I presented this analysis in the British Thoracic Society Winter Meeting 2016; the poster is available in the [Appendix B.1](#).

- ‘ever diagnosed with asthma’
- ‘had GP attendances related to asthma in the last year’, and
- ‘received inhaled steroids in the last year’.

Each of these events alone may not be sensitive or specific to the ‘true status’ of asthma. However, analysing the patterns of correlation between them these events can help identify people groups that are likely to have the disease. Using cross-tabulation, the existing patterns of correlation between these variables can be identified along with their frequencies; these patterns of correlation represent different *patient profiles*. By examining these profiles, it is possible to assign, to patients, meaningful clinical labels that explain their patterns of observed characteristics. For example, if the pattern suggests asthma, it will be assigned the label ‘asthma’, otherwise it will be labelled as ‘none’ or ‘no asthma’. In reality, different profiles may have the same clinical label, possibly each with different certainty. Eventually, these label-sharing patterns are merged together in common groups, some of which represent patients with the disease.

The aforementioned exercise allows us to understand the population distribution and facilitate the process of identifying patients based on their observed data patterns. However, with a high number of observed variables and their levels (i.e. high-dimensional data), this exercise becomes extremely complex and impractical. This is a typical problem where computational clustering methods, such as LCA, can be used. LCA can be useful in analysing and understanding complex patterns in high-dimensional data.

3.1.3.1 Concept and assumptions

LCA is a finite mixture modelling method, i.e. a method that models a mixture of sub-groups in a population. It aims to cluster a population into sub-groups related to a set of observed variables [257, 258]. LCA assumes that the patterns of correlation in these observed variables within the population can be explained by, in addition to measurement errors, a hidden categorical variable, called a *latent variable*, which has a pre-defined number of levels (Figure 3.1). The latent variable partitions the population into sub-groups called *latent classes*. These latent classes are qualitatively distinct although they are fuzzy in nature as individuals have probabilistic memberships in these classes (see Section 3.1.3.4:

“Membership probabilities” below). In this chapter, the latent variable represents a patient’s disease state at a given point of time.

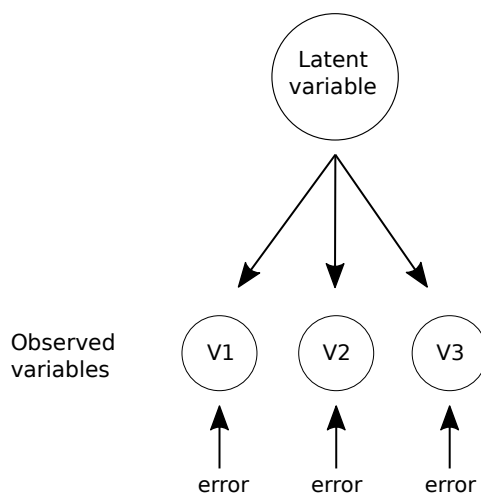


Figure 3.1: Visual representation of a latent class model. The model assumes that the observed variables are influenced by the categorical latent variable as well as measurement and random errors.

LCA is a non-parametric model which requires categorical (including dichotomous) or ordinal observed variables. Therefore, by using these types of data, assumptions about the normality, linearity, or homogeneity of data are avoided.

LCA aims to identify latent classes that explains the whole correlation between variables at the population level; that is, it assumes that any associations between the observed variables are assumed to be solely due to the influence of the latent variable. Therefore, in a perfect latent class model, the observed variables are independent of each other within each class [257, p. 44]. For example, inside the latent class of “asthma”, the presence of diagnosis codes does not *per se* influence the probability of presence of prescription codes. This assumption is called the ‘local independence’ or ‘conditional independence’ and is important in latent class modelling, although it may not hold in a large sample.

3.1.3.2 Model specification

The construction of a simple latent class model starts with the hypothesis that certain structure exists in a population. Then a set of observed variables are chosen, which could be dichotomous or categorical. Ordinal variables can be used in the modelling but they are treated as categorical. Interval variables need to be transformed into these types of variables. The observed variables should be chosen on the basis of having strong relevance to the hypothesised population

structure. More specifically, they should have high discriminatory power to distinguish between some of the hypothesised sub-groups in the population.

An empirical study has shown that adding a larger number of highly discriminating observed variables to the model had beneficial effects on the model estimation [259]. However, a larger number of observed variables may increase the possibility of sparseness in their contingency table, especially with a relatively small sample size. To avoid sparseness, larger sample sizes, when available, are preferred [259].

3.1.3.3 Parameter estimation

The parameters estimated by LCA include the following:

- Probabilities for any random individual to be in each latent class, i.e. prevalences of the latent classes.
- Probabilities of observing each level of each observed variable in each latent class—also known as *item-response probabilities*.

The model parameters are estimated using the expectation-maximisation (EM) algorithm, sometimes along with the Newton-Raphson algorithm [260], both of which iteratively search for maximum-likelihood parameter values for which the data are more likely to be observed [261]. The algorithm starts the iterations with random values for the estimated parameters and estimates the expected cell counts in the contingency table of the observed variables. Depending on how expected cell counts fit the observed ones, the model parameters are then improved in order to maximise the log-likelihood function. In each iteration, if the increase in the log-likelihood for the current solution is less than a pre-defined value, the maximum log-likelihood is considered to have been found and the current solution is chosen as the 'best solution' [257, 260]. In ideal conditions, the algorithm converges to find a best solution among all possible solutions, called the 'global maximum solution'. However, sometimes the algorithm converges to a 'local maximum' solution that is optimal only among neighbouring solutions and not among all possible solutions. This problem is more likely to happen when the number of latent classes is too high. To avoid local maxima, it is advised to repeat the estimation algorithm with different starting values so that a local maximum model is not selected by the estimation algorithm as the best solution [260]. Practically, the number of iterations is usually limited by a maximum limit, e.g. 1000 iterations,

determined by the researcher. If the expatiation-maximisation algorithm reaches this limit before convergence, no best solution is selected.

3.1.3.4 Membership probabilities

Based on observed characteristics, individuals are assigned posterior probabilities of their membership in each of the latent classes [257, p. 67].

Thus, since each individual may belong to more than one latent class, the identified latent classes have in principle a fuzzy nature rather than being mutually exclusive. However, usually, each individual is eventually assigned to the latent class of maximum membership probability [262]. Each of the identified latent classes may contain more than one patient profile. In other words, individuals in each latent class may have different, but usually similar, observed characteristics. Item-response probabilities in each latent class represent the averaged observed characteristics in that class.

3.1.3.5 Model homogeneity and separation

The aim of LCA is to identify 'distinct' and 'homogeneous' latent groups in the population. However, since each latent class may include mixed patient profiles, especially in a model with high complexity, the best solution is that which maximises the within-class similarities and the between-classes differences.

Class *homogeneity* is the degree of similarity between individuals in a given latent class. It can be evaluated by examining the item-response probabilities within that latent class independently of the other latent classes. In that latent class, if most of the item-response probabilities are high or low (i.e. close to 1 or 0), which means that either there is a single prevalent patient profile (i.e. with a prevalence close to 1) or that the patient profiles in that class are highly similar to each other, then this latent class is said to be highly homogeneous. Otherwise, if most of the item-response probabilities are marginal (i.e. close to 0.5), which indicates there is no single prevalent patient profile, then this class is said to have low homogeneity.

Latent class *separation* is the degree to which the individuals in each latent class are different from the individuals in the other latent classes. Class separation can be evaluated by comparing the item-response probabilities between the latent classes. In a model with high latent class separation, each latent class has

item-response probabilities that are clearly distinct from those in the other latent classes.

A latent class model with a high class separation necessarily implies a high degree of homogeneity within each latent class. However, a latent class model with a high degree of homogeneity within latent classes does not necessarily imply a high degree of separation between latent class.

A useful latent class model will identify highly-differentiated, well-separated, and likely more interpretable latent classes. One way to improve homogeneity and separation is to choose observed variables with high discriminating power, i.e. thought to be strongly related, both conceptually and quantitatively, to the latent variable in question. However, a latent class model that has a good fit to the data may otherwise have poor class separation and thus poor interpretability.

3.1.3.6 Model assessment and selection

In LCA, the number of latent classes is pre-specified by the researchers based on their hypothesis and assumptions about the population structure. However, the actual data may fit lower or higher number of classes than assumed by the researchers. Therefore, they often repeat the modelling with different numbers of classes and compare them in terms of model diagnostics (to assess the absolute and relative fit of the models) and clinical plausibility.

Absolute model fit

The absolute fit of a latent class model considers its fit to the data regardless of its competing models. Since a latent class model is based on a contingency table of its observed variables, the absolute model fit can be assessed using this contingency table. Therefore, Pearson's chi-square test can be performed to test the null hypothesis that the observed counts of patient profiles in the contingency table can be produced by the estimated latent class model [257]. The likelihood ratio chi-squared statistic, G^2 , is a variant of Pearson's chi-square test. In both tests, the expected counts of patient profiles, as estimated based on the model parameters, are compared to the observed counts of patient profile. The degree of freedom is the total number of observed patterns minus the number of estimated parameters minus 1. The test statistic is then compared to the reference

chi-square distribution to obtain a p -value representing the probability that the observed data can be produced by the estimated latent class model under the null hypothesis.

However, both Pearson's chi-square and the likelihood ratio chi-squared tests are not appropriate when sparseness exists in the contingency table of the observed variables, i.e. when the contingency table contains too many cells with small or zero counts. This happens in latent class models with small sample sizes and/or large number of variables. In addition, in modelling using larger samples, both tests are likely to produce smaller p -values, i.e. indicating weaker evidence of absolute model fit.

Relative model fit

The Akaike information criterion (AIC) [263] and the Schwarz's Information Criterion (often called the Bayesian information criterion, BIC) [264, 265] are two information criteria each of which provide an estimate of the information lost when a given statistical model is fitted to given data. They seek a balance between the model's goodness of fit and its parsimony. A model's parsimony is the opposite of complexity, and is represented by the number of estimated parameters, or simply by the number of latent classes. Thus, AIC and BIC are also called 'parsimony indices', as they prefer models with fewer number of estimated latent classes. To obtain the desired balance, AIC imposes a penalty on G^2 based on the number of parameters, while BIC imposes a larger penalty based on both the number of estimated parameters and the sample size.

AIC and BIC can be used to measure which one of the competing latent class models, differing by number and characteristics of classes, best fits a given dataset. The model with the lowest value represents a more optimal balance between model fit and parsimony and is preferred over its competitors. Due to the difference in imposed penalty, AIC and BIC may not agree on the 'optimal model' they suggest.

However, they do not provide a meaningful absolute measure of quality for a latent class model independently of the other competing models. This means that a competing model with the best AIC and/or BIC may still have poor absolute fit to the data.

Model interpretability

Although the best-fit model based on the above model selection methods is often chosen by researchers, it does not always contain a clinically meaningful structure. For example, variables with lower clinical relevance to the clustering exercise (e.g., have clinically lower discriminatory power) may be used by the best-fit model in the generation of latent classes more than other more clinically relevant variables. In such cases, researchers use domain knowledge to decide whether the best-fit model is clinically and/or biologically plausible. Otherwise, the researchers may need to re-specify the model, i.e. revise the choice of observed variables and their levels.

3.1.3.7 Model interpretation

The interpretation of an LCA model is based mainly on its estimated parameters. The item-response probabilities can be used to describe the latent classes and assign meaningful labels to them [257, p. 29]. For example, a class in which individuals have high probabilities (e.g., 0.8 or more) of both of 'having asthma diagnosis' and 'receiving asthma prescriptions in the last year' can be assigned the label of 'doctor diagnosed currently treated asthma'.

If the number of classes in the best fit-model is higher than that expected by the researcher, i.e. the clustering is deeper than desired, then the researcher may choose to keep this model but manually combine similar classes to form fewer but larger 'super classes' representing the desired clinical clustering.

If the model is not clinically interpretable, then the researcher may need to choose a different competing model, or even decide to re-specify the model by modifying the observed variables.

Model interpretation particularly depends on the relevance of input variables and the quality of data. The careful choice of both input variables and number of classes (the latter is partly determined by the expected clusters as based on the predictor variables) are critical for clinically meaningful clustering. Even though, meaningful interpretation of the resultant classes could be difficult; e.g., classes may not correspond with the commonly recognised clinical definitions of the disease sub-groups.

A latent class may contain individuals who do not match the clinical description assigned to the latent class as a whole. For example, the ‘no-asthma’ class, based on the model design (i.e. choice of predictor variables) may include people who can be classified by experts as people with asthma. Similarly, the ‘asthma’ class may include people who have conditions other than asthma, e.g., COPD, but share some of the common characteristics of asthma patients (receipt of bronchodilators). Therefore, LCA is usually considered exploratory rather than confirmatory.

3.1.4 Using latent class analysis to identify asthma patients in the Wales Asthma Observatory

The development of the Wales Asthma Observatory requires reliable case definitions for asthma. The Observatory is based on data in the Secure Anonymised Information Linkage (SAIL) Databank. Case identification algorithms for ‘ever asthma’ and ‘currently treated asthma’ based on diagnosis and prescription events have been validated in other datasets [102]. However, they may not necessarily retain their validity in the SAIL Databank. Assessing their validity against complete patient records is time-consuming, labour-intensive and expensive, rendering it unfeasible in this doctoral project. Data on self-reported ‘currently treated asthma’ were available in the results of the WHSs for the years 2013 and 2014 which were record-linked to the SAIL Databank. However, these data only contained responses to a simple question for ‘currently treated asthma’, without specifying a time-frame for such status. Therefore, this self-reported variable cannot be considered an accepted reference for ‘currently treated asthma’.

In the absence of a feasible, accepted reference standard for asthma, LCA is an appropriate method to identify asthma patients in the SAIL Databank. This method fits the nature of both asthma and routine data held in the SAIL Databank.

Clinically, the presence of asthma is not a binary status, and the certainty of physician’s diagnosis is not always perfect. In practice, asthma diagnosis could be classified into ‘absent’, ‘possible’, ‘likely’, or ‘confirmed asthma’. Therefore, the uncertainty of asthma diagnosis can be represented probabilistically, which corresponds to the fuzzy approach followed by LCA to assign class memberships for individuals in a population.

Asthma-related observed characteristics that can be derived from the SAIL Databank can be dichotomous (e.g., 'ever asthma diagnosis'), polytomous (e.g., 'age group at asthma diagnosis'), and continuous data (e.g., 'number of ICS prescriptions in the last year'). However, for the purpose of detecting asthma patients from such data, continuous data can be transformed into categorical (dichotomous and/or polytomous) data using appropriate cut-offs, e.g., 'no prescriptions' vs. 'one or more prescriptions'. This fits well with the types of observed variables required in LCA, namely, categorical (or ordinal) variables.

Despite guidelines on diagnosis, asthma could be clinically confused with a variety of conditions, especially COPD. Distinguishing asthma from COPD can be challenging [266, 267]. Some patients exhibit features of both conditions in what has been recognised as the asthma-COPD overlap syndrome (ACOS) [207]. Therefore, it is worthwhile to consider the misdiagnosis and overlap between asthma and COPD in latent class models that attempt to identify asthma patients.

The output of LCA in this chapter will include individuals' class membership probabilities and their assigned classes. While such results are useful to create a reference identification for asthma in the SAIL Databank, it cannot be used for this purpose elsewhere. In order to facilitate the re-use of the resulting reference identification in data sources that are similar to the SAIL Databank, a classification algorithm in the form of a decision tree can be trained from a dataset labelled by the developed latent class model.

3.2 Objectives

In this chapter, I developed an identification model for asthma in the SAIL Databank, and evaluated it against other widely used physician-reported and self-reported case definitions. Specifically, the objectives were to:

- Develop a data-driven reference identification of asthma, particularly those with currently treated asthma, using LCA of RCD in the SAIL Databank.
- Derive a classification algorithm for asthma that can be used in the SAIL Databank and similar databases.
- Assess the agreement between the classification algorithm and the following commonly used case definitions:
 - GP-reported ever-diagnosed asthma

- GP-reported currently treated asthma
- GP-reported ever-diagnosed, currently treated asthma
- Self-reported currently treated asthma.
- Self-reported currently treated COPD.

These case definitions are explained below in [Section 3.3.5.1](#). This assessment included analysis of concordance and calculating estimates of diagnostic accuracy for those case definitions using the predictions by classification algorithm as references and vice versa.

3.3 Methods

In a cross-sectional design, I used primary care data on asthma and COPD recorded in or before 2014 for a sample of the Welsh population to find, using LCA, clinically meaningful classes (i.e. clusters) related to the two conditions in that year. Based on the chosen latent class model, I then derived a classification algorithm to identify patients with asthma, including those with currently treated asthma, as well as those with COPD and ACOS in the primary care population. I compared that classification algorithm with other case definitions for asthma and COPD based on doctor-reported and patient-reported data.

[Figure 3.2](#) illustrates the methodology followed in this chapter.

3.3.1 Data sources

I used individual-level demographic and primary care data from the SAIL Databank. The SAIL Databank contains anonymised linked data datasets derived from EHRs and several non-health datasets in Wales [[140](#), [141](#)]. I used the following two datasets:

- The Welsh Demographic Service (WDS): The WDS contains de-identified demographic and administrative information for National Health Services (NHS) patients in Wales. I used this dataset to identify NHS patients who satisfied the follow-up criteria described below.
- The GP dataset: The GP dataset contains de-identified health care events, such as recorded diagnoses, clinical findings, prescriptions and monitoring as well as other events codified in Read codes by GPs. I used the GP dataset as

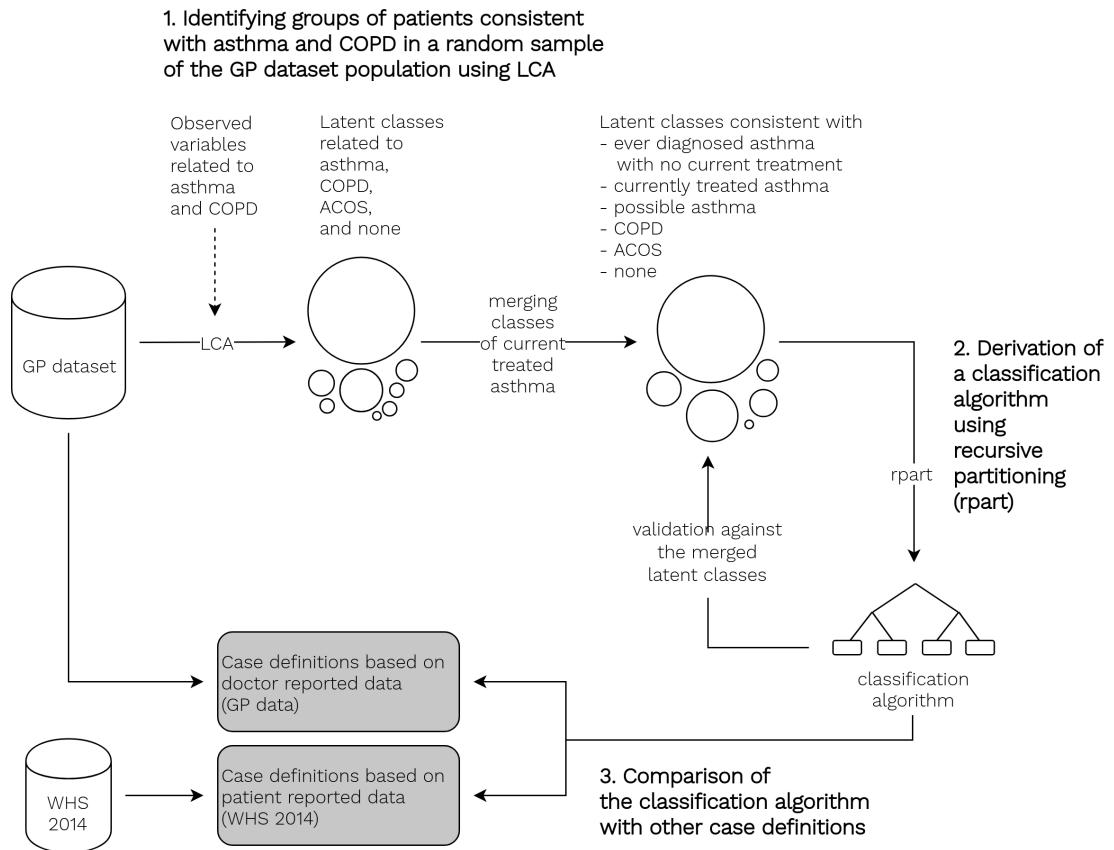


Figure 3.2: The methodology followed in [Chapter 3](#). ACOS: asthma-COPD overlap syndrome; COPD: chronic obstructive pulmonary disease; GP: General Practice; LCA: latent class analysis; WHS: Welsh Health Survey.

a main data source in my analysis since in the United Kingdom (UK) asthma is mainly treated in primary care [32]. At the time of data extraction and analysis, the most recent extract of the GP dataset was in April 2016, covering about 80% of GP surgeries in Wales, which voluntarily sent their data to the SAIL Databank. I used the GP dataset to measure the observed variables described below.

Since I only accessed anonymised health data within the SAIL Databank and did not work with humans, no ethical approval was required. This doctoral project was covered by approval of the SAIL Information Governance Review Panel (see [Appendix C.2](#)).

3.3.2 Patient population

I defined the source population of the study as every individual who satisfied the following criteria:

- Registered in at least one GP practice that contributed its data to the SAIL Databank at the time of the analysis.
- Had continuous GP registration during the analysis year, i.e. between 1-1-2014 and 31-12-2014. To calculate GP registration periods for individuals, I used an unpublished commonly used algorithm developed in-house by the analyst team of the SAIL Databank. This algorithm takes into account periods of registration with SAIL and non-SAIL participating practices as well as the volumes of data contributed by each practice over time. The algorithm uses that information to determine periods of continuous follow-up of patients in the primary care dataset in the SAIL Databank.
- Did not die on or before 31-12-2014.

I did not apply age restrictions to the source population. The study sample was randomly selected from the source population. The sampling was stratified by general practices to improve their representativeness. I determined the sample size for latent class modelling based on the available computational capacity in the SAIL Gateway.

3.3.3 Latent class modelling

3.3.3.1 Observed variables

In this modelling, the observed variables were based on GP-recorded primary care events related to asthma and COPD. The choice of these events and the dimensions of the observed variables was determined based on their usefulness, from a clinical perspective, for identifying and distinguishing between patients with asthma and/or COPD. Including observed variables of both asthma and COPD in the same latent class model allowed identification of patients with either or both conditions (i.e. ACOS). These variables included events on disease-related diagnosis, GP visits, prescriptions, and smoking. Events of GP visits and prescriptions were queried over the analysis year, while the other events were queried over any time up to the end of the analysis year. Most of the observed variables were transformed into binary variables: '0' for 'no events found' and '1' for 'one or more events found'. The 'age at asthma diagnosis' variable had three categories: '< 40 years' and '40 or more' as well as 'no diagnosis'. A full list of variables is shown in [Table 3.1](#). The lists of Read codes used in the variable definitions are available in [Table B.2.1](#).

Table 3.1: Observed variables used in the latent class model. Clinical codes are listed in [Table B.2.1](#).

Variable	Time interval for calculation	Categories
<i>Asthma related</i>		
Asthma diagnosis codes ever	ever	0, 1+
Age at asthma first diagnosis codes, if any	-	< 40, \geq 40, no diagnosis
Asthma GP visits codes in the last 12 months	last two years	0, 1+
<i>COPD related</i>		
COPD diagnosis codes ever	ever	0, 1+
COPD GP visits codes in the last 12 months	last two years	0, 1+
COPD-specific prescriptions codes**	last two years	0, 1+
<i>Prescriptions</i>		
ICS codes	last two years	0, 1+
SABA codes	last two years	0, 1+
LABA codes	last two years	0, 1+
ICS+LABA codes	last two years	0, 1+
OCS codes	last two years	0, 1+
LTRA codes	last two years	0, 1+
<i>Others</i>		
Smoking history	ever	no, yes

Abbreviations: COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; GP = general practitioner; LTRA = leukotriene receptor antagonists; LABA = long-acting β 2 agonists; OCS = oral corticosteroids; SABA = short-acting β 2 agonists.

** COPD-specific prescriptions include: glycopyrronium bromide, indacaterol, olodaterol, anticholinergic bronchodilators (ipratropium bromide, oxitropium bromide, tiotropium, aclidinium, umeclidinium), roflumilast, oxygen cylinders, and COPD rescue packs.

3.3.3.2 Number of classes and model selection

Theoretically, the expected number of latent classes is based on the observed variables used in the modelling. Since I included variables for both asthma and COPD, the minimum number of latent classes that I expected was four classes consistent

with the following labels: ‘asthma’, ‘COPD’, ‘both’, and ‘none’. However, the inclusion of asthma-related current GP visits and prescriptions (i.e. in the last 12 months) in the model was aimed at distinguishing groups of patients with currently treated asthma from those with ever diagnosed asthma without current treatment. Therefore, the number of expected classes could be increased to represent those different groups and possibly to also differentiate between patients with less and more severe diseases.

I started the modelling for two latent classes and then iteratively increased the numbers of latent classes. I aimed to select a model that satisfied both the following criteria:

- The BIC was minimum (compared to the competing models) or became ‘stabilised’ (e.g., using the elbow method, in which the researcher looks for an “elbow” in the plot of the model’s BIC against the number of classes).
- The latent classes were clinically relevant.

In the selected model, I assigned to the identified classes clinical labels consistent with ‘asthma’ (including currently treated asthma), ‘COPD’, ‘both’, and ‘none’ based on the estimated item-response probabilities in each latent class. I used the class proportions as prevalence estimates of their corresponding labels in the study population in 2014.

To simplify the model, I aimed to merge similar classes (e.g., classes of currently treated asthma differing in the probability of prescriptions) into super classes (e.g., currently treated asthma).

3.3.3.3 Statistical tool

I performed the latent class modelling using the R package *poLCA* (version 1.4.1, 2014) [260].

3.3.4 Derivation of a classification algorithm

Based on the latent class model, which was developed for the year 2014, I derived a classification algorithm which can be used to identify patients with asthma (including currently treated asthma), COPD and ACOS as well as in similar primary care datasets. To do so, I performed recursive partitioning [268] using the R package *rpart* (version 4.1-11, 2017) [269]. In this method, a decision tree is

constructed using supervised learning from a labelled dataset, called the training dataset. A decision tree is composed of nodes and branches. The nodes include *interior nodes*, each of which splits the corresponding partition of the sample into two parts based on a true/false question about one of the features, and *final nodes* (i.e. leaves) representing the labels assigned to the corresponding branches.

The construction of a decision tree is stepwise. At each step, including that of the root node (i.e. which corresponds to the whole training dataset), the dataset is split into two subsets based on a true/false question about one of the features (i.e. a predictor variable). Each split is intended to reduce the misclassification in the resulting two children nodes compared to that in their parent node. Since more than one feature can provide a reduction in misclassification, the feature that maximises such reduction is chosen for the given split. Theoretically, while the model fit may improve with further binary splits, a split may provide only small improvement in the misclassification. Allowing such a split leads to overfitting where the accuracy of the tree is high or perfect in the training dataset but low in the validation dataset. To prevent overfitting, smaller improvements in the model fit are penalised with higher costs. In addition, any split with improvement in the model fit that is smaller than a control measure, called the *complexity parameter*, is considered not worth pursuing. To determine the complexity parameter, *rpart* fits a full tree from which it extracts all the possible sub-trees and performs on each of which 10-fold cross-validation. It then shows the sub-trees for which the complexity parameters are greater than a set threshold, usually 0.01 or as desired by the researcher. It then determines the complexity parameter from the sub-tree that has the lowest cross-validation error. The determined complexity parameter can be then used to prune the full tree, giving a trimmed tree as a best solution, representing the best possible balance between model complexity and cost.

I used the sample previously used for LCA to perform recursive partitioning. I used the latent classes (after being merged into super classes as appropriate) as labels and the observed variables, used in the LCA model, as features. I randomly partitioned the sample into two subsets: a training subset (approximately 70%) for the decision tree development, and a validation subset (approximately 30%) to validate the developed decision tree. The two subsets were balanced in terms of the proportions of labels. To validate the developed decision tree, I used it to predict the labels in the validation dataset. Then, I calculated various diagnostic measures for the model using the `confusionMatrix` function of the `caret` package

(version 6.0.77, 2017) in R. These statistics included: classification accuracy (Acc, the proportion of correct predictions in the validation dataset along with its 95% confidence interval); the no information rate (NIR, also known as the no information error rate; the proportion of the largest class, which gives an idea of how useful the predictors were in predicting the classes compared with just predicting them using class proportions); the p-value of $\text{Acc} > \text{NIR}$ (a one-sided test to see if the model accuracy is better than just predicting the most prevalent class); Cohen's Kappa (for the agreement between the known labels and predictions); and statistics by class including sensitivity, specificity, positive and negative predictive values of class-specific prediction, class prevalence, detection rate (proportion of detected class members relative to the whole sample), detection prevalence (the prevalence of truly and falsely detected cases), and balanced accuracy ((sensitivity + specificity)/2).

3.3.5 Comparison of the classification algorithm with other case definitions

I compared the classification algorithm described above with other case definitions based on objective and self-reported data.

3.3.5.1 Case definition used for comparison

GP-reported ever-diagnosed asthma

The case definition 'GP-reported ever-diagnosed asthma' refers to patients who had, on a given date, asthma diagnosed by GPs and recorded using one of a set of Read codes indicating asthma diagnosis. To identify such patients in the SAIL's GP Dataset, I used the asthma diagnosis Read code set shown in [Table B.2.1](#), which were based on the Quality of Outcomes Framework (QOF)'s AST001 indicator [270].

GP-reported currently treated asthma

The case definition 'GP-reported currently treated asthma' refers to patients who were receiving asthma prescriptions on a given date. For the purpose of this thesis, this case definition was operationalised by identifying patients who had at

least one asthma prescription during the last 12 months before the end of the analysis year (i.e. between 1 January 2014, and 31 December 2014). The prescription codes are shown in [Table B.2.1](#) and were based on those used in the AST001 indicator of the QOF [270].

GP-reported ever-diagnosed, currently treated asthma

The ‘GP-reported ever-diagnosed, currently treated asthma’ case definition requires the patients to satisfy both the aforementioned case definitions. Thus, this case definition is almost identical to the AST001 indicator (but without excluding patients with ‘exception codes’).

Self-reported currently treated asthma

The ‘self-reported currently treated asthma’ case definition was based on the WHS of 2014 [54]. The WHS 2014 collected self-reported information on a range of health and health-related lifestyles from samples of the population of Wales. The WHS 2014 results dataset was already linked to the SAIL Databank [271]. However, from that dataset, I only had access to the responses of participants who were 16-year-old and above in 2014. Those participants consented to link their responses to their medical records [272]. From this dataset, the only question related to asthma, other than asthma symptoms, asked the participant whether he or she was currently treated for a number of diseases including asthma. I used responses to this question as a case definition for ‘self-reported currently treated asthma’. I considered invalid responses as negative responses.

Self-reported currently treated COPD

In the WHS 2014, participants were asked whether they were currently treated for chronic bronchitis and/or emphysema. I considered positive responses for any of these two conditions as a case definition for ‘self-reported currently treated COPD’. I treated invalid responses as negative responses.

3.3.5.2 Statistical analysis

I performed the comparisons between the classification algorithm and each of the above-mentioned case definitions in the group of WHS 2014 participants whose

responses where successfully linked to SAIL. I calculated sensitivity, specificity, PPV, NPV, and Cohen's kappa coefficient for concordance of each of the above case definitions against the classification algorithm labels as references and vice versa.

3.4 Results

3.4.1 Sampling and sample characteristics

The size of source population was 2,303,819 which equated to approximately 74.5% of the mid-year estimate of 3,092,000 for the population of Wales in mid-2014.² The study sample included 50,000 individuals, 50.3% of whom were females. [Figure 3.3](#) shows a histogram for the sample age at the beginning of the year 2014.

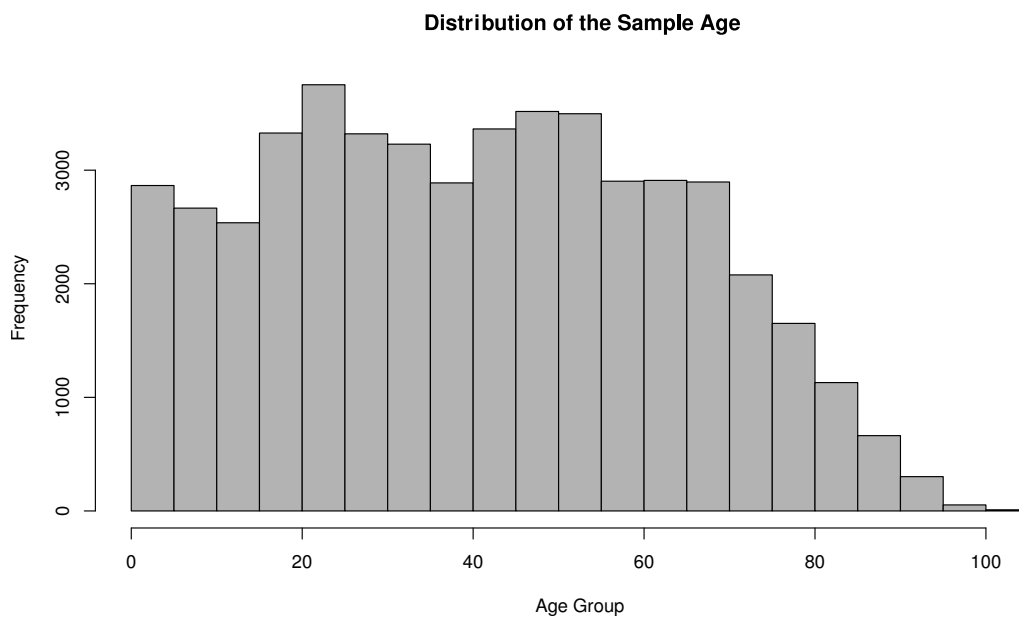


Figure 3.3: A histogram for the sample age at the beginning of year 2014.

²Based on: *Time series: Wales population mid-year estimate*. Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/timeseries/wapop/pop>

3.4.2 Summary of the competing latent class models

I started the latent class modelling for a one-class model and was able to increase the number of classes up to 12. The item-response probabilities within the classes of each of the 12 competing models are shown in diagrams in [Appendix B.3](#).

According to the item-response probabilities, in the two-class model, people with events related to asthma and/or COPD were aggregated in one class with a share of 16.2%. In the three-class model, this class was split into two classes. One of these classes had a share of 14.1% and seemed to include mostly asthma patients, although it appeared highly heterogeneous, as indicated by the marginal probabilities for the asthma GP visits and short acting beta agonist (SABA) variables. A very small proportion of people in this class showed characteristics suggesting COPD. Another class in the two-class model, with a share of 2.4% appeared to have significant marginal probabilities (i.e., close to 0.5) for events related to COPD and prescriptions, suggesting high heterogeneity. A very small proportion of people in this class had characteristics related to asthma.

With higher number of classes, the models continued to reclassify people into more refined classes with higher homogeneity. In the four-class model, the asthma-dominated class in the three-class model was further split into two classes both of which had 'asthma diagnosis ever'; one class with a share of 7% had also high probabilities for asthma-related current GP visits and prescriptions, while the other class had zero to very low probabilities of these events.

In the five-class model, there were one class (6.6%) for ever diagnosed asthma with no current treatment, another class (2.4%) appeared to be dominated by COPD characteristics with some probability for asthma-related events, and one class (84%) with no asthma or COPD characteristics. However, there were two almost-similar classes (3.7% and 3.4%) for ever-diagnosed currently treated asthma; the main differences between these two classes were that one had a very high probability for ICS and a very low probability for ICS-long-acting beta adrenoceptor agonist (LABA), while the other class had the opposite situation: a very low probability for ICS and a very high probability for ICS-LABA.

In the six-class models, those two 'ever-diagnosed currently treated asthma' were almost reunited into one class (6.9%). There was one highly homogeneous class (1.4%) suggesting currently treated COPD, one class (6.2%) for ever diagnosed

not currently treated asthma, and one class (1.0%) that showed high probabilities for both current asthma and COPD events. However, there was one class (2.2%) showing only variable probabilities for asthma prescription, with almost no recorded diagnosis of asthma or COPD.

In the seven-class model, the only significant refinement on the previous model was that a previously one large class with asthma or COPD characteristics become split into two classes (51.4% and 30.9%) with low and high probabilities of smoking history, respectively.

In the eight-class model, these two none-asthma none-COPD classes joined again. The two treated asthma classes, previously observed in the five-class model, emerged again, however, with different prevalences; the class with ICS events and no ICS-LABA combinations had a prevalence of 3.9%, while the class with ICS-LABA combinations but with no sole ICS prescriptions had a prevalence of 2.7%. There was an 'asthma and COPD' class (0.9%) with high homogeneity; it showed high probabilities for both asthma and COPD diagnosis, GP visits, and prescription events. In this class, the probability of 'smoking ever' was very high (90.9%). However, the probability of having recent asthma-related GP-recorded events visits in the last 12 months was marginal (43.0%). These item-response probabilities indicated patients in this class potentially had ACOS. Other classes included one class (6.6%) for 'ever diagnosed asthma without current treatment', one class (1.3%) for 'ever diagnosed currently treated' COPD, one class (1.4%) with high probabilities for each of SABA and ICS events and low probability for 'ever smoking', and one class (1.0%) with low to marginal probabilities for asthma prescriptions but very high probability for 'ever smoking'.

Compared to the eight-class model, the significant refinements in the nine-class model were a split of one of the 'ever-diagnosed currently treated asthma' classes into two (2.5% and 1.3%) which, however, were reunited in the 10-class model. The assignment of the sample individuals into the classes across the competing models is shown in [Figure 3.4](#), while model diagnostics for each of the competing models are shown in [Figure 3.5](#).

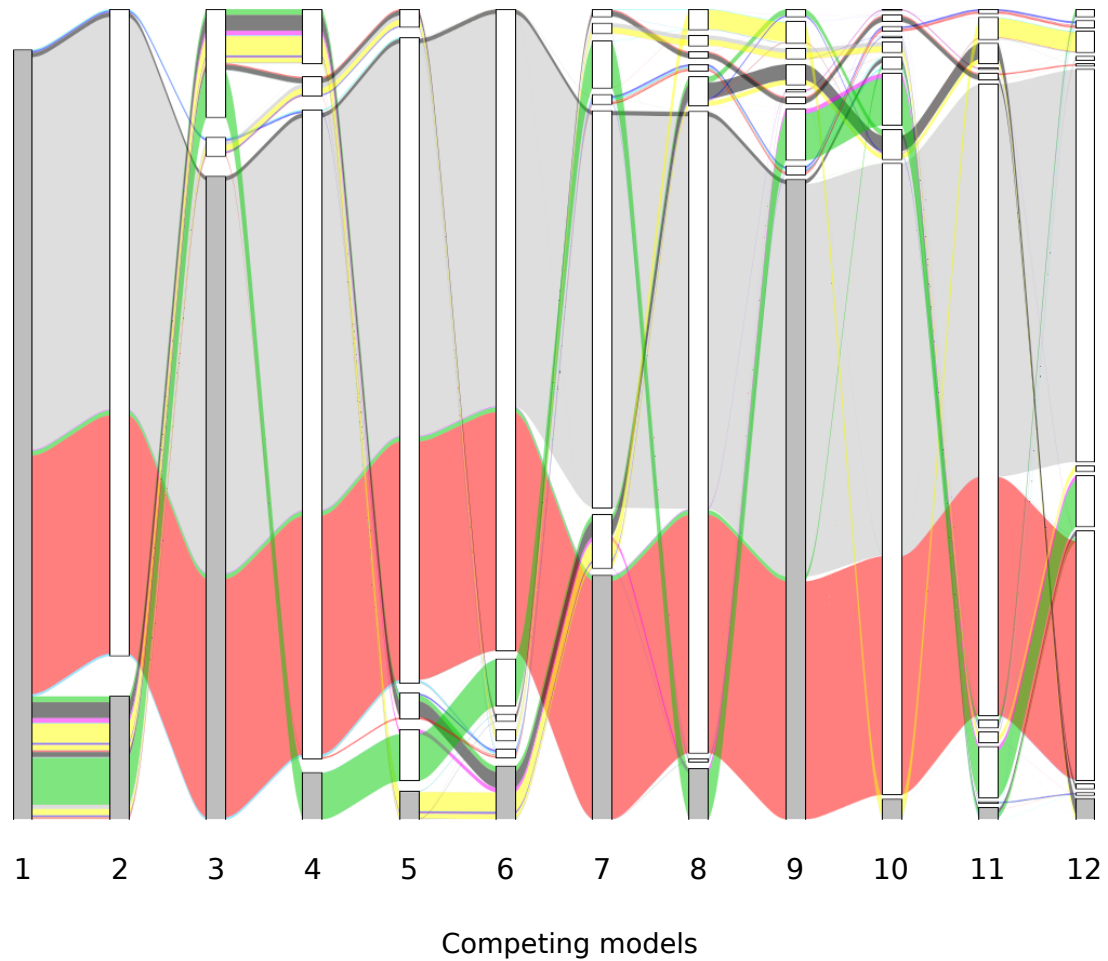


Figure 3.4: An alluvial diagram showing assignment of the sample individuals across the produced competing latent models. Each band represents a group of individuals sharing the same class in the last competing model (the 12-class model), and demonstrates how they were assigned to classes across the other competing models.

3.4.3 Model selection

Based on the model diagnostics diagram (Figure 3.5), the AIC, BIC, and G^2 diagnostics for the models with one to 12 classes declined significantly between the one-class model and the four-class model. Then, these three diagnostics continued to decline slightly until they stabilised at the eight-class and nine-class models. The Chi-square static declined abruptly between the one-class and the two-class models before it appeared to visually stabilise across the competing models. The nine-class model had the lowest BIC value, while the other diagnostics, AIC, Chi-square, and G^2 , had their lowest values at the 12-class model. However, the decline in the BIC value between the eight-class model and the nine-class model was very small and negligible, indicating very little improvement in the information gain. In addition, as described in Section 3.4.2, the structures of these two

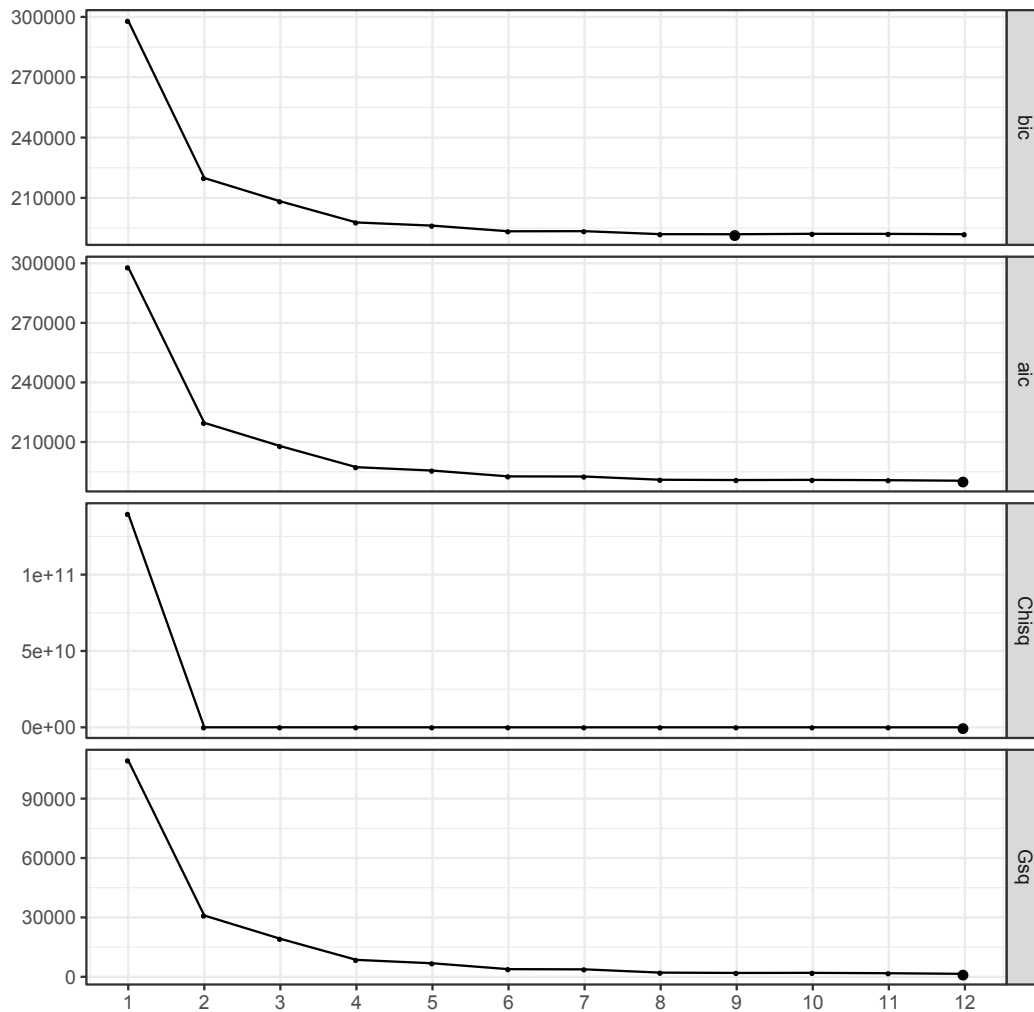


Figure 3.5: Diagnostics for the competing latent class models. For each diagnostic, the class with the minimum value was marked with a large dot. AIC = Akaike information criterion; BIC = Bayesian information criterion; Chisq = Pearson Chi-square goodness of fit statistic; Gsq = G-squared.

classes were very similar, with the eight-class model showing clinically meaningful classes. Therefore, the eight-class model appeared to be a good solution, among the other competing models, that reasonably fit with the purpose of this latent class modelling of identifying patients with asthma, including those with currently treated asthma.

3.4.4 Model interpretation: Characterising and labelling the identified classes of the best-fit model

Figure 3.6 shows the classes of the eight-class model along with their prevalences, item-response probabilities, and assigned labels.



* Queried in the last 12 months.

Figure 3.6: Class prevalences and item-response probabilities of the eight-class model. The left-most column shows the names and levels of the observed variables. The small bar-plot above each latent class demonstrates class separation: it shows the average probabilities, within each class, of membership in all the latent classes.

The eight-class latent class model which I chose as a solution consisted of the following classes:³

1. **'Ever diagnosed asthma without current treatment'**: This class has a prevalence of 6.6%. All the individuals in this class had asthma diagnosis events, recorded in 88% of the class members before the age of the 40. Only 5.0% and 6.9% of the class members had asthma-related events and SABA prescriptions in 2014. 41.0% of the class members ever smoked on or before 2014.
2. **'No asthma or COPD diagnosis, modestly currently treated, with ever smoking'**: The prevalence of this class was 1.0%. Its individuals had no asthma diagnosis events recorded by the end of 2014, although 2.4% of them had non-diagnosis asthma-related GP events (excluding prescriptions) recorded in 2014. However, 50.0% of people in this class had at least one SABA prescription, and 19.5% had ICS-LABA combination prescriptions, and 30.6% had prescriptions for oral steroids in 2014. In addition, 11% of the class members had recorded COPD diagnosis, although only 1.8% had COPD-related GP visits in 2014, and 11.8% had prescriptions usually prescribed for COPD. Smoking related events were found for 78.9% of the class members. The average age of the class members in 2014 was 61.7 years with a standard deviation of 18.6 years. The marginal item-response probabilities in this class suggests it includes heterogeneous sub-groups of individuals some of which might be COPD patients, while others were smokers without symptoms of COPD. For this class, I suggested the label 'possible/at risk of COPD'.
3. **'Neither asthma/COPD' (or 'none')**: This class, having a prevalence of 82.2%, had almost zero item-response probabilities for all the observed variables except for 'smoking ever' which had 37.8% probability.
4. **'Ever-diagnosed currently treated asthma; ICS without LABA'**: This class had a prevalence of 3.9% and showed 100% probability for asthma diagnosis events recorded before the age of 40 for 73.2% of the class members. 74.6% of the class members had recorded asthma-related GP events in 2014 and had high probabilities for ICS (75.8%) and SABA (93.9%) prescriptions and 15.6% probability for oral corticosteroids (OCS) prescriptions. 43.3%

³The numbers associated with the classes were merely numerical labels as appeared in the output of the latent class modelling, and did not imply any order.

of individuals in this class had recorded smoking-related events. This class included people from all age groups (mean = 39.5; standard deviation (SD) = 22.2).

5. **'Currently treated; no recorded asthma or COPD diagnosis; with modest smoking ever'**: This class with 1.4% prevalence was similar to the class #2 described above, as it had no recorded asthma diagnosis. However, it had zero-probability for COPD diagnosis events, very high probability (89.9%) for SABA prescriptions in 2014, and 41.8% for ICS prescriptions in that year. The class members had 24.5% probability for OCS prescriptions, and 29.8% probability of 'smoking ever'. Based on these characteristics, I assumed this class included asthma patients with no recorded diagnosis.
6. **'ACOS'**: This class with 0.9% prevalence showed almost total probabilities for asthma and COPD diagnosis events. For 75.3% of the class members, the earliest recorded diagnosis of asthma was before the age of 40. While 43.5% of the class members had asthma related events in 2014, 74.3% had COPD-related events in that year, and 70.2% had prescriptions specific to COPD in the same year. The class members had a very low probability for ICS prescriptions (11.0%), but very high probabilities for SABA (91.4%) and ICS-LABA combination prescriptions (81.7%) in 2014. Almost half of the class members (49.7%) had oral steroids in 2014 indicating severe and/or uncontrolled symptoms.
7. **'Ever-diagnosed currently-treated COPD'**: This class had a prevalence of 1.3% and showed 100% probability for COPD diagnosis events, high probabilities for COPD-related GP events (74.3%) and COPD-specific prescriptions (70.2%). The vast majority of people in this class (96.5%) had recorded smoking history. There was also a high probability for SABA prescriptions (81.2%), but a marginal probability for ICS-LABA combination prescriptions (54.4%), and a 31.7% probability for oral steroids in 2014. The class showed very low probabilities for ICS-alone (9.0%) and LABA-alone (8.8%) prescriptions.
8. **'Ever-diagnosed currently treated asthma; ICS with LABA'**: This class had a prevalence of 2.7%. It was closely related to the class #4 described above. All its members had asthma diagnosis events, recorded under the age of 40 for 60.2% of patients. 74.0% of the class members had asthma-related

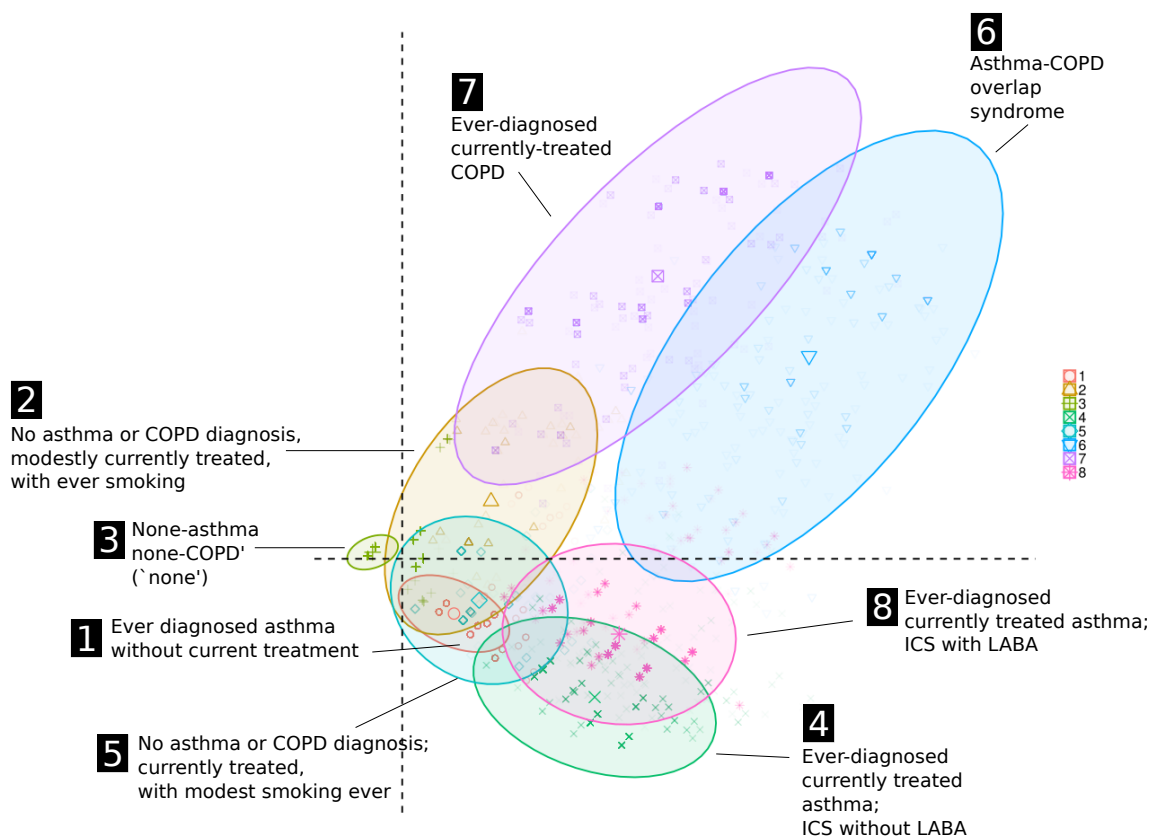


Figure 3.7: Visualisation of the eight-class latent class model using principal component analysis. ACOS: asthma-COPD overlap syndrome; COPD: chronic obstructive pulmonary disease; GP: general practitioner; ICS: inhaled corticosteroid; LABA: long-acting beta adrenoceptor agonist; SABA: short acting beta agonist.

GP events in 2014. Almost half of people in this class (53.2%) had recorded smoking events.

Figure 3.7 visualises the classes in the eight-class model using principal component analysis (PCA). The distances and overlaps between the latent classes were consistent with their clinical interpretation overall.

3.4.5 Class merging

Since I was interested in identifying patients with active asthma, I merged the classes 4, 5, and 8 into a single super class labelled “currently treated asthma”. Most patients (82.5%) in that super class had recorded asthma diagnosis (Class 4 and 8); the remaining patients (Class 5, 17.5%) had no recorded diagnosis of asthma or COPD but had a very high probability of using SABA inhalers, a marginal probability ($\approx 42\%$) of using ICS inhalers, and a low probability ($\approx 30\%$) of smoking history. Although some of those patients might have received those inhalers without actually having asthma, I kept them in the super class of ‘currently treated

Table 3.2: The latent classes and their prevalences before and after merging.

Before merging	After merging
1. Ever diagnosed asthma without current treatment (6.6%)	No change
2. possible COPD (1.0%): No asthma or COPD diagnosis, modestly currently treated, with ever smoking	No change
3. 'None' (82.2%): None-asthma none-COPD'	No change
4. Ever-diagnosed currently treated asthma; ICS without LABA (3.9%)	Currently treated asthma (8.0%)
5. No asthma or COPD diagnosis; currently treated, with modest smoking ever (1.4%)	
8. Ever-diagnosed currently treated asthma; ICS with LABA (2.7%)	
6. Asthma-COPD overlap syndrome (0.9%)	No change
7. Ever-diagnosed currently-treated COPD (1.3%)	No change

ICS: inhaled corticosteroid; LABA: long-acting beta adrenoceptor agonist; COPD: chronic obstructive pulmonary disease.

asthma'. [Table 3.2](#) shows the classes and their prevalences before and after merging. I later used the labels of the resulting simpler six-class structure in the training of a classification algorithm (see [Section 3.4.6](#) below).

3.4.6 Derivation of classification algorithm

A decision tree with 11 splits (and 12 leaves) was trained using recursive partitioning. The R package `rpart` performed 10-fold cross-validation to select the optimal size of the decision tree. The results of this cross-validation are shown in [Table 3.3](#) and [Figure 3.8](#).

Table 3.3: Results of 10-fold cross-validation for the recursive partitioning model, showing the cross-validation error for each sub-tree.

	Complexity parameter	Number of splits	Relative error	Cross-validation error rate	Cross-validation standard deviation
1	0.39935	0	1.00000	1.00000	0.01193
2	0.31750	1	0.60065	0.60065	0.00961
3	0.07331	2	0.28315	0.28315	0.00679
4	0.04409	3	0.20984	0.20984	0.00588
5	0.03589	4	0.16576	0.16576	0.00525
6	0.02324	5	0.12987	0.12987	0.00466
7	0.01162	6	0.10663	0.10663	0.00423
8	0.01025	7	0.09501	0.09501	0.00400
9	0.01003	8	0.08476	0.08630	0.00381
10	0.01000	11	0.05468	0.07536	0.00357

Root node error: $5,852/35,004 = 0.16718$

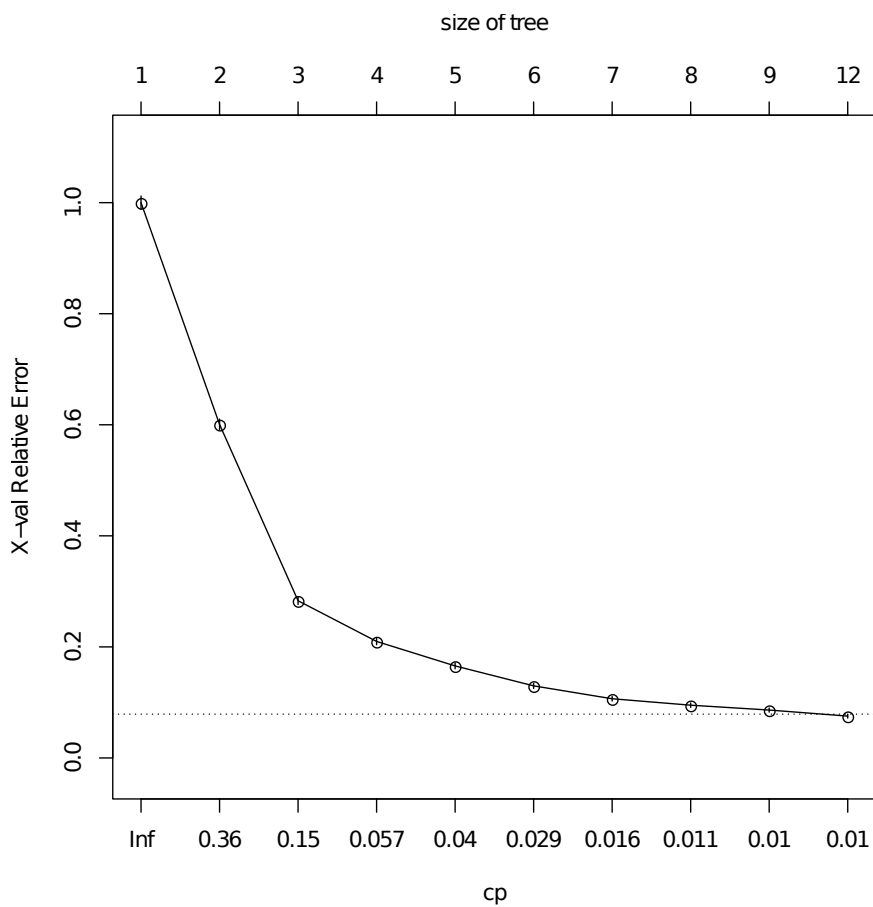


Figure 3.8: Visualisation of the 10-fold cross-validation results of the recursive partitioning. The horizontal dotted line was drawn at 1 standard deviation above the minimal cross-validation error. The tree with 12 leaves (11 splits) had the lowest cross-validation error and was therefore the desired tree. cp: complexity parameter.

In these results, the tree with 11 splits had the minimum cross-validation error rate (0.075). The complexity parameter of this tree, 0.01, was used to further prune the full tree in an attempt to remove splits, if any, that caused over-fitting. The final pruned tree is shown in [Figure 3.9](#).

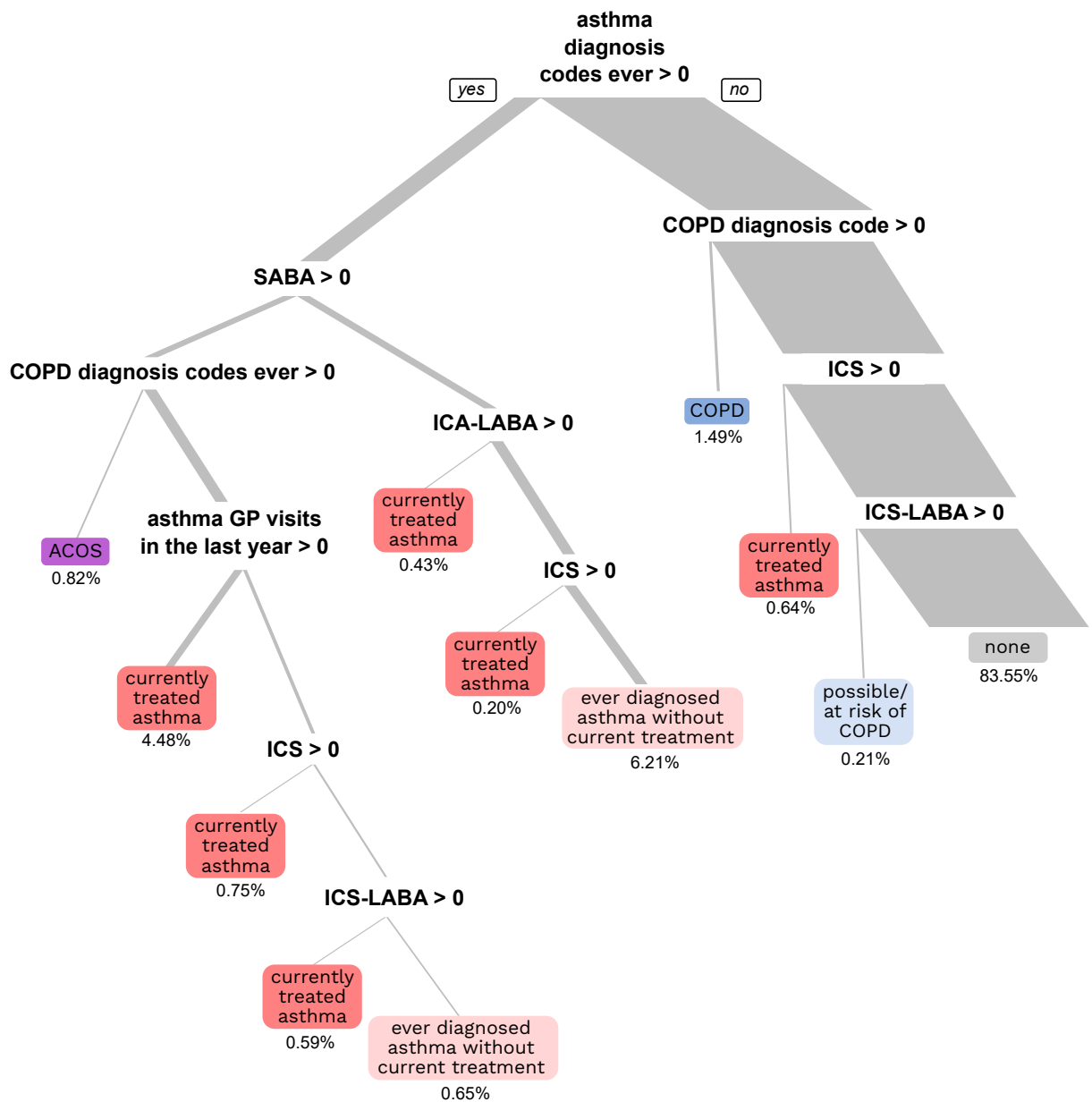


Figure 3.9: A decision tree representation of the classification algorithm. At each node, the left branch is followed when the condition is true. The width of branches is proportionate to the number of individuals in the derivation sample who followed these branches relative to the size of the entire sample. The final nodes, i.e. leaves, represent the labels to be assigned for new cases. ACOS: asthma-COPD overlap syndrome; COPD: chronic obstructive pulmonary disease; GP: general practitioner; ICS: inhaled corticosteroid; LABA: long-acting beta adrenoceptor agonist; SABA: short acting beta agonist.

This final tree represented the classification algorithm that could be used to classify individuals in new samples into the following six categories:

- ‘currently treated asthma’
- ‘ever diagnosed asthma without current treatment’
- ‘possible/at risk of COPD’
- ‘COPD’
- ‘ACOS’
- ‘none’

The actual variables that were chosen by recursive partitioning for the construction of this tree were:

- “having asthma diagnosis events ever”
- “having asthma related GP visit in the last 12 months”
- “having COPD diagnosis events ever”
- “having SABA prescriptions in the last 12 months”
- “having ICS prescriptions in the last 12 months”
- “having ICS-LABA combination prescriptions in the last 12 months”

Table 3.4 shows a confusion matrix and statistics for the cross-classification of the decision tree predictions against the labels used in the training. Overall, the predictive performance of the classification algorithm was very high, which meant it could be used to accurately classify new populations based on the latent class model that I described earlier in this chapter. An exception was that the algorithm had a low sensitivity to identify people with ‘possible/at risk of COPD’ as it misclassified 56% of them into the ‘none’ class.

Table 3.4: Confusion matrix and statistics for the cross-classification of the predicted classifications against the LCA-based labels in the validation dataset. These statistics reflected how well the classification algorithm represented the latent class model.

		Predicted				
		Ever diagnosed asthma without current treatment	Currently treated asthma	Possible/at risk of COPD	COPD	None
<i>LCA-based labels</i>						
ACOS		116	6	10	0	0
Ever diagnosed asthma without current treatment		<5*	1,004	0	0	0
Currently treated asthma		9	8	1,035	10	<5*
Possible/at risk of COPD		0	0	0	26	<5*
COPD		0	0	0	<5*	199
None		0	<5*	11	0	25
						12,455

* Values masked to comply with the SAIL Databank policy on small numbers.

Overall statistics	
Accuracy	0.9893 (95% confidence interval (CI): 0.9875, 0.9909)
No Information Error Rate	0.8354 (i.e. if we just assigned all individuals into the largest class 'none', we would be correct in 83.5% of cases)
p-value of a one-sided test that "the Accuracy is higher than the No Information Error Rate"	< 2.2 ⁻¹⁶
Cohen's Kappa	0.9635

Class-specific statistics						
	ACOS	Ever diagnosed asthma without current treatment	Currently treated asthma	Possible/at risk of COPD	COPD	None
Sensitivity	0.921	0.984	0.980	0.684	0.873	0.994
Specificity	0.999	1.000	0.995	0.997	1.000	0.985
Positive Predictive Value	0.879	0.999	0.940	0.406	0.990	0.997
Negative Predictive Value	0.999	0.999	0.998	0.999	0.998	0.971
Prevalence	0.008	0.068	0.070	0.003	0.015	0.835
Detection Rate	0.008	0.067	0.069	0.002	0.013	0.831
Detection Prevalence	0.009	0.067	0.073	0.004	0.013	0.833
Balanced Accuracy	0.960	0.992	0.988	0.841	0.936	0.989

ACOS: asthma-COPD overlap syndrome; COPD: chronic obstructive pulmonary disease; CI: confidence interval; LCA: latent class analysis.

3.4.7 Comparing the classification algorithm with other case identification methods

Comparisons between the classification algorithm and the case definitions described in [Section 3.3.5.1](#) are shown in [Table 3.5](#).

The following two comparisons had the highest Cohen's kappa value:

- The 'QOF indicator of asthma (AST001)' and the algorithm's definition of 'currently treated asthma' had a kappa value of 86.5%.
- 'Asthma diagnosis code' and the union of algorithm's definitions of 'ever diagnosed asthma without current treatment' and 'currently treated asthma' (i.e. 'any asthma') had a kappa value of 94.5%.

However, the self-reported definition of currently treated asthma had lower agreements with the algorithm classification, with kappa values of 61.5% and 56.4% for the concordance with the algorithm's definitions of 'currently treated asthma' and 'any asthma', respectively. Interestingly, the self-reported definition of currently treated asthma has a level of non-random agreement, although very low (kappa = 12.2), with the algorithm's definition of 'ever diagnosed asthma without current treatment'. Finally, the algorithm's definition of currently treated COPD had poor disagreement with the self-reported definition of currently treated COPD (kappa = 28.3).

Table 3.5: Comparison between the classification algorithm's case definitions and other case definitions based on doctor-reported and self-reported data.

		(a) The algorithm's definition as references.						
Case definition	Reference	Sensitivity	Specificity	PPV	NPV	Kappa		
1 Self-reported currently treated asthma	Predicted ever diagnosed asthma without current treatment	25.7	91.0	14.3	95.5	12.2		
2 Self-reported currently treated asthma	Predicted currently treated asthma	86.9	94.9	51.1	99.2	61.5		
3 Self-reported currently treated asthma	Predicted any asthma (ever diagnosed and/or currently treated)	57.1	96.1	65.4	94.6	56.4		
4 QOF AST001 indicator criteria	Predicted ever diagnosed asthma without current treatment	7.7	93.7	6.6	94.6	1.3		
5 QOF AST001 indicator criteria	Predicted currently treated asthma	91.7	98.9	83.4	99.5	86.5		
6 QOF AST001 indicator criteria	Predicted any asthma (ever diagnosed and/or currently treated)	50.9	99.3	90.0	94.0	61.9		
7 Ever GP diagnosed asthma	Predicted ever diagnosed asthma without current treatment	100.0	93.7	48.0	100.0	62.0		
8 Ever GP diagnosed asthma	Predicted currently treated asthma	91.7	93.5	46.5	99.5	58.5		
9 Ever GP diagnosed asthma	Predicted any asthma (ever diagnosed and/or currently treated)	95.8	99.3	94.5	99.5	94.5		
10 Currently treated asthma	Predicted ever diagnosed asthma without current treatment	7.7	90.0	4.3	94.4	-1.7		
11 Currently treated asthma	Predicted currently treated asthma	100.0	95.7	58.9	100.0	72.1		
12 Currently treated asthma	Predicted any asthma (ever diagnosed and/or currently treated)	55.1	95.9	63.2	94.4	54.0		
13 Self-reported currently treated COPD	Predicted ever diagnosed currently treated COPD	36.5	98.2	25.0	98.9	28.3		

(b) The other case definition as references.

	Case definition	Reference	Sensitivity	Specificity	PPV	NPV	Kappa
1	Predicted ever diagnosed asthma without current treatment	Self-reported currently treated asthma	14.3	95.5	25.7	91.0	12.2
2	Predicted currently treated asthma	Self-reported currently treated asthma	51.1	99.2	86.9	94.9	61.5
3	Predicted any asthma (ever diagnosed and/or currently treated)	Self-reported currently treated asthma	65.4	94.6	57.1	96.1	56.4
4	Predicted ever diagnosed asthma without current treatment	QOF AST001 indicator criteria	6.6	94.6	7.7	93.7	1.3
5	Predicted currently treated asthma	QOF AST001 indicator criteria	83.4	99.5	91.7	98.9	86.5
6	Predicted any asthma (ever diagnosed and/or currently treated)	QOF AST001 indicator criteria	90.0	94.0	50.9	99.3	61.9
7	Predicted ever diagnosed asthma without current treatment	Ever GP diagnosed asthma	48.0	100.0	100.0	93.7	62.0
8	Predicted currently treated asthma	Ever GP diagnosed asthma	46.5	99.5	91.7	93.5	58.5
9	Predicted any asthma (ever diagnosed and/or currently treated)	Ever GP diagnosed asthma	94.5	99.5	95.7	99.3	94.5
10	Predicted ever diagnosed asthma without current treatment	Currently treated asthma	4.3	94.4	7.7	90.0	-1.7
11	Predicted currently treated asthma	Currently treated asthma	58.9	100.0	100.0	95.7	72.1
12	Predicted any asthma (ever diagnosed and/or currently treated)	Currently treated asthma	63.2	94.4	55.1	95.9	54.0
13	Predicted ever diagnosed currently treated COPD	Self-reported currently treated COPD	25.0	98.9	36.5	98.2	28.3

3.5 Discussion

3.5.1 Summary and interpretation of the findings

In this chapter, I described the development of a latent class model to identify patients with ‘asthma’ (including those currently treated asthma and those with ever diagnosed asthma without current treatment), COPD, and both conditions in Wales in 2014. Based on this model, I trained a classification tree which can be used to classify new samples into the above labels.

I performed the latent class modelling for 1-12 classes on a random sample of the Welsh population. Based on model diagnostics and clinical interpretability, I chose the eight-class model as the optimal clustering in relation to the observed variables used in the modelling. The eight-class model succeeded in clustering the population into distinct, homogeneous classes. There was one large class (82.2%) characterised by the absence of almost all asthma and COPD related events, except having a modest probability of smoking history. There were also four classes consistent with asthma: one ever-diagnosed currently treated, two ever-diagnosed currently treated, and a smaller class with no diagnosis but with current asthma prescriptions. Only one small class had distinctive characteristics of COPD with a prevalence of 1.3%; almost all people this class had positive smoking history. Interestingly, the overlap between asthma and COPD was clearly represented in a distinct, homogeneous class with a prevalence of 0.9%. One class showed low to marginal probabilities for asthma prescriptions with very low probabilities for COPD diagnosis and prescriptions.

Following model interpretation, I merged classes consistent with ‘currently treated asthma’, simplifying the clustering model into six labelled groups. I then used these labelled groups along with all the observed variables, which were used in the latent class modelling, to train a classification algorithm. The best classification algorithm was a decision tree with 11 splits and 12 final nodes. This algorithm is transferable and therefore can be used in new samples in the GP dataset in the SAIL Databank and could be also tested in similar external datasets. The ‘currently treated asthma’ label predicted by the classification algorithm included all patients with current asthma prescriptions. This label had a high agreement

with the QOF definition of ever diagnosed currently treated asthma (the AST001 indicator). The union of the algorithm's definitions of 'currently treated asthma' and 'ever diagnosed asthma without current treatment' had a very high agreement with having an asthma diagnosis code ever. There was a suboptimal agreement between the algorithm's definition of 'currently treated asthma' and the WHS-based definition of 'self-reported currently treated asthma'. This can be potentially explained by the possibility that some respondents thought they had asthma while they did not, while some respondents may have not received asthma prescriptions in the last 12 months or ever. In addition, the WHS did not specify a time frame when asking the respondents whether they were "currently" treated for asthma. Accordingly, respondents might have understood the word "currently" as time frames different from the 12-month interval traditionally used by researchers. These explanations were supported by the analysis I presented in [Appendix B.1](#), in which I found discordance between the WHS definition of self-reported currently treated asthma and GP-reported asthma diagnosis and prescriptions, including the interval over which prescriptions were queried.

3.5.2 Strengths and limitations

3.5.2.1 Strengths

Latent class analysis can reveal asthma epidemiology in routine data

In the absence of an accepted reference standard for identifying asthma patients, mixture modelling methods such as LCA allow identifying likely patients using the available observed data. Since asthma in the UK is mainly managed in primary care, I performed the LCA based on the informed assumption that asthma epidemiology was reflected in primary care EHR data. LCA follows a top-down approach, unlike the bottom-up approach used in cluster analysis. The latter assesses the similarities between individuals in order to form clusters. LCA, however, utilises the distributions of the observed characteristics to identify distinct latent groups, and then assesses the membership probabilities of each individual into these hypothesised groups. Those probabilistic class memberships fit with the nature of asthma as a probabilistic rather than a binary condition. The identified latent classes reflected the heterogeneous nature of asthma as a condition with varying severity which overlapped with COPD. By computationally uncovering the popula-

tion structure in relation to asthma and COPD related variables, LCA identified the likely patient groups, some of which could be otherwise overlooked in the manual researcher-led development of case definitions.

Asthma and COPD were included in the same model

The inclusion of both asthma and COPD data in the same model is a particular strength. It allowed the exploration of the overlap between the two diseases in EHR-derived data. It makes the model useful for researchers who, for example, want to study a subset of asthma or COPD patients who also have characteristics of the other disease (i.e. ACOS), or those who have one disease without the other.

Derivation of a transferable classification algorithm

A remarkable strength in this chapter is the derivation of a classification algorithm based on the best-fit LCA model. Since the LCA model was performed on a sample of 50,000 individuals and not on the entire population in the SAIL Databank, its output cannot be directly used to identify asthma patients from outside this sample. To overcome this limitation, I derived a classification algorithm that can be used on different samples and by other researchers in the SAIL Databank. This transferable classification algorithm can be also tested in other similar primary care databases. This algorithm could be used to produce more accurate estimates of the disease prevalence compared to methods based on patient-reported data from national health surveys and those based on the Quality of Outcomes Framework's AST001 indicator.

3.5.2.2 Limitations

The analysis in this chapter has some limitations.

Limitations related to EHR-derived data

The latent class modelling described in this chapter was based on relatively limited data that were usually insufficient to establish diagnosis at the point of care. Much of the information that were usually available to GPs to establish asthma and COPD diagnoses were not available in the primary care dataset of the SAIL Databank. This was mostly due to poor recording and/or coding at the point of

care. Those low-quality data included, for example, lung function tests. Since the quality of observed variables was essential for a well-specified latent class model, I constructed these observed variables only using events that were thought to be of reasonable quality in the SAIL Databank. Such events included diagnosis, disease-related GP visits, prescriptions, and smoking history.

Despite using those observed variables which were known to be well coded in the GP dataset in the SAIL Databank, the imperfect quality of these variables could have still affected the latent class modelling. In interpreting the chosen model, I took into account the nature and limitations of EHR-derived data such as the possibility of missing or incorrect coding as well as record-linkage errors. For example, Class 5 (Figure 3.6) was characterised by current prescriptions suggesting active asthma but with no recorded diagnosis events; I assumed that patients in this class were possibly being treated for asthma as they had a low probability of ever smoking.

The clinical meanings of the latent classes were based on surrogate variables, such as GP diagnosis codes, visits, and prescriptions as well as smoking history, rather than on more direct disease markers such as clinical and laboratory findings. Nevertheless, I hypothesised that LCA of these surrogate variables can reasonably distinguish between patients with asthma, COPD, and ACOS. This provided an opportunity to assess how clustering based on these surrogate variables will perform compared with that based on asthma and COPD biomarkers [7, 273-279].

Asthma-COPD overlap syndrome was treated as a separate class

In the merging of the latent classes and the derivation of the classification algorithm, I treated ACOS as a separate label rather than merging it with asthma and/or COPD groups. This approach was in-line with the view that ACOS is a third condition, a view which was, however, subject to active debate [210, 211, 280].

Limitations of latent class analysis

Although latent class analysis was an appropriate clustering method that fit with asthma heterogeneity, it had some particular limitations.

The specification of the observed variables, model selection and interpretation all involved significant levels of subjectivity. Model interpretation and usefulness

both depend largely on the choice and configuration of the observed variables, which thus needs careful consideration. Therefore, assessing the derived classification algorithm against other GP- and self-reported measures was needed and provided useful information to assess its meaning.

The population structure identified by LCA may not exactly represent the clinical epidemiology of asthma and COPD. It has been shown that the patient groups identified using such unsupervised statistical learning techniques may partly reflect artefact from the analysis method including transformation and encoding of the observed variables [281]. Therefore, an important future work is to compare this LCA model performance with full patient record data and/or clinical assessment for a population sample.

The classification algorithm was not a superior reference

Given limitations related to data quality and provenance as well as LCA, the classification algorithm derived in this chapter was not intended to be a superior reference against which other asthma case definitions, could be assessed. Arguably, no easily implementable, gold standard operational definition for asthma exists.

3.5.3 Comparison with related works

LCA has been widely used on asthma-related data. While some studies used LCA to mainly identify asthma cases [282], the more common use was to uncover phenotypes of asthma and related wheezing and atopic disorders [274, 277, 283].

The study by Prosser et al. [282] was closely related to the aim I pursued in this chapter. However, the main aim of that study was to identify only patients with treated asthma. It used slightly different configuration of the observed variables derived from health insurance claims and hospital discharge data. The model diagnostics favoured the two-class model, which estimated the prevalence of treated asthma in British Columbia in 2001 as 9.9%. In my analysis, however, I aimed to identify treated and untreated cases of asthma, which had the prevalences of 8.0% and 6.6%, respectively, and a combined prevalence of 14.5%. In addition, the authors of that study did not take into consideration COPD-related data in their model specification. Conversely, I included events related to COPD as observed variables in order to allow the model to distinguish between asthma and COPD

patients (and to identify those with ACOS). I also used the age at asthma diagnosis (before vs. after the age of 40) in order to improve distinction between asthma and COPD patients and to improve the overall interpretation of the model.

Another LCA-based study was based on questionnaire data about respiratory symptoms of 4,000 children aged 8-12 years [277]. The authors used 11 questions from the International Study of Asthma and Allergies in Childhood (ISAAC) questionnaire. They assessed their LCA model using objective asthma markers such as allergic sensitisation and bronchial hyper-responsiveness. They identified seven latent classes labelled as ‘no respiratory symptoms’ (with prevalence of 59.4%), ‘cough during colds’ (19.1%), ‘chronic cough and phlegm’ (5.3%), ‘nocturnal breathlessness’ (4.9%), ‘wheeze only with colds’ (4.8%), ‘wheeze without colds, with cough’ (4.5%), and ‘wheeze without colds, without cough’ (2.1%). These classes were overall different from those I identified in this chapter. The authors reported that asthma diagnosis was highly reported by the parents in the ‘wheeze’ and ‘nocturnal breathlessness’ classes, leading to an 8.5% prevalence for parent-reported doctor diagnosis of asthma. In my LCA model, however, the classes consistent with ‘asthma’ (including every diagnosed asthma and currently treated asthma) had an aggregated prevalence of 14.5%. In my modelling, I did not use objective disease markers to specify or validate the model since these data were under-recorded in the SAIL’s GP dataset (see [Chapter 4](#)). Finally, that study was performed on children only, whereas my latent class modelling was based on data from all age groups. For future work, however, it would be worthwhile to perform a separate latent class modelling for each age group, or using age group as a covariate for the model in order to control its effect on the identified latent structure.

3.5.4 Future directions

Future developments of the latent class model described in this chapter include refinement of the observed variables and exploring new predictor variables from primary and secondary care data.

In addition, Wales-wide pathology data are expected to be linked to the SAIL Data-bank in 2018. These data include important asthma-related data such as peripheral eosinophil count and immunoglobulin E (IgE) levels. Such data linkage would provide an opportunity to explore asthma phenotypes in Wales in a greater depth.

Patients with asthma exhibit different profiles in terms of the disease natural history and progression. These temporal profiles may be related to clinically recognised phenotypes and underlying endotypes. A longitudinal extension of LCA, latent transition analysis (LTA), would allow modelling the temporal profiles of asthma natural history in Wales. That extended analysis would provide better understanding of the disease's changing epidemiology, and help inform service planning, resource allocation, and support more personalised disease management.

3.6 Conclusion

Accurate case definitions are critical to the development of the Wales Asthma Observatory.⁴ However, due to various sources of uncertainty in asthma-related routine data, clear identification of asthma patients using these data is challenging.

In the absence of a reliable reference standard, I used LCA of recorded primary care events in the SAIL Databank to identify clusters of likely patients with asthma and/or COPD. The model diagnostics and interpretability favoured the eight-class model which included four classes for asthma (differing by recorded asthma diagnosis and prescriptions), one for COPD, one for asthma-COPD overlap syndrome, one with scarce prescriptions probabilities, and one with no asthma or COPD related events.

Based on the latent class model, and after merging three classes of currently treated asthma, I derived a classification algorithm which could be used to classify new samples into six clinical labels: ever diagnosed asthma without treatment, currently treated asthma, COPD, ACOS, possible/at risk of COPD, and none. I assessed the classification algorithm against other objective and self-reported case definitions. The classification algorithm can be also used or tested by other researchers in similar primary care data sources.

The unsupervised machine learning approach used in this chapter relied on the assumption that despite the challenges to define asthma from RCD, these data reflected the disease clinical epidemiology. By computationally uncovering the population structure, LCA identifies all the likely patient groups that could be overlooked in the manual researcher-led development of case definitions. There-

⁴<http://www.wales-asthma-observatory.uk/>

fore, the developed LCA model could produce more reliable estimates for asthma prevalence using RCD.

Specifying the LCA model using asthma and COPD observed variables allowed identifying patients with one or both disease. This approach concurs with the current interest to understand the asthma-COPD overlap and allows defining this controversial clinical entity using RCD.

The LCA-based method to identify asthma patients from RCD that I developed in this chapter will be one of a set of asthma case definitions available in the Observatory as described in [Chapter 4](#).

Chapter 4

Development of the Wales Asthma Observatory

Purpose, design, and data quality

A main output of this doctoral project is to establish the foundations of the Wales Asthma Observatory based on a national asthma registry using the Secure Anonymous Information Linkage (SAIL) Databank. In this chapter, I discuss the purpose of the Wales Asthma Observatory as a platform for asthma research and surveillance, and its wider context in the United Kingdom and worldwide.

I describe the design of the Observatory and the underlying asthma registry, including source population, structure, content, technical logistics, and approaches to support reproducibility. I then identify important data gaps in asthma related events in the primary care database. Afterwards, I discuss the strengths and limitations of the Observatory, including the wide coverage, implications of data gaps and anonymisation, and specific challenges of assessing asthma outcomes using the SAIL Databank. I then present recommendations for better capture of asthma routine data in the SAIL Databank. I conclude by proposing further developments to the Observatory.

Chapter Contents

4.1	Introduction	109
4.2	Purpose and context	109
4.3	Methods	110
4.3.1	Ethics and information governance	112
4.3.2	Logistics and technical environment: the SAIL Databank	113
4.3.3	Data used in the Observatory	113
4.3.3.1	Data sources	113
4.3.3.2	Data anonymisation and linkage	118
4.3.4	Eligibility criteria	119
4.3.5	Registry structure and variables	120
4.3.6	Dealing with missing and invalid data	122
4.3.7	Updating the Observatory data	122
4.3.8	Support for reproducible research	123
4.3.9	Dissemination of the Observatory output	124
4.3.10	Data sharing and access to the Observatory	125
4.4	Summary statistics	125
4.5	Quality of asthma-related events in the GP database	130
4.5.1	Background	130
4.5.2	Methods	131
4.5.3	Results	132
4.5.3.1	Recording of event groups	132
4.5.3.2	Quality of values of lung function events	132
4.5.4	Interpretation	134
4.6	Discussion	135
4.6.1	Summary of the Observatory design and data quality	135
4.6.2	Strengths and opportunities	136
4.6.3	Challenges and limitations	136
4.6.3.1	Primary care-based case definition of asthma may exclude some patients	136
4.6.3.2	Inherent limitations of routine data	137
4.6.3.3	Traditional methods to define cases and outcomes may need reconsideration	137
4.6.3.4	Implications of the suboptimal quality of asthma-related routinely collected data on asthma research	138
4.6.3.5	Primary care coding in the United Kingdom (UK) will change	139
4.6.3.6	Data security and implications of anonymisation	139
4.6.4	Recommendations for better capture of asthma data	140
4.6.5	Future development	141

4.1 Introduction

A main output of this doctoral project was to establish the foundations of the Wales Asthma Observatory (<http://www.wales-asthma-observatory.uk/>) as a platform for asthma research and surveillance. In this chapter, I describe the Observatory's purpose, context, design and methodology. In addition, due to the known data quality issues in electronic health record (EHR) data [284, 285], I present quality assessment of selected asthma-related variables in routinely collected, primary care data in Wales. I then discuss the strengths and limitations of the Observatory as well as challenges related to asthma and routinely collected data (RCD). I also propose further developments to the Observatory, and suggest recommendations to improve the quality of asthma RCD.

4.2 Purpose and context

The Observatory is intended to be used as a research platform for supporting a wide range of cross-sectional and longitudinal asthma studies. It can be used as a surveillance tool, producing disease insights at both local and national levels in Wales. The Observatory includes a regularly updated, cumulative e-cohort (described in the next section), which enables near real-time disease monitoring, tracking, and forecasting. The linkage to area-based deprivation indices enables investigations into the inequalities in asthma care across Wales, as demonstrated in [Chapter 5](#).

The Wales Asthma Observatory is closely aligned with the UK Asthma Observatory (UKAO), a UK-wide platform led by the Asthma UK Centre for Applied Research (AUKCAR) [286]. Data from the four UK countries feed into the UKAO [287]. These currently include person-level and aggregate data about primary and secondary care, community medication dispensing, ambulance services, and national health surveys [43].

In the wider context, the Wales Asthma Observatory will support the AUKCAR's research endeavours towards promoting better asthma control, maximising treatment benefits, and reducing exacerbations and adverse outcomes.

The Observatory can play a vital role in the endeavours to bridge the gap between evidence and practice. The concept of *learning health systems (LHS)* has been developed over the past decade to address the crucial need to improve the re-use of health data to make continuous improvements of health care delivery [288, 289]. An LHS is a healthcare system that continuously ‘learns’ from care delivery. It requires an integrated, seamless cyclical process of collecting data generated as a by-product of care delivery, using these data to create new knowledge, and use the created knowledge to inform decision making and performance of everyday care delivery [290-293]. With plans for piloting an LHS for asthma in Wales, the Wales Asthma Observatory can be a building block in the foundations of such a vital system.

4.3 Methods

The Observatory includes a national asthma registry in Wales. I built the registry from routinely collected de-identified health data in the Secure Anonymised Information Linkage (SAIL) Databank. This registry represents a cumulative, nationwide cohort of asthma patients based on the dynamic population of Wales. It is intended to include all previous (remitted or deceased) cases of asthma in the country as well as existing and new incident cases. The cohort was assembled using variety of case definitions, mostly based on the systematic scoping review in Chapter 2, and included essential asthma outcomes such as disease severity, exacerbations, and death due to asthma as well as asthma remission.

Figure 4.1 illustrates the general process of compiling the Observatory.

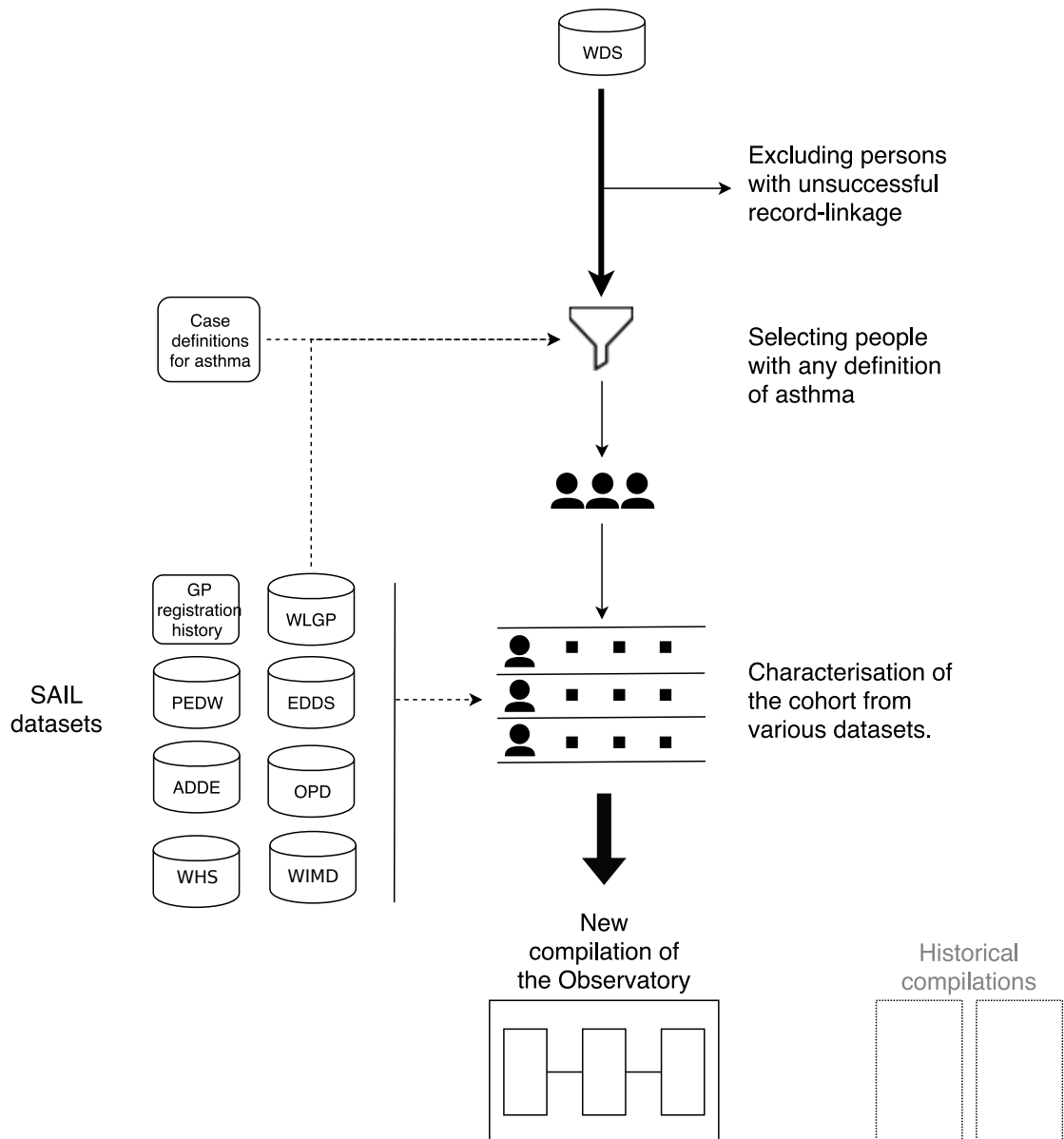


Figure 4.1: Compilation of the Wales Asthma Observatory. WDS: Welsh Demographic Service; OPD: Outpatient Dataset; WLGP: Welsh Longitudinal General Practice; PEDW: Patient Episode Database for Wales; EDDS: Emergency Department Data Set; ADDE: Annual District Death Extract.

In this section, I describe:

- the ethical and information governance requirements of the Observatory development;
- the technical environment in which the Observatory was developed;
- data sources used, including content, coverage, and data quality indicators (in addition, I consider in-depth the quality of recording of selected asthma-specific Read codes in [Section 4.5](#));
- the source population;

- case definitions used to identify people with asthma;
- the Observatory's data structure and variables;
- output dissemination plan;
- approaches to support research reproducibility; and
- data sharing and requirements to access the Observatory.

4.3.1 Ethics and information governance

The development of the Wales Asthma Observatory was approved by the SAIL Information Governance Review Panel (IGRP), which acts as a data guardian (see [Appendix C.2](#) for the approval letter).

Typically, an application to access SAIL data in a research project starts with a scoping discussion in which a SAIL analyst discusses with the applicants the suitability of their research projects and the required datasets. In the second stage, the applicants are required to complete a detailed application including a research proposal (objectives, methods, data required from SAIL datasets, and time periods and geographical and demographical distributions of data). The application is submitted to the IGRP which assesses the project's suitability and compliance with SAIL's information governance policy, and may raise issues and questions based on the application. After these questions are resolved and the IGRP is satisfied with the project, it sends a letter of approval to the applicants. Following approval, and before the applicants can start accessing SAIL data, they need to demonstrate appropriate information governance knowledge and skills by undertaking an accepted training course.¹ In addition, users are required to sign the SAIL's Data Access Agreement which details operational and information governance rules. Detailed information about the application process can be found on the SAIL Databank website.²

An approval from a research ethics committee was not required for the Observatory development since it only used de-identified data, which was consistent with the current National Health Services (NHS) Health Research Authority guidance [294].

¹Examples of accepted training course are the "Research, GDPR (General Data Protection Regulation) and confidentiality Quiz", which is run by the Medical Research Council, and the "Safe User of Research data Environments (SURE) Training" course, which is run by the Office for National Statistics (ONS), the UK Data Service, and the Administrative Data Research Network (ADRN). More details are regularly published on the SAIL Databank website: <https://saildatabank.com/application-process/following-approval/>

²<https://saildatabank.com/application-process/>

4.3.2 Logistics and technical environment: the SAIL Databank

I developed the Observatory using data from the SAIL Databank.³ The SAIL Databank is a repository of de-identified, linked health datasets in Wales. Data providers include general practices, hospitals, and the Office for National Statistics (ONS).

The Observatory was developed within the SAIL Gateway. The SAIL Gateway is a privacy-protecting safe haven for anonymised person-level data [295]. Approved users can access data through a secure remote access system.

Data in the SAIL Databank are organised in *database schemas*, which are logical structures that group database elements such as tables, views, and procedures. I currently maintain the Observatory data in a dedicated SAIL schema (SAILW0317V), the content of which is described below in Section 4.3.5. These data can be interrogated using the Structured Query Language (SQL). I used the SQL and R programming languages to extract data from SAIL and to organise the Observatory schema.

4.3.3 Data used in the Observatory

4.3.3.1 Data sources

I used the following SAIL datasets in the development of the Wales Asthma Observatory. Table 4.1 summarises the coverage and content of these datasets, while Table C.1.1 (in the Appendix) shows their data fields and Table C.1.2 and Figure C.1.1 show the frequency of events and unique patients in each dataset by calendar year. Each of these datasets cover all of Wales, except the Welsh Longitudinal General Practice (WLGP) dataset which currently covers about 80% of general practices. New extracts of these datasets are regularly received by the SAIL Databank every few months to over a year.

Welsh Demographic Service dataset

The Welsh Demographic Service (WDS) contains de-identified demographic and administrative data about people who use the NHS in Wales. These data include

³<https://saildatabank.com/>

Table 4.1: Datasets used for the development of the Wales Asthma Observatory. Details on data fields, data missingness, and record counts can be found in Table C.1.1 and Table C.1.2.

Name	Cumulative population (million)*	Time period	Coverage	Content	Data collection method	Record linkage matching percentage†
Welsh Demographic Service dataset	~5.3	1992 - current	All Wales	Demographics, encrypted addresses, and general practices registration history of NHS-registered people	Data is collected by GPs when patients register at their practices	100%
Welsh Longitudinal General Practice dataset	~4.0	1993 - current	~80% of general practices in Wales	Primary care events, such as diagnoses, clinical and laboratory findings, prescriptions, and follow-ups	Data is collected by GPs during routine practices and then are coded in Read codes	99.58%
Emergency Department Data Set	~2.5	2009 - current	All Wales	Attendances to A&E departments at NHS hospitals in Wales	Data is routinely collected and coded using a limited alphanumeric coding system	96.8%
Patient Episode Database for Wales	~3.3	1991 - current	All Wales	Admissions to NHS hospitals in Wales	Data is routinely collected	94.7%
Outpatient Dataset	~3.2	2004 - current	All Wales	Outpatient attendances to NHS hospitals in Wales	Data is routinely collected and coded by outpatient departments	99.1%
Annual District Death Extract	~7.0	1996 - current	All Wales	Up to eight causes of death, coded in ICD-10	Data automatically or manually coded from Medical Certificates of Cause of Death (MCCD) certified by a medical practitioner or a coroner.‡	94.9%

* Based on the latest extract of the dataset.

† Record linkage percentage represent the percentage of successful linkage of a dataset to the Welsh Demographic Service dataset.

‡ Source: *Mortality Statistics: Metdata*, July 2015, Office for National Statistics (see https://www.ons.gov.uk/file?uri=/aboutus/transparencyandgovernance/freedomofinformationfoi/howyouobtaindetailsofthenumberofpeopleofficersuicides/mortality/metdata2014_tcm77-241077.pdf).

gender, week of birth (defined by date of the first Monday after birth), date of death (for dead people), registration history at general practices, and Lower Layer Super Output Areas (LSOAs) of residence as reported by the individual upon registration. The WDS data has been recorded since 1992 and covers all of Wales. The most recent extract of the WDS in SAIL was created in April 2018 and included data of a cumulative population of ~5.3 million people.

The NHS is a free to use service. However, the WDS normally does not capture persons who do not use the NHS. Those may include healthy people (particularly young men for whom there is no health screening), disengaged people who do not use the NHS unless in emergencies, some prisoners, and those who used private GPs [296].

Welsh Longitudinal General Practice dataset

The Welsh Longitudinal General Practice (WLGP) Dataset contains de-identified health care events, such as recorded diagnoses, clinical findings, prescriptions and monitoring as well as other events. Data are collected by GPs during patient visits and are then coded into Read codes by GPs or clinical coders. The most recent extract of the GP dataset was created in April 2018, covering about 80% of GP surgeries in Wales, and including data for a cumulative population of about four million people. This dataset is of paramount importance for the Observatory since in the UK asthma is mainly treated in primary care [32].

The dataset has 99.58% matching with the WDS. The WLGP dataset shows increased recording of primary care events since the introduction of the Quality of Outcomes Framework (QOF) in 2004-2005 (see [Table C.1.2](#)). However, the lack of standardised coding practices leads to variations and inconsistencies in data recording. In [Section 4.5](#), I consider the quality of recording of selected asthma-specific Read codes.

Emergency Department Dataset for Wales

The Emergency Department Data Set (EDDS) was created in 2009 and captures visits to accident & emergency (A&E) departments as well as minor injury units (MIUs) in NHS hospitals across Wales. The most recent extract of the EDDS was created in April 2018 and included data on a cumulative population of about 2.5

million people. Data collected during each A&E attendance include investigations performed, diagnosis made, anatomical areas involved, treatment provided, as well as other administrative data related to the attendance. Diagnosis is coded using a three-digit alphanumeric code chosen from a list of 83 possible codes representing broad diagnostic categories. In addition, to the primary diagnosis, there are further five positions to record additional or secondary diagnoses. Due to the nature of emergency attendances, recorded diagnoses may be uncertain or unconfirmed. Practices of recording and coding of data vary between the different A&E departments and MIUs. The EDDS currently receives data on all emergency attendances in Wales. However, in the earlier years (2008-2011), some A&E departments were not able to submit their data to the EDDS, and therefore data in that period may be incomplete (see [Table C.1.2](#)). Therefore, caution should be exercised when using this dataset for epidemiological and research analyses.

Patient Episode Database for Wales

The Patient Episode Database for Wales (PEDW) dataset was created in 1991 and includes records for all planned and emergency inpatient admissions in addition to day case admissions to all NHS Wales hospitals as well as most admissions of Welsh residents to hospitals in England. The most recent extract of the PEDW dataset was created in May 2018 and included data on a cumulative population of about 3.3 million people. Recorded data are captured during the inpatient episode and includes admission diagnosis, procedures and operations performed, as well as length of stay (LOS), Healthcare Resource Group (HRG), and other administrative data. Admission diagnosis is recorded using ICD-10. In addition to a mandatory primary diagnosis code for a hospital episode, the database allows recording of 1-13 secondary diagnosis codes. The PEDW is generally considered to be high quality [297, 298]. However, it is mainly an administrative database which was created as a tool to track hospital financial activity rather than for epidemiological and research purposes. The database also suffers from between-hospital variations in practices of coding admission diagnosis in the available 14 diagnosis positions [297]. Further discussion about the quality of the PEDW is in [Section 5.5.4.1](#).

Outpatient Dataset

The Outpatient Dataset (OPD) includes data on outpatient appointments in all NHS hospitals in Wales. These data include attendance date, specialities of care and the treating physician, and site of treatment. In addition, fields for diagnosis and procedures performed in outpatient settings are available, but data on these items are poorly recorded. The most recent extract of the OPD was created in June 2018, including data for about 3.2 million people since 2004 across all of Wales.

Annual District Death Extract

The Annual District Death Extract (ADDE) dataset is produced and maintained by the ONS and is linked to the SAIL Databank. This dataset contains mortality data since 1996 including up to eight causes of death from Medical Certificates of Cause of Death (MCCD) certified by a medical practitioner or a coroner.⁴ Causes of death are automatically or manually coded in ICD-10 from the MCCD. I used it to ascertain whether death was recorded as due to or related to asthma in deceased asthma patients.

Welsh Health Survey

The Welsh Health Survey (WHS) has been conducted in 1995, 1998, and annually since 2003 before it ceased in 2015 [53]. It collected self-reported information on a range of health and health-related lifestyles from samples of the population of Wales. The WHS 2013 and 2014 results datasets for respondents aged 16-year-old and above are already linked to the SAIL Databank [271]. Those participants consented to link their responses to their medical records [272]. The only question related to asthma, other than asthma symptoms, asked the participant whether he or she was currently treated for a number of diseases including asthma. I used responses to this question as a case definition for ‘self-reported currently treated asthma’. I considered invalid responses as negative responses.

⁴Mortality Statistics: Metadata, July 2015, Office for National Statistics ([link](#)).

Welsh Index of Multiple Deprivation

The Welsh Index of Multiple Deprivation (WIMD) is the official tool to assess the level of multiple deprivation in small areas in Wales. It consists of the following weighted eight domains: Income, Employment, Health, Education, Geographical Access to Services, Housing, Physical Environment, and Community Safety. I discuss the WIMD index in depth in [Section 5.1.3.2](#).

4.3.3.2 Data anonymisation and linkage

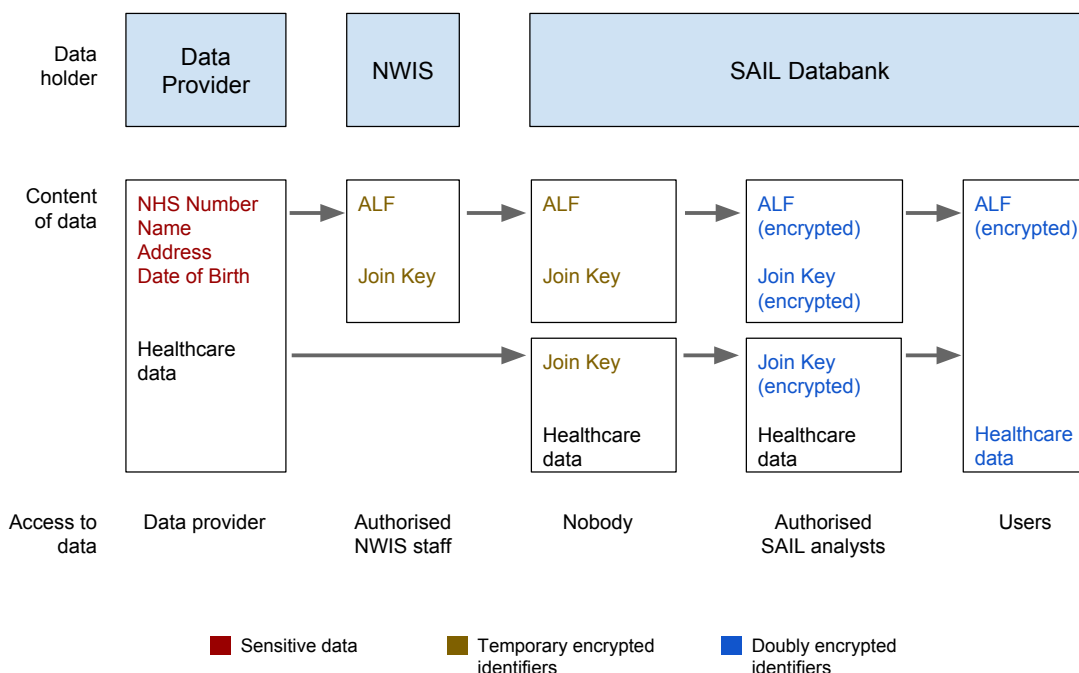


Figure 4.2: Split-file approach to data anonymisation by a trusted third party (adapted from Ford et al. [141]). ALF: Anonymised Linking Field; NWIS: National Health Services Wales Informatics Service; SAIL: Secure Anonymised Information Linkage.

Data anonymisation and linkage on the aforementioned data sources is performed by the National Health Services Wales Informatics Service (NWIS), which acts as a trusted third party (see [Figure 4.2](#)) [140, 141]. A data provider splits its data in two files: File 1 which contains demographic data, and File 2 which contains clinical data. The data provider assigns a join key for these two files. File 1 is securely transferred to NWIS. Then, NWIS replaces the demographic data with an Anonymised Linking Field (ALF), which is designed to be a unique identifier across different data providers. NWIS then creates File 3 which contains the ALF as well as the data provider’s join key. Then, it sends this File 3 to the SAIL Databank. Separately, the data provider sends File 2 to the SAIL Databank. In the

Table 4.2: Case definitions used in the Wales Asthma Observatory. Clinical codes used in the GP-data-based case definitions are listed in [Appendix E](#)

Case definition	Description
Ever-diagnosed asthma	One or more asthma diagnosis codes any time before a given date.
Ever-diagnosed and currently treated asthma	One or more asthma diagnosis codes any time before a given date, and one or more asthma prescription codes in the last 12 months; this case definition corresponded to the Quality of Outcomes Framework indicator AST001 without considering exceptions.
Currently treated asthma	One or more asthma prescription codes in the last 12 months.
Ever-treated asthma	One or more asthma prescription codes any time before a given date.
Self-reported currently treated asthma	Based on the Welsh Health Surveys; only available for a small number of patients. The survey question on whether the participant was ‘currently being treated for asthma’ did not specify a time frame. However, based on my own analysis (see Appendix B.1) I used a period of 26 months ending with the end of the survey year.

SAIL Databank, File 2 and File 3 are re-joined using the join key to produce a de-identified dataset, containing encrypted ALF and clinical data, which can be linked to datasets from other data providers.

4.3.4 Eligibility criteria

The Observatory aims to cover the entire dynamic population of Wales. Therefore, I defined the source population as all individuals for whom records exist in the WDS dataset.

Since the Observatory aims to include all potential and confirmed asthma patients in Wales, I included in the Observatory people who met any of a set of case definitions for asthma, from the most inclusive to the most specific ones. [Table 4.2](#) lists case definitions currently included in the Observatory. Each person in the Observatory satisfies at least one case definition ever.

The use of these multiple case definitions of asthma allows capturing most patients, ranging from those with uncertain diagnosis to those with the more strictly defined currently-treated asthma. This approach facilitates studying diverse subgroups of asthma patients differing in diagnosis certainty and disease activity. It also provides flexibility for researchers to choose the appropriate case definitions for their studies. The use of broad case definitions to capture patients with any indication of asthma was previously adopted by a similar project in the United States (US), the Population-Based Effectiveness in Asthma and Lung Diseases (PEAL) Network [[137](#)].

Each of the used case definition has its own meaning and uses in research. Nonetheless, I considered the 'ever GP-diagnosed currently-treated asthma' the main case definition of asthma in the Observatory as it reflects the active disease, which is often an essential criterion in many asthma studies.

The case definitions in Table 4.2 can be represented as state variables, i.e. assessed over time periods. I assessed the case definitions over patient-specific periods (Figure 4.3) rather than fixed periods (e.g., start and end of year) for all patients. This allowed an accurate start and end of disease states. For example, the definition of "currently treated asthma" ("there was at least one asthma prescription in the last 12 months") is true for any given date within the state period (see Figure 4.3).

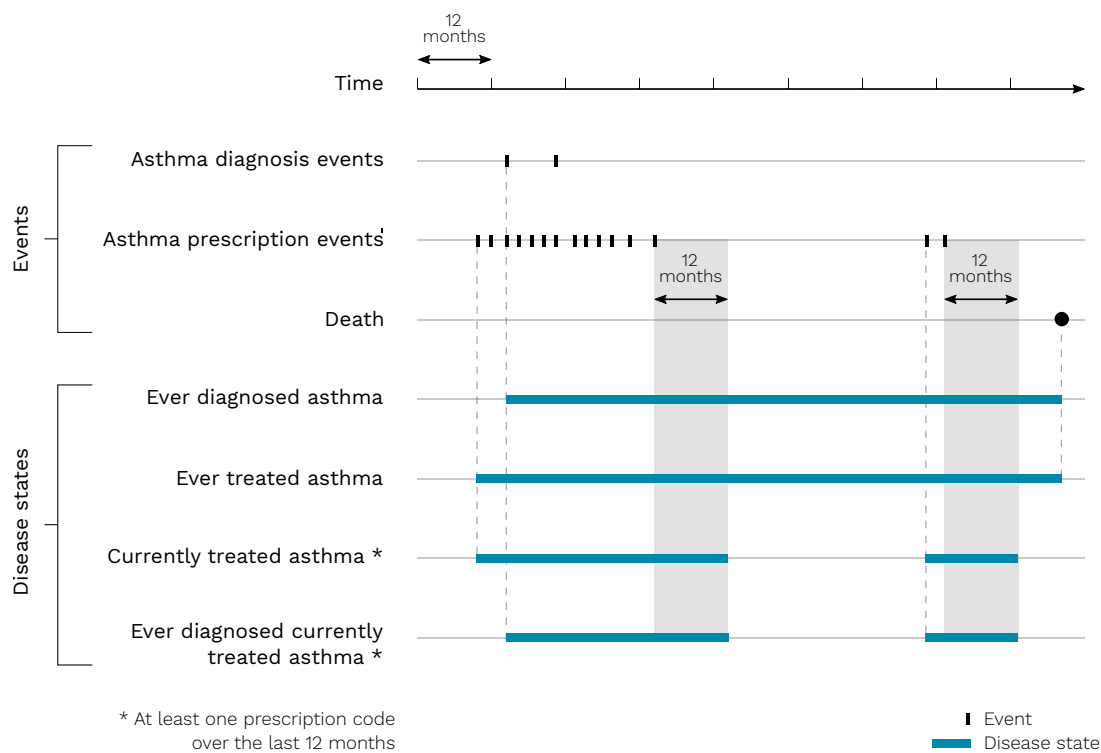


Figure 4.3: Examples of the assessment of case definitions as disease states (state variables) in the Wales Asthma Observatory. The diagram shows asthma diagnosis and prescription events for an imaginary patient in addition to four case definitions assessed over patient-specific time intervals based on those events.

4.3.5 Registry structure and variables

Data for each person in the Observatory currently include demographics, which asthma case definitions ("asthma states") were satisfied over which periods, as well as asthma-related outcomes and variables. These outcomes and variables

include asthma treatment step, disease severity, and exacerbations, and periods of follow-up based on GP registration history (see Table 4.3). Metadata about the Observatory compilations and dataset versions used in each compilation are stored in a separate database table (WAO_data_sources_versions).

Table 4.3: Data tables in the asthma registry

Name	Description	Fields
Demographics	Basic demographic data	ALF WOB DOD
Patient follow-up	Periods of follow-up for patients based on GP registrations from the WDS dataset. Those GP registration records were filtered using an in-house algorithm (FNC.CLEAN_GP_REG, developed by the SAIL Analytical Services Team) that excludes periods over which some GP practices, despite being participating in SAIL, did not send reasonably adequate amounts of data to SAIL.	ALF start date end date
Asthma State	Whether an individual satisfied a case definition of asthma over a specific period or a calendar year. Examples of asthma case definitions include “ever diagnosed asthma” and “currently treated asthma” (see Table 4.2).	ALF case definition/state start date end date
Treatment Step	Treatment step from 1 to 5 based on the 2016 BTS/SIGN asthma guidelines, in addition to ‘SABA as needed’, over a period of up to six months.	ALF start date start end treatment step
Asthma Severity	Disease severity classified as intermittent, mild, moderate, or severe based on prescriptions [299] over a period of up to six months. <i>Intermittent</i> = SABA as needed; <i>mild</i> : low-dose ICS or other low-intensity therapies; <i>moderate</i> : low/moderate-dose ICS with LABA (or other additional therapies); <i>severe</i> : high-intensity therapies (high-dose ICS with LABA, oral corticosteroids, or other additional therapies).	ALF start date end date asthma severity
Asthma Exacerbation	Records for asthma exacerbation defined based on primary and secondary care utilisation. An asthma exacerbation is defined by short course of oral corticosteroids, asthma-related emergency admission, or asthma-related hospitalisation, with periods less than 4 weeks apart being merged into single exacerbation period.	ALF start date end date
Asthma-related death	Record of death in which asthma was an underlying cause of death from the ONS’s ADDE dataset.	ALF DOD position of asthma code in death record
Data source versions	Shows the available compilations of the Observatory and the names and versions/extracts of SAIL datasets used in each compilation (table name: WAO_data_sources_versions).	Observatory compilation number and date (e.g., WAO_2_20181005) Source data table name and version (e.g., SAILWLGPV.ALF_GP_EVENTS_CLEANSED_20180820)

ALF: Anonymous Linking Field; BTS: British Thoracic Society; DOD: date of death; GP: general practitioner; LABA: long-acting beta adrenoceptor agonist; SABA: short acting beta agonist; SIGN: Scottish Intercollegiate Guidelines Network; WOB: week of birth.

Additional data that could be added in the future include:

- Timeline of laboratory test results such as total immunoglobulin E (IgE), blood eosinophil count, as well as lung function measurements.
- Asthma phenotypes (e.g., eosinophilic asthma, adult-onset asthma, asthma with fixed airflow limitation, and poorly steroid-responsive asthma); pheno-

types can be identified using clustering methods guided by the relevant literature [4, 5, 7].

- Environmental data including air pollution, housing quality, calculated for small areas and linked through the Residential Anonymous Linking Fields (RALF) [300].
- Patient reported outcome measures (PROMs), such as Asthma Control Questionnaire (ACQ) responses.

4.3.6 Dealing with missing and invalid data

In the Observatory compilation script, I excluded persons with an invalid ALF (i.e. had a value of NULL). In addition, the Observatory includes periods of follow-up based on GP registration history, which can be used for censoring in time-to-event analyses.

Non-existence of a health event in an event-based dataset, such as most SAIL datasets, does not imply non-occurrence of that event; it may rather due to non-recording of such an event in categorical codes and may have been recorded in narrative fields not available within SAIL. The nature of such event-based datasets means that it was impossible to identify such unrecorded events. However, in the Observatory development, the case definitions were based on events that are assumed to be well recorded.

4.3.7 Updating the Observatory data

The Observatory data are based on the SAIL Databank. Data in the SAIL Databank are not collected in real-time but are rather collected and updated with a variable lag time ranging from few months to over a year. Subsequently, the Observatory data are not real-time, but are intended to be updated following updates to any of the source SAIL datasets.

The updating process can be performed using the same data extraction and programming script used in the initial compilation of the dynamic cohort. For each update, names of newer dataset extracts should be used as input in that script. This process will create a new version of the Observatory, including an updated patient cohort and variables (see [Figure 4.1](#)).

4.3.8 Support for reproducible research

Reproducibility is important for epidemiological studies [157, 218]. It requires full and clear documentation of the methods used, including algorithms to define health variables and extract data as well as programming code used for analysis.

For studies using routine data, certain considerations are needed to address reproducibility. In such studies, interpretation of findings largely depends on the clinical codes and data extraction methods used to identify patients and define outcomes. Therefore, these codes should be accessible and reusable by the wider research community in order to support transparency, reproducibility, and comparability of findings [102, 157, 301].

A key part of an EHR-based disease registry is a library of clinical code lists used by studies. In the Wales Asthma Observatory, I developed a technical platform where clinical code lists could be collaboratively maintained, shared, and reused by researchers and analysts (see Figure C.3.1). Interrogation of routine databases often involves repetitive programming tasks, such as manually constructing and modifying complex database queries. These tasks generally require significant time from an experienced data analyst. The above-mentioned platform enables users to collaboratively develop and reuse study-specific data extraction procedures. To minimise the need to write manual and repetitive queries, the platform automates significant parts of data extraction from the Observatory and the GP dataset. It automatically generates and executes the required SQL queries. Automating data extraction is aimed to support scalability, save significant time by analysts, and reduce human error. In addition, the platform has a graphical user interface which allows researchers with no programming skills to develop code to subsequently interrogate the data.

At the time of writing, the platform is maintained inside the SAIL Gateway at the address <http://gpact.chi.swan.ac.uk>. However, requirements are being discussed with the SAIL Databank team to make the query building platform available for the public outside the SAIL Gateway. A similar public repository of clinical codes is being maintained by the University of Manchester [216].⁵ Compared to that repository, the platform that I developed will enable public sharing not only of clinical code lists, but also of data extraction procedures that are used inside

⁵<http://ClinicalCodes.org>

the SAIL Gateway, but not the underlying patient level data. When the platform will be publicly available, each code set and data extraction procedure will have a permanent citable Internet address. This platform is intended to support research transparency and reproducibility as stated in the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) [153] and the REporting of studies Conducted using Observational Routinely collected health Data (RECORD) statements [157].

Another threat to reproducibility is that the data sources used by the asthma registry are regularly updated. A new data extract (i.e. version) usually contains more recent data captured since the previous extract. However, it is possible that a new data extract also includes additional historical data not sent before from data providers to SAIL. Reproducibility of analyses is not guaranteed if repeated using different extracts of data. To minimise this limitation, the registry data tables are versioned based on the updates of the source dataset. This allows epidemiological estimates to be reproduced using the same data used in the previous calculations.

4.3.9 Dissemination of the Observatory output

The Observatory can be used to perform epidemiological analyses on a regular basis or on demand. Examples of these analyses include basic epidemiological parameters for asthma such as disease incidence, life-time and annual disease prevalences, incidence of exacerbations, emergency visits, hospitalisation, as well as disease burden. In addition, prevalence of asthma phenotypes and temporal profiles of disease activity can be explored.

Consumers of the Observatory output are intended to include several user groups such as service planners and managers, policy makers, scientists and academics, health care professionals, asthma patients, and other members of the public. It is therefore important for the published output to consider the needs of this wide spectrum of users. The Observatory output will be published on a dedicated public website using appropriate format and state-of-the-art visualisation techniques. Users will be able to subscribe with newsletters and alerts about output of their interests. Dissemination will also utilise the infrastructure of AUKCAR as well as Asthma UK's dissemination channels, allowing wider reach to people with asthma.

4.3.10 Data sharing and access to the Observatory

Since the Observatory is based on the SAIL Databank, researchers who wish to access the Observatory need to seek approval from the SAIL's IGRP (see 4.3.1). The Observatory can be queried by approved SAIL projects and can be linked to internal SAIL data (e.g., about other health conditions) or external data (e.g., for a bespoke cohort) that are linked into SAIL.

4.4 Summary statistics

This section presents statistics from the Observatory describing the database records of asthma case definitions, incidence and prevalence of selected case definitions, and asthma-related health care utilisation.

[Table 4.4](#) shows the all-time number of records and unique patients for each of the case definitions defined in [Table 4.2](#) based the most recent versions of SAIL datasets. The current version of the Observatory data includes a cumulative cohort of 541,159 patients with ever diagnosed asthma for whom there are 6,456,786.3 years of follow-up data available in the primary care dataset (WLGP, 2018-08-20 extract).

Table 4.4: All-time number of records and unique patients for each of the case definitions. These records belong to all patients in the WDS, WLGP, and WHS datasets, including living and deceased patients.

Case definition	Number of records	Number of unique patients	Patient-years of follow-up in SAIL*
Ever-diagnosed asthma	541,159	541,159	6,456,786.3
Ever-treated asthma	1,174,389	1,174,389	11,531,260.3
Currently treated asthma	2,220,979	1,175,621	5,518,568.2
Ever-diagnosed, currently treated asthma	1,046,819	476,546	3,594,309.5
Self-reported currently treated asthma (based on the WHS)	1,199	1,173	-

GP = general practitioner; WDS = Welsh Demographic Service (WDS); WHS = Welsh Health Survey (WHS); WLGP = Welsh Longitudinal General Practice (WLGP). * Only available for case definitions based on the GP dataset (WLGP).

Using these records, [Table 4.5](#) shows the period prevalences of asthma case definitions in the calendar year 2017 at national and health board levels.

In addition, cumulative incidences and period prevalences of the asthma case definitions between 2000 and 2017 are shown in [Figure 4.4](#). For cumulative incidence of each of the case definitions in each year, the denominator was the number of people with continuous registration at GP practices and complete data in the WLGP (extract 2018-08-20) in the respective year, excluding people who already satisfied the case definition at the beginning of the year. The numerator included people in the denominator who satisfied the case definition during the respective year for the first time in their life. For period prevalence, I defined the denominator was the same used for incidence without excluding people with the condition at the beginning of the year. The numerator was the number of people in the denominator who satisfied the case definition for any period in that year.

Prevalences of lifetime and current asthma showed a steady although slow increase between 2000 and 2017, except for the prevalence of lifetime asthma treatment which showed a steeper increase from 15% to 30%. However, incidences of asthma diagnosis asthma treatment showed an overall decreasing trend, starting in 2000 at 7.4% and 18.8% for diagnosis and treatment, respectively, with a slight

Table 4.5: Period prevalences of asthma case definitions at national and health board levels in Wales in 2017.

Health Board	Population (2017)	Ever-diagnosed asthma		Ever-treated asthma		Currently-treated asthma		Ever-diagnosed, currently-treated asthma	
		Numerator	Prevalence (%)	Numerator	Prevalence (%)	Numerator	Prevalence (%)	Numerator	Prevalence (%)
Abertawe Bro Morgannwg University Health Board	478,673	75,467	15.8	158,250	33.1	47,574	9.9	35,209	7.4
Aneurin Bevan Health Board	378,875	52,978	14.0	119,498	31.5	36,199	9.6	26,341	7.0
Betsi Cadwaladr University Health Board	501,093	70,933	14.2	158,340	31.6	49,168	9.8	35,457	7.1
Cardiff & Vale University Health Board	342,358	49,922	14.6	101,874	29.8	28,821	8.4	22,180	6.5
Cwm Taf Health Board	273,098	38,767	14.2	91,628	33.6	27,536	10.1	19,458	7.1
Hywel Dda Health Board	266,362	38,957	14.6	81,352	30.5	24,834	9.3	18,040	6.8
Powys Teaching Health Board	53,791	7,910	14.7	16,792	31.2	4,903	9.1	3,640	6.8
All Wales	2,353,730	340,338	14.5	739,309	31.4	222,859	9.5	162,958	6.9

increase between up to 2002, before declining significantly between 2003 and 2006-2007 (which might be, in part, due to a change in the recording of asthma diagnosis during that period), followed by stabilisation at 3.1-3.4% and 15.0-16.6% for the incidence of diagnosis and treatment, respectively.

Figure 4.5 shows statistics about asthma-related primary and secondary care utilisation by patients with GP diagnosed asthma patients who received at least one asthma prescription in 2017. The figure shows percentages of those patients who had specific asthma-related events including specific asthma prescriptions, A&E events, and hospitalisations in the same year (2017).

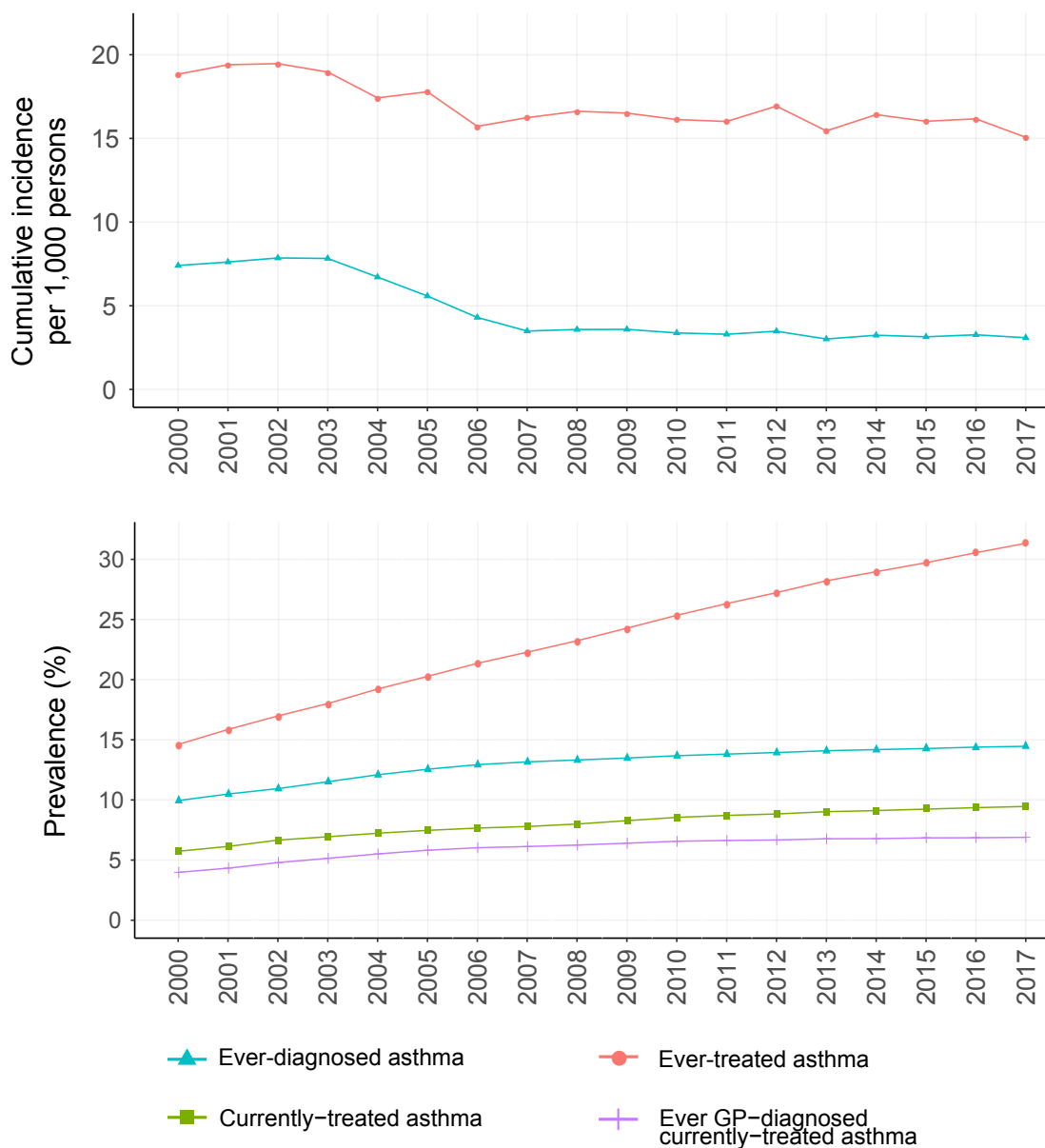


Figure 4.4: Cumulative incidences and period prevalences of asthma case definitions in Wales between 2000 and 2017.

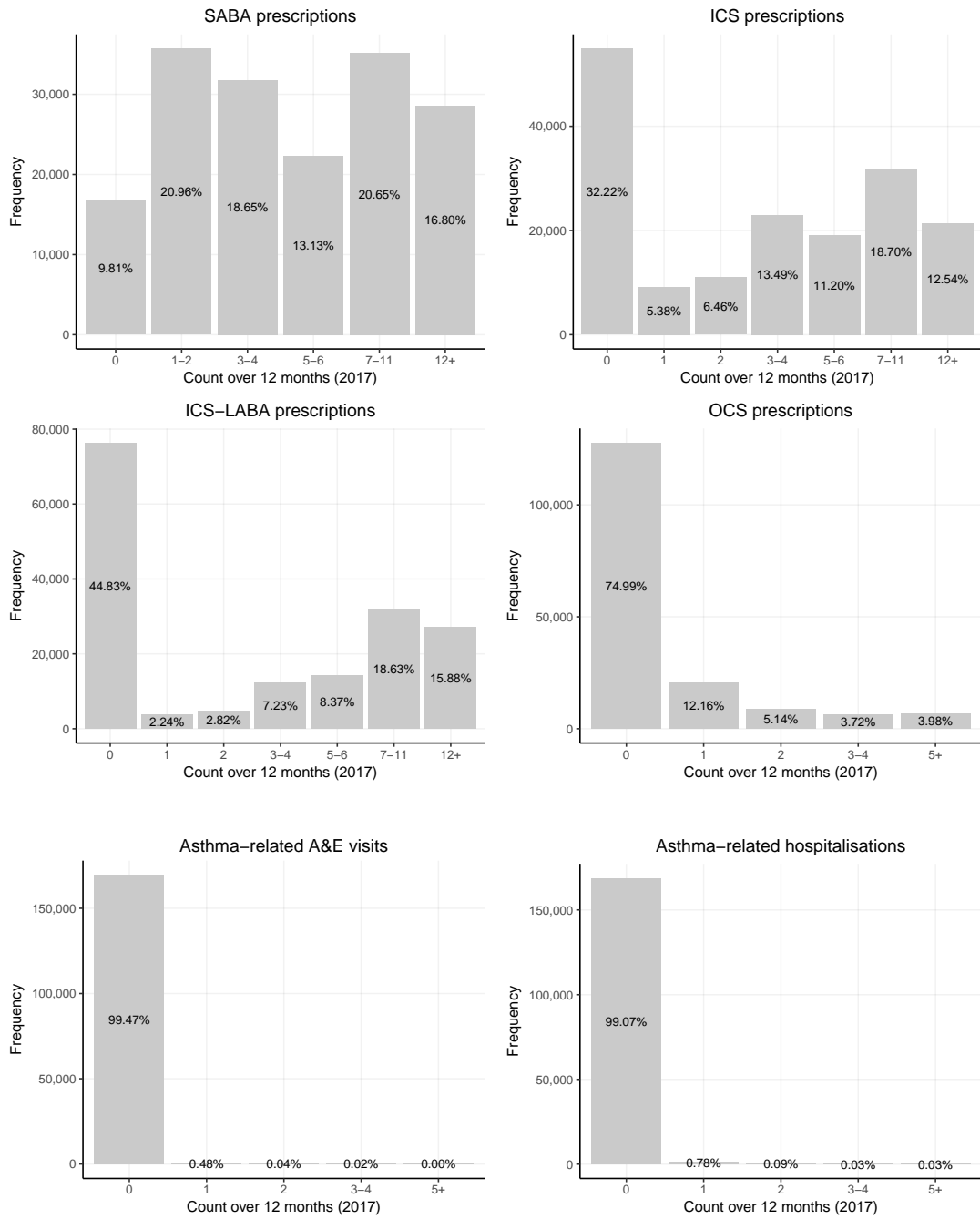


Figure 4.5: Distributions of asthma-related primary and secondary care utilisation events by patients with GP diagnosed asthma patients who received at least one asthma prescription in 2017.

Lastly, in the most recent extraction of the ADDE dataset, there were 3,180 people for which asthma was recorded among the conditions related to death. Among those people, 1,273 (40.0%) had no GP-recorded asthma diagnosis, 859 (27.0%) had no GP-recorded asthma prescriptions ever, and 823 (25.9%) had neither diagnosis nor prescription records for asthma.

4.5 Quality of asthma-related events in the GP database

4.5.1 Background

Health care data are mainly event-based and are mostly captured in a narrative format [302]. Clinical coding is often motivated by clinical and administrative purposes. However, even when perfect, clinical coding often involves information loss since coding schemes do not cover all aspects of health and health care. In practice, both electronic capture and coding of patient data are suboptimal [285]. Examples of barriers include cost, training needs, inefficient design and negative attitudes to EHR systems as well as lengthy lists of codes to choose from [233, 302]. Subsequently, many healthcare events may not be captured or only partially recorded or coded at the point of care. Since only coded data are usually routinely collected, those healthcare events would be missing in central repositories.

For most primary care events, minimal event attributes include code and date. This also applies to events involving measurement of health parameters such as body mass index (BMI) and blood eosinophil count. For such measurement events, a GP can use an informative code (e.g., 42K1.: Eosinophil count normal) to describe the measurement. Alternatively, he or she can use a declarative code (e.g., 42K..: Eosinophil count) with the measurement value recorded in a separate field.

The incidence of non-recording may differ between different types of events; for example, while some primary care events are known to be well-recorded (e.g., diagnosis codes that are required for the QOF indicators), others are less frequently coded, partially coded, or are even completely not coded [284]. Regarding events for which numerical values are expected besides the codes, it is possible to directly calculate the frequency of missing or invalid values for the recorded codes. However, where the code itself is absent, it is impossible to ascertain, at the individual level, whether the event did not happen at all or it was simply not coded. Nonetheless, some insights into the levels of missingness of these events could be potentially still obtained by calculating their recording frequency in the asthma

population. This is particularly feasible for essential health care events that are required by national guidelines (e.g., lung function testing to confirm diagnosis), which are expected to be recorded for large proportions of patients.

To investigate the quality of recording of asthma-related events in the GP dataset of the SAIL Databank, I examined the percentage frequency of recording for selected asthma-related event codes and event values for a sub-cohort of the asthma population.

4.5.2 Methods

I included in the sub-cohort all patients who had an asthma diagnosis Read code between 1-1-2006 and 1-1-2012, with no “asthma resolved” in the four years following the diagnosis date. I did not include patients diagnosed before 1-1-2016 (one and nine months after the introduction of the QOF in April 2004) to ensure adequate level of recording of GP data. The Read codes used for patient inclusion are listed in [Table C.4.1](#).

I chose the following six groups of asthma-related events: Triggers of asthma in the patient, disease severity, steps undertaken by GPs to manage asthma control, spirometry tests to assess lung function, serum eosinophil count, and level of total IgE. The Read code definitions of these events are shown in [Table C.4.3](#). For each event group, I calculated the proportion of patients in the sub-cohort who had at least one code over their follow-up period. For the first three groups, the follow-up period was four years from the diagnosis date. This was an appropriate period to allow equal follow-up for all patients in the cohort within the date range of the available GP data at the time of data extraction. For the events of lung function, eosinophil count, and total IgE, the follow-up period was similar except that it also included three months before the diagnosis date. This was because these diagnostic procedures could have been performed before the diagnosis was confirmed and recorded.

I also examined the recorded values for 54 Read codes for lung function testing (see [Table 4.7](#)). I calculated the percentage of missing values and inspected the distributions of the recorded values. Usually, event values only include numerical data. For values represented as percentage, there was no percent sign (%) and

therefore they were not directly distinguishable from other data formats. Nonetheless, the intended values can be often easily inferred from the event description.

4.5.3 Results

The sub-cohort included 127,303 asthma patients, 55.1% of whom were females.

4.5.3.1 Recording of event groups

Table 4.6 shows percentages of patients with at least one recording of key asthma-related GP events over the follow-up period. 81.6% of the patients had at least one event code for lung function testing. In addition, 52.5% of the patients had at least one event code for serum eosinophil count. However, for the other event groups, the proportions were smaller: 9.7% for asthma control steps, 4.8% for asthma triggers, 1.8% for asthma severity, and 1.2% for total serum IgE.

Table 4.6: Percentages of patients with at least one recording of key asthma-related GP events over specified periods from diagnosis.

GP Event	% of patients (95% confidence interval)
Asthma triggers *	4.8 (4.7, 4.9)
Asthma severity *	1.8 (1.8, 1.9)
Asthma control steps *	9.7 (9.6, 9.9)
Lung function test **	81.6 (81.4, 81.8)
Serum eosinophil count **	52.5 (52.2, 52.7)
Serum total IgE **	1.2 (1.1, 1.2)

GP = general practitioner; IgE = immunoglobulin E.

* In the four years after diagnosis date.

** From three months before to four years after diagnosis date.

4.5.3.2 Quality of values of lung function events

For the recorded lung function events, values were missing in 11.4% of these events. The lowest proportion of missingness was for predicted peak expiratory flow rate (PEFR) using the 13826 European Standard⁶ (0.3% [95% confidence interval (CI): 0.2-0.4%]), while the highest proportion was for percentage of predicted vital capacity (93.5% [81.1-98.3%]). Results for the rest of the codes are shown in Table 4.7).

Figure C.5.1 includes visualisation of lung function event values using beanplots [304] which show distribution density and makes it easy to spot anomalies in data

⁶This standard specifies the requirements for peak expiratory flow meters which is designed to be used to evaluate lung function in humans with spontaneous breathing [303].

Table 4.7: Percentages of missing values for lung function event codes.

Event code	Event description	Number of all events	Number of events with missing values	Percentage of events with missing values	
				Percentage	95% confidence interval
3395.	Peak exp. flow rate: PEFR/PFR	320,621	5,869	1.8	(1.8-1.9)
339A.	PFR - before bronchodilation	45,657	750	1.6	(1.5-1.8)
339H.	Predicted peak flow	35,741	127	0.4	(0.3-0.4)
339p.	Predict PEFR using EN13826 std	25,121	73	0.3	(0.2-0.4)
339g.	Serial peak expiratory flow rate	23,325	2,371	10.2	(9.8-10.6)
339M.	FEV1/FVC ratio	19,962	361	1.8	(1.6-2.0)
339R.	FEV1/FVC percent	18,974	1,503	7.9	(7.5-8.3)
339S.	Percent predicted FEV1	14,866	1,920	12.9	(12.4-13.5)
339o.	PEFR using EN 13826 device	14,206	112	0.8	(0.7-1.0)
339B.	PFR - after bronchodilation	13,140	437	3.3	(3.0-3.7)
339P.	Expected FEV1	7,186	55	0.8	(0.6-1.0)
339C.	PFR - expected	6,436	144	2.2	(1.9-2.6)
339D.	PFR - best ever	6,284	65	1.0	(0.8-1.3)
339Q.	Expected FVC	5,086	31	0.6	(0.4-0.9)
339i.	FVC/Expected FVC percent	4,735	1,655	35.0	(33.6-36.3)
339b.	FEV1 after bronchodilation	3,843	148	3.9	(3.3-4.5)
339n.	Serial PEFR abnormal	3,570	2,349	65.8	(64.2-67.4)
339N.	Expected FEV1/FVC ratio	3,212	105	3.3	(2.7-4.0)
745D4	Post bronchodilator spirometry	3,137	2,668	85.0	(83.7-86.3)
339a.	FEV1 before bronchodilation	2,680	114	4.3	(3.5-5.1)
339E.	PFR >80% of predicted	2,405	1,174	48.8	(46.8-50.8)
339m.	FEV1/FVC ratio after bronchodilator	2,344	74	3.2	(2.5-4)
339T.	FEV1/FVC > 70% of predicted	1,811	810	44.7	(42.4-47.1)
339V.	Recorded/predicted PEFR ratio	1,392	<5	<0.4	*
339d.	PEFR post steroids	1,307	338	25.9	(23.5-28.3)
339c.	PEFR pre steroids	1,262	169	13.4	(11.6-15.4)
339F.	PFR 60-80% of predicted	1,171	553	47.2	(44.3-50.1)
339l.	FEV1/FVC ratio before bronchodilator	1,083	59	5.4	(4.2-7.0)
66Yc.	Num consecutive days <80% PEFR	956	83	8.7	(7.0-10.7)
339U.	FEV1/FVC < 70% of predicted	838	314	37.5	(34.2-40.9)
339u.	Peak inspiratory flow rate	799	642	80.4	(77.4-83)
33950	Diurnal variation of PEFR	753	302	40.1	(36.6-43.7)
339G.	PFR <60% of predicted	640	251	39.2	(35.4-43.1)
339O1	FEV1/vital capacity ratio	344	7	2.0	(0.9-4.3)
339L.	Expected peak flow rate x 80%	329	<5	<1.5	*
339I.	Expected peak flow rate x 50%	327	<5	<1.5	*
339K.	Expected peak flow rate x 30%	325	<5	<1.5	*
339O0	FEV1 reversibility	170	7	4.1	(1.8-8.6)
339X.	Percentage of best ever PEFR	162	<5	<3.1	
339Y.	Percentage of PEFR variability	155	23	14.8	(9.8-21.6)
339r.	FEV1/VC percent	118	86	72.9	(63.8-80.5)
339f.	FEV1 post steroids	69	<5	<7.2	*
339S0	Percentage predicted FEV1 after bronchodilation	69	8	11.6	(5.5-22.1)
339e.	FEV1 pre steroids	48	<5	<10.4	*

* Value masked to comply with the SAIL Databank's disclosure policy.

Table 4.7: Percentages of missing values for lung function event codes. (cont'd)

Event code	Event description	Number of all events	Number of events with missing values	Percentage of events with missing values	
				Percentage	95% confidence interval
339t.	Percentage of predicted VC	46	43	93.5	(81.1-98.3)
339Z.	Respiratory flow rates NOS	39	27	69.2	(52.3-82.5)
339s.	FVC before bronchodilation	34	16	47.1	(30.2-64.6)
339k.	FEV1/FVC ratio post steroids	21	<5	<23.8	*
339J.	Optimal peak flow rate	21	<5	<23.8	*
339W.	Worst peak flow rate	12	<5	<41.7	*
33972	FEV1 after change of bronchodilator	7	<5	<71.4	*
339j.	FEV1/FVC ratio pre steroids	5	<5	*	*
33951	PEFR after exercise	<5	<5	*	*

* Value masked to comply with the SAIL Databank's disclosure policy.

and multimodal distributions. For most of the lung function testing events, the distribution of the recorded values appeared to be consistent with the expected units and ranges. For example, event values such as the forced expiratory volume in the first second (FEV1) and expected forced vital capacity (FVC) appeared to be recorded mostly in litres as expected, but with few apparent percentage values. For event values that were expected to be recorded as percentages, most of the values appeared to be percentages, with few values, for some event types, recorded as simple ratios. Examples included percent predicted FEV1, percent of actual to expected FVC, FEV1/FVC ratios, FEV1/vital capacity (VC) ratio, recorded/predicted PEFr ratio, and percentage of best ever PEFr. FEV1 reversibility had a large peak between 1 and 10, with a small peak at 100. Post-bronchodilator spirometry (745D4) values distributed mostly between 200 and 500, likely representing the change in FEV1 in millilitres from before and after bronchodilator administration.

4.5.4 Interpretation

Based on the above-studied asthma-related events, the recording of events and their values varied widely between event groups. Events that document asthma triggers, severity, and steps to manage the disease control were occasionally recorded. These events are important for asthma studies concerned in disease activity and management. Blood eosinophil count is usually a part of the full blood count test which can be performed for many indications other than asthma (e.g., for women in pregnancy). Blood eosinophil count can be used to predict severe exacerbations

and poor asthma control [305]. However, this test was only available in the GP dataset for about half of the studied patients. Future developments in automatic reporting of results may change this.

Lung function tests, particularly PEFr, were relatively better recorded. However, codes for airway obstruction reversibility tests were underrecorded. They are examples of events for which numerical measurements are supposed to be recorded along with the event code. However, this analysis demonstrated that the values of these events showed variable levels of missingness and inconsistency. Bimodal distributions were common among these event values. One apparent reason is the different ways test results were recorded by healthcare professionals. Many of the values that were supposed to be recorded as either percentages or simple ratios were recorded in both formats, one of which was often dominant. For events such as FEV1 before bronchodilation, a possible explanation of the bimodal distribution is that GPs had different understanding of the unit in which the event values should be recorded (e.g., litres vs. percent or change). A longitudinal between-GP practice analysis of these values could be helpful in evidencing these potential explanations.

4.6 Discussion

4.6.1 Summary of the Observatory design and data quality

The Wales Asthma Observatory represents a regularly updated asthma registry. It also offers a platform for various types of asthma epidemiological research and a surveillance tool to inform health policy and service planning. While traditional disease registries use *de novo* data collection, the Observatory represents an untraditional approach to disease registry as it mainly uses RCD to identify and describe cases.

The Observatory included patients who satisfied one or more of multiple case definitions for asthma, including the one developed in Chapter 3. Patients were longitudinally characterised using a number of key asthma outcomes. Improving efficiency and reproducibility of data extraction was considered in the Observatory structure and user interface. The Observatory data are versioned, and the user interface allows rapid, reproducible, reusable, and shareable data extraction.

However, I demonstrated a traditional problem of using RCD: suboptimal quality of data. There were various patterns of missingness and inconsistency in asthma data in Wales. Many lung function tests were recorded without measurements. When recorded, measurements of some tests were inconsistent.

4.6.2 Strengths and opportunities

The Wales Asthma Observatory project has several strengths. It relies on inexpensive, sustainable, and regularly updated sources of routine data in the SAIL Data-bank. The wide-to-complete national coverage of the SAIL datasets enables performing representative, population-based studies with large number of patients. This opportunity is usually not available through other traditional sources of data such as national surveys and primary data collected by researchers. The availability of several case definitions to identify asthma patients provides researchers with flexibility and ability to compare their findings with studies performed elsewhere using various case definition. In addition, the Observatory benefits from a collaborative platform to share clinical code lists and data extraction procedures within research teams and, in the future, with the wider research community and the public as well. This collaborative platform is intended to save analyst time, to improve collaboration and sharing of methods, as well as to support research documentation, reproducibility and transparency.

4.6.3 Challenges and limitations

4.6.3.1 Primary care-based case definition of asthma may exclude some patients

The asthma case definitions used in the Observatory, including the inclusive and strict ones, were based on primary care data only. This was justified as asthma in the UK is managed mostly in primary care [32].

However, it is possible that some people with asthma may not be captured by the primary care dataset (WLGP). The WLGP dataset currently covers only ~80% of GP practices (see [Section 4.3.3](#)). This means that people with asthma who never registered at those participating practices were not included in the Observatory. In addition, it is possible that some people had presented with acute asthma symp-

toms at A&E departments and/or were hospitalised for asthma without being captured by GP data-based case definitions of asthma used in the Observatory.

Therefore, the Observatory would benefit from using secondary care data in identifying asthma patients, which will be considered in future developments.

4.6.3.2 Inherent limitations of routine data

Limitations of routine data for use in research are discussed in Chapters 2 and 3. The use of these data in disease registries is challenged by a range of limitations. For example, these data are usually collected for clinical and administrative purposes in mind and therefore may not be readily appropriate for secondary uses such as disease surveillance, service planning, or research. Despite the wide use of standardised clinical coding schemes such as ICD-10 or Read codes, EHR-derived data collected over many years often lack standardisation as the same piece of clinical information may be recorded in different forms [233, 306].

Case definitions are a core part of a disease registry or observatory. However, developing accurate case definitions based on routine data is challenging. Disease registries that are built using active reporting of individual cases have the advantage of individual-level assessment of eligibility, often using confirmed diagnosis by clinicians. In contrast, a routine data-based disease registry is populated by applying the same eligibility criteria *en masse* to all people in a large population in a database [213], which can introduce high risk of misclassification. In addition, it has been shown that methods to estimate asthma prevalence from routine data may be inaccurate [307].

4.6.3.3 Traditional methods to define cases and outcomes may need re-consideration

In Chapter 2, I found significant heterogeneity in the definitions of asthma and asthma outcomes. There were variations, not only in the types of health events used to assess the disease, but also in the time interval over which these health events are queried. These query intervals should be chosen based on stability of disease statuses over time [251]. Longer intervals may conceal important temporal variations of the measured disease status. Conversely, shorter intervals may introduce unrealistic temporal variations. The case definitions currently supported

by the Observatory use 12-month intervals. This has been traditionally the most frequently used interval in research and for clinical and administrative purposes (e.g., for the QOF). However, evidence is lacking about whether 12 months is the best interval over which asthma activity, severity, and control are assessed. Therefore, further studies are needed to choose the most appropriate and meaningful interval for each of these variables.

4.6.3.4 Implications of the suboptimal quality of asthma-related routinely collected data on asthma research

Data gaps undermine the ability of routinely collected EHR data to inform asthma care [308] and support research. Significance and implications of data gaps are specific to how data are used. For example, gaps in lung function data make it difficult to assess asthma severity, which could be alternatively assessed by asthma medications [309]. Similarly, unavailability of medication dispensing data is a significant limitation in adherence studies. In contrast, such gaps are unlikely to be an issue in prevalence studies that rely on physician's diagnosis codes.

Guidelines in the UK recommend performing airway obstruction reversibility tests when asthma diagnosis is uncertain [16]. An essential objective part of the clinical diagnosis of asthma is to confirm the airway obstruction reversibility, which often strongly indicates an asthma diagnosis and rules out chronic obstructive pulmonary disease (COPD). Reversibility data would allow identification of a patient subgroup with very high certainty of asthma diagnosis. However, I found that codes of these tests were under-recorded in the GP dataset. One explanation is that a significant number of lung function tests were performed in secondary care settings. A similar gap between the number of diagnosis codes and spirometry codes were observed in Alberta, Canada [308].

The QOF, arguably, appeared to improve the recording of healthcare events that are required by its quality indicators (see Table C.1.2). However, the asthma indicators AST001 and AST002, for example, only require event codes to be recorded; they do not assess the quality and completeness of recorded event measurements *per se* [310].

4.6.3.5 Primary care coding in the UK will change

The primary care data used in the Observatory are coded using Read codes. Read codes are a clinical coding scheme which covers wide aspects of primary care encounters. It is the main coding scheme used in primary care in the UK. However, GP practices in the UK are expected to transition from Read codes to Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) in 2018 [311]. Therefore, to ensure sustainability of the Observatory, the methods used to identify patients and assess disease outcomes need to be modified to support SNOMED-CT coding in due time.

4.6.3.6 Data security and implications of anonymisation

There are important considerations when using de-identified data, such as those used in the Observatory. Despite replacing people's identifiers in the SAIL Data-bank with multiply encrypted unique identifiers (i.e. ALFs), re-identifiability of patient data is possible unless additional steps are taken. For example, very small groups of patients with rare combinations of characteristics related to their health or health care usage may be re-identified if the data are presented in particular formats. This may be an issue for less common events, such as hospital admissions or day cases, or complex combinations of multiple broad criteria. For instance, there could be fewer than five patients who were in a specific age group at a specific date, lived in a small city, had asthma diagnosis made in a specific year, and received a very high number of inhalers in a specific year; these combinations may increase the risk of patient re-identifiability.

To avoid the risk re-identifiability of patients, all outputs from the Observatory must conform to the SAIL information governance policy [295]. For outputs to be available outside the Gateway, they must be first reviewed by senior analysts who assess the output for re-identifiability risk and compliance with the approved project proposals. In the output data, groups with frequencies smaller than a certain limit, usually five, must be suppressed from reporting or aggregated with other groups, and dates are aggregated into time periods.

Risk of patient re-identifiability can be high when person-level information from difference sources is publicly available. However, SAIL controls all individual link-

ages within the databank and does not export individual level data to avoid this possibility.

Although the use of anonymised data in the Observatory protects patient identities, it places limits on how these data can be used. For example, a prediction model can be developed within the anonymised registry using the power of multiple, linked, large volume datasets. Within such an anonymised setting, individual predictions for anonymous persons can be obtained. Due to anonymisation, however, it is impossible to link these predictions back to patient records at the point of care, e.g., a GP surgery. Instead, to use that prediction model at the point of care, ideally all predictor variables from other care settings (e.g., emergency, inpatient, and/or outpatient care) that are used in the model development need to be available at the point of care. However, this is often not feasible. A possible workaround approach is to develop a prediction model, within the anonymised registry, using only the types of data that will be available at the point of care. Then, using sensitivity analysis, the performance of this prediction model can be compared with a model that is based on all the data sources available in the registry.

4.6.4 Recommendations for better capture of asthma data

Based on the analysis of data quality that I presented earlier in this chapter, I recommend that effort needs to be made to ensure that better data on asthma care are captured and recorded in EHRs. Wider standardisation of the ways health events are recorded is particularly needed. This could be achieved through better clinical coding training of healthcare professionals. Motivating GPs to improve coding when their time is very short could be potentially achieved by demonstrating the value of projects using complete coding (e.g., those already performed using the SAIL Databank⁷). EHR systems should facilitate standardisation of clinical coding by incorporating on-screen coding advice and better validation rules, which insure the right data are recorded in the right place and in the expected format. Natural language processing (NLP) techniques, which are increasingly implemented in EHR systems to codify narrative data, should consider validity of the produced coded data. Capture of more accurate and complete data at the point of care will result in better value and utility of the asthma registry and the Observatory. Pro-

⁷Examples of studies that used the SAIL Databank can be found in the following links: <https://sail-databank.com/saildata/uses-for-sail-data/> and <https://saildatabank.com/saildata/sail-publications/>

professional societies representing primary and secondary care respiratory medicine could be the most appropriate groups to lead efforts to improve the quality of captured data. In addition, implication of data quality and the lack thereof on patient care, service planning, and research needs to have more presence in venues of continued professional education and research meetings to make clinicians more data conscious. Furthermore, payment for performance schemes such as the QOF need to consider the quality and completeness of the recorded data in addition to their quantity [284] as added incentives.

4.6.5 Future development

Although the Wales Asthma Observatory benefits from rich sources of routine data in the SAIL Databank, data about several aspects of healthcare are still missing. These include clinical, laboratory, and medication prescribing data from secondary care, as well as community medication dispensing data. However, endeavours are currently under way to link the all-Wales pathology data to the SAIL Databank as well as to increase the depth of coding of clinical correspondences using NLP techniques. Using these data in the Observatory would enable highly useful research applications such as improving accuracy of asthma case definitions and disease phenotyping. Environmental data including data on housing quality, pollution, greenness, use of outdoor spaces, commuting routes, and modes of transport can be also linked to the Observatory in order to answer questions about the effect of various environmental factors on asthma outcomes [300, 312, 313].

Emerging sources of data such as data from smart inhalers and wearable technologies, despite being currently of limited use, can be later used to enrich the asthma registry with important variables about disease activity and medication usage and adherence over time. Asthma-related PROMs can be of high significance to clinical care and research [314]. Future development of the asthma registry can include developing platforms to collect asthma-related PROMs and link them to asthma-related routine data. This will enable investigating the relationship between doctor-reported and self-reported asthma outcomes and will allow assessing the association of PROMs with each of health services utilisation patterns, health care quality, and health care inequalities.

An LHS of asthma in Wales is needed to close the gap between evidence and practice. By facilitating near real-time disease surveillance, the Wales Asthma

Observatory can be a building block which would help this ambitious project to materialise.

4.7 Conclusion

The Wales Asthma Observatory represents an untraditional approach to a disease registry, and a platform for research and surveillance. In this chapter, I described development of the Observatory, including purpose, source population, structure, content, and technical logistics.

The quality of asthma related data in Wales is suboptimal. There were various patterns of missingness and inconsistency in these data. Many lung function tests were recorded without measurements. When recorded, measurements of some tests were inconsistent. To improve the capture of asthma data, I proposed enhanced EHR data entry quality checks, data quality awareness training for health-care professionals, and data quality based incentivisation of health care providers.

I described approaches to improve efficiency and reproducibility of studies that will use the Observatory. I developed an easy-to-use user interface that supports shareable, reusable, and scalable data extraction from the Observatory and the GP dataset.

Further developments to the Observatory will provide linkage to additional RCD sources and PROMs, and adaptation to the upcoming clinical coding system, SNOMED-CT.

Chapter 5

Inequalities in asthma care and outcomes in Wales

In the previous chapters, I have discussed the development of the Wales Asthma Observatory. In this chapter, I demonstrate an example of utilising the Observatory to inform health policy. Variations in asthma outcomes between population groups have been widely reported worldwide. These inequalities can be assessed using area-based deprivation indices. An important application of the Wales Asthma Observatory in supporting health policy is to investigate whether inequalities in asthma care and outcomes exist between socioeconomic groups. In this chapter, I investigated the variations in the incidence of asthma-related healthcare utilisation in primary and secondary care among asthma patients in Wales across the quintiles of the Welsh Index of Multiple Deprivation. I found wide social gradient in asthma where patients in the most deprived areas had remarkably more asthma-related hospitalisations indicating poorer outcomes. I also discuss the implications of these findings on health policy.

Chapter Contents

5.1	Introduction	147
5.1.1	Asthma variations are common	147
5.1.2	Inequality and inequity in health and health care	148
5.1.3	Area-based socioeconomic measures	149
5.1.3.1	Overview	149
5.1.3.2	The Welsh Index of Multiple Deprivation	150
5.2	Aims and Objectives	154
5.3	Methods	154
5.3.1	Data sources	154
5.3.2	The source population and study cohorts	156
5.3.3	Socioeconomic status	156
5.3.4	Outcome variables	157
5.3.4.1	Number of asthma-related GP visits	157
5.3.4.2	Number of asthma reviews	157
5.3.4.3	Asthma-related emergency department visits	157
5.3.4.4	Asthma-related hospital admissions	157
5.3.5	Statistical analysis	158
5.3.5.1	Descriptive statistics	158
5.3.5.2	Variation of age between deprivation quintiles	158
5.3.5.3	Count regression	158
5.4	Results	160
5.4.1	Descriptive statistics of the source population and study cohorts	160
5.4.2	Zero-inflated negative binomial regression models	168
5.4.2.1	Incidence rate ratios of study outcomes across deprivation quintiles	169
5.4.2.2	Incidence rate ratios of study outcomes for age groups and gender	171
5.4.2.3	Model fit	174
5.4.2.4	Sensitivity analysis	175
5.5	Discussion	177
5.5.1	Summary of findings	177
5.5.2	Interpretation in the light of previous studies	178
5.5.2.1	Comparison with previous studies	178
5.5.2.2	WIMD mainly describes areas, and to a lesser extent, individuals	179
5.5.2.3	Asthma-related emergency department visits and hospitalisations usually indicate worse asthma severity and control	179
5.5.2.4	Why did the most deprived asthma patients have more GP visits?	180
5.5.3	Study strengths	181

5.5.4 Study limitations	182
5.5.4.1 Case definitions for asthma-related A&E visits and hospitalisations were not validated	182
5.5.4.2 Possible residual confounders	183
5.5.4.3 WIMD overall index and asthma exacerbations: a possible circular relationship	184
5.5.5 Implications for health policy	185
5.5.6 Future work	187
5.6 Conclusion	188

5.1 Introduction

5.1.1 Asthma variations are common

The epidemiology of asthma and asthma outcomes exhibits variations around the world. The International Study of Asthma and Allergies in Childhood (ISAAC) revealed wide geographical variations in the prevalence of self-reported asthma diagnosis. There were twenty-fold variations in the prevalence of self-reported asthma between the study centres around the world [48, 315]. The variations were not only seen between countries but also within countries and within cities such as Mexico City and New York city [37]. In the United Kingdom (UK), moderate variations have existed between its member countries in the prevalence and incidence of asthma, based on self-reported and doctor-reported data [43]. However, there were limited variations between UK metropolitan and non-metropolitan areas, with the latter having slightly higher prevalences [316].

Studying the variations in asthma epidemiology can potentially help understand the aetiological factors and determinants of the disease. It has been suggested that the geographical variations in asthma epidemiology result from complex interaction of numerous factors. The effect of environmental determinants on asthma has been extensively studied. Air pollution has been linked to asthma epidemiology. There is contradictory evidence on whether air pollution is associated with increased asthma incidence and prevalence [317, 318], with a suggestion that adverse effects of traffic-related air pollution tend to be close to major roads [318]. It is more evident, however, that air pollution increases the incidence of asthma exacerbations among people who already have the disease [318, 319]. Climate has been also suggested to influence the prevalence of asthma symptoms. Data from

146 centres of the ISAAC study showed that in Western Europe the prevalence of asthma symptoms in school children was positively associated with indoor humidity, and negatively associated with temperature, outdoor humidity, and altitude [320].

Socioeconomic determinants have been also found to influence asthma epidemiology. In 1973, Mitchell et al. found that in Scotland, severe asthma was more often observed in children of semi-skilled and unskilled manual worker parents and children of larger families [321]. A systematic review has found that lower socioeconomic status is associated with higher asthma prevalence [322]. In England, a study in the early 1990s found that asthma hospital admissions rates were higher in areas with high deprivation where most admissions came via Accidents and Emergency departments rather than referrals from general practitioners (GPs) [323]. Another study in Cardiff found that hospital admission rates for asthma were correlated with the level of social deprivation [324]. Those hospital admission rates were, however, not correlated with the prevalence of asthma or wheezing but with the prevalence of chronic phlegm and the exposure to second-hand smoke at home. Low socioeconomic status was associated with less treatment in wheezy children [325] and poorer asthma control and persistent airway obstruction in adults [326].

Variations in asthma epidemiology, especially those ascribed to socioeconomic factors, highlight inequalities in health and health care and represent important challenges to health policy.

5.1.2 Inequality and inequity in health and health care

Health inequalities have been defined by the World Health Organisation [327] as: “differences in health status or in the distribution of health determinants between different population groups”. Uneven distributions of health status and its determinants may result from numerous factors creating advantages and disadvantages that accumulate over the course of life [328]. The term *health inequalities* is a descriptive term that is often used to describe those uneven distributions in health status and determinants and that do not *per se* imply moral judgement [329]. Some forms of health inequalities are practically unavoidable and do not represent injustice. Examples include health disparities that result from biological differences between population groups or external factors, such as natural

environment, over which they usually do not have control. Arguably, individuals' 'free choices' may also contribute to health inequalities [327], although this is debatable since individuals' 'free choices' can be influenced by the environment in which they live [330]. However, other disparities in health result from unnecessary, unfair, and unjust variations in health determinants, in which case they are called *health inequities* [327, 331]. The concept of health inequity has attracted hot debate and controversy [329], and since it is based on value judgement, it is not easy to determine which health inequalities are universally inequitable. Braveman et al. (2003) presented an operationalised definition of health equity: "*Equity in health is the absence of systematic disparities in health (or in the major social determinants of health) between groups with different levels of underlying social advantage/disadvantage—that is, wealth, power, or prestige*" [332].

In epidemiological studies, health inequalities are often assessed by comparing health status and determinants between social groups in the population. In large epidemiological studies, however, individual-level data on socioeconomic status may not be available, and collecting these data from the study individuals is impractical. In these studies, area-based measures of socioeconomic status and deprivation have been widely used to study health inequalities.

5.1.3 Area-based socioeconomic measures

5.1.3.1 Overview

The different socioeconomic factors can have accumulating, interactive effects on the individual's health over the course of life [328, 333]. Some of these factors, such as income, employment, and educational attainment, act on the individual level. On the other hand, factors related to the community and environment act on the group and area levels and may have effects on the person's health status independent from individual health determinants [334, 335]. To account for those complex and interacting factors, area-based socioeconomic measures have been developed, usually using census data, to provide 'simple' socioeconomic profiles and ranks for geographic areas [336]. These measures can be based on a single or, more commonly, multiple components representing different socioeconomic factors [333]. Area-based socioeconomic measures have been widely used in epidemi-

ological studies to assess the effect of socioeconomic status on health, although they are most commonly used as control variables and confounders [333].

In the UK, examples of area-based socioeconomic measures include:

- Townsend's Index [337],
- Index of Multiple Deprivation (England) [338],
- the Scottish Index of Multiple Deprivation [339], and
- the Welsh Index of Multiple Deprivation [340].

5.1.3.2 The Welsh Index of Multiple Deprivation

The Welsh Index of Multiple Deprivation (WIMD) is the official area-based measure of relative socioeconomic deprivation in Wales. The WIMD is based on socioeconomic indicators that represent aggregate characteristics of residents in the area and/or describe the area itself. The WIMD was commissioned by the Welsh Government to create a measure to understand relative differences in deprivation, based on several domains measured at a small area level across Wales. The WIMD was designed a tool to inform the development of policies and allocation of funding so that they target the most disadvantaged communities [341]. The WIMD index is updated every few years, with versions released in the years 2000, 2005, 2008, 2011, and 2014.

The WIMD 2011 index is constructed from weighted sum of eight deprivation domains, each is composed of several deprivation related indicators. According to the WIMD 2011 Technical Report [342], those deprivation domains include the following, ordered by weighting: Income, Employment, Health, Education, Geographical Access to Services, Housing, Physical Environment, and Community Safety. I provide an overview for these deprivation domains including how they were constructed.

1. *Income domain* This domain is based on the proportion of residents in a given area with low income or those who claim income-related benefits, and has a 23.5% weighting in the overall WIMD 2011 index.
2. *Employment domain* This domain represents the proportion of residents in the working age in a given area who have employment-related deprivation (i.e. receiving benefits related to employment). This domain has a 23.5% weighting in the overall WIMD 2011 index.

3. *Health domain* This domain captures the health-related deprivation, and is constructed from four indicators including limiting long-term illness, death rate in the area from all causes, incidence of cancer, and low birth weight. This domain has a weighting of 14% in the overall WIMD 2011 index.
4. *Education domain* This domain reflects the deprivation relating to educational attainment in a given area among children and young residents as well as the lack of educational qualifications and skills among adults. It is constructed from average school scores of children, proportion of residents not in higher education at the age of 18 or 19, proportion of residents aged 25 or above with no educational qualifications, and proportions of half day absence among children in primary and secondary schools. This domain has a weighting of 14% in the overall WIMD 2011 index.
5. *Geographical Access to Services domain* This domain captures the deprivation relating to inaccessibility of necessary services to each household in a given area. Inaccessibility to a service is measured by the average time needed to reach it using the shortest trips by bus and/or by walking. The services include National Health Service dentists, food shops, GPs, Post Office, primary and secondary schools, leisure centres, and transport nodes. This domain has a weighting of 10% in the overall WIMD 2011 index.
6. *Housing domain* This domain represents the level of disadvantage due to lack of adequate housing, and is constructed from indicators including proportion of residents who lack central heating in their households, and proportion of residents who live in overcrowded households. This domain has a weighting of 5% in the overall WIMD 2011 index.
7. *Physical Environment domain* This domain represents the disadvantage from environmental factors in a given area that can affect the quality of life. These factors include air quality and pollution, emissions, risk of flooding, and distance to waste disposal and industrial sites. This domain has a weighting of 5% in the overall WIMD 2011 index.
8. *Community Safety domain* This domain reflects the level of safety and protection from crimes in a given area. It is constructed from indicators including the proportions of offenders among adults and young people, numbers of

burglaries, criminal damages, thefts, violent crimes, and fire incidents. This domain has a weighting of 5% in the overall WIMD 2011 index.

The WIMD index is produced for all Lower Layer Super Output Areas (LSOAs) in Wales. LSOAs were outlined by the Office for National Statistics of the UK for census related purposes [343]. LSOAs vary widely in spatial size but they are intended to have comparable population sizes; according to the WIMD 2011 index, the average population in those small areas was $\approx 1,600$ people in that year [341]. The WIMD index gives a rank from 1 (most deprived) to 1,896 (least deprived) to the 1,896 LSOAs in Wales.

The most deprived areas in Wales are distributed mostly in the southern areas such as Rhondda Cynon Taf, Blaenau Gwent, as well as the east and north of Swansea, pockets in Newport, and the south and east of Cardiff (Figure 5.1).

Limitations of the WIMD index

A relative ranking measure

WIMD is a ranking system in which areas are ordered according to their sum of weighted deprivation scores. However, it does not quantify the level of deprivation, and therefore it does not quantify the differences in deprivation between areas [341]. That is, the WIMD index can tell that an area has a higher or lower multiple deprivation than another area, but it does not tell by how much.

It describes areas, not residents

Being an area-based measure, the WIMD is intended to describe the relative multiple deprivation in the area as a whole based on average scores of individuals. Therefore, it does not imply that all the residents have the same multiple deprivation. For example, it is possible that different residents in an area have different types and levels of deprivation. A consequence of this limitation is that not all deprived individuals live in the most deprived areas, and not all least deprived individuals live in the least deprived areas [341]. Rather, it is possible that a number of individuals with very low deprivation live in areas with overall high deprivation. This limitation mainly concerns the deprivation domains that are based on individual-level data. However, it almost does not apply to the other deprivation

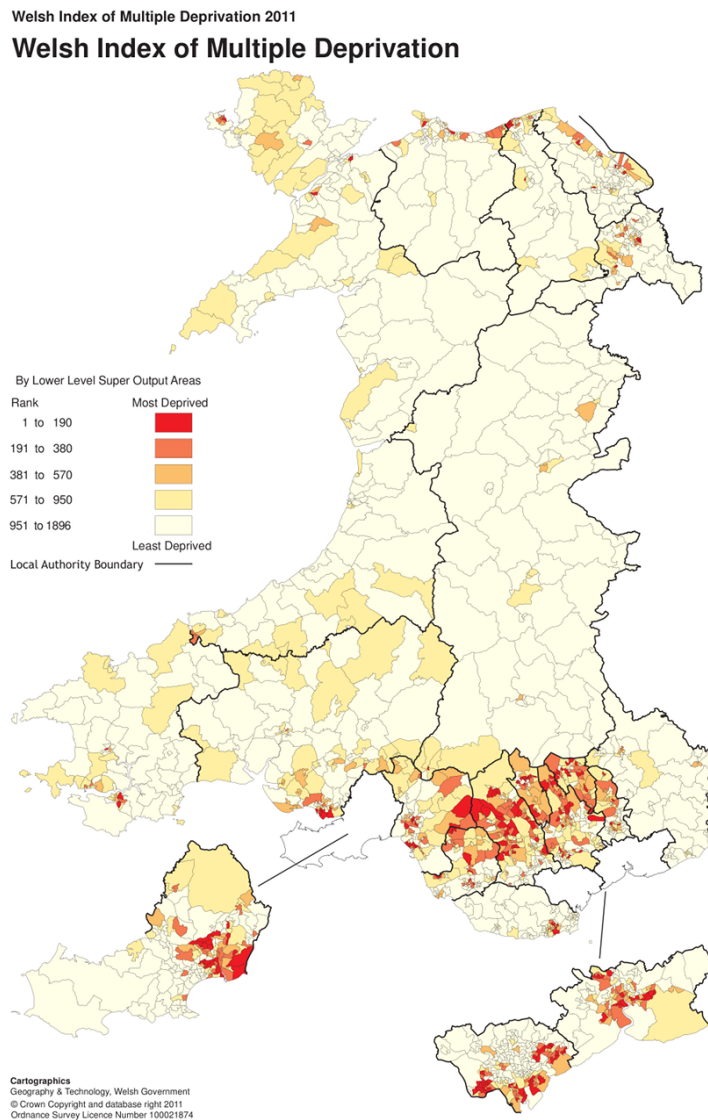


Figure 5.1: Map of Wales showing ranks of the 2011 Welsh Index of Multiple Deprivation for Lower-level Super Output Areas. Source: StatsWales, Welsh Government (<https://statswales.gov.wales/Download/File?fileId=91>). © Crown Copyright and database right 2011.

domains that use data on the areas themselves, namely the Geographical Access to Services, Physical Environment, and the Community Safety domains.

It is incomparable with indices in other UK countries

Other UK countries have their own multiple deprivation indices. However, it is not possible to directly compare these indices with the WIMD index due to the differences in the deprivation domains and the ways they are calculated [341].

5.2 Aims and Objectives

Variations in asthma outcomes between socioeconomic groups represent a significant challenge to health policy. While such variations were previously reported on small, localised populations in Wales [324], a country-wide analysis is needed to assess the scale of these inequalities. The aim in this chapter is to investigate the variations of asthma outcomes across the socioeconomic deprivation spectrum in Wales.

The objectives were as follows:

- To develop count regression models for asthma-related outcomes against the WIMD index quintiles, adjusted for age group and gender. The count regression models will be performed in a cohort of ever-diagnosed asthma (regardless of current treatment) and in a cohort of ever-diagnosed currently-treated asthma ('current asthma').

The asthma-related outcomes included the following:

- Asthma-related GP events: Any asthma-related visits to GPs
- Asthma routine reviews
- Asthma-related visits to Accident and Emergency departments
- Asthma-related hospital admissions
- To interpret the models in the light of previous studies and strengths and limitations of the routinely collected data used.
- To reflect on the implications of the findings on health policy in Wales.

5.3 Methods

Using the Wales-wide Secure Anonymised Information Linkage (SAIL) Databank, I accessed anonymised data on patients with a GP diagnosis of asthma in or before 2009 and continuous GP registration between 2010 and 2014 and linked those data to the quintiles of the 2011 version of the WIMD. I define the follow-up period as five calendar years from 2010-1-1 to 2014-12-31.

5.3.1 Data sources

In this analysis, I used the following datasets in the SAIL databank:

- **The Welsh Demographic Service (WDS):** The WDS contained de-identified demographic and administrative information for National Health Services (NHS) patient in Wales.
- **The WIMD 2011 dataset:** I used the 2011 version of the WIMD index which was the latest version available in the SAIL Databank. This dataset included rank quintiles of the overall WIMD index for all small areas (i.e. LSOAs) in Wales.
- **The Welsh Longitudinal General Practice dataset:** I described the Welsh Longitudinal General Practice (WLGP) dataset in [Section 3.3.1](#). I used the “2018-08-20” version of the dataset.
- **The Emergency Department Data Set (EDDS) for Wales:** The EDDS dataset was created in 2009 and captured visits to Accident & Emergency (A&E) departments as well as minor injury units (MIUs) in NHS hospitals in Wales. Recorded data for each attendance include investigations performed, diagnosis made, anatomical areas involved, treatment provided, as well as other administrative data related to the attendance. Diagnosis is coded using a three-digit code chosen from a list of 83 possible codes representing broad diagnostic categories. In addition, to the primary diagnosis, there are five further positions to record additional or secondary diagnoses. Due to the nature of emergency attendances, recorded diagnoses may be uncertain or unconfirmed. Practices of recording and coding of data vary between the different A&E departments and MIUs. The EDDS currently receives data on all emergency attendances in Wales. However, in the earlier years, some A&E departments were not able to submit their data to the EDDS, and therefore data in that period may be incomplete. Therefore, caution should be exercised when using this dataset for epidemiological and research analyses.
- **Patient Episode Database for Wales:** The Patient Episode Database for Wales (PEDW) database was created in 1991 and includes records for all planned and emergency inpatient admissions in addition to day case admissions to NHS Wales hospitals as well as most admissions of Welsh residents to hospitals in England. Recorded data include admission diagnoses, procedures and operations performed during admissions, as well as length of stay (LOS), Healthcare Resource Groups (HRGs), and other administrative data. Admission diagnoses are recorded using the 10th revision of the International Classification of Disease (ICD-10). In addition to a mandatory

primary diagnosis code for a hospital episode, the database allows recording of an additional subsidiary code and up to 12 secondary diagnosis codes. The PEDW database is considered of a high quality.¹ However, it is mainly an administrative database which was created as a tool to track hospital financial activity rather than for epidemiological or research purposes. The database also suffers from between-hospital variations in practices of coding of admission diagnosis in the available fourteen diagnosis positions.

5.3.2 The source population and study cohorts

The source population included people who met all the following criteria:

- Had records in the WDS dataset (version 2018-04-10).
- Had records in the WLGP dataset (version 2018-08-20).
- Lived at least to 2014-12-31.
- Were successfully linked to a valid WIMD 2011 ranking.
- Had continuous GP registration in the period between 2010-1-1 and 2014-12-31, which includes the period over which the outcome events are queried in addition to one year before it. To calculate GP registration periods for individuals, I used an unpublished algorithm developed in-house by the analyst team of the SAIL Databank. I assessed the effect of requiring continuous GP registration in a sensitivity analysis in [Section 5.4.2.4](#).

From the source population defined above, I created the following two cohorts:

- Cohort 1 included people with asthma diagnosis recorded before 2010-01-01.
- Cohort 2 was a sub-cohort of Cohort 1 in which people received at least one asthma prescription in any year between 2010 and 2014 (i.e. the follow-up period), in addition to having asthma diagnosis before 2010-01-01.

Asthma diagnosis was defined using the Read codes "H33%", "H3120", "102..".

Asthma prescriptions were defined using the Read code sets in [Appendix E](#).

5.3.3 Socioeconomic status

I linked each patient to the WIMD quintile of their area of residence during the follow-up period of 2010-2014. Where a patient had more than one address during

¹ See <http://www.publichealthwalesobservatory.wales.nhs.uk/PEDW>

the follow-up period, I selected the address with the longest duration within that period. This WIMD quintile variable was coded with 1 (the most deprived) to 5 (the least deprived).

5.3.4 Outcome variables

The outcome variables were counts of asthma-related events in primary and secondary care in the period from 2010-1-1 to 2014-12-31. The code lists used in the construction of these variables are shown in [Appendix E](#). The following are description of each of the outcome variables:

5.3.4.1 Number of asthma-related GP visits

For this analysis, I defined an ‘asthma-related GP visit’ by any Read code that indicates an asthma-related contact with a GP. Where more than one relevant code occurred on the same date, I treated them as a single visit.

5.3.4.2 Number of asthma reviews

An asthma review is a special, scheduled visit to GP, in which disease control is assessed and management plan including prescriptions and self-management advice is reviewed. Asthma reviews are ideally arranged regularly on at least an annual basis [16]. To identify asthma reviews from the GP dataset, I used a list of codes for annual review, medication review, follow-up, monitoring by nurse, and review using the Royal College of Physicians’ three questions [344].

5.3.4.3 Asthma-related emergency department visits

I identified asthma-related emergency department (ED) visits from the EDDS dataset using the code 14A (“asthma”). I treated ED visits as asthma-related if it contained this code in any of the primary or secondary diagnosis positions.

5.3.4.4 Asthma-related hospital admissions

I identified asthma-related hospital admissions from the PEDW dataset by looking for episode records in which an ICD-10 code for asthma (J45) or status asthmaticus

(J46) was in the primary diagnosis position or any of the remaining 13 secondary diagnosis positions.

5.3.5 Statistical analysis

I performed the statistical analyses described below for each of the two study cohorts.

5.3.5.1 Descriptive statistics

The descriptive statistics described the source population and the two asthma cohorts. For the source population, I calculated statistics about age and gender in addition to the prevalence of ever-diagnosed asthma at 2010-01-01 and the prevalence of ever-diagnosed currently treated asthma over 2010.

For the two asthma cohorts, I calculated the distributions of specific characteristics in relation to the WIMD quintiles. These characteristics included age, gender, receiving specific types of asthma prescriptions over the follow-up period, and the four outcomes variables (see above). In addition, I calculated asthma medication ratio, which represents the ratio of controller to controller-and-rescuer asthma medications [345]. I included inhaled corticosteroids (ICSs) and ICS-LABA (long-acting beta adrenoceptor agonist) combination inhalers as controller prescriptions, and included short acting beta agonist (SABA) inhalers as the rescuer medications. The formula was $\frac{ICS + ICS_LABA}{ICS + ICS_LABA + SABA}$ calculated over the five follow-up years. In averaging the ratio, I excluded those who received none of these three inhaler categories over that period.

5.3.5.2 Variation of age between deprivation quintiles

I used the Kruskal-Wallis rank sum test to test the differences in age distribution between quintiles of the overall WIMD index.

5.3.5.3 Count regression

To test the effect of multiple deprivation on each of the four outcome asthma variables, which were count variables, I developed a count regression model for each of them. The independent variable was the quintile of the overall WIMD index. I

tested the distribution of counts for each of the outcome variables and found that their variances were larger than their means. In addition, there were excessive numbers of patients with zero counts for each of the outcome variables. Therefore, I used zero-inflated negative binomial (ZINB) regression, which allows for over dispersion and models the excess in the zero counts.

A ZINB model assumes that count data are generated by two processes. One of these is a Bernoulli process which determines whether an individual is theoretically eligible to have a non-zero count [346]. Accordingly, there are individuals who are not eligible to have non-zero counts and therefore should have no events. In my study, those individuals were asthma patients in whom the disease was mild or remitted and therefore they needed no visits to GPs. Those patients would also have had no need for asthma-related ED visits or hospital admissions. Patients with more severe but well-controlled disease would also have no asthma-related ED visits or hospital admissions. On the other hand, for individuals who are eligible to have non-zero counts, the counts are assumed to be determined by a negative binomial distribution which expects some individuals to have no events and others to have one or more events. In my study, this applied to asthma patients who had active disease and, depending on disease severity and control as well as other non-asthma-related factors, might or might not need to have contact with primary and/or secondary care. To model the above described two processes, a ZINB model fits two regressions: a logistic regression to model the probability of having non-zero count, and a negative binomial regression to model the magnitude of counts.

In this chapter, I used ZINB regression models to model the counts of the above-mentioned outcome variables in relation to the WIMD index quintile. I considered the least deprived areas (i.e. the fifth quintile) the reference group. Therefore, in the resulting model, the exponentiated coefficients for each of the other four quintiles (1 to 4) represented the incidence rate ratio (IRR) of the relevant events for that quintile compared with the least deprived areas. I also calculated the 95% confidence intervals (CIs) for these IRRs. I adjusted the model for 5-year age groups and gender.

I used the `zeroinfl` function from the R package `pscl` version 1.4.9 to perform the zero-inflated negative binomial modelling [347].

I examined the model fit with quantile-quantile (Q-Q) plots of the raw residuals as well as with rootograms. Rootograms are graphical representation of both the observed counts as bars, and the expected counts, which are predicted by the model, as a curve [348]. The axis that represents the counts (i.e. usually the vertical axis) has a square-root scale. By including both the predicted and observed values in the same graph, a rootogram helped show the deviation of the predicted counts from the observed counts. A *hanging* rootogram has the bars of observed counts “hanged” on the curve of predicted counts. The deviations of the predicted counts from the observed counts were shown as deviations from the horizontal axis, and provided a visualisation of the goodness of fit of the model.

5.4 Results

5.4.1 Descriptive statistics of the source population and study cohorts

Table 5.1 shows characteristics of the source population. A flowchart of case selection for both asthma cohorts is shown in Figure 5.2.

Table 5.1: Characteristics of the source population in the year 2010 across the WIMD quintiles.

	WIMD 1	WIMD 2	WIMD 3	WIMD 4	WIMD 5	no valid WIMD	All
Gender (% of females)	50.1	50.2	50.3	50.2	50.4	45.7	50.2
Age mean	36.7	38.9	40.7	41.3	41.9	45.6	39.9
SD	(22.0)	(22.2)	(22.4)	(22.5)	(22.5)	(21.8)	(22.4)
Prevalence (%) of ever-diagnosed asthma	12.0	11.4	11.0	10.9	11.0	9.6	11.2
Prevalence (%) of ever-diagnosed asthma, currently treated asthma	7.4	7.0	6.8	6.6	6.5	6.2	6.8

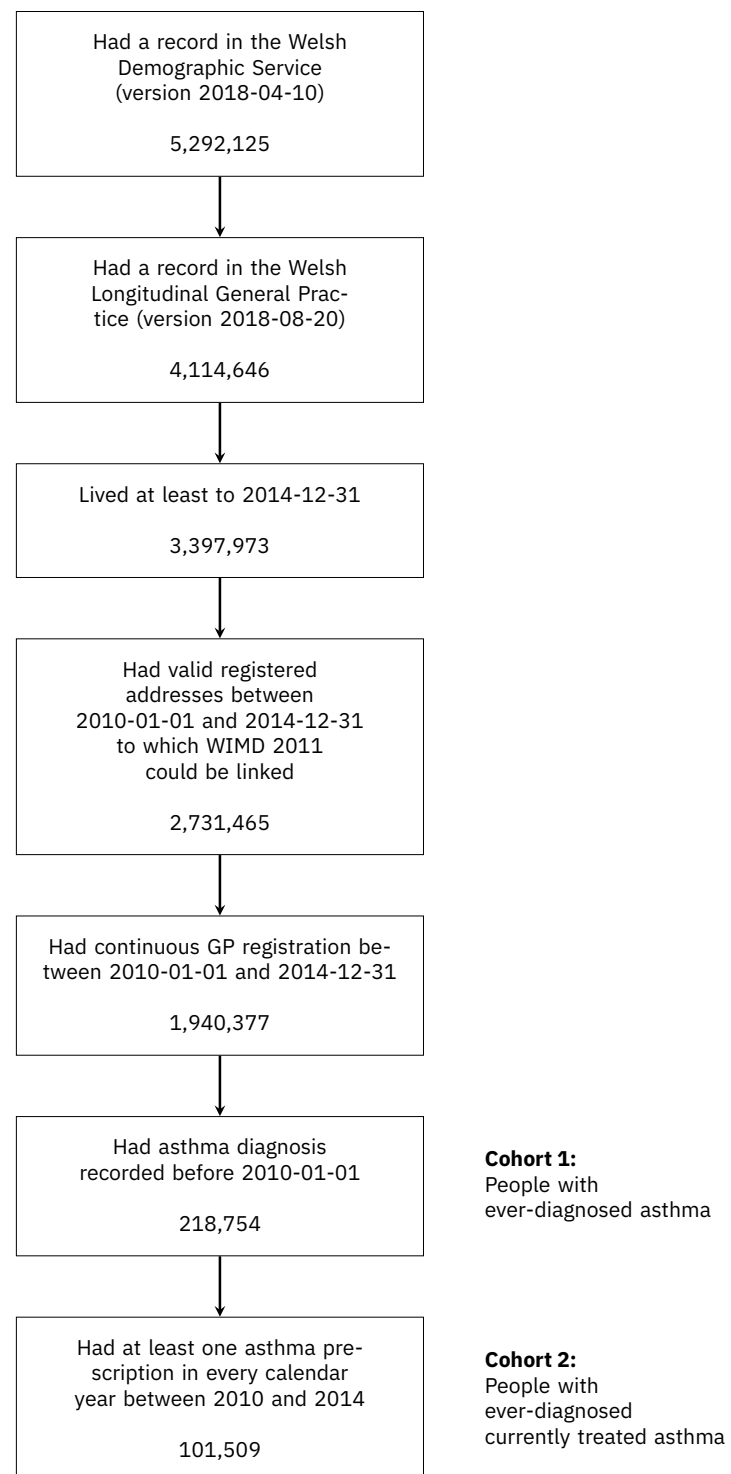


Figure 5.2: A flowchart of case selection.

The first cohort included 218,754 patients with ever-diagnosed asthma, while the second cohort included 101,509 patients with ever-diagnosed currently treated asthma. [Table 5.2](#) shows basic characteristics of both cohorts.

Table 5.2: Characteristics of the study cohorts: **Cohort 1.**

		WIMD 1	WIMD 2	WIMD 3	WIMD 4	WIMD 5	All
Number of patients	N	49,597	43,681	43,570	37,228	44,678	218,754
	%	22.7	20.0	19.9	17.0	20.4	100.0
Gender	% Females	53.6	52.5	51.8	51.2	50.5	52.0
Age	mean	36.7	38.1	39.3	39.8	40.4	38.8
	SD	20.1	20.4	20.7	20.7	20.6	20.5
	median	33.9	35.9	37.4	38.1	39.2	37.0
	IQR	19.8-51.5	20.7-53.7	21.6-55.4	22.2-56.0	22.8-56.1	21.3-54.4
<i>Study outcomes</i>							
Asthma-related GP visits	N	179,149	165,094	198,431	176,769	150,756	870,199
	mean count	4.00	4.05	4.11	4.05	3.70	3.98
	% with count ≥ 1	68.6	69.0	69.3	68.8	67.3	68.6
Asthma reviews	N	111,065	97,277	97,656	82,053	98,721	486,772
	mean count	2.24	2.21	2.24	2.23	2.20	2.25
	% with count ≥ 1	62.4	63.1	63.3	63.0	62.6	62.9
Asthma related A&E visits	N	1,011	848	808	702	621	3,990
	mean count	0.020	0.019	0.019	0.019	0.014	0.018
	% with count ≥ 1	1.5	1.4	1.4	1.4	1.1	1.3
Asthma related hospitalisations	N	2,390	1,568	1,351	1,037	980	7,326
	mean count	0.048	0.036	0.031	0.028	0.022	0.033
	% with count ≥ 1	2.4	2.1	1.8	1.6	1.4	1.9
<i>Prescriptions</i>							
SABA inhalers	mean count	19.8	17.1	15.2	13.9	11.8	15.7
	% with count ≥ 1	71.8	71.4	71.1	70.6	69.6	70.9
ICS inhalers	mean count	6.5	5.8	5.5	5.4	5.0	5.7
	% with count ≥ 1	38.4	38.0	37.8	38.6	37.8	38.1
ICS-LABA combination inhalers	mean count	11.7	10.8	10.0	9.1	8.2	10.0
	% with count ≥ 1	35.9	34.9	34.3	32.7	31.1	33.8
Asthma medication ratio	mean	0.42	0.43	0.44	0.44	0.45	0.43
Theophylline	mean count	0.7	0.6	0.5	0.4	0.3	0.5
	% with count ≥ 1	1.9	1.7	1.4	1.3	0.8	1.4
Leukotriene receptor antagonists	mean count	2.2	2.1	1.8	1.6	1.5	1.8
	% with count ≥ 1	8.4	8.0	7.6	7.0	6.6	7.6
Oral corticosteroids	mean count	1.9	1.8	1.8	1.7	1.4	1.7
	% with count ≥ 1	31.8	30.6	30.7	29.7	26.6	29.9
% of patients with ≥ 1 asthma prescriptions in every <i>N</i> years of the 5-year follow-up period	0 years	28.5	28.6	28.4	28.9	30.0	28.9
	1 year	6.0	6.1	6.2	6.2	6.5	6.2
	2 years	5.4	5.5	5.6	5.8	5.7	5.6
	3 years	5.5	5.5	5.6	5.7	5.9	5.6
	4 years	5.1	5.1	5.3	5.7	5.7	5.4
	5 years *	49.4	49.1	49.0	47.8	46.3	48.3

* Patients in this row represent Cohort 2 (patients with ever-diagnosed currently treated asthma).

Table 5.2: Characteristics of the study cohorts (continued): **Cohort 2.**

		WIMD 1	WIMD 2	WIMD 3	WIMD 4	WIMD 5	All
Number of patients	N	19,760	23,574	20,657	20,471	17,047	101,509
	%	19.5	23.2	20.3	20.2	16.8	100
Gender	% Females	59.0	57.2	55.4	54.7	53.7	56.1
Age	mean	45.1	46.3	47.0	47.6	47.6	46.6
	SD	20.0	20.2	20.6	20.5	20.6	20.4
	median	46.5	47.8	48.6	48.7	48.8	48.0
	IQR	30.4-60.7	31.6-62.3	32.2-63.3	33.4-63.7	33.6-63.5	32.1-62.7
<i>Study outcomes</i>							
Asthma-related GP visits	N	155,206	137,387	138,218	114,315	125,401	670,527
	mean count	6.58	6.65	6.75	6.71	6.35	6.61
	% with count ≥ 1	98.0	98.0	98.3	98.2	98.7	98.2
Asthma reviews	N	90,765	79,034	78,574	65,497	78,302	392,172
	mean count	3.9	3.8	3.8	3.8	4.0	3.9
	% with count ≥ 1	93.6	94.5	94.6	95.0	96.3	94.7
Asthma related A&E visits	N	825	694	694	552	506	3,271
	mean count	0.035	0.034	0.034	0.032	0.026	0.032
	% with count ≥ 1	2.4	2.3	2.4	2.3	1.9	2.3
Asthma related hospitalisations	N	2,193	1,381	1,214	895	874	6,557
	mean count	0.093	0.067	0.059	0.053	0.044	0.065
	% with count ≥ 1	4.3	3.8	3.3	3.0	2.6	3.4
<i>Prescriptions</i>							
SABA inhalers	mean count	38.1	33.0	29.3	27.3	23.7	30.7
	% with count ≥ 1	98.6	98.1	97.4	97.5	97.5	97.8
ICS inhalers	mean count	12.2	10.9	10.4	10.4	9.9	10.8
	% with count ≥ 1	53.9	53.1	53.8	55.4	55.2	54.2
ICS-LABA combination inhalers	mean count	23.7	21.9	20.5	19.0	17.7	20.7
	% with count ≥ 1	64.9	63.5	62.3	60.3	59.3	62.2
Asthma medication ratio	mean	0.48	0.50	0.52	0.52	0.54	0.51
Theophylline	mean count	1.5	1.3	1.0	0.9	0.6	1.1
	% with count ≥ 1	3.8	3.5	2.9	2.7	1.8	3.0
Leukotriene receptor antagonists	mean count	4.5	4.3	3.6	3.4	3.2	3.8
	% with count ≥ 1	15.9	15.2	14.3	13.4	13.1	14.4
Oral corticosteroids	mean count	3.5	3.3	3.2	3.1	2.5	3.1
	% with count ≥ 1	52.0	49.7	49.1	48.6	44.7	49.0

* All patients in this cohort had one or more prescriptions in every year between 2010-2014.

Asthma prevalences in the source population across WIMD quintiles

The prevalence of ever-diagnosed asthma in the source population at the beginning of 2010 was 11.2%, ranging from 10.9 in the next least deprived areas (WIMD 4) and 11.0 in the least deprived areas (WIMD 5) to 12.0% in the most deprived areas (WIMD 1). The prevalence of ever-diagnosed, currently treated asthma during the same year was 6.8%, ranging from 6.2 in the least deprived areas (WIMD 5) to 7.4% in the most deprived areas (WIMD 1).

Distribution of WIMD quintiles

In both cohorts, the quintiles of the WIMD rank had comparable shares ([Table 5.2](#)), with the next least deprived quintile (WIMD 4) having the least proportion of individuals (17.0% and 16.8% in Cohorts 1 and 2) and the most deprived quintile (WIMD 1) having the highest proportion (22.7% and 23.2% in Cohorts 1 and 2).

Distribution of gender

Females were 52.0% of Cohort 1 and 56.1% of Cohort 2. In both cohorts, the more deprived areas had higher proportions of females—the gradient in Cohort 1 ranged from 50.5% in WIMD 5 areas to 53.6% in WIMD 1 areas and in Cohort 2 from 53.7% to 59.0% in those areas, respectively.

Distribution of age

Cohort 1 was younger than Cohort 2. The mean age in Cohort 1 was 38.8 years with a standard deviation (SD) of 20.5, a median of 37.0, and an inter-quartile range of 21.3-54.4 years. In Cohort 2, the mean age was 46.6 years (SD = 20.4), and the median was 48.0 with an inter-quartile range of 32.1-62.7 years.

There were more young people in areas of higher deprivation. The mean age ranged in Cohort 1 from 36.7 in the most deprived areas (WIMD 1) to 40.4 in the least deprived areas (WIMD 5) and in Cohort 2 from 45.1 in WIMD 1 areas to 47.6 in WIMD 5 areas. Kruskal-Wallis rank sum test confirmed the unequal distribution of age ranks across the five deprivation groups in both cohorts ($\chi^2 = 923.8$ and 231.0 for Cohorts 1 and 2, with p -values < 0.0001). [Figure 5.3](#) shows the distribution of age for each of the WIMD quintiles.

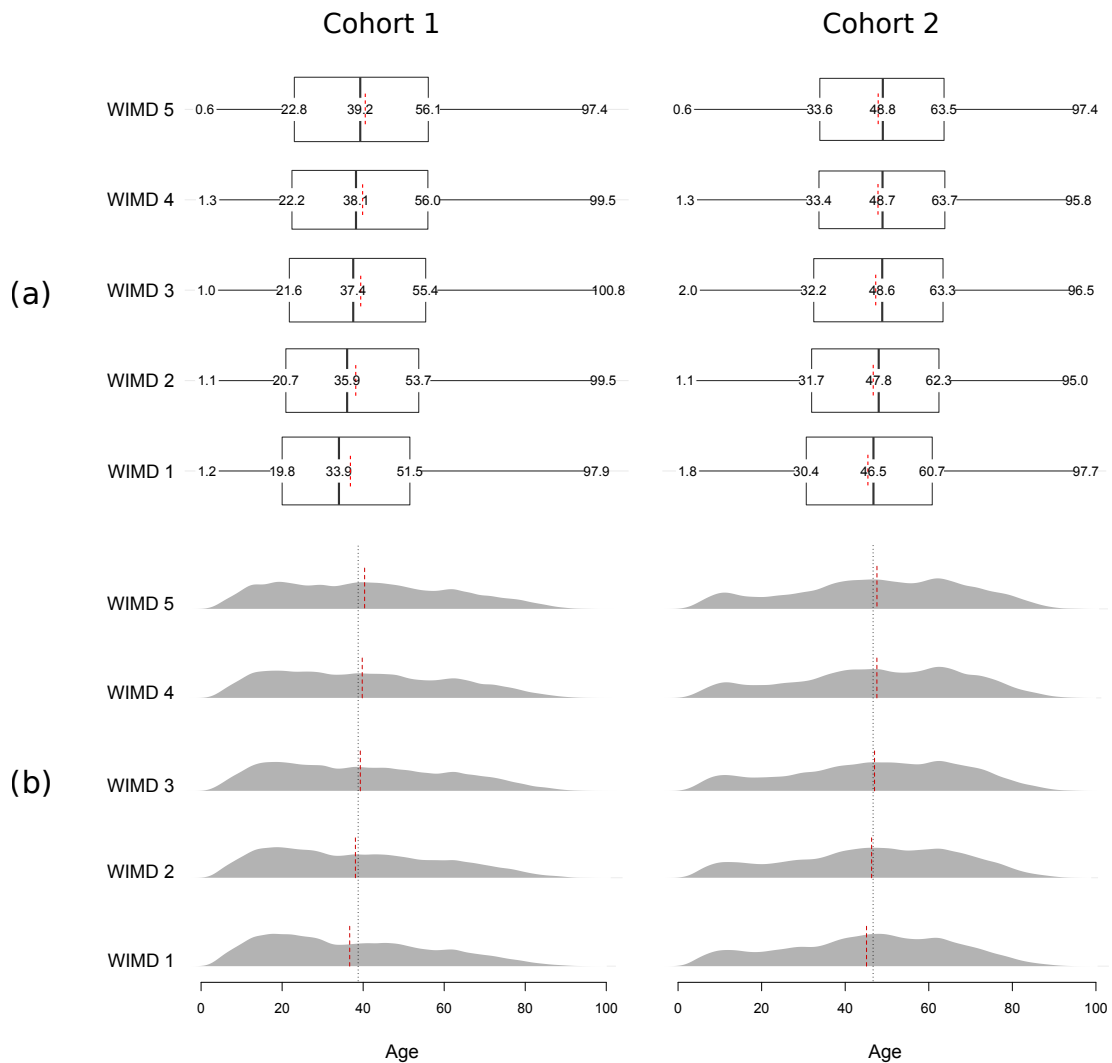


Figure 5.3: Distribution of age across WIMD overall rank quintiles in the study cohorts. (a) Boxplots showing the minimum, maximum, first and third quartiles, median (black lines), and mean (red dotted lines) of age for each WIMD quintile. (b) Beanplots showing the density of age distribution for each WIMD quintile. The large black dotted line represents the overall mean, and the smaller red dotted lines represent the per WIMD quintile means.

Distribution of asthma prescriptions across the deprivation quintiles

In Cohort 1, the more affluent areas had more patients with intermittent or no asthma treatment over the five follow-up years (Table 5.2). Conversely, the higher the deprivation, the higher the percentage of patients with continuous asthma prescriptions.

In Cohort 2, where all patients received asthma prescriptions in each of the follow-up years, there was a remarkable gradient of more prescriptions with higher deprivation. For example, in the most deprived areas the average number of SABA inhalers over five years per patient was 38.1 compared with 23.7 in the most affluent areas, while the gap for ICS inhalers was 12.1 to 9.9, for ICS-LABA combi-

nation inhalers was 23.7 to 17.7, and for oral corticosteroids (OCS) was 3.5 to 2.5, respectively, between those areas. It is worth noting that the average controller to controller-and-rescuer medication ratio in the most deprived areas (0.48) was lower than that in the least deprived areas (0.54).

Percentages of patients with outcome events overall

68.6% and 62.9% of Cohort 1 patients had recorded asthma GP visits and asthma reviews, respectively, in the 5-year follow-up period. In comparison, 98.2% and 94.7% of patients in Cohort 2 had no recorded asthma GP visits and asthma reviews, respectively. In both cohorts, however, only very few patients had recorded asthma-related A&E visits and hospital admissions (1.3% and 1.9% in Cohort 1, and 2.3% and 3.4% in Cohort 2). Histograms of the outcome variable counts illustrate the skewed data in both cohorts ([Figure 5.4](#)).

Average number of outcome events overall

The average counts of outcome events over the five-year follow-up period were significantly higher in Cohort 2 than in Cohort 1; on averages there were 4.0 and 6.6 asthma-related GP visits, 2.2 and 3.9 asthma reviews, 0.018 and 0.032 asthma-related A&E visits, and 0.033 and 0.065 asthma-related hospital admissions in Cohorts 1 and 2, respectively.

Average counts of outcome events in each WIMD quintile

Without adjustment to age group and gender, in Cohort 1, the most deprived areas had on average more outcome events per patient than the least deprived areas ([Table 5.2](#)): 4.00 vs. 3.70 asthma-related visits to GPs, 2.24 vs. 2.21 asthma reviews, 0.020 vs. 0.014 asthma-related visits to A&E, and 0.048 vs. 0.022 asthma-related hospital admissions.

In Cohort 2, on average, the most deprived areas had also more asthma-related visits to GPs (6.58 vs. 6.35), more asthma-related visits to A&E (0.035 vs. 0.026), more asthma-related hospital admissions (0.093 vs. 0.044), but less asthma reviews (3.85 vs. 3.96) per patient than the least deprived areas.

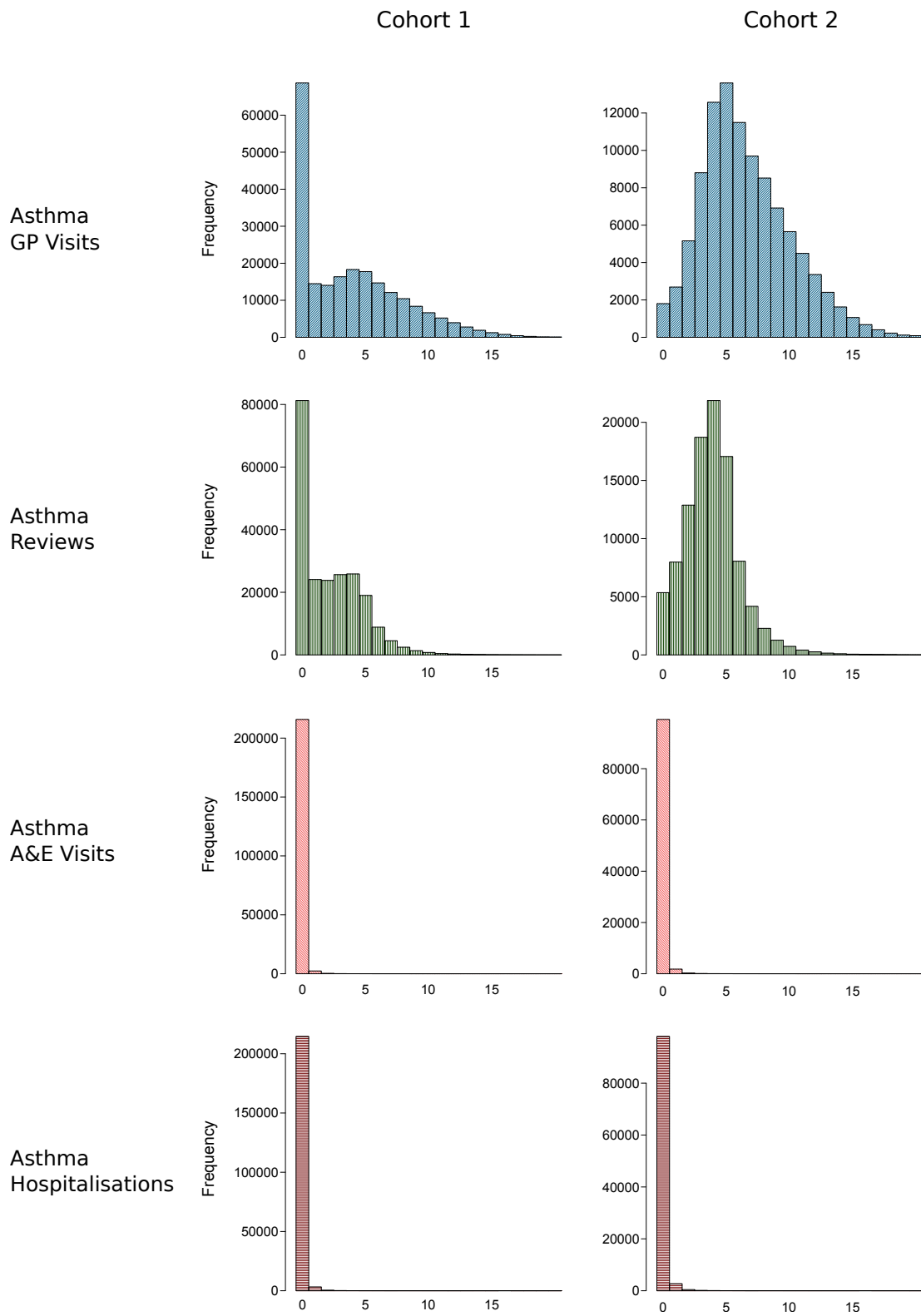


Figure 5.4: Histograms of outcome event counts in the study cohorts.

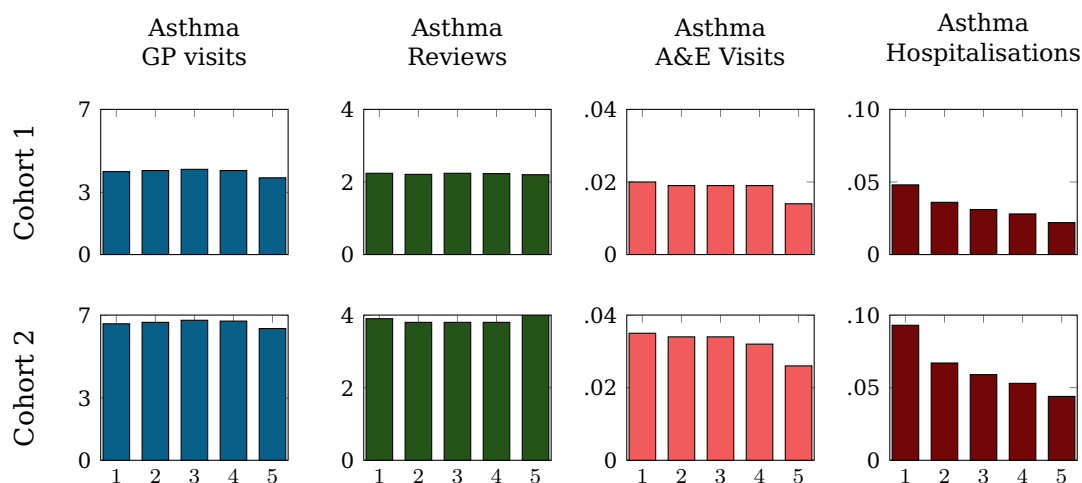


Figure 5.5: Average counts of the outcome events between 2010 and 2014 per WIMD rank quintiles in Cohort 1 and Cohort 2.

Gradients

In both cohorts, despite the gap in asthma GP visits and reviews on the two sides of the deprivation scale, there were no consistent gradients across the whole deprivation scale (Figure 5.5). The three middle deprivation areas (WIMD 2, 3, and 4) had on average more asthma GP visits than both the most (WIMD 1) and least deprivation (WIMD 5) areas, with the middle deprivation areas having the highest averages in both cohorts. In contrast, for asthma reviews, the WIMD 2, 3 and 4 areas had smaller average counts than WIMD 1 and 5 areas in Cohort 2, with no clear pattern in Cohort 1.

In both cohorts, although the average counts of asthma A&E visits decreased with less deprivation, there was no gradient since the least deprived areas (WIMD 5) had significantly smaller average counts than the other four more deprived areas.

Finally, there were clear gradients of more asthma hospitalisations with higher deprivation in both cohorts.

5.4.2 Zero-inflated negative binomial regression (ZINB) models

The outputs of the four ZINB models for each study cohort are shown in Table 5.3.

5.4.2.1 Incidence rate ratios of study outcomes across deprivation quintiles

Following adjustments for age group and gender, I found statistically significant differences in incidence rates of the outcome events between the WIMD quintiles.

For asthma-related GP visits, all the first four WIMD quintiles had statistically significant IRRs compared to the fifth quintile. Over the five-year follow-up period, there were 7.8% more predicted events per patient in the most deprived areas (WIMD 1) than in the least deprived areas (WIMD 5), with predicted counts of 3.94 and 3.65, respectively. In Cohort 2, the difference in predicted events between the WIMD 1 and 5 areas was smaller (3.6%), with predicted counts of 6.37 and 6.15 in those areas, respectively. The WIMD 2, 3, and 4 areas had slightly higher IRRs than WIMD 1 areas in both cohorts.

For asthma reviews, the only statistically significant IRR at the 0.05 level in Cohort 1 was 1.037 (1.025-1.048) for WIMD 1; there were 3.7% more predicted events in the least deprived areas (WIMD 1, 2.34 events) compared to the least deprived areas (WIMD 5, 2.25). In Cohort 2, all the IRRs were statistically significant and slightly less than 1, with an IRR of 0.99 (0.978-0.999) for the most deprived areas compared to the least deprived areas (with predicted number of asthma reviews of 3.81 and 3.86, respectively).

For asthma A&E visits, the IRRs for WIMD 1 to 4 areas in Cohort 1 were all statistically significant, ranging from 1.368 for the most deprived areas (WIMD 1) to 1.298 for the middle deprivation areas (WIMD 3). There were 36.8% more predicted asthma-related A&E visits in the most deprived areas than in the least deprived areas (0.023 vs. 0.016 visits). In Cohort 2, the IRRs were slightly lower than those in Cohort 1, ranging from 1.229 for the most deprived areas (WIMD 1) to 1.257 for the middle deprivation areas (WIMD 3). There were 22.9% more predicted asthma-related A&E visits in the most deprived areas than in the least deprived areas (0.043 vs. 0.035 visits).

In both cohorts, the clear gaps in the predicted asthma-related GP visits and A&E visits were between the four more deprived quintiles (WIMD 1-4) together, which showed relatively similar estimates, and the least deprived quintile (WIMD 5).

Lastly, a clear and steep social gradient existed for asthma-related hospital admissions. In Cohort 1, in the most deprived areas, there were 123.2% more pre-

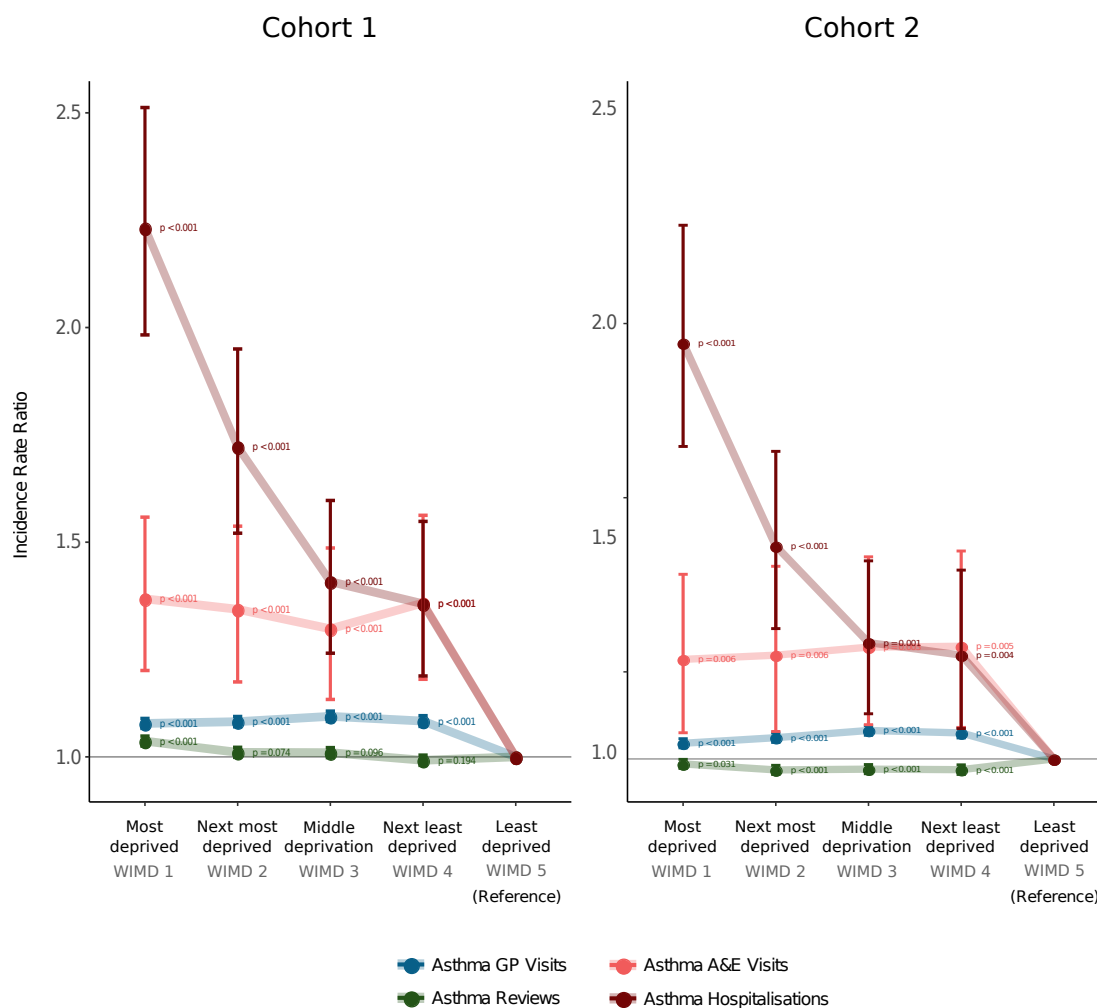


Figure 5.6: Incidence rate ratios with 95% confidence intervals in both study cohorts for each of the outcome variables in each of the deprivation quintiles relative to the least deprived quintile, controlled for age group and gender.

dicted asthma-related hospital admissions than in the least deprived areas (0.074 vs 0.033 predicted asthma admissions, IRR = 2.232 [1.983-2.512]). The gap was slightly smaller in Cohort 2, where there were 95.5% more predicted asthma-related hospital admissions in the most deprived areas compared to the least deprived areas (0.139 vs. 0.071 predicted asthma admissions, IRR = 1.955 [1.718-2.226]).

For Cohort 1, all p-values for the IRRs of the outcome variables were less than 0.0001, except those for in the asthma reviews model which were above 0.05. For Cohort 2, most of the p-values were less than 0.001; the IRRs of asthma reviews and A&E visits between WIMD 1 vs. WIMD 5 had p-values of 0.031 and 0.006, respectively.

A visualisation of IRRs of the outcome variables for the WIMD rank quintiles is shown in [Figure 5.6](#).

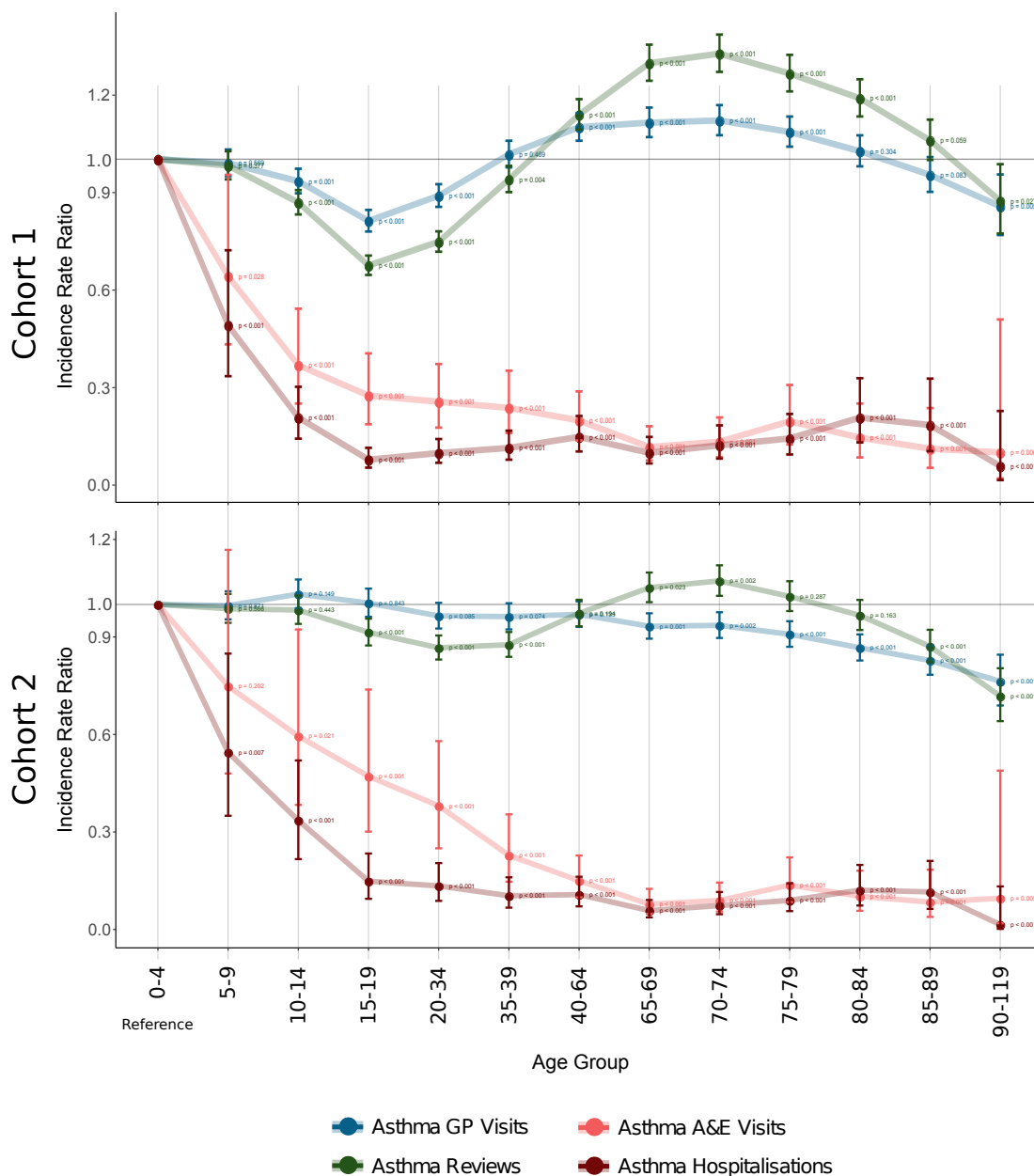


Figure 5.7: Incidence rate ratios with 95% confidence intervals for each of the outcome variables in age groups relative to the youngest age group (0-4 years), controlled for deprivation quintile and gender.

5.4.2.2 Incidence rate ratios of study outcomes for age groups and gender

IRRs of the outcome variables across age groups are shown in [Figure 5.7](#).²

In both study cohorts, the variations of IRRs, relative to the reference age group 0-4 years, between age groups were relatively small for asthma GP visits and asthma reviews but were large for A&E visits and admissions.

²It is worth noting that since age was calculated at the beginning of the follow-up period, IRRs for patients in a given age group covered their next five years of life; e.g., for patients in the 15-19 age group, the IRRs covered periods starting at the age of 15 to 19 and ending at the age of 20 to 24, depending on patient's age at the beginning of the follow-up period.

Table 5.3: Outputs of the zero-inflated negative binomial models. The values shown include the incidence rate ratios and their 95% confidence intervals.

Cohort 1

Outcome Variable ~ WIMD Quintile + Age Group + Gender 1											
Outcome variables											
Asthma-related GP visits			Asthma reviews			Asthma-related A&E visits			Asthma-related hospitalisations		
IRR	sig	95% CI	IRR	sig	95% CI	IRR	sig	95% CI	IRR	sig	95% CI
<i>Deprivation Quintile</i>											
<i>(reference level: WIMD 5)</i>											
WIMD1	1.078 ***	1.066-1.090	1.037 ***	1.025-1.048	1.368 ***	1.201-1.558	2.232 ***	1.983-2.512			
WIMD2	1.082 ***	1.070-1.094	1.011 .	0.999-1.022	1.344 ***	1.175-1.537	1.722 ***	1.521-1.950			
WIMD3	1.094 ***	1.082-1.107	1.010 .	0.998-1.022	1.298 ***	1.134-1.487	1.408 ***	1.241-1.597			
WIMD4	1.083 ***	1.071-1.096	0.992	0.980-1.004	1.358 ***	1.181-1.563	1.357 ***	1.188-1.548			
<i>Age Group</i>											
<i>(reference level: 0-4 years)</i>											
5-9	0.988	0.948-1.030	0.980	0.938-1.024	0.641 *	0.431-0.952	0.490 ***	0.334-0.720			
10-14	0.932 ***	0.895-0.971	0.867 ***	0.830-0.905	0.367 ***	0.249-0.541	0.206 ***	0.142-0.301			
15-19	0.810 ***	0.778-0.844	0.674 ***	0.645-0.704	0.274 ***	0.186-0.404	0.077 ***	0.053-0.113			
20-34	0.888 ***	0.854-0.923	0.746 ***	0.716-0.778	0.255 ***	0.176-0.371	0.097 ***	0.068-0.140			
35-39	1.015	0.975-1.057	0.938 **	0.899-0.979	0.236 ***	0.159-0.350	0.113 ***	0.077-0.166			
40-64	1.098 ***	1.057-1.142	1.137 ***	1.091-1.185	0.198 ***	0.136-0.287	0.147 ***	0.102-0.211			
65-69	1.112 ***	1.068-1.159	1.295 ***	1.241-1.352	0.116 ***	0.075-0.180	0.098 ***	0.066-0.147			
70-74	1.119 ***	1.074-1.166	1.324 ***	1.268-1.383	0.132 ***	0.085-0.207	0.121 ***	0.081-0.183			
75-79	1.084 ***	1.039-1.131	1.263 ***	1.208-1.321	0.195 ***	0.124-0.307	0.142 ***	0.093-0.217			
80-84	1.025	0.978-1.073	1.187 ***	1.131-1.246	0.145 ***	0.084-0.249	0.206 ***	0.130-0.327			
85-89	0.952 .	0.900-1.007	1.058 .	0.998-1.122	0.111 ***	0.052-0.236	0.184 ***	0.103-0.326			
90-119	0.855 **	0.767-0.953	0.872 *	0.772-0.985	0.099 **	0.019-0.508	0.057 ***	0.014-0.227			
<i>Gender</i>											
<i>(reference level: male)</i>											
Female	1.042 ***	1.034-1.049	1.100 ***	1.092-1.109	1.708 ***	1.568-1.860	2.310 ***	2.132-2.503			
Intercept	5.051 ***	4.857-5.252	2.994 ***	2.873-3.121	0.042 ***	0.029-0.061	0.088 ***	0.061-0.127			
Log likelihood		-520934.8		-415837.7		-17497.8		-24249.9			

Significance codes: . p<0.1; * p<0.05; ** p<0.01; *** p<0.001; IRR: incidence rate ratio.

Table 5.3: Outputs of the zero-inflated negative binomial models (continued). The values shown include the incidence rate ratios, statistical significance, and their 95% confidence intervals.

Cohort 2

Outcome Variable ~ WIMD Quintile + Age Group + Gender 1											
Outcome variables											
Asthma-related GP visits			Asthma reviews			Asthma-related A&E visits			Asthma-related hospitalisations		
IRR	sig	95% CI	IRR	sig	95% CI	IRR	sig	95% CI	IRR	sig	95% CI
<i>Deprivation Quintile</i>											
<i>(reference level: WIMD 5)</i>											
WIMD1	1.036 ***	1.025-1.046	0.989 *	0.978-0.999	1.229 **	1.060-1.424	1.955 ***	1.718-2.226			
WIMD2	1.049 ***	1.038-1.060	0.975 ***	0.964-0.985	1.238 **	1.063-1.442	1.489 ***	1.299-1.706			
WIMD3	1.065 ***	1.054-1.076	0.977 ***	0.966-0.987	1.257 **	1.079-1.464	1.267 ***	1.104-1.455			
WIMD4	1.060 ***	1.048-1.071	0.975 ***	0.964-0.986	1.259 **	1.072-1.478	1.238 **	1.070-1.434			
<i>Age Group</i>											
<i>(reference level: 0-4 years)</i>											
5-9	0.996	0.954-1.041	0.987	0.943-1.032	0.749	0.480-1.168	0.545 **	0.350-0.849			
10-14	1.032	0.989-1.077	0.983	0.940-1.027	0.595 *	0.383-0.923	0.335 ***	0.216-0.520			
15-19	1.004	0.962-1.049	0.914 ***	0.874-0.956	0.471 **	0.301-0.738	0.148 ***	0.094-0.234			
20-34	0.965	0.926-1.005	0.867 ***	0.830-0.904	0.380 ***	0.250-0.579	0.134 ***	0.088-0.204			
35-39	0.963	0.923-1.004	0.876 ***	0.839-0.915	0.228 ***	0.147-0.354	0.104 ***	0.067-0.160			
40-64	0.969	0.931-1.009	0.973	0.933-1.014	0.150 ***	0.099-0.227	0.107 ***	0.071-0.162			
65-69	0.933 **	0.895-0.973	1.052 *	1.007-1.098	0.078 ***	0.048-0.125	0.058 ***	0.037-0.091			
70-74	0.936 **	0.897-0.976	1.072 **	1.027-1.120	0.089 ***	0.055-0.144	0.073 ***	0.047-0.115			
75-79	0.908 ***	0.870-0.949	1.025	0.980-1.072	0.137 ***	0.084-0.222	0.089 ***	0.056-0.142			
80-84	0.867 ***	0.828-0.908	0.966	0.921-1.014	0.102 ***	0.057-0.181	0.121 ***	0.074-0.199			
85-89	0.828 ***	0.784-0.874	0.871 ***	0.822-0.922	0.084 ***	0.039-0.184	0.116 ***	0.063-0.211			
90-119	0.763 ***	0.689-0.845	0.718 ***	0.641-0.804	0.096 **	0.019-0.489	0.015 ***	0.002-0.132			
<i>Gender</i>											
<i>(reference level: male)</i>											
Female	1.031 ***	1.025-1.038	1.061 ***	1.054-1.069	1.597 ***	1.448-1.760	2.135 ***	1.954-2.334			
Intercept	6.536 ***	6.275-6.808	4.068 ***	3.900-4.244	0.085 ***	0.056-0.129	0.208 ***	0.137-0.315			
Log likelihood		-267255.4		-218843.2		-12495.4		-18616.1			

Significance codes: • p<0.1; * p<0.05; ** p<0.01; *** p<0.001; IRR: incidence rate ratio.

The IRRs patterns across age groups for asthma reviews, A&E visits, and admissions were similar between the two cohorts, which, however, differed in the patterns of asthma GP visit IRRs.

In Cohort 1, the rates of asthma GP visits gradually decreased with age in the youngest age groups with a minimum of 0.81 in the 15-19 group (relative to the reference age group 0-4) year before increasing again towards a maximum of 1.12 in the 70-74 group and decreasing again to 0.85 in the 90-119 group. In contrast, in Cohort 2 the estimates peaked at 1.03 in the 10-14 before decreasing slightly and steadily with older age, reaching 0.76 for the age group 90-119.

In Cohort 1, asthma review IRRs pattern across age groups was similar to that of asthma GP visits, although it had a higher magnitude of variation. In Cohort 2, the pattern was similar to that in Cohort 1 but with less variation between age groups.

For asthma-related A&E visits, there was a general gradient of decreasing IRRs with older age in both cohorts. In Cohort 1, the gradient was steep in the younger age groups, i.e. under 19-year old, before flattening between the age groups 15-18 and 35-39 years and decreasing slightly further in the older age groups. However, in Cohort 2, the gradient was less steep in the younger groups and steadily decreased up to the 65-68 age group.

The two cohorts had almost similar patterns of asthma-related hospitalisation IRRs, showing steeper, decreasing gradients with age in children, before stabilising at the 15-18 age group with slight fluctuations over the older groups.

Females had overall 4% more predicted asthma-related GP visits, 10% more asthma reviews, 71% more asthma related A&E visits, and 131% more asthma-related hospitalisations — p-values for gender differences were all less than 0.001.

5.4.2.3 Model fit

In both study cohorts, the quantile-quantile plots for the raw residuals for the four ZINB models showed that the raw residuals overall followed a normal distribution (Figure 5.8). However, while there was a little right skewness for the asthma-related GP visit models, the right skewness was clearer in the models of the other three outcome variables—asthma reviews, asthma-related A&E visits and asthma-related hospitalisations—especially for hospitalisations.

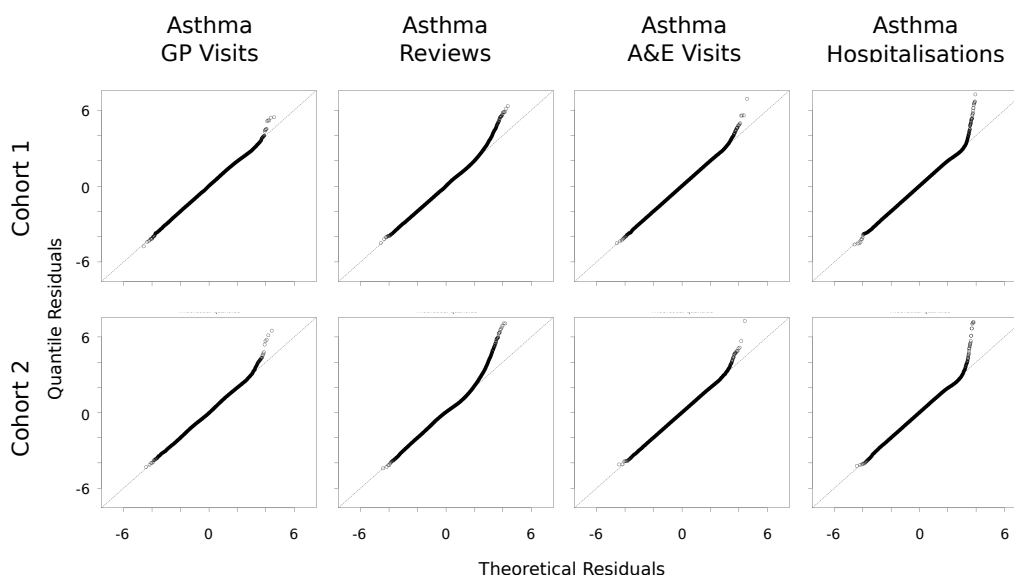


Figure 5.8: Quantile-quantile plot for the model residuals.

Rootograms showed an alternative visualisation of the model fit, focusing of the deviation between the observed and predicted frequencies for the counts of the outcome variables (Figure 5.9). The deviations between the observed and predicted frequencies were relatively small, indicating a good overall fit for each of the four models in both cohorts.

5.4.2.4 Sensitivity analysis

The findings presented above in both cohorts were produced for patients with complete GP registration in Wales between 2010 and 2014. However, among 297,976 and 111,253 patients with “ever-diagnosed asthma” and “ever-diagnosed currently treated asthma over the follow-up period” in the source population, 79,222 (26.6%) and 9,744 (8.8%) patients, respectively, had incomplete GP registration over the five-year follow-up period.

By including those previously excluded patients to Cohort 1 and Cohort 2, the overall patterns of IRRs across the WIMD quintiles did not change significantly in both cohorts. Between the most and least deprived areas, most IRRs became slightly higher than but still close to those in the original cohorts with complete GP registrations. The IRRs of asthma GP visits, reviews, A&E visits and hospitalisations became 1.102 [1.090-1.114], 1.064 [1.052-1.075], 1.542 [1.377-1.726], and 2.254 [2.033-2.498] in the ever diagnosed asthma cohort (Cohort 1), and 1.032 [1.022-1.042], 0.987 [0.977-0.997], 1.242 [1.081-1.427], and 1.889 [1.671-2.135] in the ever-diagnosed, currently-treated asthma cohort (Cohort 2), respectively.

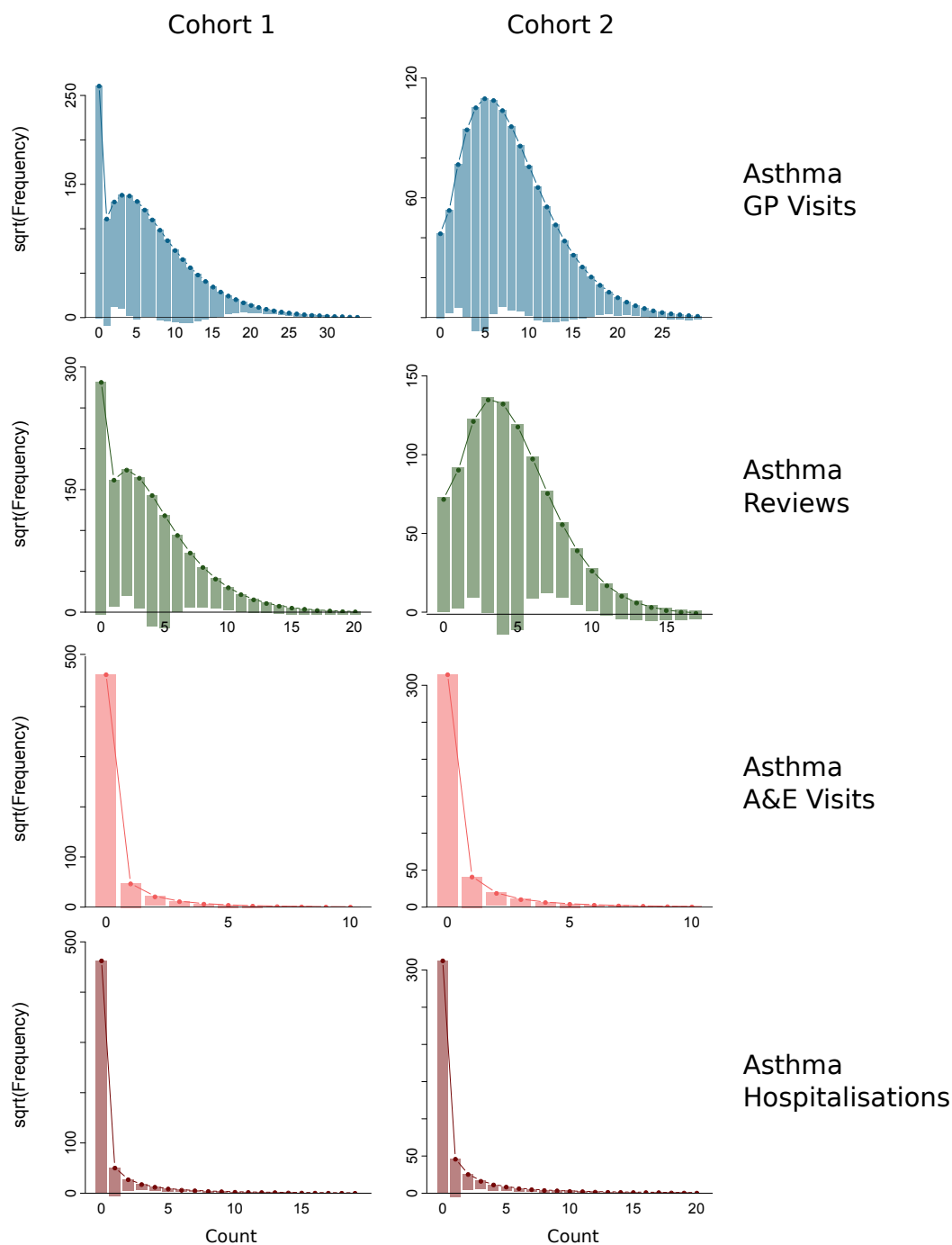


Figure 5.9: Rootograms illustrating the goodness of fit for the zero-inflated negative binomial models for the four outcome variables in both study cohorts.

The sensitivity analysis confirmed that the observed differences in asthma-related primary and secondary care utilisation between the most and the least deprived areas existed regardless of the continuity of GP registrations.

5.5 Discussion

5.5.1 Summary of findings

I identified patterns of variation in asthma-related healthcare utilisation across the deprivation scale as well as between age and gender groups in Wales.

Compared to the least deprived areas, there were wide gaps in asthma-related ED visits (36.8%) and hospital admissions (123.2%) in the most deprived areas despite a small excess in asthma-related primary care use, including asthma reviews, in these areas. However, these gaps were slightly smaller (22.9% and 95.5%, respectively) among asthma patients with continuous asthma prescriptions throughout the study period.

There was a gradient of increasing asthma-related hospital admissions by increased deprivation. However, there were no consistent gradients for asthma-related GP visits, reviews, and A&E department visits across the deprivation scale. Instead, for asthma-related GP visits and A&E department visits, the contrast was between the more affluent areas, which had the lowest incidence rates, and the other more deprived areas together, which had higher incidence rates.

Notably, there were also significant differences by gender in asthma-related healthcare utilisation. Although females with asthma had a modest excess in disease-related primary care contacts including asthma reviews, they were far more likely to attend A&E departments and to be admitted to hospitals due to asthma.

Age groups also showed wide variations in asthma-related healthcare utilisation. Among patients who continuously received asthma treatments, asthma-related GP visits decreased with older ages. However, asthma patients between the age of 10 and 35 were less likely to undergo asthma reviews compared to those in the age groups between 40 and 85. Finally, younger patients were more likely to visit A&E departments and to be admitted for asthma than older patients.

There was a remarkable gradient of more prescriptions but lower proportion of asthma controller medications with higher deprivation.

The above findings were not affected by whether patients were treated for asthma throughout the follow-up period or by their continuity of GP registrations.

5.5.2 Interpretation in the light of previous studies

5.5.2.1 Comparison with previous studies

The associations that I found between socioeconomic status and asthma outcomes in Wales were consistent with other studies in Wales, England, and elsewhere. Asthma outcomes were shown to be associated with both area-based and individual-level deprivation [349, 350].

Asthma severity has been widely linked to socioeconomic status, where severe disease was found to be more prevalent in people with lower socioeconomic status [351]. Lower socioeconomic status also increased the risk of poor asthma control and the incidence of exacerbation, independent of disease severity [349, 352, 353]. Also, asthma patients with low socioeconomic groups, measured mainly through education level, had higher asthma morbidity, including higher incidence of exacerbations [352] and worse asthma control independently from the disease severity. One explanation of the poorer asthma control among patients with low socioeconomic status is that this group has poorer medication adherence [354].

In my study, the prevalences of both ever-diagnosed asthma and ever-diagnosed, currently treated asthma increased in the more deprived areas. Similar findings were reported in England [350]. However, the poorer asthma outcomes associated with higher deprivation were contrasted, in other studies, with the reports of increased disease incidence with lower deprivation [324, 355]. In one study, asthma-related hospitalisation rates were correlated with the prevalence of chronic phlegm and indoor exposure of second-hand smoke but the variations of these rates between socioeconomic groups were not explained by variations in the prevalences of asthma or wheezing [324].

Another study, however, reported contradicting findings that no statistically significant difference was found in the number of asthma exacerbations between patients in low and medium/high socioeconomic status groups [356]. That study reported that the less deprived people had lower secondary care utilisation after exacerbations, which might suggest they received better management and had better self-management than the most deprived groups [356].

5.5.2.2 WIMD mainly describes areas, and to a lesser extent, individuals

The Welsh Index of Multiple Deprivation is an area-based measure, i.e. an ecological variable. Therefore, associations studied using the WIMD index may not necessarily hold at the individual level [357, 358]. Essentially, the findings in this chapter highlighted variations between small areas of different deprivation ranks. While deprivation indicators for a given small area represent the average levels of disadvantage of persons living there, those persons may widely differ in their own characteristics. Therefore, depending on the extent of those variations, group-level associations might not be valid for individuals who were far from the average characteristics.

While the Income, Employment, Health, Education, and Housing domains of the WIMD are constructed based on average scores of individuals living in a given area, the other domains, i.e. Access to Services, Community Safety, Physical Environment domains, are by definition roughly equal for all residents in that area. Therefore, the findings in this chapter could be carefully extended to the individual level, i.e. asthma patients living in the most deprived areas had worse asthma control despite having slightly more primary care contact.

5.5.2.3 Asthma-related emergency department visits and hospitalisations usually indicate worse asthma severity and control

Asthma-related ED visits and hospital admissions have been widely used as proxy measures for asthma severity, control, and exacerbations [102]. Asthma control is the extent to which the disease symptoms and future risks are sufficiently eliminated or reduced through treatment to an acceptable target [1, 2, 16]. Suboptimal asthma control may result from inadequate secondary and tertiary prevention for asthma, such as inappropriate prescribing, inadequate asthma review, and suboptimal self-management and patient education. Accordingly, the findings in this chapter suggest that asthma patients in the least deprived areas could have not only more severe disease (i.e. requiring a higher level of medication to achieve control) but also that disease control (i.e., achieving satisfactory symptom control) was poorer overall (e.g., due to inadequate treatment). The difference in the average asthma medication ratio and asthma reviews between the least and the

most deprived areas may partly explain the gap in asthma-related A&E visits and admissions between those areas.

However, not all the asthma-related visits to A&E departments indicate real medical emergencies or worsening in the disease control. Instead, in areas of higher deprivation levels in particular, the higher number of such visits might not be completely due to genuine need to visit these facilities. There are a wide range of reasons why patients may bypass their GPs to visit A&E departments, i.e. using A&E departments as primary care facilities [359]. Inaccessibility to GP practices and pharmacies, which were accounted for in the Access to Services domain of the WIMD index, might play a role in the increased A&E visits in the more deprived areas.

A study in the United States found that patients with insufficient health literacy made more return visits in 14 days to A&E departments than those with adequate health literacy [360]. Subsequently, inadequate health literacy in areas of higher deprivation could be contributing to the higher asthma-related A&E visits found in this chapter.

Overcrowded GP practices in the more deprived areas may also contribute to the increased A&E visits in these areas. A recent report investigating the pressure on general practices in the UK had found that general practices in areas with higher levels of deprivations faced higher pressure than those in lower deprivation areas [361]. In that report, residents of higher deprivation areas used health services more likely and frequently than those in lower deprivation areas. In East London, for a given age group after the third decade of age, areas with higher deprivation had higher rates of GP consultations per patient compared to areas with lower deprivation [362]. This suggests that the IRR of 1.31 in asthma A&E visits between the most deprived and the least deprived areas may overestimate the differences in asthma control between them.

5.5.2.4 Why did the most deprived asthma patients have more GP visits?

The small excess of 8% in primary care contacts in the most deprived groups compared to the least deprived might be due to several factors. The intrinsic disease severity might be contributing to this gap. It has been found that asthma patients in the most deprived areas often had more severe disease than those in the least deprived areas [351, 355].

The excess could be also, in part, due to poorer disease control in the most deprived areas, leading to higher use of primary and secondary care. Unscheduled GP visits for asthma often indicate suboptimal disease control and/or exacerbations. However, in the SAIL Databank, it was not possible to directly distinguish between scheduled and unscheduled GP visits.

Another potential factor for the excess in asthma-related GP visits in the most deprived areas is lower health literacy. This might have hindered successful asthma self-management, and reduced patient's adherence to treatment and engagement in decision making [363]. Asthma patients with lower levels of education were found to use more short-acting bronchodilators and less controller medications, leading to poorer asthma control [364]. Lower health literacy can therefore lead to higher dependency on GPs and an increasing tendency to request appointments.

5.5.3 Study strengths

The main strength of this work was the use of objective real-world data at the individual patient level. The GP dataset extract that I used had a coverage of almost 80% of practices in Wales at the time of analysis. Through this dataset, I was able to identify and study a very large proportion of all asthma patients living in Wales between 2010 and 2014. In addition, I used an in-house developed algorithm that identified patients with gaps of more than 30 days that might result from either de-registration by the patient or the practice not sending data to the SAIL Databank. By excluding those patients from the study population, I minimised the possibility of missing GP events due to these reasons. However, it was possible that patients with gaps in their GP registrations were less concerned about their health, which could be due to inadequate health literacy which was in turn associated with lower education attainment and worse health outcomes [365]—characteristics suggesting higher deprivation levels. Alternatively, it was possible that those patients could have had milder asthma, or simply did not reside in Wales for the whole follow-up period. I demonstrated through a sensitivity analysis that the exclusion of those patients led to only small reductions to the IRRs that did not alter the conclusions.

Another strength is that the 2011 WIMD index incorporated a comprehensive range of deprivation domains for small areas in Wales, representing a helpful tool for multifaceted measurement of socioeconomic status.

In Wales, the National Health Services provides free-of-charge medical services, including prescriptions, to Welsh residents. This could be an advantage in this study by avoiding possible bias that could have happened if those services were paid for, completely or partly, by patients. Direct contributions to health care costs required by patients, especially with low income, may lead to less healthcare utilisation in order to avoid costs [366, 367]. For example, a study in the United States (US) suggested that higher proportions of consumed out-of-pocket asthma medication costs to household income were associated with higher exacerbation risk in children [353].

5.5.4 Study limitations

5.5.4.1 Case definitions for asthma-related A&E visits and hospitalisations were not validated

To my knowledge, there are currently no validated algorithms that would have allowed me to accurately identify asthma-related A&E visits and hospitalisations in the corresponding datasets in the SAIL Databank. Data recording and coding practices at inpatient and A&E departments may vary across Wales [297, 368]. For ED visits, the recorded diagnosis may not be final, and may not always correspond to the reason of the visit [368]. In the PEDW dataset, hospital episodes had 14 available fields for recording admission diagnoses: one primary diagnosis, one subsidiary diagnosis, and 12 secondary diagnoses [369].

In my analysis, asthma-related hospital admissions included any hospital episodes in which asthma was recorded as an admission diagnosis in any of the 14 available diagnosis positions. However, this definition of asthma-related hospital admission may have overestimated the number of actual events. In some cases, it was possible that asthma was recorded as a secondary indication for admission. Alternatively, asthma could have been recorded despite not being a reason for admission (e.g., it was well controlled prior or during the admission, and did not need specialist care hospital stay). Therefore, it is possible that some counted hospital admissions in my analysis were not related to asthma and therefore should have been excluded. Nonetheless, I assumed that the algorithm that I used to identify asthma-related hospitalisations equally inflated, if any, the estimated counts

across the five WIMD rank quintiles and therefore would not significantly affect the estimated IRRs.

In contrast, querying asthma diagnosis codes from the first position only may exclude relevant hospital episodes in which asthma was a main or a contributing reason for admission. In the PEDW episode data, primary diagnosis field was inconsistently used. Diagnosis codes were sometime found only following non-diagnosis codes.³ This means that the actual main medical diagnosis was not always recorded in the primary diagnosis field. Such a data quality issue was addressed in a recent study where asthma-related hospital episodes in Wales were identified using an asthma code being recorded in the first diagnosis position after ignoring the aforementioned non-diagnosis codes [43]. In my own analysis, I found that this approach increased the number of identified asthma-related hospital episodes by about 64% compared with using the primary diagnosis field only. Further work is needed to develop and validate accurate case definitions for asthma-related A&E visits and hospitalisations, possibly using a machine learning approach.

5.5.4.2 Possible residual confounders

Several variables that might have affected both the deprivation level and asthma outcomes were not included in the regression models presented in this Chapter. These variables were not available or were sub-optimally coded in routine data.

For example, I did not adjust the regression models for smoking status. Nor did I investigate the effect of smoking status on the outcome variables. Cigarette smoke is a known and strong trigger of asthma exacerbations among first-hand or second-hand smoker asthma patients. A systematic review and meta-analysis found that exposure to second-hand tobacco smoke among children with asthma almost doubled the risk of hospitalisation for asthma exacerbation and worsened pulmonary function [370].

Smoking was also shown to be positively associated with deprivation [371]. In England, an analysis by the Office of National Statistics found a strong association between the proportion of current smokers and the level of deprivation [372]. In that analysis, people living in the areas at the highest deprivation quintile were

³Such codes include symptoms, signs, abnormal clinical or laboratory findings, medical, surgical and allergy history, and other miscellaneous health-related statuses or events that exist in the R and Z chapters of the ICD-10 classification.

more than twice likely to smoke than those living in areas categorised as the lowest deprivation quintile. Part of the effect of the level of multiple deprivation on the number of asthma-related ED visits and hospital admissions might be influenced by differential smoking status.

On the other hand, smoking is a leading cause of premature death, and a risk factor for numerous health problems, including cancer, low birth weight, and limiting health conditions [373], all of which are part of the Health domain of the WIMD index. Accordingly, smoking status might have, to some extent, confounded the relationship between multiple deprivation level and the four asthma-related events that I studied. Therefore, if recorded in high quality, smoking status should have been added to the regression models used in my analysis to reduce the potential confounding bias.

Health literacy is also a potential confounder. The Institute of Medicine [374] defined health literacy as *“the degree to which individuals have the capacity to obtain, process and understand basic health information and services needed to make appropriate health decisions”*. Health literacy affects asthma outcomes both directly and indirectly, though inadequate patient knowledge of asthma, suboptimal asthma self-management and improper use of inhalers [375, 376]. Health literacy also affects the WIMD through the Education domain [377]. However, data on health literacy were not available in the routinely collected electronic health record (EHR) data in the SAIL Databank.

5.5.4.3 WIMD overall index and asthma exacerbations: a possible circular relationship

An important consideration when studying the associations between area-based socioeconomic measures and health outcomes is when the former include a health component.

The overall WIMD index includes the Health domain, which includes limiting long-term illness, death rate in the area from all causes, incidence of cancer, and low birth weight. Asthma mortality in Wales in 2011 was 1.8 per 100,000 population, a small contribution to the all-cause crude death rate of 990 per 100,000 population [378, 379]. However, asthma, especially when uncontrolled, is likely to be a limiting condition with significant and profound effects on patients' quality of life [380–382]. A recent survey of 4,650 asthma patients in the UK by Asthma UK

found that 46.4% of the respondents reported sleep difficulties due to asthma, and 45.2% reported that asthma interfered with their daily activities [383]. Low birth weight and other adverse perinatal outcomes, such as pre-term delivery and small for gestational age infants, have been linked to the severity of maternal asthma and/or asthma medications used by mothers during pregnancy [384–386].

Asthma could also affect the Education domain of the WIMD index. Suboptimal asthma control among children, as well as urgent or emergent asthma-related healthcare utilisation, have been shown to be associated with school absenteeism [387]. In addition, asthma could affect the Employment domain of the WIMD index by increasing job absenteeism and hindering job retention. A Danish study found that people with current asthma were more likely to miss work and lose their jobs [388].

With the above links between asthma and the Health, Education, and Employment domains of the overall WIMD index, using the latter as a predictor of asthma outcomes is challenging, and caution should be exercised in the interpretation of results. However, the effect of such circular relationship could be limited. A study in England found that removing the health domain from the Index of Multiple Deprivation had small, practically unimportant effect on the measured socioeconomic disparities in census measures of health [389].

5.5.5 Implications for health policy

Health inequalities are of paramount importance to health policy. The work presented in this chapter is an example of the utility of routinely collected, linked data from EHRs and the WIMD in Wales in assessing health inequalities, namely those in asthma outcomes, and in informing health policy.

The presented findings could improve our understanding of the social gradient in asthma in Wales and inform the development and redesign of policies to reduce inequalities in asthma outcomes. These findings suggest that several aspects of health care services for asthma patients in the most deprived areas in Wales could be targeted for improvement. These aspects may include quality of primary care services, including early diagnosis and optimal prescribing, the lack of which could lead to poor disease control.

To reduce health literacy gap and the associated gaps in asthma outcomes, the most deprived asthma population requires active efforts and further resources to ensure effective education on asthma and asthma self-management. This includes training in proper inhaler technique, adequate adherence to medications, and avoiding exacerbation triggers. The variation in quality of secondary care for asthma patients across the socioeconomic spectrum could also identify further potential to reduce the gap in asthma outcomes.

Avoidable health inequalities at any level create concerns and are socially and politically unacceptable. They may also result in wasted resources. With the high prevalence of asthma in Wales, even modest inequalities in asthma outcomes are likely to result in likely avoidable, significant disease costs at the country level. The overall cost of asthma in Wales in the fiscal year 2011-2012 was estimated as £74.7 million pounds (approximately US\$104.7 million) [43]. An IRR of 2.23 for asthma hospitalisations between the extreme WIMD quintiles means that there were 1,340 hospitalisations⁴ in the most deprived areas that would have not happened if the hospitalisation rate there was equal to that in the most affluent areas. Assuming a conservatively estimated average of £1,000 for the cost of an asthma hospital episode (approximated from cost data presented by Mukherjee et al. [43]), those extra 1,340 asthma hospitalisations in the most deprived areas costed NHS Wales at least £1,340,000 over the five-year follow-up. This calculation demonstrates the potential for avoiding significant, unnecessary costs of asthma healthcare utilisation in the most deprived areas. However, it does not take into consideration the costs of increased asthma prescriptions, A&E visits, and GP visits in the more deprived areas. In fact, while hospital admissions accounted for 13.1% of asthma-related costs to the NHS in Wales in 2011-2012, prescriptions accounted for two third and asthma-related GP visits, ambulance trips, and A&E visits together accounted for around 16% [43]. A comprehensive cost analysis is therefore needed to estimate the variations in the overall asthma financial burden across deprivation levels in Wales and the potential savings by reducing these variations. Such an analysis should include the costs of asthma-related visits to GPs, A&E and outpatient departments, ambulance trips, and prescriptions and hospital admissions as well as other wider societal costs such as Disability Living Allowance and costs resulting from school and work absenteeism.

⁴ $(2.23 - 1) \times 0.022 \times 49,597$; where 0.022 is the hospitalisation rate in WIMD 5, and 49,597 is the number of patients in WIMD 1 areas based on Cohort 1 characteristics (Table 5.2).

Identifying inequalities in asthma outcomes and identifying potential targets to reduce them aligns with the Welsh Respiratory Health Implementation Group's vision towards reducing inappropriate variations in respiratory outcomes across Wales [390]. Reducing health inequalities is also a key objective of the Welsh Government. The use of routinely collected data in this exercise to explore asthma outcomes aligns with political landscape in Wales to maximise the use of these data to support respiratory health policy and care delivery [96, 391].

5.5.6 Future work

The limitations and methodological challenges encountered in this study warrant further work. For example, algorithms to identify asthma-related A&E visits and hospital admissions in the SAIL databank need to be externally validated. If available in high quality, measures of potential confounders such as smoking status and health literacy could be added to the statistical model.

Further analysis would be needed to improve the understanding of the asthma social gradient in Wales and explore possible explanations to the observed inequalities, taking into account the complex inter-relationships between the relevant variables. Including additional data on healthcare utilisation and asthma outcomes across the deprivation levels would provide additional insights into the differential severity of asthma exacerbations and the associated cost of avoidable outcomes. Such data may include detailed hospitalisation data, such as LOS and HRG, as well as death due to asthma.

Further work on associations between asthma outcomes and each of the eight domains of the overall WIMD index would be useful to understand the individual contribution of each deprivation domain to asthma inequalities.

Asthma is associated with a wide range of comorbidities such as rhinitis, sinusitis, obstructive sleep apnoea, gastroesophageal reflux disease, obesity, anxiety, and depression [392-394]. More comorbidities are in general associated with lower socioeconomic status [395]. Therefore, the impact of comorbidities on asthma inequalities and the associated avoidable cost and burden of these comorbidities is worth exploring.

Ethnicity may in general influence both socioeconomic status and health outcomes [396-398]. Wales is generally a homogeneous country with about 5% of the pop-

ulation identified themselves as being from non-white background [399]. However, the percentages of non-white people are higher in large urban areas such as Cardiff (17.2%), Swansea (10.6%), and Newport (7.7%) [400]. In these areas, the individual role of ethnicity in the social gradient of asthma is worth investigating. Finally, future work will explore the trends and costs of asthma inequalities over the last two decades to see whether they are changing over time.

5.6 Conclusion

This chapter demonstrated an important application of the Wales Asthma Observatory in supporting health policy regarding equality in health care. I found wide inequality gaps in asthma outcomes between the extremes of socioeconomic deprivation spectrum in Wales and showed a social gradient in asthma-related hospital admissions. Compared to the least deprived areas, the most deprived areas had slightly more primary care contacts, including annual asthma reviews, per asthma patient. However, asthma patients in the most deprived areas were 37% more likely to have asthma-related ED visits and more than twice as likely to be admitted to hospitals due to asthma than those in the least deprived areas, although those gaps decreased slightly among patients with continuous asthma prescriptions throughout the study period.

These wide gaps in asthma healthcare utilisation were possibly due to higher severity and poorer control of the disease in the least deprived areas. Possible underlying factors such as suboptimal asthma prescribing, inadequate health literacy, poor asthma self-management, and wider non-health related socioeconomic determinants, such as income, employment, education, air pollution, might be contributing to the observed gaps and require further investigation.

Chapter 6

General Discussion

Reflection and future directions

In this chapter, I summarise the work I performed in this thesis about the development and utilisation of the Wales Asthma Observatory, highlighting my original contributions. I critically review the strengths and limitations of the Observatory as a platform for asthma research and surveillance, and as a tool to improve our understanding of asthma in Wales and to inform health policy, service planning and delivery. I then compare my findings with related works and studies. I then discuss the opportunities and challenges encountered during the Observatory development and utilisation. I reassert the high potential of using asthma-related routinely collected data to improve asthma patients' lives, and the pressing need to reduce avoidable harm and waste from the suboptimal re-use of these data. I discuss the potential role of the Observatory in the national efforts to improve asthma outcomes. I then propose a future research agenda to improve the Observatory's methodology and to answer further questions about the social gradient of asthma in Wales, I also propose further technical and content developments to the Observatory.

Chapter Contents

6.1	Summary of findings	193
6.2	Original contributions	195
6.3	Strengths and limitations	196
6.3.1	Data sources: pros and cons	196
6.3.2	Case definitions: Flexibility and data driven approach	197
6.3.3	Longitudinally assessed disease outcomes	198
6.3.4	Supporting research reproducibility	199
6.4	Interpretation of findings in the light of related literature	199
6.4.1	Methods to define complex disease entities using routinely collected data	199
6.4.2	Various approaches to asthma registries, surveillance systems, and research platforms	200
6.4.2.1	Asthma surveillance systems used routinely collected and/or self-reported data	201
6.4.2.2	Asthma registries generally target the problematic cases	202
6.4.2.3	Routinely collected data (RCD) for disease registries	203
6.4.2.4	Data acquisition, management, and quality	204
6.4.2.5	Facilitating data interrogation is an increasingly recognised need	205
6.4.3	Social gradient of asthma: consistent findings and methodological challenges	207
6.5	Challenges	208
6.5.1	Asthma heterogeneity complicates case identification and comparability of studies	208
6.5.2	Data from important care domains are still missing	209
6.5.3	Quality of routinely collected data is imperfect	209
6.5.4	Routinely collected data suffer from time lags	210
6.5.5	Routinely collected data does not reflect precise disease timeline	211
6.5.6	Lack of valid methods to assess outcomes	211
6.5.7	Public attitudes to data re-use are mixed	211
6.6	Implications and potential uses of the Observatory	212
6.6.1	Implications on health policy and wider societal impact	212
6.6.2	Implications on service planning and delivery	212
6.6.3	Implications on clinical practice and patient outcomes	213
6.6.4	Implications on asthma research	214
6.7	Towards maximising population data benefits to improve asthma outcomes	214
6.7.1	Improving data capture	214
6.7.2	Reducing waste from underuse of data	215
6.7.3	Supportive data-intensive research environment	216
6.7.4	Potentials of asthma big data	217
6.8	Future work	218
6.8.1	Improving methods to define asthma patients and assess disease outcomes	218

6.8.2 Understanding inequalities in asthma outcomes	219
6.8.3 Monitoring and forecasting asthma trends	219
6.8.4 Linking additional data sources	220
6.8.5 Improving the Observatory's technical platform	221
6.8.6 Data quality reports	221
6.8.7 Getting ready for SNOMED-CT	221
6.9 Conclusions	222

6.1 Summary of findings

In this doctoral project, I demonstrated how routinely collected electronic health record (EHR) data can be used to develop a data-intensive platform for asthma research and surveillance—the Wales Asthma Observatory. The Observatory aims to maximise the benefit of these data and is based on a regularly updated, national cohort for asthma.

Due to the inherent limitations of routinely collected data (RCD) [88–93], special attention should be paid to methods of defining diseases and health outcomes using these data. Therefore, to inform the Observatory development, I systematically reviewed the contemporary methods to define and assess asthma using routinely collected data and the ways these methods have been described (Chapter 2). I found a wide variation in these methods and suboptimal reporting on their implementation and validity. I highlighted the challenges of standardising methods to define and assess asthma, and the need to develop and validate database-specific methods.

In Chapter 3, in the light of the literature review findings, and considering asthma heterogeneity, data limitations, and the absence of a gold standard to define asthma, I justified the use of data-driven approaches to identify people with asthma. I demonstrated the appropriateness and benefits of using latent class analysis (LCA) on recorded asthma-related primary care data to identify clusters of asthma patients, including those with current asthma. My latent class modelling was based on healthcare utilisation data related to asthma and chronic obstructive pulmonary disease (COPD) for a large, random sample of the population of Wales. I chose the eight-class model as the best-fit model based on its model diagnostics and clinical interpretability. I assigned clinical labels to the latent classes. I then

reduced the model complexity by merging the 'asthma' classes into two classes representing 'ever diagnosed asthma without current treatment' and 'currently treated asthma'. I then applied recursive partitioning (a supervised machine learning technique) to derive a decision tree which could identify patients with asthma, including whether they were currently treated, as well as those with COPD and asthma-COPD overlap syndrome (ACOS) in primary care data.

I used this case identification algorithm, in addition to several other case definitions for asthma, in the development of the Observatory, described in [Chapter 4](#). I included in the Observatory a number of essential disease outcomes and variables such as disease severity, treatment step, and asthma exacerbations. I also developed and described a technical platform to improve the efficiency, reporting, and reproducibility of data extraction of studies that use the Observatory.

I investigated the quality of selected asthma-related event groups, and described variable patterns and levels of data missingness. Notably, many lung function tests were recorded without their measurements. When recorded, measurements were inconsistent for many of the lung function event codes. To improve the capture of asthma data, I recommended improved data entry quality checks by EHR systems, data-quality awareness training for clinicians, and the inclusion of data quality to receive a greater focus in payment-for-performance schemes.

In [Chapter 5](#), to demonstrate the Observatory's utility for health policy, I investigated the inequalities of asthma outcomes across the socioeconomic spectrum in Wales. I used count regression models to compare asthma-related primary and secondary care events between asthma patients living in areas with different deprivation levels. I found that, compared to asthma patients who lived in the least deprived areas, those in the most deprived areas had slightly more primary care contact (7.8% more general practitioner (GP) visits per patient), yet they had significantly more asthma-related emergency department visits (31.1%) and hospitalisations (123.2%). There was a clear gradient of more asthma-related hospital admissions in the more deprived areas in Wales. The inequality gaps were slightly smaller among patients who continuously received asthma prescriptions over the five-year follow-up period. I discussed the implications of these inequality gaps, and outlined future research directions to improve the modelling and account for possible confounders. I then proposed potential measures to reduce and bridge the inequality gaps in asthma outcomes.

6.2 Original contributions

I summarise the original contributions in this thesis as follows:

Chapter 2

1. I found a wide variation in the contemporary methods to define asthma and assess asthma outcomes in observational studies conducted using routinely collected EHR data.
2. I found that reporting on the implementation and validity of these methods was poor overall.
3. I identified 10 practices of reporting or justifying the validity of these methods (Table 2.1). These practices varied widely from performing validity assessment in the same study, to relying on clinical guidelines or the validity of database coding. The majority of EHR-based asthma studies reported no information on the methods' validity.

Chapter 3

4. I described a probabilistic approach, using LCA of primary care data of a large population sample, to identify groups of people with asthma (including those with current asthma), distinguish them from those with COPD, and to identify people with asthma-COPD overlap.
5. I described the groups of people with asthma based on their asthma-related health care utilization.

Chapter 4

6. To identify people with asthma in the Wales Asthma Observatory, I used the above-mentioned data-driven probabilistic model as well as commonly used deterministic case definitions.
7. I described different levels and patterns of missingness and inconsistencies in asthma-recorded data, apparently due to different approaches used by GPs to record similar data items and using the same clinical codes for different purposes.

Chapter 5

8. I demonstrated the extent of inequality gaps in asthma health care utilisation in Wales: Compared to the least deprived areas, the most deprived areas had 8% more asthma related GP contact, but 123.3% more asthma hospitalisations; the inequality gaps were slightly smaller among patients who continuously received asthma prescriptions over the five-year follow-up period. I also found wide variations in these outcomes across age and gender groups.

6.3 Strengths and limitations

The Wales Asthma Observatory is the first of its kind in Wales as a platform for asthma research and surveillance using routinely collected data. The work presented in this thesis has a number of strengths and limitations related to the data used in the Observatory, the methods used to identify asthma patients and assess disease outcomes, and the Observatory design and structure.

6.3.1 Data sources: pros and cons

The type and sources of data used in the Observatory are one of its major strengths. The Observatory currently utilises key nationwide clinical datasets in Wales, including data on primary care, secondary care, area-based deprivation as well as causes of death. These data are already collected from EHRs, anonymised, and linked in the Secure Anonymised Information Linkage (SAIL) Databank. This offers a unique opportunity to answer a wide range of research questions that is not feasible with *de novo* data collection.

Throughout the previous chapters, I highlighted the advantages of EHR-based routinely collected data over purpose-specific data collection. Briefly, RCD are inexpensive, person-level streams of data that reflect the real-world picture of people's health status and clinical care. They are mainly recorded from the perspective of healthcare professionals rather than patients. Thus, their validity does not rely on patients' memory or health literacy. These data are routinely collected in huge volumes across Wales. This allows obtaining nationally representative epidemiological estimates, and enables conducting high-power studies and investigating rare outcomes. These opportunities are usually not present with small-sized primary data collected first hand by investigators.

However, RCD suffer from a wide range of problems such as missingness, miscoding, under-recording, and linkage errors. There was no information about the version of coding system used to codify primary care data. Primary care events in the United Kingdom (UK) are usually coded with the Read code vocabulary, namely the second and third versions. The second version is a hierarchical vocabulary, while the third version is a radically developed version which, for example, supports poly-hierarchy (i.e., a code can have multiple parent codes), includes additional concepts, and has the codes changed for some of the existing concepts. It was difficult within the SAIL Databank to ascertain the vocabulary version in which a GP event was coded. The vast majority of GP practices in Wales, however, use the second version of Read Code vocabulary [401]. Therefore, in the Observatory development, I used only the second version for data extraction. However, this might have introduced a misclassification bias. This may happen, for example, if the GP practices that used Read Code version 3 differed from the rest of practices in their population characteristics or in the quality of care. Nonetheless, given the small number of those GP practices, this limitation is unlikely to undermine the national representativeness of the Observatory.

6.3.2 Case definitions: Flexibility and data driven approach

The Observatory is empowered by the availability of multiple case definitions of asthma. Thereby, it allows capturing most cases of asthma including those with uncertain diagnosis. At the same time, it includes more strict case definitions such as currently treated asthma. This flexible approach facilitates studying diverse groups of asthma patients. It also allows researchers to choose, for their studies, the appropriate case definitions that are comparable with other particular studies.

Those case definitions, whether they are based on a single diagnosis code or more complex deterministic algorithms, are based on clinical guidelines, clinical knowledge, or epidemiological judgement. Each of those case definitions has a specific meaning and is intended to be used to identify a specific group of people with 'asthma'. In particular, the case definition of *ever-diagnosed currently-treated asthma* is the most useful one as it allows selecting people with active asthma at a certain point in time. This case definition has been commonly used as an essential eligibility criterion in the contemporary EHR-based asthma studies [102] and is also the basis of the main asthma indicator in the Quality of Outcomes Frame-

work (QOF). Therefore, the Observatory will commit to use the *ever-diagnosed currently-treated asthma* as the main case definition of asthma for surveillance and research.

The Observatory also benefits from a latent class model which can be used as an ‘internal’ data-driven reference to identify asthma patients. Unlike cluster analysis, which uses distances between individuals, LCA uses a top-down approach to understand the population structure: it utilises the distributions of the observed data to identify the likely population latent classes. By computationally uncovering the population structure, LCA identifies all the likely patient groups, some of which could be overlooked in the manual researcher-led development of case definitions. LCA probabilistically determines to which latent class each person belongs. This probabilistic approach fits well with the nature of asthma as a heterogeneous condition represented by a continuous spectrum of various pathogeneses, overlapping phenotypes, and variable severity and natural history, which may coexist with other conditions (COPD, for example). This approach to patient identification allows researchers to select patients, not only based on their most likely classes, but also by preferred probability thresholds, or based on overlap patterns of interest (e.g., asthma-COPD overlap). The decision tree that I derived from the LCA model allows researchers to use this model to identify asthma patients in the SAIL Databank and similar databases.

However, unsupervised approaches such as LCA has limitations. The output of LCA depends on the quality of the input data and their relevance to the desired classification. It also involves a level of subjectivity in the model specification (i.e. choosing features), selection of the best-fit model, and interpretation of the latent classes. Therefore, while data-driven approaches can be useful to understand the population structure, they need to be coupled by knowledge about the disease pathophysiology, clinical course, and epidemiology as well as about data provenance and quality [142].

6.3.3 Longitudinally assessed disease outcomes

The Observatory includes longitudinally calculated key disease outcomes. For each patient in the Observatory, key disease states and outcomes such as treatment step and disease severity and exacerbations are ascertained longitudinally as state variables along the patient’s follow-up period. This allows both cross-

sectional and longitudinal analysis of these variables. The definition of these variables was informed by the algorithms to assess asthma outcomes that I found in the systematic scoping review in [Chapter 2](#). While the validity of those algorithms was assessed elsewhere, it should not be assumed to hold in the SAIL Databank. However, assessing the validity of these definitions was not feasible within the time-frame of this doctoral project.

6.3.4 Supporting research reproducibility

In the Observatory's design and implementation, I took into consideration the challenges of data extraction reproducibility. Therefore, I equipped the Observatory with a clinical code set library and data extraction platform, with an easy-to-use graphical interface, which allows researchers with no programming skills to interrogate the Observatory and the SAIL's primary care dataset. This platform is also intended to save time for experienced analysts by reducing unnecessarily repetitive programming code writing and database query development. This platform is aimed to support and promote research transparency and reproducibility as well as sharing and re-use of clinical code sets and data extraction procedures.

I designed the Observatory data structure in such a way that it can be seamlessly updated when the source datasets in the SAIL Databank are updated, using a programming script I built for this purpose. Since updated data may include historical changes, versioning of the Observatory data allows reproducing studies performed on historical versions.

6.4 Interpretation of findings in the light of related literature

6.4.1 Methods to define complex disease entities using routinely collected data

The findings in the [Chapter 2](#) were in line with previous studies. A related systematic review identified wide variation in the categorisation of asthma severity using health insurance claim data [222]. Similar variations in case definitions and the need for standardisation have been recognised in other conditions such as heart

diseases [402], osteoarthritis [403], and immune thrombocytopenic purpura (ITP) [404, 405].

Standardisation of methods to define and assess asthma from RCD has been faced with a number of challenges including the disease heterogeneity, lack of consensus on its clinical definitions, variations in populations' characteristics, and cross-country differences and limitations of RCD resources.

Clinical coding systems help standardise the documentation of health and health care concepts in EHRs. However, the clinical meaning of those concepts should be standardised in the first place. Standardisation of asthma terminology requires clear understanding of the disease's aetiology, genetic and molecular pathogenesis including gene-environment interaction, and how the underlying disease mechanisms manifest in different pathophysiological and clinical phenotypes. However, from a precision medicine perspective, a fixed terminology to describe a heterogeneous continuum of diseases may be insufficient to provide personalised diagnosis and management. Instead, Agusti et al. have suggested a label-free precision medicine strategy for chronic lung diseases in which treatable traits, rather than encompassing labels such as asthma and COPD, form the basis for diagnosis and management [406]. This contemporary clinical perspective has been epitomised in a recent editorial with Oscar Wilde's quote "*to define is to limit*" [208].

Wherever the debate on clinical definitions of asthma and asthma outcomes might move, standardisation and harmonisation of the corresponding operationalised RCD-based definitions are needed. Algorithms to define particular clinical concepts (e.g., 'asthma', asthma endotypes or phenotypes, or treatable traits) should be ideally validated wherever they are used. Subsequently, the optimal method to measure the same clinical concept may differ across databases and populations.

6.4.2 Various approaches to asthma registries, surveillance systems, and research platforms

Asthma surveillance systems and registries around the world have various purposes. They have been also established using different approaches to defining source populations and cases of interest. They also differ in their data sources, content, and usability, and data security models.

6.4.2.1 Asthma surveillance systems used routinely collected and/or self-reported data

To be used for surveillance, data sources should have high geographical coverage and representation, sufficient data quality, and sustainability. RCD usually satisfy these requirements and are widely used for asthma surveillance. These include data on asthma-related accident and emergency (A&E) visits, urgent care, and hospital admissions, medication dispensing, and health insurance claims [134, 407–409].

The Ontario Asthma Surveillance Information System (OASIS) is closely related to the Observatory in terms of the purpose and the use of RCD [134, 242, 410]. It was established as a platform for asthma surveillance and epidemiological research in the Canadian province of Ontario. OASIS uses administrative and health insurance data about out-of-hospital, emergency, and inpatient asthma care. The Observatory, however, use EHR data which are richer and more comprehensive than administrative and health insurance data.

Self-reported data are commonly used in the United States (US) asthma surveillance systems to estimate the disease prevalence [407–409]. These data are collected as part of the Behavioral Risk Factor Surveillance System (BRFSS) [411] telephone survey. The BRFSS survey contains questions about whether the respondent ever had and still has asthma, the age at diagnosis, symptoms frequency, exacerbation history, routine asthma check-up, number of preventer and rescuer asthma medications used, and the impact of asthma on the quality of life [412]. People who report having asthma are invited to the Asthma Call-back Survey (ACBS) [413]. The ACBS collects more detailed information on the disease history, healthcare utilisation, knowledge of asthma and management plan, patient's behaviour towards environmental risk factors, medication use, medical self-management, personal cost of asthma, asthma effects on work and/or school attendance, work-related asthma, asthma comorbidities, and use of complementary and alternative therapy [414, 415].

Compared to the rich asthma-related self-reported data collected in the US, asthma-related self-reported data in the UK nations' annual health surveys are much more limited. In particular, the Welsh Health Survey (WHS) only asked whether the respondent (or their child) was currently being treated for asthma or wheezing or

had recent shortness of breath, tightness of chest, or wheezing [416]. In addition to their scarce asthma-related details, the WHS data had small sample sizes and suboptimal geographical representation, let alone the biases of self-report. Therefore, their role in the Observatory development was limited to being used as an external data source for the evaluation of the RCD-based case identification model.

6.4.2.2 Asthma registries generally target the problematic cases

Unlike asthma surveillance systems which generally target the whole asthma population, asthma registries are mainly dedicated to the more severe or complicated cases. One example is the British Thoracic Society (BTS) Difficult Asthma Network (DAN) registry [126], which was succeeded by the UK Severe Asthma Registry.¹ The DAN registry included people who fulfilled the American Thoracic Society (ATS) definition of refractory asthma [417]. The UK Paediatric Difficult Asthma Network Registry² comprises four specialist asthma specialist centres in the UK and aligns with efforts to incentivise the appropriate identification and management of problematic asthma cases. Both registries receive data entered by health professionals into secure portals as well as routinely collected data.

The Belgian Severe Asthma Registry (BSAR) is dedicated to difficult asthma cases [418]. It collects data on asthma diagnostics, such as lung functions, fractional exhaled nitric oxide (FeNO), blood eosinophil count, serum immunoglobulin E (IgE), sputum inflammatory cell profile, skin prick test, medication use, comorbidities, as well as smoking status. It also collects asthma-related patient reported outcome measures (PROMs) including Asthma Quality of Life Questionnaire (AQLQ), Asthma Control Questionnaire (ACQ), and Asthma Control Test (ACT).

The Italian Registry of Severe and Uncontrolled Asthma (abbreviated in Italian as RIItA) offers secure web-based access to a database of asthma-related clinical data, risk factors, and exacerbations for patient with severe and/or uncontrolled asthma [419].

In comparison to those asthma registries, the Observatory is designed to be both a surveillance and research platform and a disease registry. Thanks to its nationwide

¹<https://cl2.n3-dendrite.com/csp/asthma/frontpages/index.html>

²<http://rs2.e-dendrite.com/csp/paedasthma/frontpages/index.html>

data sources, the Observatory is a comprehensive disease registry that targets the whole asthma population across Wales, regardless of disease severity.

Algorithms based on domain expert knowledge (e.g., researcher's clinico-epidemiological judgement and/or clinical guidelines) to identify eligible cases have been the conventional approach in asthma registries. Those algorithms are often validated against clinical reassessment or review of the full medical record, neither of which is a universal gold standard for asthma diagnosis.

The Observatory's multi-approach to case identification, based on domain expert knowledge and data-driven methods, offers the flexibility needed to study different asthma populations. LCA helped *see* the likely population structure behind the recorded data, while recursive partitioning produced a corresponding transferable algorithm to identify patients with asthma and/or COPD. Deriving a simple classification algorithm from a complex clustering model was previously described by Moore et al. [273]. They identified asthma phenotypes from cluster analysis of 34 variables, and derived a simpler three-variable classification tree which later identified similar clusters in a different population [420].

6.4.2.3 Routinely collected data (RCD) for disease registries

For disease registries, the approach of using RCD to develop a disease registry is relatively new. Instead, traditional disease registries are set up to include cases of interest that are managed in health care facilities within particular geographical areas. Cases are usually included in a registry by healthcare professionals based on defined criteria which are assessed on a case by case basis. Inclusion of cases in the registry is often carried out using detailed clinical information available in the doctor-patient encounter and the full patient record. However, there are a number of disadvantages of the traditional approach of asthma registries:

- the denominator is often not defined [124];
- the number of patients in the registry is relatively small due to the often-limited geographical coverage and the strict case definitions;
- case ascertainment is labour-intensive (which makes traditional disease registries more suitable for rare conditions);
- the inclusion in the registry often require the patient's consent; and,
- data collection is subject to the experimenter bias.

For a prevalent condition such as asthma, individual case identification at a national scale would consume significant time and resources.

RCD offer inexpensive, accessible, wide-coverage, and rich alternative data sources for disease registry development. In this approach, cases of interest are automatically identified and characterised *en masse* from large datasets [127]. Larger numbers of cases can be identified. RCD usually have defined denominators, allowing estimation of the condition's epidemiology and burden. RCD are usually de-identified, and therefore no individual patient consents are needed [421]. However, misclassification of cases and missing variables are among the disadvantages of using RCD to create a disease registry.

6.4.2.4 Data acquisition, management, and quality

Sariyar et al. proposed a framework to evaluate medical registries purposes, data acquisition and management processes, and data quality [422]. The framework included assessment of data accuracy, trustworthiness, consistency, granularity, timeliness, completeness, security, and privacy. These criteria should be assessed along the flow of data in the registry: from acquisition, through storage, to presentation [422]. The authors argued that a registry should only include high quality variables that fit with its purpose(s). The Observatory, however, is built for generic purposes of asthma research and surveillance. This requires continuously expanding sets of case definitions and research-ready variables to satisfy the growing surveillance needs and the emerging research questions. However, RCD used in the Observatory inherently suffer from accuracy, trustworthiness, consistency, and completeness issues (see Section 4.5). Therefore, further assessment of case definitions validity and variables quality should be high priority in the future developments of the Observatory.

Modern implementations of registries for asthma and other conditions are increasingly web-based, where data are entered by patients and/or healthcare professionals through secure online user interfaces [126, 130, 418]. Patient consent is normally needed before their data are included in a registry. The stored data are usually de-identified. In contrast, the Observatory uses already-linked de-identified data from the SAIL Databank. The Observatory's person-level data are maintained within the secure environment of the SAIL Gateway. In the future,

however, secure online or mobile-based data collection could be implemented to capture asthma-related PROMs and link them to RCD in the Observatory.

6.4.2.5 Facilitating data interrogation is an increasingly recognised need

I designed the Observatory's user interface so that it improves the workflow of data interrogation including automation, reproducibility, reusability, and shareability. With the growing use of EHR-derived data for research elsewhere, the need for data extraction automation and code set engineering has been recognised in many EHR-derived data resources [301]. Facilitating data interrogation from EHR data resources can be achieved by variety of approaches such as providing 'research-ready' variables, developing tools to automate common data extraction tasks, and maintaining clinical code set libraries.

The Observatory provides essential asthma-related 'research-ready' variables including disease state (i.e. current case definition) and key outcomes and variables such as treatment steps, asthma severity, and exacerbations. 'Research-ready' variables have been provided in EHR-derived data resources elsewhere. An example is the Clinical research using Linked Bespoke studies and Electronic health Records (CALIBER), a UK-based platform that provides access to 'research-ready' variables derived from data linked across EHRs, disease registries, bespoke cohort studies, and other routine data sources through a common data model [129].

The Observatory also provides an easy-to-use platform to design, share, and reuse complex data extraction procedures as well as manage clinical code sets. With this platform, researchers can create additional study-specific Read code sets and complex variables derived from the SAIL's GP dataset. The graphical interface mimics the process of creating a data table and populating its fields in the Structured Query Language (SQL). It can be used by users with no programming skills. In addition, the visual interface accepts inserting SQL pieces of codes so that users with programming skills can design more advanced data extraction procedures. A related tool is rEHR, an R package which provides functions for advanced data interrogation from the Clinical Practice Research Datalink (CPRD) [423]. However, as a programming library, rEHR can only be used by members of a research team who have programming skills. In contrast, the Observatory interface allows collaborative development of code sets and data extraction procedures by anyone in a multidisciplinary research team. Another difference is that rEHR works on data

files exported from the CPRD, whereas data extraction procedures designed with the Observatory's platform can be exported to be run on a database connection, such as the case in the SAIL Gateway. Lastly, an rEHR-based data extraction code can be shared and published as a computer file. By contrast, a data extraction procedure designed in the Observatory is maintained centrally on a web address where it can be (re)used, shared, cited, and exported as an SQL or an interoperable JavaScript Object Notation (JSON) file.³

The need for supporting the collaboration in data interrogation has been previously recognised. eLab is a web-based environment which allows multidisciplinary research teams including researchers and healthcare professionals to access health dataset and collaboratively develop methods to analyse and visualise the results [424].

eLab is based on the concept of Research Objects. Research Objects have been proposed as a generic, comprehensive approach to representing the research process and outcomes as semantically linked, reusable, shareable, entities [425, 426]. eLab has been used as 'Asthma eLab' in the Study Team for Early Life Asthma Research (STELAR) consortium [427]. Asthma eLab provides web-based platform for collaborative management and analysis of asthma-related data from five birth cohorts in the UK. It allows research teams to model relationships between pathological and physiological processes in graphical and computable forms.

The Observatory's approach of using simpler variable types to specify complex data extraction procedures ('building blocks' approach) that are reusable, extensible, and shareable roughly corresponds to the concept of Research Objects.

ClinicalCodes.org is another related tool which provides repository for clinical codes used in EHR studies [216]. This public web-based repository is similar to the code set library provided in the Observatory. Whereas it only archives already used code sets, the Observatory's platform allows users to collaboratively create, edit, and revise code sets and then easily use them in data extraction procedures hosted in the same platform.

In summary, an efficient and effective user interface for an EHR-based data resource such as the Observatory should ideally satisfy a number of principles [426] including methods versioning, repeatability, auditability, reusability, repurposeabil-

³JSON is an open-standard file format to exchange human-readable data in the form of arrays and attribute-value pairs.

ity, shareability, referenceability, and interoperability, as well as results reproducibility.

6.4.3 Social gradient of asthma: consistent findings and methodological challenges

The findings in [Chapter 5](#) were consistent with local and international studies. While some studies found that societal affluence was associated with higher asthma prevalence, others found the disease more severe in poorer areas [[428](#)]. Low socioeconomic status was associated with less treatment in wheezy children [[325](#)], poorer asthma control and persistent airway obstruction in adults [[326](#)], and higher asthma hospitalisation rates [[323](#), [324](#)]. Watson et al. found that higher asthma hospitalisation rates among the most deprived could not be explained solely by readmissions; instead, more asthma patients from the poorer areas were hospitalised [[323](#)]. The route of admission to hospitals for asthma was not considered in my study, but were previously found to differ by deprivation level. In the West Midlands, England, the proportion of asthma admissions through A&E departments was higher in the poorest than in the richest areas [[323](#)], but rates of GP referrals for asthma were not associated with deprivation level. The environment plays a role in inequalities. Air pollution induces asthma exacerbations [[318](#), [319](#)], whereas persistent asthma was associated with poor housing [[326](#), [429-431](#)].

Health inequalities including those in asthma outcomes has been traditionally ecologically assessed. For example, Watson et al. assessed the association between asthma age-standardised admission rate in the whole population of a geographical area with the area's Townsend Deprivation Index [[323](#)]. Theoretically, an ecological variable such as asthma prevalence in an area might have affected the admission rate in that area. Yet, the authors ruled out an increased asthma prevalence in poorer communities based on previous surveys. In my study, I assessed the person-level association between area-based deprivation level and asthma hospitalisations, among other outcomes. Therefore, my findings were independent from asthma prevalence. That study used the deprivation level of the hospital area rather than that of the person's address. In my analysis, I used the deprivation index associated with the patient address, which eliminated the bias from admissions in hospitals located in areas with different deprivation levels.

When both the explanatory and outcome variables being measured at the group or area levels, associations are threatened by ecological fallacy. Replacing aggregated data, person-level RCD are increasingly used to measure health outcomes in health inequality studies [432]. In this thesis, the availability of nation-wide person-level data on asthma outcomes in the SAIL Databank reduced the risk of ecological bias. This bias was further reduced by using a deprivation index, the Welsh Index of Multiple Deprivation (WIMD), that was calculated for relatively small areas (average population \approx 1,600).

Researching health inequalities is challenging. It is important to calculate the pure effects of socioeconomic factors on health outcomes and to determine causality direction [433]. However, these are not straightforward exercises in an indeterminate space of complexly interrelated factors. Important social determinants of health as well as confounders are often missing or indirectly assessed. RCD often lack these variables; EHRs usually do not capture sufficient data on health literacy, disease self-management, and wider social determinants of health. In order to advance health inequality research, those vital data need to be routinely collected [434].

6.5 Challenges

I identified several challenges towards the development of the Observatory. These were mainly related to the complex nature of asthma, data limitations, lack of locally validated methods to assess asthma outcomes, and the public's attitudes to reusing health data.

6.5.1 Asthma heterogeneity complicates case identification and comparability of studies

There is an increasing recognition that asthma is a heterogeneous condition, comprising distinct phenotypes and endotypes [4, 5]. In addition, there is no consensus on the clinical definitions of asthma and its key outcomes such as disease severity, control, exacerbations [147, 148]. This is probably reflected in the wide heterogeneity in the methods in which asthma and asthma outcomes have been defined from routinely collected data, as I found in Chapter 2. The lack of standardisation of methods to define and assess asthma hinders the comparability of studies and

evidence of synthesis. For flexibility, however, the Observatory users can choose from several case definitions of asthma, in addition to the one I developed locally using LCA and recursive partitioning. This enable researchers who use the Observatory to choose a method to identify people with asthma so that their study can be compared to studies that used the same case definition.

6.5.2 Data from important care domains are still missing

Despite the availability of various datasets in the SAIL Databank, data from important healthcare domains are still not available. Dispensing data contain information needed to assess medication adherence. However, they are yet to be linked into the SAIL Databank. In addition, treatments and prescriptions given to asthma patients during hospital episodes are not collected into the SAIL Databank. These data can be potentially useful to improve the sensitivity and specificity of methods to identify asthma-related hospitalisations. These data, especially pathology test results, could also improve methods of asthma phenotyping in the Observatory.

In [Chapter 5](#), I demonstrated the utility of the Observatory to support health policy by investigating inequalities in asthma outcomes across the socioeconomic groups. The study had high statistical power and provided useful insights into the magnitude of the asthma social gradient in Wales. However, as with all observational studies, there were limitations with potentially residual confounders, many of which were not readily available or directly measurable in the SAIL Databank. An example of such confounders was health literacy which has significant effects on health status, disease prevention, early diagnosis, adherence to treatment, and disease self-management. Direct, patient-level data on health literacy was not available in the SAIL Databank. Instead, possible proxies include individual or area-based data on education attainment.

6.5.3 Quality of routinely collected data is imperfect

Data quality has implications in almost all uses of data. Quality of RCD can significantly influence the internal validity of studies using these data. Data quality is potentially compromised by a variety of factors at different stages of their flow from points of care to data safe havens. These factors include, for example, poor capture and record linkage errors.

At the point of care, recording and coding of clinical data is often incentivised by payment-for-performance schemes such as the UK's QOF. This means that only a small set of essential health events are well recorded. I showed in [Section 4.5](#) examples of data quality issues where many asthma related data were missing from the SAIL's GP dataset, including lung functions recordings, disease severity stratification, and measures to manage disease control.

Practices of data recording and coding into EHRs differ between health organisations and potentially between healthcare professionals in the same organisation. These practices have been influenced by the type of EHR systems used [[435](#), [436](#)]. They may also change over time due to administrative requirements (e.g., introduction and changes in the QOF indicators) and in response to changes in clinical guidelines and practice protocols.

Record linkage errors also compromise quality of linked data. Despite the high matching rate in the SAIL Databank [[140](#)], linkage error may still happen if identifiers are incorrectly recorded or missing [[234](#)]. Record linkage errors have been associated with several individual and population factors such as gender, race, geographical location, health status, and socioeconomic status [[437](#)]. If not properly addressed, record linkage errors may introduce random and/or systematic errors to study findings [[234](#), [438](#)].

6.5.4 Routinely collected data suffer from time lags

Many routinely collected data are not available for secondary uses in real-time [[80](#), [439](#), [440](#)]. Rather, considerable lead time is usually needed before they are made available in usable form [[80](#)]. This time, ranging from weeks to several months or years [[441](#)], is needed for preparation, transfer, anonymisation, record linkage and encryption as well as quality checks of data [[141](#)]. This time lag limits the usability of routinely collected data for applications that require timely data such as producing real-time epidemiological estimates and follow-up of outcomes in prospective studies and clinical trials [[442](#)]. Therefore, advances in data collection and transfer operations as well as in infrastructures are needed for seamless and faster production of usable RCD in order to facilitate applications that need timely access to data [[440](#)].

6.5.5 Routinely collected data does not reflect precise disease timeline

Data derived from EHRs often do not reflect precise timeline of chronic disease development. For example, it is practically impossible to accurately determine the date on which a disease starts to develop without intensive follow-up [443, p. 11]. Instead, EHRs usually record the dates at which patients report symptoms to their physicians and when physicians make and record the diagnosis. This leaves implications on epidemiological studies and timely surveillance of asthma. For instance, asthma prevalence at a certain date or during a certain period, may be underestimated unless future data for the denominator population are considered in the estimation.

6.5.6 Lack of valid methods to assess outcomes

There were no standardised operational definitions for asthma outcomes in the SAIL Databank. For example, to my knowledge, there were no validated and standardised methods to identify asthma-related emergency and secondary care use. Using different diagnosis positions in hospital admission records to ascertain asthma-related hospitalisation may have significant impact on sensitivity and specificity. Lack of standardisation hinders comparability of studies and evidence synthesis.

6.5.7 Public attitudes to data re-use are mixed

The Wales Asthma Observatory is based on linked, anonymised routinely collected data held in the SAIL Databank. The public's awareness of secondary use, anonymisation, and linkage of person level health data is currently limited [444]. In addition, attitudes towards these important concepts are mixed, although a minority of people in the United Kingdom are thought to have concerns about them [444, 445]. Public and patient involvement and partnership, strict information security and governance, and transparency [446] are all needed to win and maintain the trust of data safe haven stakeholders including the public, patients, and data providing organisations. Data safe havens should satisfy high level of competency in safe-guarding data and must have strict protocols to ensure the use of data for the

public's good [447]. These requirements are fulfilled in the operation model of the SAIL Databank [295]. The mixed public's attitudes toward reusing health data for research may have implications on the plan to collect PROMs into the Observatory from the asthma population in Wales.

6.6 Implications and potential uses of the Observatory

The Observatory can be utilised for asthma research as well as to support asthma care in Wales at national, organisational, and patient levels. The design of the Observatory facilitates answering a wide range of questions about asthma in Wales, ranging from prevalence studies to incidence and retrospective longitudinal studies as well as assessment of quality and equality of care.

6.6.1 Implications on health policy and wider societal impact

With near complete geographical coverage, the Observatory can support health policy and service planning across Wales. For example, the Observatory can be used to identify variations in the asthma outcomes between patient groups differing by socioeconomic status or age as demonstrated in [Chapter 5](#). The Observatory can also be used to analyse the trends of these variations and their implication on the disease burden. Linking the Observatory to data on air pollution and housing quality would allow generating insights about the effects of environmental factors data on asthma outcomes. These insights could be used to support healthy urban planning and assess housing regeneration interventions [312, 448, 449].

6.6.2 Implications on service planning and delivery

The Observatory can be used to monitor the trends of asthma incidence, prevalence, and estimating the disease burden on the National Health Services (NHS) at regional and national levels and across patient groups on a regular basis. By including linked data on asthma management and disease outcomes across the levels of care, the Observatory provides an ideal platform for Health Boards to assess their performance in asthma care and evaluate the impact of asthma ser-

vice level interventions. The quality of primary care services, including asthma prescribing and reviews, can be assessed against national guidelines.

6.6.3 Implications on clinical practice and patient outcomes

The Observatory has potential applications in clinical practice to improve the outcomes of patients with asthma. The availability of longitudinally assessed asthma outcomes, which are linkable to person level EHR-derived data in the SAIL Data-bank, facilitates research aiming to improve patient outcomes.

The Observatory can be used in the identification of risk factors, including modifiable ones, of asthma and asthma adverse outcomes. This allows development of algorithms to predict the risk of asthma exacerbations and assess asthma prognosis [70]. I used the Observatory data and its technical platform in the validation of the asthma risk prediction algorithm that has been developed in the “At-Risk Registers Integrated into primary care to Stop Asthma crises in the United Kingdom” (ARRISA-UK) study [450].

The Observatory can also facilitate pharmacovigilance research and studies on how the effectiveness of different interventions differs based on patient characteristics. Such interventions include therapeutic regimens; primary care interventions such as routine and proactive review of asthma status, medications, and action plans; asthma self-management approaches.

The outputs of asthma studies that will use the Observatory can be translated into clinical decision support tools that can be used by healthcare professionals to improve patient care. For example, the ARRISA-UK risk-finding algorithm will be used in GP practice EHR systems to flag records of high-risk asthma patients so that they receive the appropriate attention, disease management, and prevention [451]. In addition, comparative effectiveness studies can be used to develop EHR-tools that provide clinicians with patient-tailored asthma prescribing recommendations. The development and validation of such clinical decision support tools is an important potential application of the Observatory in providing stratified and personalised care.

The Observatory can also contribute to improvement of patient outcomes by facilitating implementation research and informing care pathway development. Linking the Observatory data to asthma-related healthcare utilisation from primary

and secondary care in the SAIL Databank can be useful for the assessment and auditing of guideline implementation and in the improvement of asthma care pathways in primary and secondary care settings.

6.6.4 Implications on asthma research

The Observatory can be used as a platform to conduct various types of person-level observational studies including cross-sectional and longitudinal studies. By linking the Observatory to other data sources in the SAIL Databank, such as education or pollution data, using the linking field, many more research questions about patients in the Observatory can be answered. The Observatory facilitates interrogation of data on asthma patients in a way that increases research efficiency and reproducibility.

6.7 Towards maximising benefits from population-based data to improve asthma outcomes

6.7.1 Improving data capture

In [Chapter 4](#), I proposed recommendations to improve the capture of asthma data in routine health care. These data should not only include clinical data but also wider societal determinants of health. It has been argued that “every doctor writing in the medical record is an information designer and is responsible for making the data recorded easy to find and interpret” [\[452\]](#). There are increasing calls to include health informatics education in medical curricula [\[453\]](#), which could improve data quality awareness among health professionals. However, with short clinical encounters, doctors have limited time to spend on data recording. Facilitating valid, accessible, and timely recording of data is one of the core functions of EHR systems. Informatics approaches including natural language processing (NLP), machine learning, and medical knowledge engineering promise automated capture of data that are locked in narrative clinical documentation [\[454\]](#). EHR design should consider secondary uses of data such as research, service planning, and health policy [\[455, 456\]](#).

6.7.2 Reducing waste from underuse of data

Huge volumes of asthma-related RCD are collected every day across Wales. However, a small amount of these data is actually utilised to advance medical knowledge and improve health care delivery and patient outcomes. It has been argued that *“the biggest waste in the healthcare system is not unnecessary treatment or duplicated test results; it is that we collect data and never use it again.”*⁴ Suboptimal utilisation of these data can arguably lead to unnecessary waste in resources, repeated care mistakes and, most importantly, avoidable adverse outcomes for patients [457].

Despite being a preventable disease, asthma adverse outcomes such as exacerbations and deaths still unnecessarily happen. The National Review of Asthma Deaths (NRAD) report *“Why asthma still kills”* which was published in 2014, found that over two-thirds of asthma deaths were potentially avoidable by better health care and adoption of clinical guidelines as well as better patient adherence to medical advice and treatment [51]. Among its recommendations, the report called for a national audit of asthma. It recommended that asthma audits should be performed on an ongoing basis with involvement and collaboration of patient organisations and commissioners as well as clinicians. A national audit for asthma for Wales and England is currently being scoped and developed in a project led by the Royal College of Physicians [458]. It will focus on helping clinicians improve the documentation of asthma reviews in order to improve patient outcomes. It may also cover several aspects of asthma care such as diagnosis, prescribing, personalised action plan, disease triggers, emergency and secondary care, patient monitoring [459]. Many of these care events can be assessed through the Wales Asthma Observatory on an ongoing basis. Therefore, with its aforementioned strengths, the Observatory is well positioned to play a vital role to support this forthcoming audit programme [458]. Bringing additional asthma-related data such as community prescribing and pathology data to the Observatory will further augment its capability for regularly performed asthma audit and surveillance.

The NRAD was a crucial inquiry into reasons of asthma deaths, which received publicity and attention among respiratory health professional societies and patient organisations. Nonetheless, it is important to evaluate the report’s impact since

⁴Chris Lehmann, MD, Vanderbilt University Medical Center.
<https://www.healthcare-informatics.com/blogs/david-raths/promise-structured-data-capture>

its publication in 2014. Asthma continues to kill and exacerbate. Similar repeated inquiries are therefore likely to be needed in order to explore the avoidable factors. The NRAD was based mainly on manual review of clinical records from health care providers—a burdensome, expensive and time-consuming process. Such an inquiry can potentially be instead performed using routinely collected data. This will allow rapid and timely investigation into asthma adverse outcomes, which can be regularly repeated, possibly as a part of the forthcoming national asthma audit. The overriding and growing need to bridge the gap between research and care has led to developing the concept of *learning health systems (LHS)*. An LHS aims to maximise learning from delivered care on an ongoing basis which can seamlessly inform future care delivery [291]. Ideally, in a cyclical process, data is converted into knowledge, which in turn informs performance, from which new data is generated and feedback to create new knowledge [292]. The opposite case of a learning health system has been dubbed as a ‘forgetting health system’ in which “today’s mistakes are forgotten quickly and are repeated tomorrow” [460]. In the UK, an initiative to develop a learning health system for asthma has started in Scotland [461]. In Wales, the Wales Asthma Observatory is well-suited to be a building block in a future LHS for asthma.

6.7.3 Supportive data-intensive research environment

The Observatory benefits from a supportive, unique research environment and atmosphere in Wales. This doctoral project was funded by Welsh Government’s Health and Care Research Wales (HCRW)⁵ and the Abertawe Bro Morgannwg University Health Board. This funding came in line with the Welsh Government’s vision to extend the investment in novel research applications of routinely collected data [462]. The Government’s report titled “*Maximising the Use of Routine Data for Research in Wales*” described its plans to support research based on routinely collected data that can be translated into actionable knowledge and direct benefits to the residents of Wales [96].

Partnerships involving government bodies, healthcare providers, research community, and funders are vital to maximise benefits from routinely collected data [77]. The SAIL Databank is funded mainly by the Welsh Government’s Health and Care Research Wales and receives support from Farr Institute of Health Infor-

⁵Previously named as the National Institute for Social Care and Health Research (NISCHR).

matics Research which is in turn funded by the Medical Research Council (MRC). Governmental and political support and the cooperation of data providing organisations are all crucial for sustainable routine data collection [77]. This facilitates disease surveillance and enables undertaking up-to-date studies. Continuous collection of evidence is also a core requirement of a successful learning health system [289, 291].

In the SAIL Databank, anonymisation and record linkage of routine health data are performed with support of the National Health Services Wales Informatics Service (NWIS) [140, 141]. The SAIL Databank regularly seeks to link new datasets to the existing ones. For example, the Welsh Result Report Service dataset, which includes pathology data, is expected to be available in the SAIL Databank in 2018. Medication dispensing data is another vital dataset which are expected to be linked to the SAIL Databank. Linked data from various domains of health care in the SAIL Databank are key to study and monitor a chronic condition that is managed at various levels of care such as asthma. The prospect of the Observatory, including further developments and wider utilisation, is highly contingent on the continuous support of routinely collected data research in Wales.

6.7.4 Potentials of asthma big data

Successful experiences of learning from big data to offer personalised services and insights for organisations have been already happening in non-health care sectors. These sectors include, for example, personal banking, marketing, retailing, social networking, and digital personal assistants. Health care is already in a significant lag behind other industries towards unlocking the full potentials of linkable big data [463]. Utilisation of asthma big data is still at a nascent stage. Patient data are scattered across multiple healthcare providers (e.g., GP practices, hospitals, and pharmacies). While medical record linkage in the UK is more than half a century old, not all patient data are currently linked together. Despite being a single organisation, the UK National Health Services effectively has been acting as disconnected providers. Experience of US healthcare providers such as Kaiser Permanente demonstrates promising case of rapid learning from patient data [288].

Asthma data are rapidly expanding in volumes and complexity including data from emerging and non-clinical paradigms. Advanced biomedical technologies such as

smart inhalers, wearable sensors, and internet-enabled devices, and gene analysis enable the collection of rich, granular information about symptoms, breathomics, disease self-management, medication use and adherence, environmental factors, and PROMs [464-468]. Large volume of these data could enable better understanding of the disease aetiology and mechanism, and developing precise risk prediction and decision support tools. Publicly available data on people's internet information seeking behaviour and online behaviour can be used to forecast asthma epidemiology and health care resource utilisation [469-472]. Ubiquitous collection, linkage, analysis of asthma-related big data can unlock substantial advantages for individuals and populations.

6.8 Future work

6.8.1 Improving methods to define asthma patients and assess disease outcomes

The latent class model to identify asthma patients from the GP data, described in [Chapter 3](#), was based on a single calendar year (2014). However, due to the change in data capture practices over time, this model may not be valid for different years. It is therefore important to compare this model with similar models developed in different years.

The change in disease status (e.g., new diagnosis, change in severity, and development of comorbidities) can be tracked over time. Latent transition analysis (LTA) can be used for this purpose. It aims to identify latent statuses of individuals, defined over multiple time points or intervals, which explain the changes in the observed characteristics in the population. LTA can be thought of as an LCA repeated for the same cohort of patients over several intervals. LTA will allow exploration of common disease trajectories of asthma patients and their transition, if any, between different disease subgroups.

In [Chapter 2](#), I found that, in the contemporary asthma literature, asthma severity and control were most often assessed over a 12-month interval. An expert report proposed that asthma control should be assessed over an interval of 2 to 4 weeks for adults and at least 4 weeks for children [2]. A 12-month interval for assessing a disease state is arguably a traditional artefact that makes assessment easier for

investigators. However, this interval may not be the optimal one for assessment of asthma severity and control. For example, a high disease severity inferred from the number of high-dose prescriptions over 12 months does not necessarily mean the disease was 'severe' all over the year. An unjustified interval may lead to misclassification of disease statuses and may undermine the study validity. Therefore, it is important to explore the optimal intervals for disease statuses, possibly using event sequence analysis and related visualisation techniques [473, 474].

6.8.2 Understanding inequalities in asthma outcomes

The analysis of inequalities of asthma outcomes, presented in [Chapter 5](#), revealed a wide gradient in asthma outcomes over five-year follow-up across the socioeconomic groups in Wales. In that analysis, I used the WIMD index quintiles as an explanatory variable. However, it is important to investigate the distribution of asthma outcomes across each of the individual domains that make up the overall deprivation index.

An extension of the study should also investigate whether the inequality gaps can be partly explained by the individual deprivation domains and a number of potential factors. These include patient's health literacy, education attainment, self-management, inhaler technique, environmental smoke exposure, housing conditions, air pollution, comorbidities, ethnicity, as well as quality of primary care and prescribing and proximity to GP practices and emergency departments. The time trends of the social gradient of asthma should be also assessed at national and regional levels.

6.8.3 Monitoring and forecasting asthma trends

The availability of longitudinal data from about two decades in the SAIL Databank allows understanding and forecasting the seasonal and annual trends of asthma. This can be performed by time series analysis and forecasting techniques such as the autoregressive integrated moving average (ARIMA) of counts of healthcare events over time intervals.

6.8.4 Linking additional data sources

Linking additional healthcare datasets to the Observatory will allow answering more research questions. Among such person-level datasets is the Welsh Results Reports Service (WRRS) dataset [475]. This dataset includes laboratory test results across all Wales, and is currently in the process of being transferred into the SAIL Databank. While laboratory test results can be recorded in the general practice EHR system, they suffer from missingness and inconsistency across general practices. Linking the WRRS data to the Observatory will potentially allow improvement in the identification and phenotyping of asthma patients. For example, more accurate data on peripheral eosinophil counts, total and specific IgE can help in the identification of patients with eosinophilic and atopic asthma.

Person-level medication dispensing data are also available in Wales, but are not currently available in the SAIL Databank. These data can be used to complement and crosscheck prescription data in the General Practice (GP) dataset. Prescription data in the GP dataset currently do not include the quantity of the total prescribed dose prescribed to the patient. Dispensing data ideally contain these pieces of information, which can be used, for example, to assess asthma severity and control (e.g., using a more accurate number of actually used short-acting beta agonist inhalers) and identify exacerbations (e.g., by calculating the total supplied dose and duration of administration of oral corticosteroids). Dispensing data also contain the evidence that prescriptions issued by GPs are taken by patients to pharmacies where they are dispensed. This information would provide an indicative picture, although not certain, about patient's adherence to medications. It is not currently known, however, when dispensing data will be available in the SAIL Databank.

Collecting asthma-related PROMs is important to understand patient's perspective about the disease, treatment, disease control, and quality of life. These data include standardised tools such as the AQLQ, ACQ, and ACT, which can be collected during the clinical encounter or at patients' home. The incentives and barriers towards collecting such data need to be explored. Linking those data to RCD in the SAIL Databank will enable identifying healthcare interventions that are most important from patients' perspectives. It will also allow the assessment

of the concurrent validity of objective measures of asthma outcomes against the correspondent PROMs (e.g., RCD-based asthma control definition vs. ACQ).

6.8.5 Improving the Observatory's technical platform

The Observatory's data interrogation interface needs further work to improve the workflow of users. Planned improvements include support to additional data extraction methods, and better documentation of data extraction. I also work with the SAIL technical team towards making this interface available to the public.⁶ This will promote transparency, sharing, and reproducibility of studies that use the Observatory.

6.8.6 Data quality reports

The Observatory could provide reports for the quality of its data. Examples of these reports are those on the quality of recorded lung function data presented in [Chapter 4](#). It is important for users of the Observatory to be aware of data quality issues beforehand. This will inform their study design and analysis, and makes it easier for them to communicate data quality issues in their reports [[157](#)].

6.8.7 Getting ready for SNOMED-CT

In 2018, the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) will replace Read code vocabulary as the terminology of primary care in the UK [[311](#)]. This transition is intended to support standardisation of clinical data capture across the National Health Service. It is unclear how long this transition will take across Wales. However, it will be crucial to ensure compatibility of the Observatory with the new data that will feed into the SAIL Databank from general practices. This requires updating the methods used to identify asthma patients and assess disease outcomes to capture data coded in SNOMED-CT. This can be performed with the help of the Data Migration package, originally provided by the NHS's Health and Social Care Information Centre (NHS Digital) to guide the transition. In addition to recognising the new data coded with SNOMED-CT, the Observatory will still support backward compatibility with the historical Read Code data held in the SAIL Databank.

⁶The interrogation interface outside the SAIL Gateway does not allow access to patient data.

6.9 Conclusions

In this thesis, I described the establishment of the Wales Asthma Observatory using routinely collected data in Wales. The Observatory represents a non-traditional, cost-effective approach to a patient registry and a platform for asthma surveillance and research.

RCD offer unique opportunities to understand asthma, inform health policy and service planning, and improve patient outcomes. However, the inherent limitations of these data impose challenges on those endeavours.

Among these challenges, defining a heterogeneous disease such as asthma using RCD is fraught with pitfalls. In a systematic scoping review of the contemporary literature, I found wide variation in methods to define asthma and its key outcomes using RCD, let alone suboptimal reporting on implementation and validity of those methods. These findings reflected the lack of consensus on the clinical definitions of asthma and its outcomes as well as the wide differences in data resources. The findings highlight the need to reach a consensus on clinical definition of asthma and its outcomes and to harmonise operational definitions in RCD studies.

With the absence of a gold standard for asthma definition, unsupervised analysis of asthma-related RCD coupled with clinico-epidemiological knowledge can identify likely asthma patients. Clustering methods seek to identify the most likely population structure behind the recorded data. Using latent class analysis, I identified fuzzy clusters of asthma and COPD patients, based on which I derived a classification algorithm to identify patients with any or both diseases. The probabilistic case definition of asthma using RCD fits with the probabilistic approach to diagnose asthma in clinical practice. In addition to the LCA-based case definition, the Observatory offers other commonly used asthma case definitions.

Quality of asthma-related RCD in Wales is suboptimal. I described various patterns of missingness and inconsistencies in the recorded asthma data. I recommended measures to improve the capture of asthma data including data quality awareness training of healthcare professionals, improved data entry checks, and data quality-based incentives. NLP promises to capture clinical data that are otherwise locked in narrative documentations.

Facilitating data interrogation is a growing requirement in RCD resources. The Observatory supports shareable, reusable, and scalable data extraction, which promotes research efficiency and reproducibility.

Health inequalities indicate unjust outcome variations within the population and lead to wasted resources. They are a key challenge to health policy. I demonstrated the Observatory's value to health policy by exploring the social gradient of asthma in Wales. I described a wide social gradient in asthma outcomes; despite an excess in asthma-related primary care contact in the most deprived areas, asthma patients there were more than twice likely to be hospitalised for asthma than those in the least deprived areas. This suggested a wide gap in asthma control, that should be further investigated to identify avoidable contributing factors.

There is a growing attention to the waste and harm caused by the underuse of health data. The Observatory is a promising endeavour to maximise the use of asthma data in Wales in research and surveillance. It is well-positioned to play a vital role in the upcoming national asthma audit programme in Wales. Learning health systems effectively learn from experience in order to improve services. The Observatory could be a building block in a future learning health system for asthma in Wales.

References

- [1] Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention (2017 update). 2017.
- [2] Bousquet J, Mantzouranis E, Cruz AA, Ait-Khaled N, Baena-Cagnani CE, et al. Uniform definition of asthma severity, control, and exacerbations: document presented for the World Health Organization Consultation on Severe Asthma. *J Allergy Clin Immunol* 126.5 (2010), 926-938.
- [3] Wenzel SE. Asthma: defining of the persistent adult phenotypes. *Lancet* 368.9537 (2006), 804-813.
- [4] Lötvall J, Akdis CA, Bacharier LB, Bjermer L, Casale TB, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *J Allergy Clin Immunol* 127.2 (2011), 355-360.
- [5] Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med* 18.5 (2012), 716-725.
- [6] Subbarao P, Mandhane PJ, and Sears MR. Asthma: epidemiology, etiology and risk factors. *Can Med Assoc J* 181.9 (2009), E181-E190.
- [7] Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, et al. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 178.3 (2008), 218-224.
- [8] Hekking PPW and Bel EH. Developing and emerging clinical asthma phenotypes. *J Allergy Clin Immunol Pract* 2.6 (2014), 671-80, quiz 681.
- [9] Holgate ST, Wenzel S, Postma DS, Weiss ST, Renz H, et al. Asthma. *Nature Reviews Disease Primers* (2015), 15025.
- [10] Maskell N and Millar A. Oxford Desk Reference: Respiratory Medicine. Oxford University Press, 2009. 471 pp. ISBN: 0199239126.
- [11] Kasper DL, Fauci AS, Hauser SL, Longo DL, and Jameson JL. Harrison's Principles of Internal Medicine, 19 ed. McGraw-Hill Education, LLC CoreSource, 2015.
- [12] Ricciardolo FL, Folkerts G, Folino A, and Mognetti B. Bradykinin in asthma: Modulation of airway inflammation and remodelling. *European Journal of Pharmacology* 827 (2018), 181-188.
- [13] Spiro S. Clinical Respiratory Medicine. Elsevier Health Sciences, 2012. 1000 pp.
- [14] Chapman DG and Irvin CG. Mechanisms of airway hyper-responsiveness in asthma: the past, present and yet to come. *Clinical & Experimental Allergy* 45.4 (2015), 706-719.
- [15] JA Bernstein and ML Levy, eds. Clinical Asthma: Theory and Practice. CRC Press, 2014. 337 pp.
- [16] British Thoracic Society, Scottish Intercollegiate Guidelines Network. British guideline on the management of asthma : A national clinical guideline. 2016.

- [17] Weinberger M and Fischer A. Differential diagnosis of chronic cough in children. *Allergy Asthma Proc* 35.2 (2014), 95–103.
- [18] Miravittles M, Andreu I, Romero Y, Sitjar S, Altés A, et al. Difficulties in differential diagnosis of COPD and asthma in primary care. *Br J Gen Pract* 62.595 (2012), e68–e75.
- [19] Third Expert Panel on the Diagnosis and Management of Asthma. Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma. National Asthma Education, Prevention Program. National Heart, Lung, and Blood Institute (US), Bethesda (MD), 2007.
- [20] Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, et al. Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. *European Respiratory Journal* 40.6 (2012), 1324–1343.
- [21] Melbye H, Drivenes, Dalbak, Leinan, Ostrem, et al. Asthma, chronic obstructive pulmonary disease, or both? Diagnostic labeling and spirometry in primary care patients aged 40 years or more. *International Journal of Chronic Obstructive Pulmonary Disease* (2011), 597.
- [22] Brouwer AFJ, Roorda RJ, and Brand PLP. Home spirometry and asthma severity in children. *European Respiratory Journal* 28.6 (2006), 1131–1137.
- [23] Bacharier LB, Strunk RC, Mauger D, White D, Lemanske RF, et al. Classifying Asthma Severity in Children. *American Journal of Respiratory and Critical Care Medicine* 170.4 (2004), 426–432.
- [24] Irvin CG. Pulmonary function testing in asthma. UpToDate. 2018. URL: <https://www.uptodate.com/contents/pulmonary-function-testing-in-asthma>.
- [25] Calverley PMA. Bronchodilator reversibility testing in chronic obstructive pulmonary disease. *Thorax* 58.8 (2003), 659–664.
- [26] A Harver and H Kotses, eds. *Asthma, Health and Society*. Springer-Verlag GmbH, 2010. ISBN: 0387782842.
- [27] Saydain G, Beck KC, Decker PA, Cowl CT, and Scanlon PD. Clinical Significance of Elevated Diffusing Capacity. *Chest* 125.2 (2004), 446–452.
- [28] Dweik RA. Exhaled nitric oxide analysis and applications. UpToDate. 2018. URL: <https://www.uptodate.com/contents/exhaled-nitric-oxide-analysis-and-applications>.
- [29] Rees J, Kanabar D, and Pattani S. *ABC of Asthma*. BMJ Books, 2010. ISBN: 978-1-4051-8596-7.
- [30] Asthma: diagnosis, monitoring and chronic asthma management [NG80]. 2017. URL: <https://www.nice.org.uk/guidance/ng80/chapter/recommendations>.
- [31] British Thoracic Society, Scottish Intercollegiate Guidelines Network. *British guideline on the management of asthma : A national clinical guideline*. 2014.
- [32] Lenney W, Clayton S, Gilchrist FJ, Price D, Small I, et al. Lessons learnt from a primary care asthma improvement project. *npj Primary Care Respiratory Medicine* 26.1 (2016).
- [33] British National Formulary. URL: <https://www.bnf.org> (visited on 02/07/2018).
- [34] Chee C, Sellahewa L, and Pappachan JM. Inhaled Corticosteroids and Bone Health. *The Open Respiratory Medicine Journal* 8.1 (2015), 85–92.
- [35] Wechsler ME, Wong DA, Miller MK, and Lawrence-Miyasaki L. Churg-Strauss Syndrome in Patients Treated With Omalizumab. *Chest* 136.2 (2009), 507–518.
- [36] *The Global Asthma Report 2014*. The Global Asthma Network, 2014.
- [37] Mallol J, Crane J, von Mutius E, Odhiambo J, Keil U, et al. The International Study of Asthma and Allergies in Childhood (ISAAC) Phase Three: A global synthesis. *Allergol Immunopathol (Madr)* 41.2 (2013), 73–85.

- [38] Masoli M, Fabian D, Holt S, and Beasley R. The global burden of asthma: executive summary of the GINA Dissemination Committee Report. *Allergy* 59.5 (2004), 469–478.
- [39] Anandan C, Nurmatov U, Van Schayck O, and Sheikh A. Is the prevalence of asthma declining? Systematic review of epidemiological studies. *Allergy* 65.2 (2010), 152–167.
- [40] Bousquet J, Bousquet PJ, Godard P, and Daures JP. The public health implications of asthma. *Bull World Health Organ* 83.7 (2005), 548–554.
- [41] Bahadori K, Doyle-Waters MM, Marra C, Lynd L, Alasaly K, et al. Economic burden of asthma: a systematic review. *BMC Pulm Med* 9.1 (2009), 24.
- [42] Fletcher M, Jha A, Dunlop W, Heron L, Wolfram V, et al. Patient Reported Burden of Asthma on Resource Use and Productivity Across 11 Countries in Europe. *Adv Ther* 32.4 (2015), 370–380.
- [43] Mukherjee M, Stoddart A, Gupta RP, Nwaru BI, Farr A, 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* 14.1 (2016).
- [44] Ivanova JI, Bergman R, Birnbaum HG, Colice GL, Silverman RA, et al. Effect of asthma exacerbations on health care costs among asthmatic patients with moderate and severe persistent asthma. *J Allergy Clin Immunol* 129.5 (2012), 1229–1235.
- [45] O’Neill S, Sweeney J, Patterson CC, Menzies-Gow A, Niven R, et al. The cost of treating severe refractory asthma in the UK: an economic analysis from the British Thoracic Society Difficult Asthma Registry. *Thorax* 70.4 (2015), 376–378.
- [46] Pawankar R. Allergic diseases and asthma: a global public health concern and a call to action. *World Allergy Organ J* 7.1 (2014), 12.
- [47] Asher MI, Montefort S, Björkstén B, Lai CK, Strachan DP, et al. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *Lancet* 368.9537 (2006), 733–743.
- [48] Patel SP, Järvelin MR, and Little MP. Systematic review of worldwide variations of the prevalence of wheezing symptoms in children. *Environ Health* 7 (2008), 57.
- [49] Lai CKW, Beasley R, Crane J, Foliaki S, Shah J, et al. Global variation in the prevalence and severity of asthma symptoms: Phase Three of the International Study of Asthma and Allergies in Childhood (ISAAC). *Thorax* 64.6 (2009), 476–483.
- [50] Asthma UK. URL: <http://www.asthma.org.uk/cymru> (visited on 03/31/2016).
- [51] Levy M, Andrews R, Buckingham R, Evans H, Francis C, et al. Why asthma still kills: The National Review of Asthma Deaths (NRAD). Royal College of Physicians, 2014.
- [52] Porta M, Greenland S, Hernán M, Santos Silva I dos, and Last JM. A dictionary of epidemiology. 6th Edition. Oxford University Press, USA, 2014.
- [53] Welsh Assembly Government. Welsh Health Survey. URL: <http://gov.wales/statistics-and-research/welsh-health-survey/> (visited on 02/06/2018).
- [54] Welsh Assembly Government. Welsh Health Survey 2014: Health status, illnesses, and other conditions. 2015.
- [55] Welsh Assembly Government. Welsh Health Survey 2014: Health of children. 2015.
- [56] Doyle M, Dixon J, and Sadler K. Welsh Health Survey Evaluation. National Centre for Social Research, 2010.
- [57] Weekly Returns Service Annual Report 2011. Royal College of General Practitioners - Research & Surveillance Centre, 2011.

- [58] GPOne. QOF. URL: <http://www.gpone.wales.nhs.uk/qof> (visited on 02/06/2018).
- [59] GMS Contract. StatsWales. URL: <https://statswales.wales.gov.uk/Catalogue/Health-and-Social-Care/NHS-Primary-and-Community-Activity/GMS-Contract> (visited on 05/09/2016).
- [60] Kupczyk M, Haahtela T, Cruz AA, and Kuna P. Reduction of asthma burden is possible through National Asthma Plans. *Allergy* 65.4 (2010), 415-419.
- [61] Haahtela T. A 10 year asthma programme in Finland: major change for the better. *Thorax* 61.8 (2006), 663-670.
- [62] Burr ML, Davies BH, Hoare A, Jones A, Williamson IJ, et al. A confidential inquiry into asthma deaths in Wales. *Thorax* 54.11 (1999), 985-989.
- [63] Cowie RL, Revitt SG, Underwood MF, and Field SK. The Effect of a Peak Flow-Based Action Plan in the Prevention of Exacerbations of Asthma. *Chest* 112.6 (1997), 1534-1538.
- [64] Ducharme FM, Zemek RL, Chalut D, McGillivray D, Noya FJD, et al. Written Action Plan in Pediatric Emergency Room Improves Asthma Prescribing, Adherence, and Control. *Am J Respir Crit Care Med* 183.2 (2011), 195-203.
- [65] Honkoop PJ, Taylor DR, Smith AD, Snoeck-Stroband JB, and Sont JK. Early detection of asthma exacerbations by using action points in self-management plans. *Eur Respir J* 41.1 (2012), 53-59.
- [66] Turner S, Burden A, Thomas M, Murray C, and Price D. Predicting asthma exacerbations in children - A real life observational study. Vol. 46. suppl 59. European Respiratory Society (ERS), 2015, PA4511.
- [67] Pinart M, Smit HA, Keil T, Bousquet J, Antó JM, et al. Systematic review of childhood asthma prediction models. *Eur Respir J* 46.suppl 59 (2015), PA4513.
- [68] O'Connor RD, Bleecker ER, Long A, Tashkin D, Peters S, 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* 47.4 (2010), 422-428.
- [69] van der Mark LB, van Wonderen KE, Mohrs J, van Aalderen WM, ter Riet G, et al. Predicting asthma in preschool children at high risk presenting in primary care: development of a clinical asthma prediction score. *Prim Care Respir J* 23.1 (2014), 52-59.
- [70] Blakey JD, Price DB, Pizzichini E, Popov TA, Dimitrov BD, et al. Identifying Risk of Future Asthma Attacks Using UK Medical Record Data: A Respiratory Effectiveness Group Initiative. *J Allergy Clin Immunol Pract* 5.4 (2017), 1015-1024.e8.
- [71] Dimond B. Exploring the legal status of healthcare documentation in the UK. *Br J Nurs* 14 (9 2005), 517-518.
- [72] Hansell A and Aylin P. Routine Data and Health Impact Assessment: A Review of Epidemiological Studies of Socio-economic Influence on Health and Evaluation of Outcome Indicators Derived from Routine Health Data for Health Impact Assessment. 2000.
- [73] NHS Wales Informatics Service. ICD-10. URL: <http://www.nwisinformationstandards.wales.nhs.uk/icd-10> (visited on 02/07/2018).
- [74] NHS Digital. Read Codes. URL: <https://digital.nhs.uk/article/1104/Read-Codes> (visited on 02/07/2018).
- [75] NHS Digital. SNOMED-CT. URL: <https://digital.nhs.uk/snomed-ct> (visited on 02/07/2018).
- [76] NHS Digital. National Clinical Coding Standards OPCS-4 (2017). (Visited on 02/07/2018).

- [77] Morrato EH, Elias M, and Gericke CA. Using population-based routine data for evidence-based health policy decisions: lessons from three examples of setting and evaluating national health policy in Australia, the UK and the USA. *J Public Health (Oxf)* 29.4 (2007), 463–471.
- [78] Raftery J, Roderick P, and Stevens A. Potential use of routine databases in health technology assessment. *Health Technol Assess* 9.20 (2005), 1–92, iii–iv.
- [79] Husain MJ, Brophy S, Macey S, Pinder LM, Atkinson MD, et al. HERALD (health economics using routine anonymised linked data). *BMC Med Inform Decis Mak* 12 (2012), 24.
- [80] Kane R, Wellings K, Free C, and Goodrich J. Uses of routine data sets in the evaluation of health promotion interventions: opportunities and limitations. *Health Educ* 100.1 (2000), 33–41.
- [81] Hemkens LG, Langan SM, and Benchimol EI. Better research reporting to improve the utility of routine data for making better treatment decisions. *J Comp Eff Res* (2016).
- [82] Hemkens LG, Contopoulos-Ioannidis DG, and Ioannidis JPA. Routinely collected data and comparative effectiveness evidence: promises and limitations. *Can Med Assoc J* (2016).
- [83] Anandan C, Simpson CR, Fischbacher C, and Sheikh A. Exploiting the potential of routine data to better understand the disease burden posed by allergic disorders. *Clin Exp Allergy* 36.7 (2006), 866–871.
- [84] Furnham A. Response bias, social desirability and dissimulation. *Pers Individ Dif* 7.3 (1986), 385–400.
- [85] Choi BC and Pak AW. A Catalog of Biases in Questionnaires. *Prev Chronic Dis* 2.1 (2005).
- [86] Cord KAM, Salman RAS, Treweek S, Gardner H, Strech D, et al. Routinely collected data for randomized trials: promises, barriers, and implications. *Trials* 19.1 (2018).
- [87] Fralick M, Kesselheim AS, Avorn J, and Schneeweiss S. Use of Health Care Databases to Support Supplemental Indications of Approved Medications. *JAMA Internal Medicine* 178.1 (2018), 55.
- [88] Whitelaw FG, Nevin SL, Milne RM, Taylor RJ, Taylor MW, et al. Completeness and accuracy of morbidity and repeat prescribing records held on general practice computers in Scotland. *Br J Gen Pract* 46 (404 1996), 181–186.
- [89] Gray J. Use of Read codes in diabetes management in a south London primary care group: implications for establishing disease registers. *BMJ* 326.7399 (2003), 1130.
- [90] Majeed A, Car J, and Sheikh A. Accuracy and completeness of electronic patient records in primary care. *Fam Pract* 25.4 (2008), 213–214.
- [91] Quan H and Williamson T. Guiding the reporting of studies that use routinely collected health data. *Can Med Assoc J* (2016).
- [92] Loke YK. Use of databases for clinical research. *Arch Dis Child* 99.6 (2014), 587–589.
- [93] de Lusignan S. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Fam Pract* 23.2 (2005), 253–263.
- [94] Stammers JG, Kuo A, Hart AJ, Smeeth L, and Skinner JA. Registry Data—Valuable Lessons But Beware the Confounders. *J Arthroplasty* 32.9 (2017), S63–S67.
- [95] Nathan H and Pawlik TM. Limitations of Claims and Registry Data in Surgical Oncology Research. *Ann Surg Oncol* 15.2 (2007), 415–423.
- [96] The National Institute for Social Care and Health Research (NISCHR). Maximising the Use of Routine Data for Research in Wales. 2013.
- [97] Enabling Data Linkage to Maximise the Value of Public Health Research Data: Summary. Wellcome Trust, 2015.

- [98] Linking and sharing routine health data for research in England. PHG Foundation, 2017.
- [99] Schatz M and Zeiger RS. Improving asthma outcomes in large populations. *J Allergy Clin Immunol* 128.2 (2011), 273-277.
- [100] Stempel DA, McLaughlin TP, Stanford RH, and Fuhlbrigge AL. Patterns of asthma control: a 3-year analysis of patient claims. *J Allergy Clin Immunol* 115.5 (2005), 935-939.
- [101] Schatz M, Nakahiro R, Crawford W, Mendoza G, Mosen D, et al. Asthma quality-of-care markers using administrative data. *Chest* 128.4 (2005), 1968-1973.
- [102] Al Sallakh MA, Vasileiou E, Rodgers SE, Lyons RA, Sheikh A, et al. Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. *Eur Respir J* 49 (6 2017).
- [103] Schatz M, Zeiger RS, Yang SJ, Chen W, Crawford W, et al. Change in asthma control over time: predictors and outcomes. *J Allergy Clin Immunol Pract* 2.1 (2014), 59-64.
- [104] Klomp H, Lawson JA, Cockcroft DW, Chan BT, Cascagnette P, et al. Examining asthma quality of care using a population-based approach. *CMAJ* 178.8 (2008), 1013-1021.
- [105] Peled R, Tal A, Pliskin JS, and Reuveni H. A computerized surveillance system for the quality of care in childhood asthma. *J Healthc Qual* 27.5 (2005), 28-33.
- [106] Hsiao HJ, Wang LC, Yang YH, Lee JH, Yu HH, et al. A nationwide survey of the severity, comorbidity, and mortality of hospitalized patients with asthma in Taiwan. *Pediatr Neonatol* 54.4 (2013), 254-260.
- [107] Dombkowski KJ, Wasilevich EA, and Lyon-Callo SK. Pediatric asthma surveillance using Medicaid claims. *Public Health Rep* 120.5 (2005), 515-524.
- [108] Anandan C, Gupta R, Simpson C, Fischbacher C, and Sheikh A. Epidemiology and disease burden from allergic disease in Scotland: analyses of national databases. *J R Soc Med* 102.10 (2009), 431-442. eprint: <http://jrs.sagepub.com/content/102/10/431.full.pdf+html>.
- [109] Simpson CR and Sheikh A. Trends in the epidemiology of asthma in England: a national study of 333,294 patients. *J R Soc Med* 103.3 (2010), 98-106.
- [110] Almqvist C, Wettermark B, Hedlin G, Ye W, and Lundholm C. Antibiotics and asthma medication in a large register-based cohort study - confounding, cause and effect. *Clin Exp Allergy* 42.1 (2012), 104-111.
- [111] Metsälä J, Lundqvist A, Virta LJ, Kaila M, Gissler M, et al. Prenatal and post-natal exposure to antibiotics and risk of asthma in childhood. *Clin Exp Allergy* 45.1 (2015), 137-145.
- [112] Decker WW, Campbell RL, Manivannan V, Luke A, St Sauver JL, et al. The etiology and incidence of anaphylaxis in Rochester, Minnesota: a report from the Rochester Epidemiology Project. *J Allergy Clin Immunol* 122.6 (2008), 1161-1165.
- [113] Watson L, Turk F, James P, and Holgate ST. Factors associated with mortality after an asthma admission: A national United Kingdom database analysis. *Respir Med* 101.8 (2007), 1659-1664.
- [114] Lee T, Kim J, Kim S, Kim K, Park Y, et al. Risk factors for asthma-related healthcare use: longitudinal analysis using the NHI claims database in a Korean asthma cohort. *PLoS One* 9.11 (2014), e112844.
- [115] Peters D, Chen C, Markson LE, Allen-Ramey FC, and Vollmer WM. Using an asthma control questionnaire and administrative data to predict health-care utilization. *Chest* 129.4 (2006), 918-924.

- [116] Dombkowski KJ, Leung SW, and Gurney JG. Prematurity as a predictor of childhood asthma among low-income children. *Ann Epidemiol* 18.4 (2008), 290-297.
- [117] Thomas M, Cleland J, and Price D. Database studies in asthma pharmacoconomics: uses, limitations and quality markers. *Expert Opin Pharmacother* 4.3 (2003), 351-358.
- [118] Price D, Chisholm A, van der Molen T, Roche N, Hillyer EV, et al. Reassessing the evidence hierarchy in asthma: evaluating comparative effectiveness. *Curr Allergy Asthma Rep* 11.6 (2011), 526-538.
- [119] Krishnan JA, Schatz M, and Apter AJ. A call for action: Comparative effectiveness research in asthma. *J Allergy Clin Immunol* 127.1 (2011), 123-127.
- [120] Labrèche F, Kosatsky T, and Przybysz R. Childhood asthma surveillance using administrative data: consistency between medical billing and hospital discharge diagnoses. *Can Respir J* 15.4 (2008), 188-192.
- [121] Dombkowski KJ, Wasilevich EA, Lyon-Callo S, Nguyen TQ, Medvesky MG, et al. Asthma surveillance using Medicaid administrative data: a call for a national framework. *J Public Health Manag Pract* 15.6 (2009), 485-493.
- [122] Travers D, Lich KH, Lippmann SJ, Weinberger M, Yeatts KB, et al. Defining emergency department asthma visits for public health surveillance, North Carolina, 2008-2009. *Prev Chronic Dis* 11 (2014), E100.
- [123] Larsson S, Lawyer P, Garellick G, Lindahl B, and Lundstrom M. Use Of 13 Disease Registries In 5 Countries Demonstrates The Potential To Use Outcome Data To Improve Health Cares Value. *Health Aff (Millwood)* 31.1 (2011), 220-227.
- [124] Newton J and Garner S. Disease registers in England. Institute of Health Sciences, University of Oxford, 2002.
- [125] Metzger J. Using computerized registries in chronic disease care. California HealthCare Foundation, 2004.
- [126] Heaney LG, Brightling CE, Menzies-Gow A, Stevenson M, Niven RM, et al. Refractory asthma in the UK: cross-sectional findings from a UK multicentre registry. *Thorax* 65.9 (2010), 787-794.
- [127] Navaneethan SD, Jolly SE, Schold JD, Arrigain S, Saupe W, et al. Development and validation of an electronic health record-based chronic kidney disease registry. *Clin J Am Soc Nephrol* 6.1 (2011), 40-49.
- [128] Mowat F, Lau, Whyte, Kelsh, Legg, et al. Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. *Clinical Epidemiology* (2011), 259.
- [129] Denaxas SC, George J, Herrett E, Shah AD, Kalra D, et al. Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* 41.6 (2012), 1625-1638.
- [130] Ford DV, Jones KH, Middleton RM, Lockhart-Jones H, Maramba ID, et al. The feasibility of collecting information from people with Multiple Sclerosis for the UK MS Register via a web portal: characterising a cohort of people with MS. *BMC Med Inform Decis Mak* 12.1 (2012).
- [131] Hemmings J and Wilkinson J. What is a public health observatory? *J Epidemiol Community Health* 57.5 (2003), 324-326.
- [132] Ashton JR. Public Health Observatories—the key to timely public health intelligence in the new century. *J Epidemiol Community Health* 54.10 (2000), 724-725.

- [133] Liverpool Public Health Observatory. URL: <https://www.liverpool.ac.uk/psychology-health-and-society/research/public-health-observatory/about/>.
- [134] The Ontario Asthma Surveillance Information System (OASIS). URL: <http://lab.research.sickkids.ca/oasis> (visited on 05/12/2016).
- [135] Meredith S, Taylor V, and McDonald J. Occupational respiratory disease in the United Kingdom 1989: a report to the British Thoracic Society and the Society of Occupational Medicine by the SWORD project group. *Occup Environ Med* 48.5 (1991), 292-298.
- [136] Kopferschmitt-Kubler M, Ameille J, Popin E, Calastreng-Crinquand A, Vervloet D, et al. Occupational asthma in France: a 1-yr report of the Observatoire National de Asthmes Professionnels project. *Eur Respir J* 19.1 (2002), 84-89.
- [137] Desai JR, Wu P, Nichols GA, Lieu TA, and O'Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. *Med Care* 50 Suppl (2012), S30-S35.
- [138] Morris RD, Naumova EN, Goldring J, Hersch M, Munasinghe RL, et al. Childhood asthma surveillance using computerized billing records: a pilot study. *Public Health Rep* 112.6 (1997), 506-512.
- [139] Karakis I, Blumenfeld M, Yegev Y, Goldfarb D, Bolotin A, et al. A computerized surveillance system for asthma. *Int J Health Care Qual Assur* 24.4 (2011), 308-313.
- [140] Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 9 (2009), 3.
- [141] Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 9.1 (2009), 157.
- [142] Belgrave D, Henderson J, Simpson A, Buchan I, Bishop C, et al. Disaggregating asthma: Big investigation versus big data. *J Allergy Clin Immunol* 139.2 (2017), 400-407.
- [143] Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention (2015 update). 2015.
- [144] Hargreave FE and Nair P. The definition and diagnosis of asthma. *Clin Exp Allergy* 39.11 (2009), 1652-1658.
- [145] A plea to abandon asthma as a disease concept. *Lancet* 368.9537 (2006), 705.
- [146] Reddel HK, Bateman ED, Becker A, Boulet LP, Cruz AA, et al. A summary of the new GINA strategy: a roadmap to asthma control. *Eur Respir J* 46.3 (2015), 622-639.
- [147] van den Akker IL, van der Zeijden H, and Verheij TJ. Is spirometry essential in diagnosing asthma? Yes. *Br J Gen Pract* 66.650 (2016), 484-484.
- [148] Levy ML. Is spirometry essential in diagnosing asthma? No. *Br J Gen Pract* 66.650 (2016), 485-485.
- [149] Toelle BG, Peat JK, Salome CM, Mellis CM, and Woolcock AJ. Toward a Definition of Asthma for Epidemiology. *Am Rev Respir Dis* 146.3 (1992), 633-637.
- [150] Pekkanen J and Pearce N. Defining asthma in epidemiological studies. *Eur Respir J* 14.4 (1999), 951-957.
- [151] Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. *Public Health Res Pract* 25.4 (2015).
- [152] Ioannidis JP. Why Most Published Research Findings Are False. *PLoS Med* 2.8 (2005), e124.

- [153] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 61.4 (2008), 344-349.
- [154] Samaan Z, Mbuagbaw L, Kosa D, Borg Debono V, Dillenburg R, et al. A systematic scoping review of adherence to reporting guidelines in health care literature. *J Multidiscip Healthc* 6 (2013), 169-188.
- [155] Pouwels KB, Widyakusuma NN, Groenwold RHH, and Hak E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. *J Clin Epidemiol* 69 (2016), 217-224.
- [156] Hemkens LG, Benchimol EI, Langan SM, Briel M, Kasenda B, et al. The reporting of studies using routinely collected health data was often insufficient. *J Clin Epidemiol* 79 (2016), 104-111.
- [157] Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 12.10 (2015), e1001885.
- [158] Arksey H and OMalley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 8.1 (2005), 19-32.
- [159] Use of appropriate medications for people with asthma. *HEDIS 2003*. Vol. 2. Washington, DC: National Committee for Quality Assurance, 2003, 25-28.
- [160] Wu CL, Andrews AL, Teufel RJ, and Basco WT. Demographic predictors of leukotriene antagonist monotherapy among children with persistent asthma. *J. Pediatr.* 164.4 (2014), 827-831.e1.
- [161] Zeiger RS, Schatz M, Li Q, Chen W, Khatriy DB, et al. High blood eosinophil count is a risk factor for future asthma exacerbations in adult persistent asthma. *J Allergy Clin Immunol Pract* 2.6 (2014), 741-50.
- [162] Schatz M, Zeiger RS, Yang SJ, Chen W, Crawford W, et al. Change in asthma control over time: predictors and outcomes. *J Allergy Clin Immunol Pract* 2.1 (), 59-64.
- [163] Jena AB, Ho O, Goldman DP, and Karaca-Mandic P. The Impact of the US Food and Drug Administration Chlorofluorocarbon Ban on Out-of-pocket Costs and Use of Albuterol Inhalers Among Individuals With Asthma. *JAMA Intern Med* 175.7 (2015), 1171-9.
- [164] McRoy L, Weech-Maldonado R, and Kilgore M. The relationship between direct to consumer advertising (DTCA) and asthma-related emergency department use among Medicaid-enrolled children. *J Asthma* 51.9 (2014), 922-6.
- [165] Wu AC, Butler MG, Li L, Fung V, Kharbanda EO, et al. Primary adherence to controller medications for asthma is poor. *Ann Am Thorac Soc* 12.2 (2015), 161-6.
- [166] Tomasallo CD, Hanrahan LP, Tandias A, Chang TS, Cowan KJ, et al. Estimating Wisconsin asthma prevalence using clinical electronic health records and public health data. *Am J Public Health* 104.1 (2014), e65-73.
- [167] Mukherjee M, Gupta R, Farr A, Heaven M, Stoddart A, 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* 4.11 (2014), e006647.
- [168] Laforest L, Licaj I, Devouassoux G, Chatte G, Martin J, et al. Asthma drug ratios and exacerbations: claims data from universal health coverage systems. *Eur. Respir. J.* 43.5 (2014), 1378-86.

- [169] Lemke LD, Lamerato LE, Xu X, Booza JC, Reiners JJ, et al. Geospatial relationships of air pollution and acute asthma events across the Detroit-Windsor international border: study design and preliminary results. *J Expo Sci Environ Epidemiol* 24.4 (2014), 346-57.
- [170] Jian ZH, Huang JY, Lin FCF, Nfor ON, Jhang KM, et al. The use of corticosteroids in patients with COPD or asthma does not decrease lung squamous cell carcinoma. *BMC Pulm Med* 15 (2015), 154.
- [171] Garne E, Hansen AV, Morris J, Zaupper L, Addor MC, et al. Use of asthma medication during pregnancy and risk of specific congenital anomalies: A European case-malformed control study. *J Allergy Clin Immunol* 136 (6 2015), 1496-502.e1-7.
- [172] Tan NC, Nadkarni NV, Lye WK, Sankari U, et al. Ten-year longitudinal study of factors influencing nocturnal asthma symptoms among Asian patients in primary care. *NPJ Prim Care Respir Med* 25 (2015), 15064.
- [173] Kenyon CC, Rubin DM, Zorc JJ, Mohamad Z, Faerber JA, et al. Childhood Asthma Hospital Discharge Medication Fills and Risk of Subsequent Readmission. *J. Pediatr.* 166.5 (2015), 1121-7.
- [174] Rust G, Zhang S, Holloway K, and Tyler-Hill Y. Timing of emergency department visits for childhood asthma after initial inhaled corticosteroid use. *Popul Health Manag* 18.1 (2015), 54-60.
- [175] Bülow A von, Kriegbaum M, Backer V, and Porsbjerg C. The prevalence of severe asthma and low asthma control among Danish adults. *J Allergy Clin Immunol Pract* 2.6 (2014), 759-67.
- [176] Martin RJ, Price D, Roche N, Israel E, Aalderen WMC van, et al. Cost-effectiveness of initiating extrafine- or standard size-particle inhaled corticosteroid for asthma in two health-care systems: a retrospective matched cohort study. *NPJ Prim Care Respir Med* 24 (2014), 14081.
- [177] Schatz M, Meckley LM, Kim M, Stockwell BT, and Castro M. Asthma exacerbation rates in adults are unchanged over a 5-year period despite high-intensity therapy. *J Allergy Clin Immunol Pract* 2.5 (2014), 570-4.e1.
- [178] Capo-Ramos DE, Duran C, Simon AE, Akinbami LJ, and Schoendorf KC. Preventive asthma medication discontinuation among children enrolled in fee-for-service Medicaid. *J Asthma* 51.6 (2014), 618-26.
- [179] Nordlund B, Melén E, Schultz ES, Grönlund H, Hedlin G, et al. Prevalence of severe childhood asthma according to the WHO. *Respir Med* 108.8 (2014), 1234-7.
- [180] Ismaila A, Corriveau D, Vaillancourt J, Parsons D, Stanford R, et al. Impact of adherence to treatment with fluticasone propionate/salmeterol in asthma patients. *Curr Med Res Opin* 30.7 (2014), 1417-25.
- [181] Fung V, Graetz I, Galbraith A, Hamity C, Huang J, et al. Financial barriers to care among low-income children with asthma: health care reform implications. *JAMA Pediatr* 168.7 (2014), 649-56.
- [182] Dilokthornsakul P, Chaiyakunapruk N, Schumock GT, and Lee TA. Calendar time-specific propensity score analysis for observational data: a case study estimating the effectiveness of inhaled long-acting beta-agonist on asthma exacerbations. *Pharmacoepidemiol Drug Saf* 23.2 (2014), 152-64.
- [183] Adimadhyam S, Schumock GT, Walton S, Joo M, McKell J, et al. Risk of arrhythmias associated with ipratropium bromide in children, adolescents, and young adults with asthma: a nested case-control study. *Pharmacotherapy* 34.4 (2014), 315-23.

- [184] Blais L, Kettani FZ, and Forget A. Associations of maternal asthma severity and control with pregnancy complications. *J Asthma* 51.4 (2014), 391-8.
- [185] Chang J, Freed GL, Prosser LA, Patel I, Erickson SR, et al. Comparisons of health care utilization outcomes in children with asthma enrolled in private insurance plans versus medicaid. *J Pediatr Health Care* 28.1 (2014), 71-9.
- [186] Sullivan PW, Campbell JD, Ghushchyan VH, and Globe G. Outcomes before and after treatment escalation to Global Initiative for Asthma steps 4 and 5 in severe asthma. *Ann. Allergy Asthma Immunol.* 114.6 (2015), 462-9.
- [187] Ali AK, Hartzema AG, Winterstein AG, Segal R, Lu X, et al. Application of multicategory exposure marginal structural models to investigate the association between long-acting beta-agonists and prescribing of oral corticosteroids for asthma exacerbations in the Clinical Practice Research Datalink. *Value Health* 18.2 (2015), 260-70.
- [188] Wu AC, Li L, Fung V, Kharbanda EO, Larkin EK, et al. Use of leukotriene receptor antagonists are associated with a similar risk of asthma exacerbations as inhaled corticosteroids. *J Allergy Clin Immunol Pract* 2.5 (), 607-13.
- [189] Tan CC, McDowell KM, Fenchel M, Szczesniak R, and Kerckmar CM. Spirometry use in children hospitalized with asthma. *Pediatr. Pulmonol.* 49.5 (2014), 451-7.
- [190] Keast SL, Thompson D, Farmer K, Smith M, Nesser N, et al. Impact of a prior authorization policy for montelukast on clinical outcomes for asthma and allergic rhinitis among children and adolescents in a state Medicaid program. *J Manag Care Spec Pharm* 20.6 (2014), 612-21.
- [191] Kim S, Kim J, Park SY, Um HY, Kim K, et al. Effect of pregnancy in asthma on health care use and perinatal outcomes. *J Allergy Clin Immunol* 136 (5 2015), 1215-23.e1-6.
- [192] Confino-Cohen R, Brufman I, Goldberg A, and Feldman BS. Vitamin D, asthma prevalence and asthma exacerbations: a large adult population-based study. *Allergy* 69.12 (2014), 1673-80.
- [193] Tunceli O, Williams SA, Kern DM, Elhefni H, Pethick N, et al. Comparative effectiveness of budesonide-formoterol combination and fluticasone-salmeterol combination for asthma management: a United States retrospective database analysis. *J Allergy Clin Immunol Pract* 2.6 (2014), 719-26.
- [194] Bhattacharjee R, Choi BH, Gozal D, and Mokhlesi B. Association of adenotonsillectomy with asthma outcomes in children: a longitudinal database analysis. *PLoS Med.* 11.11 (2014), e1001753.
- [195] Nanchal R, Kumar G, Majumdar T, Taneja A, Patel J, et al. Utilization of mechanical ventilation for asthma exacerbations: analysis of a national database. *Respir Care* 59.5 (2014), 644-53.
- [196] Tse SM, Charland SL, Stanek E, Herrera V, Goldfarb S, et al. Statin use in asthmatics on inhaled corticosteroids is associated with decreased risk of emergency department visits. *Curr Med Res Opin* 30.4 (2014), 685-93.
- [197] Sumino K, O'Brian K, Bartle B, Au DH, Castro M, et al. Coexisting chronic conditions associated with mortality and morbidity in adult patients with asthma. *J Asthma* 51.3 (2014), 306-14.
- [198] Li L, Vollmer WM, Butler MG, Wu P, Kharbanda EO, et al. A comparison of confounding adjustment methods for assessment of asthma controller medication effectiveness. *Am. J. Epidemiol.* 179.5 (2014), 648-59.
- [199] Hagiwara M, Delea TE, and Stanford RH. Health-care utilization and costs with fluticasone propionate and fluticasone propionate/salmeterol in asthma patients at risk for exacerbations. *Allergy Asthma Proc* 35.1 (2014), 54-62.

- [200] Blais L, Kettani FZ, Forget A, Beauchesne MF, and Lemièrre C. Asthma exacerbations during the first trimester of pregnancy and congenital malformations: revisiting the association in a large representative cohort. *Thorax* 70.7 (2015), 647-52.
- [201] Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, et al. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 64.8 (2011), 821-829.
- [202] Schneeweiss S and Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 58.4 (2005), 323-337.
- [203] Sheikh A, Cornford T, Barber N, Avery A, Takian A, et al. Implementation and adoption of nationwide electronic health records in secondary care in England: final qualitative results from prospective national evaluation in "early adopter" hospitals. *BMJ* 343.oct17 1 (2011), d6054-d6054.
- [204] Frenkel LD. Electronic health records-Applications for the allergist/immunologist: All that glitters is not gold. *Allergy Asthma Proc* 37.4 (2016), 273-278.
- [205] Huzel L, Roos LL, Anthonisen NR, and Manfreda J. Diagnosing asthma: the fit between survey and administrative database. *Can Respir J* 9.6 (2002), 407-412.
- [206] Tinkelman DG, Price DB, Nordyke RJ, and Halbert RJ. Misdiagnosis of COPD and Asthma in Primary Care Patients 40 Years of Age and Over. *J Asthma* 43.1 (2006), 75-80.
- [207] Postma DS and Rabe KF. The Asthma-COPD Overlap Syndrome. *N Engl J Med* 373.13 (2015), 1241-1249.
- [208] McDonald VM and Gibson PG. To define is to limit: perspectives on asthma-COPD overlap syndrome and personalised medicine. *Eur Respir J* 49.5 (2017), 1700336.
- [209] Miravittles M. Diagnosis of asthma-COPD overlap: the five commandments. *Eur Respir J* 49.5 (2017), 1700506.
- [210] Bateman ED, Reddel HK, Zyl-Smit RN van, and Agusti A. The asthma-COPD overlap syndrome: towards a revised taxonomy of chronic airways diseases? *Lancet Respir Med* 3.9 (2015), 719-728.
- [211] Benfante A, Sorino C, and Scichilone N. The asthma-COPD overlap syndrome (ACOS): hype or reality? That is, a curiosity for the media or an opportunity for physicians? *Shortness of Breath* 3.4 (2014), 165-174.
- [212] Delgado-Rodriguez M. Bias. *J Epidemiol Community Health* 58.8 (2004), 635-641.
- [213] Manuel DG, Rosella LC, and Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ* 341 (2010), c4226.
- [214] Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 21 (2 2014), 221-230.
- [215] Ehrenstein V, Petersen I, Smeeth L, Jick S, Benchimol EI, et al. Helping everyone do better: a call for validation studies of routinely recorded health data. *Clinical Epidemiology* (2016), 49.
- [216] Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PLoS One* 9 (6 2014), e99825.

- [217] Herrett E, Thomas SL, Schoonen WM, Smeeth L, and Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 69.1 (2010), 4-14.
- [218] Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, et al. Transparency and Reproducibility of Observational Cohort Studies Using Large Healthcare Databases. *Clin Pharmacol Ther* 99.3 (2016), 325-332.
- [219] Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet (London, England)* 383 (9913 2014), 267-276.
- [220] Bel EH, Sousa A, Fleming L, Bush A, Chung KF, et al. Diagnosis and definition of severe refractory asthma: an international consensus statement from the Innovative Medicine Initiative (IMI). *Thorax* 66.10 (2011), 910-917.
- [221] Chung KF, Wenzel SE, Brozek JL, Bush A, Castro M, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. *Eur Respir J* 43.2 (2014), 343-373.
- [222] Jacob C, Haas JS, Bechtel B, Kardos P, and Braun S. Assessing asthma severity based on claims data: a systematic review. *Eur J Health Econ* (2016).
- [223] Ford ES. The epidemiology of obesity and asthma. *J Allergy Clin Immunol Pract* 115 (5 2005), 897-909, quiz 910.
- [224] Kotz D, Simpson CR, and Sheikh A. Incidence, prevalence, and trends of general practitioner-recorded diagnosis of peanut allergy in England, 2001 to 2005. *J Allergy Clin Immunol* 127.3 (2011), 623-630.e1.
- [225] Custovic A and Nicolaou N. Peanut allergy: overestimated in epidemiology or underdiagnosed in primary care? *J Allergy Clin Immunol: In Practice* 127 (3 2011), 631-632.
- [226] Panesar S, Javad S, Silva Dd, Nwaru B, Hickstein L, et al. The epidemiology of anaphylaxis in Europe: a systematic review. *Allergy* 68.11 (2013), 1353-1361.
- [227] Nwaru BI, Mukherjee M, Gupta RP, Farr A, Heaven M, et al. Challenges of harmonising data from UK national health surveys: a case study of attempts to estimate the UK prevalence of asthma. *J R Soc Med* (2015).
- [228] Mukherjee M, Wyatt JC, Simpson CR, and Sheikh A. Usage of allergy codes in primary care electronic health records: a national evaluation in Scotland. *Allergy* (2016).
- [229] Howie L, Hirsch B, Locklear T, and Abernethy AP. Assessing The Value Of Patient-Generated Data To Comparative Effectiveness Research. *Health Aff (Millwood)* 33.7 (2014), 1220-1228.
- [230] Corren J. Asthma phenotypes and endotypes: an evolving paradigm for classification. *Discov Med* 15.83 (2013), 243-249.
- [231] Naish J, Sturdy P, and Toon P. Appropriate prescribing in asthma and its related cost in east London. *BMJ* 310.6972 (1995), 97-100.
- [232] Coleman N, Halas G, Peeler W, Casaclang N, Williamson T, et al. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. *BMC Family Practice* 16.1 (2015).
- [233] Lusignan S de. The optimum granularity for coding diagnostic data in primary care: report of a workshop of the EFMI Primary Care Informatics Working Group at MIE 2005. *Inform Prim Care* 14 (2 2006), 133-137.

- [234] Harron K, Wade A, Gilbert R, Muller-Pebody B, and Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med Res Methodol* 14.1 (2014), 36.
- [235] Denney MJ, Long DM, Armistead MG, Anderson JL, and Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. *Int J Med Informatics* 94 (2016), 271–274.
- [236] Weiskopf NG and Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 20.1 (2013), 144–151.
- [237] Nissen F, Quint J, Wilkinson S, Müllerova H, Smeeth L, et al. Validation of asthma recording in electronic health records: a systematic review. *Clinical Epidemiology* Volume 9 (2017), 643–656.
- [238] Reitsma JB, Rutjes AWS, Khan KS, Coomarasamy A, and Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J Clin Epidemiol* 62.8 (2009), 797–806.
- [239] Nissen F, Morales DR, Mullerova H, Smeeth L, Douglas IJ, et al. Validation of asthma recording in the Clinical Practice Research Datalink (CPRD). *BMJ open* 7 (8 2017), e017474.
- [240] Hansen S, Strøm M, Maslova E, Mortensen EL, Granström C, et al. A comparison of three methods to measure asthma in epidemiologic studies: results from the Danish National Birth Cohort. *PLoS One* 7.5 (2012), e36328.
- [241] Blais L, Lemièrre C, Menzies D, and Berbiche D. Validity of asthma diagnoses recorded in the Medical Services database of Quebec. *Pharmacoepidemiol Drug Saf* 15.4 (2006), 245–252.
- [242] Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, et al. Identifying patients with physician-diagnosed asthma in health administrative databases. *Can Respir J* 16.6 (2009), 183–188.
- [243] Moth G, Vedsted P, and Schiøtz P. Identification of asthmatic children using prescription data and diagnosis. *Eur J Clin Pharmacol* 63.6 (2007), 605–611.
- [244] Bai JR, Mukherjee DV, Befus M, Apa Z, Lowy FD, et al. Concordance between medical records and interview data in correctional facilities. *BMC Med Res Methodol* 14.1 (2014).
- [245] Hoffmann F and Glaeske G. Prescriptions as a proxy for asthma in children: a good choice? *Eur J Clin Pharmacol* 66.3 (2010), 307–313.
- [246] Veninga C, Denig P, Pont LG, and Haaijer-Ruskamp FM. Comparison of indicators assessing the quality of drug prescribing for asthma. *Health Serv Res* 36 (1 Pt 1 2001), 143–161.
- [247] Barry DM, Burr ML, and Limb ES. Prevalence of asthma among 12 year old children in New Zealand and South Wales: a comparative survey. *Thorax* 46.6 (1991), 405–409.
- [248] Weiland SK, Björkstén B, Brunekreef B, Cookson WOC, Mutius E von, et al. Phase II of the International Study of Asthma and Allergies in Childhood (ISAAC II): rationale and methods. *Eur Respir J* 24.3 (2004), 406–412.
- [249] Ward DG, Halpin DM, and Seamark DA. How accurate is diagnosis of asthma in a general practice database? A review of patients notes and questionnaire-reported symptoms. *Br J Gen Pract* 54.507 (2004), 753–758.
- [250] Muggah E, Graves E, Bennett C, and Manuel DG. Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report. *BMC Public Health* 13 (2013), 16.

- [251] Schatz M, Zeiger RS, Yang SJT, Chen W, Crawford WW, et al. Persistent asthma defined using HEDIS versus survey criteria. *Am J Manag Care* 16.11 (2010), e281–e288.
- [252] Nelson EC, Eftimovska E, Lind C, Hager A, Wasson JH, et al. Patient reported outcome measures in practice. *BMJ* 350 (2015), g7818.
- [253] Kaleta D, Polańska K, Dzionkowska-Zaborszczyk E, Hanke W, and Drygas W. Factors influencing self-perception of health status. *Cent Eur J Public Health* 17 (3 2009), 122–127.
- [254] Al Sallakh M, Rodgers S, Lyons R, Sheikh A, and Davies G. P148 Making sense of patient-reported currently treated asthma using routinely collected data. *Thorax* 71.Suppl 3 (2016), A163.2–A164.
- [255] Peat JK, Toelle BG, Marks GB, and Mellis CM. Continuing the debate about measuring asthma in population studies. *Thorax* 56 (5 2001), 406–411.
- [256] Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, and Bossuyt PMM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 11.50 (2007), iii, ix–iii, 51.
- [257] Linda M. Collins STL. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. John Wiley & Sons Inc, 2010. 285 Seiten. ISBN: 0470228393.
- [258] Howard R, Rattray M, Prosperi M, and Custovic A. Distinguishing Asthma Phenotypes Using Machine Learning Approaches. *Curr Allergy Asthma Rep* 15.7 (2015), 38.
- [259] Wurpts IC and Geiser C. Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. *Front Psychol* 5 (2014).
- [260] Linzer DA and Lewis JB. polCA: An R package for polytomous variable latent class analysis. *J Stat Softw* 42.10 (2011), 1–29.
- [261] Dempster AP, Laird NM, and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol* (1977), 1–38.
- [262] McLachlan G and Peel D. Finite Mixture Models. 1st ed. Wiley Series in Probability and Statistics. Wiley-Interscience, 2000. ISBN: 9780471006268.
- [263] Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 19.6 (1974), 716–723.
- [264] Schwarz G. Estimating the dimension of a model. *Ann Stat* 6.2 (1978), 461–464.
- [265] Nylund KL, Asparouhov T, and Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Struct Equ Modeling* 14.4 (2007), 535–569.
- [266] Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention (2016 update). Accessed: Jun 19, 2017. 2016. URL: <http://ginasthma.org>.
- [267] Fattahi F, Vonk JM, Bulkman N, Fleischeuer R, Gouw A, et al. Old dilemma: asthma with irreversible airway obstruction or COPD. *Virchows Arch* 467.5 (2015), 583–593.
- [268] Strobl C, Malley J, and Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* 14 (4 2009), 323–348.
- [269] Therneau TM and Atkinson EJ. An Introduction to Recursive Partitioning Using the RPART Routines (2015).
- [270] HSCIC - QOF Business Rules team. New GMS Contract QOF Implementation Dataset and Business Rules - Asthma Indicator Set. 2014.

- [271] Atkinson MD, Kennedy JI, John A, Lewis KE, Lyons RA, 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* 17.1 (2017).
- [272] Consent to link. URL: <http://gov.wales/statistics-and-research/making-better-use-existing-data/consent-link/?lang=en>.
- [273] Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* 181.4 (2010), 315–323.
- [274] Garcia-Aymerich J, Benet M, Saeys Y, Pinart M, Basagaña X, et al. Phenotyping asthma, rhinitis and eczema in MeDALL population-based birth cohorts: an allergic comorbidity cluster. *Allergy* (2015).
- [275] Weatherall M, Travers J, Shirtcliffe PM, Marsh SE, Williams MV, et al. Distinct clinical phenotypes of airways disease defined by cluster analysis. *Eur Respir J* 34.4 (2009), 812–818.
- [276] Mäkikyrö EMS, Jaakkola MS, and Jaakkola JJK. Subtypes of asthma based on asthma control and severity: a latent class analysis. *Respir Res* 18.1 (2017).
- [277] Weinmayr G, Keller F, Kleiner A, Prel JB du, Garcia-Marcos L, et al. Asthma phenotypes identified by latent class analysis in the ISAAC phase II Spain study. *Clin Exp Allergy* 43.2 (2013), 223–232.
- [278] Burgel PR, Paillasseur JL, Caillaud D, Tillie-Leblond I, Chanez P, et al. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J* 36.3 (2010), 531–539.
- [279] Ghebre MA, Bafadhel M, Desai D, Cohen SE, Newbold P, et al. Biological clustering supports both Dutch and British hypotheses of asthma and chronic obstructive pulmonary disease. *J Allergy Clin Immunol* 135.1 (2015), 63–72.e10.
- [280] Rodrigo GJ, Neffen H, and Plaza V. Asthma-chronic obstructive pulmonary disease overlap syndrome: a controversial concept. *Curr Opin Allergy Clin Immunol* 17 (1 2017), 36–41.
- [281] Prospero MCF, Sahiner UM, Belgrave D, Sackesen C, Buchan IE, et al. Challenges in Identifying Asthma Subgroups Using Unsupervised Statistical Learning Techniques. *Am J Respir Crit Care Med* 188.11 (2013), 1303–1312.
- [282] Prosser RJ, Carleton BC, and Smith MA. Identifying persons with treated asthma using administrative data via latent class modelling. *Health Serv Res* 43.2 (2008), 733–754.
- [283] Depner M, Fuchs O, Genuneit J, Karvonen AM, Hyvärinen A, et al. Clinical and epidemiologic phenotypes of childhood asthma. *Am J Respir Crit Care Med* 189.2 (2014), 129–138.
- [284] Strong M, South G, and Carlisle R. The UK Quality and Outcomes Framework pay-for-performance scheme and spirometry: rewarding quality or just quantity? A cross-sectional study in Rotherham, UK. *BMC Health Serv Res* 9.1 (2009).
- [285] Jordan K. Quality of morbidity coding in general practice computerized medical records: a systematic review. *Fam Pract* 21.4 (2004), 396–412.
- [286] Asthma UK Centre for Applied Research (AUKCAR). URL: <http://www.aukcar.ac.uk/> (visited on 04/15/2015).
- [287] UK Asthma Observatory. URL: <https://www.aukcar.ac.uk/asthma-observatory> (visited on 01/09/2018).
- [288] Etheredge LM. A Rapid-Learning Health System. *Health Aff (Millwood)* 26.2 (2007), w107–w118.

- [289] Friedman C, Rubin J, Brown J, Buntin M, Corn M, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J Am Med Inform Assoc* (2014).
- [290] Ovretveit J, Nelson E, and James B. Building a learning health system using clinical registers: a non-technical introduction. *J Health Organ Manag* 30.7 (2016), 1105-1118.
- [291] Potts J, Thompson R, Merchant R, Ciemins EL, Bush RW, et al. Learning: Contemplating the unexamined core of Learning Health Systems. *Learn Health Syst* 1.4 (2017), e10036.
- [292] Nwaru BI, Friedman C, Halamka J, and Sheikh A. Can learning health systems help organisations deliver personalised care? *BMC Med* 15.1 (2017).
- [293] Friedman CP, Wong AK, and Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med* 2.57 (2010), 57cm29-57cm29.
- [294] UK Health Departments Research Ethics Service NHRA. Standard Operating Procedures for Research Ethics Committees, Version 7.2 January 2017. 2017. URL: <https://www.hra.nhs.uk/documents/7/standard-operating-procedures-version-7-2.pdf>.
- [295] Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation. *J Biomed Inform* 50 (2014), 196-204.
- [296] Office for National Statistics. Patient Register: quality assurance of administrative data used in population statistics, Dec 2016. 2016. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/patientregisterqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016>.
- [297] Public Health Wales Observatory. Patient Episode Database for Wales (PEDW). URL: <http://www.publichealthwalesobservatory.wales.nhs.uk/pedw>.
- [298] NHS Wales Informatics Service. Annual PEDW Data Tables - Notes & Definitions. URL: <http://www.infoandstats.wales.nhs.uk/docopen.cfm?orgid=869&id=142683>.
- [299] Taylor DR, Bateman ED, Boulet LP, Boushey HA, Busse WW, et al. A new perspective on concepts of asthma severity and control. *Eur Respir J* 32.3 (2008), 545-554.
- [300] Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, et al. Residential Anonymous Linking Fields (RALFs): a novel information infrastructure to study the interaction between the environment and individuals' health. *J Public Health* (2009), fdp041.
- [301] Williams R, Kontopantelis E, Buchan I, and Peek N. Clinical code set engineering for reusing EHR data for research: A review. *J Biomed Inform* 70 (2017), 1-13.
- [302] Walsh SH. The clinicians perspective on electronic health records and how they can affect patient care. *BMJ* 328.7449 (2004), 1184-1187.
- [303] Standard: DS/EN 13826 Peak Expiratory Flow Meters. URL: <http://standards.globalspec.com/std/498042/ds-en-13826>.
- [304] Kampstra P et al. Beanplot: A boxplot alternative for visual comparison of distributions (2008).
- [305] Price DB, Rigazio A, Campbell JD, Bleecker ER, Corrigan CJ, et al. Blood eosinophil count and prospective annual asthma disease burden: a UK cohort study. *Lancet Respir Med* 3.11 (2015), 849-858.
- [306] Andrews JE, Richesson RL, and Krischer J. Variation of SNOMED CT Coding of Clinical Research Concepts among Coding Experts. *Journal of the American Medical Informatics Association* 14.4 (2007), 497-506.

- [307] Yang CL, Simons E, Foty RG, To T, and Dell SD. Administrative Databases Useful For Asthma Surveillance But Overestimate Asthma Prevalence. *B61. PEDIATRIC ASTHMA*. American Thoracic Society, 2012.
- [308] LaBranche J, Crossley J, Gara CD, and Vethanayagam D. Failure of Administrative Data to Guide Asthma Care. American Thoracic Society International Conference Abstracts. doi:10.1164/ajrccm-conference.2015.191.1MeetingAbstracts.A4196. American Thoracic Society, 2015, A4196-A4196.
- [309] Yu TH, Fu PK, and Tung YC. Using medication utilization information to develop an asthma severity classification model. *BMC Med Inform Decis Mak* 17.1 (2017).
- [310] NHS Wales Informatics Service. New GMS Contract QOF Implementation Dataset and Business Rules - Asthma Indicator Set - Wales. 2015.
- [311] NHS Digital, ed. SNOMED CT implementation in primary care. URL: <https://digital.nhs.uk/SNOMED-CT-implementation-in-primary-care>.
- [312] Rodgers SE, Heaven M, Lacey A, Poortinga W, Dunstan FD, et al. Cohort Profile: The Housing Regeneration and Health Study. *Int J Epidemiol* 43.1 (2012), 52-60.
- [313] Rodgers SE. Green-blue space exposure changes and impact on individual-level wellbeing and mental health: a population-wide record-linked natural experiment. 2017. URL: <https://www.journalslibrary.nihr.ac.uk/programmes/phr/160707/#/>.
- [314] Worth A, Hammersley V, Knibb R, Flokstra-de-Blok B, DunnGalvin A, et al. Patient-reported outcome measures for asthma: a systematic review. *NPJ Prim Care Respir Med* 24.1 (2014).
- [315] Beasley R, Asthma TIS of, et al. Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. The International Study of Asthma and Allergies in Childhood (ISAAC) Steering Committee. *Lancet* 351.9111 (1998), 1225-1232.
- [316] Kaur B, Anderson HR, Austin J, Burr M, Harkins LS, et al. Prevalence of asthma symptoms, diagnosis, and treatment in 12-14 year old children across Great Britain (international study of asthma and allergies in childhood, ISAAC UK). *BMJ* 316.7125 (1998), 118-124.
- [317] Bowatte G, Lodge C, Lowe AJ, Erbas B, Perret J, et al. The influence of childhood traffic-related air pollution exposure on asthma, allergy and sensitization: a systematic review and a meta-analysis of birth cohort studies. *Allergy* 70.3 (2014), 245-256.
- [318] GOWERS AM, CULLINAN P, AYRES JG, ANDERSON HR, STRACHAN DP, et al. Does outdoor air pollution induce new cases of asthma? Biological plausibility and evidence: a review. *Respirology* 17.6 (2012), 887-898.
- [319] Sofianopoulou E, Rushton SP, Diggle PJ, and Pless-Mullooli T. Association between respiratory prescribing, air pollution and deprivation, in primary health care. *J Public Health (Oxf)* 35.4 (2013), 502-509.
- [320] Weiland SK, Hüsing A, Strachan DP, Rzehak P, and Pearce N. Climate and the prevalence of symptoms of asthma, allergic rhinitis, and atopic eczema in children. *Occup Environ Med* 61.7 (2004), 609-615.
- [321] Mitchell RG and Dawson B. Educational and social characteristics of children with asthma. *Arch Dis Child* 48.6 (1973), 467.
- [322] Uphoff E, Cabieses B, Pinart M, Valdés M, Antó JM, et al. A systematic review of socioeconomic position in relation to asthma and allergic diseases. *Eur Respir J* 46.2 (2014), 364-374.
- [323] Watson JP, Cowen P, and Lewis RA. The relationship between asthma admission rates, routes of admission, and socioeconomic deprivation. *Eur Respir J* 9.10 (1996), 2087-2093.

- [324] Burr M, Verrall C, and Kaur B. Social deprivation and asthma. *Respir Med* 91.10 (1997), 603-608.
- [325] Anderson H, Cooper J, Bailey P, and Palmer J. INFLUENCE OF MORBIDITY, ILLNESS LABEL, AND SOCIAL, FAMILY, AND HEALTH SERVICE FACTORS ON DRUG TREATMENT OF CHILDHOOD ASTHMA. *Lancet* 318.8254 (1981), 1030-1032.
- [326] Connolly CK, Chan NS, and Prescott RJ. The influence of social factors on the control of asthma. *Postgrad Med J* 65 (763 1989), 282-285.
- [327] World Health Organisation. Health Impact Assessment (HIA). Glossary of terms used. URL: <http://www.who.int/hia/about/glos/en/index1.html>.
- [328] Marmot M and Bell R. Social inequalities in health: a proper concern of epidemiology. *Ann Epidemiol* 26.4 (2016), 238-240.
- [329] Kawachi I, Subramanian SV, and Almeida-Filho N. A glossary for health inequalities. *J Epidemiol Community Health* 56.9 (2002), 647-652.
- [330] Woodward A. Why reduce health inequalities? *J Epidemiol Community Health* 54.12 (2000), 923-929.
- [331] Whitehead M. The Concepts and Principles of Equity and Health. *Int J Health Serv* 22.3 (1992), 429-445.
- [332] Braveman P and Gruskin S. Defining equity in health. *J Epidemiol Community Health* 57.4 (2003), 254-258.
- [333] Braveman PA, Cubbin C, Egerter S, Chideya S, Marchi KS, et al. Socioeconomic Status in Health Research. *JAMA* 294.22 (2005), 2879.
- [334] Yen IH and Syme SL. The Social Environment and Health: A Discussion of the Epidemiologic Literature. *Annu Rev Public Health* 20.1 (1999), 287-308.
- [335] Robert SA. SOCIOECONOMIC POSITION AND HEALTH: The Independent Contribution of Community Socioeconomic Context. *Annu Rev Sociol* 25.1 (1999), 489-516.
- [336] Denny K and Davidson MJ. Area-based socio-economic measures as tools for health disparities research, policy and planning. *Can J Public Health* 103 (8 Suppl 2 2012), S4-S6.
- [337] Townsend P, Phillimore P, and Beattie A. Health and Deprivation: Inequality and the North. Croom Helm, 1988. ISBN: 9780709943518.
- [338] Department for Communities and Local Government. English indices of deprivation 2015. URL: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>.
- [339] Scottish Government. The Scottish Index of Multiple Deprivation. URL: <http://www.gov.scot/Topics/Statistics/SIMD>.
- [340] Welsh Assembly Government. The Welsh Index of Multiple Deprivation (WIMD).
- [341] Welsh Assembly Government. The Welsh Index of Multiple Deprivation (WIMD) 2011 - Summary Report. Cardiff, 2011.
- [342] Welsh Assembly Government. The Welsh Index of Multiple Deprivation (WIMD) 2011: Technical Report. Cardiff, 2011.
- [343] Office for National Statistics. Census geography: An overview of the various geographies used in the production of statistics collected via the UK census. URL: <https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography>.
- [344] Pearson M. Measuring clinical outcome in asthma: a patient-focused approach. Royal College of Physicians. 2000.

- [345] Schatz M, Zeiger RS, Vollmer WM, Mosen D, Mendoza G, et al. The Controller-to-Total Asthma Medication Ratio Is Associated With Patient-Centered As Well As Utilization Outcomes. *Chest* 130.1 (2006), 43-50.
- [346] Lord D, Washington SP, and Ivan JN. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accid Anal Prev* 37.1 (2005), 35-46.
- [347] Zeileis A, Kleiber C, and Jackman S. Regression Models for Count Data in R. *J Stat Softw* 27.8 (2008).
- [348] Kleiber C and Zeileis A. Visualizing count data regressions using rootograms. *Am Stat* 70.3 (2016), 296-303.
- [349] Bacon SL, Bouchard A, Loucks EB, and Lavoie KL. Individual-level socioeconomic status is associated with worse asthma morbidity in patients with asthma. *Respir Res* 10.1 (2009).
- [350] Gupta RP, Mukherjee M, Sheikh A, and Strachan DP. Persistent variations in national asthma mortality, hospital admissions and prevalence by socioeconomic status and region in England. *Thorax* 73.8 (2018), 706-712. eprint: <https://thorax.bmj.com/content/73/8/706.full.pdf>.
- [351] Mielck A, Reitmeir P, and Wjst M. Severity of childhood asthma by socioeconomic status. *Int J Epidemiol* 25.2 (1996), 388-393.
- [352] Laurent O, Filleul L, Havard S, Deguen S, Declercq C, et al. Asthma attacks and deprivation: gradients in use of mobile emergency medical services. *J Epidemiol Community Health* 62.11 (2008), 1014-1016.
- [353] Ungar WJ, Paterson JM, Gomes T, Bikangaga P, Gold M, et al. Relationship of asthma management, socioeconomic status, and medication insurance characteristics to exacerbation frequency in children with asthma. *Ann Allergy Asthma Immunol* 106.1 (2011), 17-23.
- [354] Barr RG, Somers SC, Speizer FE, and Camargo CA. Patient Factors and Medication Guideline Adherence Among Older Women With Asthma. *Arch Intern Med* 162.15 (2002), 1761.
- [355] Poyser M, Nelson H, Ehrlich R, Bateman E, Parnell S, et al. Socioeconomic deprivation and asthma prevalence and severity in young adolescents. *Eur Respir J* 19.5 (2002), 892-898.
- [356] Mazalovic K, Jacoud F, Dima AL, Ganse EV, Nolin M, et al. Asthma exacerbations and socioeconomic status in French adults with persistent asthma: A prospective cohort study. *J Asthma* (2017), 1-9.
- [357] Piantadosi S, Byar DP, and Green SB. The ecological fallacy. *Am J Epidemiol* 127.5 (1988), 893-904.
- [358] Piantadosi S. Invited commentary: ecologic biases. *Am J Epidemiol* 139.8 (1994), 761-764.
- [359] Rieffe C, Oosterveld P, Wijkel D, and Wiefferink C. Reasons why patients bypass their GP to visit a hospital emergency department. *Accid Emerg Nurs* 7 (4 1999), 217-225.
- [360] Griffey RT, Kennedy SK, McGownan L, Goodman M, and Kaphingst KA. Is low health literacy associated with increased emergency department utilization and recidivism? *Acad Emerg Med* 21.10 (2014), 1109-1115.
- [361] Baird B, Charles A, Honeyman M, Maguire D, and Das P. Understanding pressures in general practice. The King's Fund, 2016.
- [362] Boomla K, Hull S, and Robson J. GP funding formula masks major inequalities for practices in deprived areas. *BMJ (Clinical research ed.)* 349 (2014), g7648.

- [363] Thai AL and George M. The effects of health literacy on asthma self-management. *J Asthma Allergy Educ* 1.2 (2010), 50-55.
- [364] Arnlind MH, Wettermark B, Sjöborg B, Dahlén E, Loikas D, et al. Socioeconomic status and the quality of prescribing asthma drugs in Sweden. *J Asthma* 50.8 (2013), 842-849.
- [365] ND B, SL S, KE D, DJ H, and K C. Low health literacy and health outcomes: An updated systematic review. *Ann Intern Med* 155.2 (2011), 97-107. eprint: /data/journals/aim/20235/0000605-201107190-00005.pdf.
- [366] Bender BG and Bender SE. Patient-identified barriers to asthma treatment adherence: responses to interviews, focus groups, and questionnaires. *Immunol Allergy Clin North Am* 25.1 (2005), 107-130.
- [367] Soumerai SB, Ross-Degnan D, Avorn J, McLaughlin TJ, and Choodnovskiy I. Effects of Medicaid Drug-Payment Limits on Admission to Hospitals and Nursing Homes. *N Engl J Med* 325.15 (1991), 1072-1077.
- [368] Public Health Wales Observatory. Emergency Department Data Set (EDDS). URL: <http://www.publichealthwalesobservatory.wales.nhs.uk/edds>.
- [369] NHS Wales Data Dictionary - Admitted Patient Care Data Set (APC Ds) - Data Set Structure. Accessed: 21.11.2017. URL: <http://www.datadictionary.wales.nhs.uk/#!WordDocuments/datasetstructure.htm>.
- [370] Wang Z, May SM, Charoenlap S, Pyle R, Ott NL, et al. Effects of secondhand smoke exposure on asthma morbidity and health care utilization in children: a systematic review and meta-analysis. *Ann Allergy Asthma Immunol* 115.5 (2015), 396-401.e2.
- [371] Dolman R, Gibbon R, and Roberts C. Smoking in Wales: current facts. Wales Centre for Health, 2007.
- [372] Do smoking rates vary between more and less advantaged areas? Office for National Statistics. 2014. URL: <http://webarchive.nationalarchives.gov.uk/20160105204521/http://www.ons.gov.uk/ons/rel/disability-and-health-measurement/do-smoking-rates-vary-between-more-and-less-advantaged-areas-/2012/sty-smoking-rates.html>.
- [373] Surgeon General. The health consequences of smoking—50 years of progress: a report of the Surgeon General. *US Department of Health and Human Services*. 2014.
- [374] Health Literacy: A Prescription to End Confusion (2004).
- [375] Mancuso CA and Rincon M. Impact of health literacy on longitudinal asthma outcomes. *J Gen Intern Med* 21.8 (2006), 813-817.
- [376] Apter AJ, Wan F, Reisine S, Bender B, Rand C, et al. The association of health literacy with adherence and outcomes in moderate-severe asthma. *J Allergy Clin Immunol* 132.2 (2013), 321-327.
- [377] van der Heide I, Wang J, Droomers M, Spreeuwenberg P, Rademakers J, et al. The relationship between health, education, and health literacy: results from the Dutch Adult Literacy and Life Skills Survey. *J Health Commun* 18.sup1 (2013), 172-184.
- [378] Office for National Statistics. Asthma deaths in England and Wales, 2001 to 2015 occurrences. 2016. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/005955asthmadeathsinenglandandwales2001to2015occurrences>.
- [379] Danielis J, Clarke L, Swain K, and Studley R. Mortality Statistics in Wales. Welsh Government, 2013.

- [380] Newacheck PW and Halfon N. Prevalence, impact, and trends in childhood disability due to asthma. *Arch Pediatr Adolesc Med* 154 (3 2000), 287–293.
- [381] Lloyd A, Price D, and Brown R. The impact of asthma exacerbations on health-related quality of life in moderate to severe asthma patients in the UK. *Prim Care Respir J* 16 (1 2007), 22–27.
- [382] GBD 2015 Chronic Respiratory Disease Collaborators. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Respir Med* 5 (9 2017), 691–706.
- [383] Annual Asthma Survey 2016 report. Asthma UK, 2016.
- [384] Stenius-Aarniala BS, Hedman J, and Teramo KA. Acute asthma during pregnancy. *Thorax* 51.4 (1996), 411–414. eprint: <http://thorax.bmj.com/content/51/4/411.full.pdf>.
- [385] Murphy VE, Gibson P, Talbot PI, and Clifton VL. Severe asthma exacerbations during pregnancy. *Obstet Gynecol* 106 (5 Pt 1 2005), 1046–1054.
- [386] Namazy JA, Murphy VE, Powell H, Gibson PG, Chambers C, et al. Effects of asthma severity, exacerbations and oral corticosteroids on perinatal outcomes. *Eur Respir J* 41.5 (2013), 1082–1090. eprint: <http://erj.ersjournals.com/content/41/5/1082.full.pdf>.
- [387] Hsu J, Qin X, Beavers SF, and Mirabelli MC. Asthma-related school absenteeism, morbidity, and modifiable factors. *Am J Prev Med* 51.1 (2016), 23–32.
- [388] Hansen CL, Baelum J, Skadhauge L, Thomsen G, Omland Ø, et al. Consequences of asthma on job absenteeism and job retention. *Scand J Soc Med* 40.4 (2012), 377–384.
- [389] Adams J and White M. Removing the health domain from the Index of Multiple Deprivation 2004—effect on measured inequalities in census measure of health. *J Public Health* 28.4 (2006), 379–383.
- [390] Group RHI. Respiratory Health Delivery Plan 2018–2020: Reducing inappropriate variation and sharing best practice. Welsh Government.
- [391] Davies G, Akbari A, Sallakh MA, Barry S, Boycott K, et al. Consensus Position Statement: Harnessing Big Data to improve Respiratory Health. 2017.
- [392] Soriano JB, Visick GT, Muellerova H, Payvandi N, and Hansell AL. Patterns of Comorbidities in Newly Diagnosed COPD and Asthma in Primary Care. *Chest* 128.4 (2005), 2099–2107.
- [393] Boulet LP and Boulay MÈ. Asthma-related comorbidities. *Expert Rev Respir Med* 5.3 (2011), 377–393.
- [394] Cazzola M, Segreti A, Calzetta L, and Rogliani P. Comorbidities of asthma: current knowledge and future research needs. *Curr Opin Pulm Med* 19.1 (2013), 36–41.
- [395] Macleod U, Mitchell E, Black M, and Spence G. Comorbidity and socioeconomic deprivation: an observational study of the prevalence of comorbidity in general practice. *Eur J Gen Pract* 10.1 (2004), 24–26.
- [396] Marmot M. Social determinants of health inequalities. *Lancet* 365.9464 (2005), 1099–1104.
- [397] Campbell C and McLean C. Ethnic identities, social capital and health inequalities: factors shaping African-Caribbean participation in local community networks in the UK. *Social Science & Medicine* 55.4 (2002), 643–657.
- [398] Evandrou M, Falkingham J, Feng Z, and Vlachantoni A. Ethnic inequalities in limiting health and self-reported health in later life revisited. *J Epidemiol Community Health* 70.7 (2016), 653–662.

- [399] Office for National Statistics. Ethnicity and National Identity in England and Wales: 2011. URL: <https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/ethnicityandnationalidentityinenglandandwales/2012-12-11>.
- [400] StatsWales. Ethnicity by area and ethnic group (2017). URL: <https://statswales.gov.wales/Catalogue/Equality-and-Diversity/Ethnicity/ethnicity-by-area-ethnicgroup>.
- [401] Migration to consistent use of coding and terminology in NHS Wales. Informing Healthcare - National Architecture Design Board. NHS Wales.
- [402] Rumsfeld JS, Joynt KE, and Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol* 13.6 (2016), 350-359.
- [403] Kraus V, Blanco F, Englund M, Karsdal M, and Lohmander L. Call for standardized definitions of osteoarthritis and risk stratification for clinical trials and clinical use. *Osteoarthritis Cartilage* 23.8 (2015), 1233-1241.
- [404] Ruggeri M, Fortuna S, and Rodeghiero F. Heterogeneity of terminology and clinical definitions in adult idiopathic thrombocytopenic purpura: a critical appraisal from a systematic review of the literature. *Haematologica* 93.1 (2008), 98-103.
- [405] Rodeghiero F, Stasi R, Gernsheimer T, Michel M, Provan D, et al. Standardization of terminology, definitions and outcome criteria in immune thrombocytopenic purpura of adults and children: report from an international working group. *Blood* 113.11 (2008), 2386-2393.
- [406] Agusti A, Bel E, Thomas M, Vogelmeier C, Brusselle G, et al. Treatable traits: toward precision medicine of chronic airway diseases. *Eur Respir J* 47.2 (2016), 410-419.
- [407] Malomo M. Pennsylvania Asthma Prevalence Report 2015. Bureau of Epidemiology, Pennsylvania Department of Health, 2015.
- [408] Milet M, Lutzker L, Flattery J, and Wohl-Sanchez L. Asthma in California: A surveillance report. 2013, 47.
- [409] The Burden of Asthma in Mississippi. Office of Health Data and Research, Mississippi State Department of Health, 2014.
- [410] To T, Dell S, Dick PT, Cicutto L, Harris JK, et al. Case verification of children with asthma in Ontario. *Pediatr Allergy Immunol* 17.1 (2006), 69-76.
- [411] Behavioral Risk Factor Surveillance System. Accessed: 21.3.2018. URL: <https://www.cdc.gov/brfss/>.
- [412] Behavioral Risk Factor Surveillance System Questionnaire (2017).
- [413] BRFSS Asthma Call-back Survey. Accessed: 21.3.2018. URL: <https://www.cdc.gov/brfss/acbs>.
- [414] BRFSS Asthma Survey Adult Questionnaire 2013. URL: <https://www.cdc.gov/brfss/acbs> (visited on 03/12/2018).
- [415] BRFSS Asthma Survey Child Questionnaire 2013.
- [416] Welsh Health Survey 2014 - Variable List. 2014.
- [417] Proceedings of the ATS Workshop on Refractory Asthma. *Am J Respir Crit Care Med* 162.6 (2000), 2341-2351.
- [418] Schleich F, Brusselle G, Louis R, Vandenplas O, Michils A, et al. Heterogeneity of phenotypes in severe asthmatics. The Belgian Severe Asthma Registry (BSAR). *Respir Med* 108.12 (2014), 1723-1732.
- [419] Maio S, Baldacci S, Bresciani M, Simoni M, Latorre M, et al. RIItA: The Italian severe/uncontrolled asthma registry. *Allergy* (2017).

- [420] Patrawalla P, Kazeros A, Rogers L, Shao Y, Liu M, et al. Application of the asthma phenotype algorithm from the Severe Asthma Research Program to an urban population. *PLoS One* 7.9 (2012), e44540.
- [421] Data Protection & Medical Research. Parliamentary Office of Science and Technology. UK Parliament., 2005.
- [422] Sariyar M, Borg A, Heidinger O, and Pommerening K. A practical framework for data management processes and their evaluation in population-based medical registries. *Inform Health Soc Care* 38.2 (2013), 104-119.
- [423] Springate DA, Parisi R, Olier I, Reeves D, and Kontopantelis E. rEHR: An R package for manipulating and analysing Electronic Health Record data. *PLoS ONE* 12.2 (2017). Ed. by J Harezlak, e0171784.
- [424] Ainsworth J, Cunningham J, and Buchan I. eLab: Bringing Together People, Data and Methods to Enhance Knowledge Discovery in Healthcare Settings. *Stud Health Technol Inform* 175. HealthGrid Applications and Technologies Meet Science Gateways for Life Sciences (2012), 39-48.
- [425] Bechhofer S, De Roure D, Gamble M, Goble C, and Buchan I. Research objects: Towards exchange and reuse of digital knowledge (2010).
- [426] Bechhofer S, Buchan I, Roure DD, Missier P, Ainsworth J, et al. Why linked data is not enough for scientists. *Future Gener Comput Syst* 29.2 (2013), 599-611.
- [427] Custovic A, Ainsworth J, Arshad H, Bishop C, Buchan I, et al. The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab: team science bringing data, methods and investigators together: Figure 1. *Thorax* 70.8 (2015), 799-801.
- [428] Strachan DP, Anderson HR, Limb ES, O'Neill A, and Wells N. A national survey of asthma prevalence, severity, and treatment in Great Britain. *Arch Dis Child* 70.3 (1994), 174-178.
- [429] Zock JP, Jarvis D, Luczynska C, Sunyer J, and Burney P. Housing characteristics, reported mold exposure, and asthma in the European Community Respiratory Health Survey. *J Allergy Clin Immunol* 110.2 (2002), 285-292.
- [430] Williamson IJ, Martin CJ, McGill G, Monie RD, and Fennerty AG. Damp housing and asthma: a case-control study. *Thorax* 52.3 (1997), 229-234.
- [431] Hughes HK, Matsui EC, Tschudy MM, Pollack CE, and Keet CA. Pediatric Asthma Health Disparities: Race, Hardship, Housing, and Asthma in a National Survey. *Academic Pediatrics* 17.2 (2017), 127-134.
- [432] Cookson R, Mondor L, Asaria M, Kringos DS, Klazinga NS, et al. Primary care and health inequality: Difference-in-difference study comparing England and Ontario. *PLoS One* 12.11 (2017). Ed. by H Zeeb, e0188560.
- [433] FRANK J and HAW S. Best Practice Guidelines for Monitoring Socioeconomic Inequalities in Health Status: Lessons from Scotland. *Milbank Q* 89.4 (2011), 658-693.
- [434] Carney TJ and Kong AY. Leveraging health informatics to foster a smart systems response to health disparities and health equity challenges. *J Biomed Inform* 68 (2017), 184-189.
- [435] Waize T, Anandarajah S, Dhoul N, and de Lusignan S. Variation in clinical coding lists in UK general practice: a barrier to consistent data entry? *J Innov Health Inform* 15.3 (2007), 143-150.

- [436] Kontopantelis E, Buchan I, Reeves D, Checkland K, and Doran T. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UKs quality and outcomes framework. *BMJ Open* 3.8 (2013), e003190.
- [437] Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, et al. Data Linkage: A powerful research tool with potential problems. *BMC Health Serv Res* 10.1 (2010).
- [438] Lim FJ, Blyth CC, Klerk N de, Valenti B, Rouhiainen OJ, et al. Optimization is required when using linked hospital and laboratory data to investigate respiratory infections. *J Clin Epidemiol* 69 (2016), 23–31.
- [439] Bottle A, Jarman B, and Aylin P. Strengths and weaknesses of hospital standardised mortality ratios. *BMJ* 342.jan21 1 (2010), c7116–c7116.
- [440] Powell GA, Bonnett LJ, Tudur-Smith C, Hughes DA, Williamson PR, et al. Using routinely recorded data in the UK to assess outcomes in a randomised controlled trial: The Trials of Access. *Trials* 18.1 (2017).
- [441] Steenkamp M, Frazier L, Lipskiy N, DeBerry M, Thomas S, et al. The National Violent Death Reporting System: an exciting new tool for public health surveillance. *Inj Prev* 12.suppl2 (2006), ii3–ii5.
- [442] Appleyard S and Gilbert D. Innovative Solutions for Clinical Trial Follow-up: Adding Value from Nationally Held UK Data. *Clin Oncol* 29.12 (2017), 789–795.
- [443] Pearce N, Beasley R, Burgess C, and Crane J. *Asthma Epidemiology: Principles and Methods*. Oxford University Press, 1998, 11. ISBN: 0195080165.
- [444] Riordan F, Papoutsi C, Reed JE, Marston C, Bell D, et al. Patient and public attitudes towards informed consent models and levels of awareness of Electronic Health Records in the UK. *Int J Med Informatics* 84.4 (2015), 237–247.
- [445] Haddow G, Bruce A, Sathanandam S, and Wyatt JC. ‘Nothing is really safe’: a focus group study on the processes of anonymizing and sharing of health data for research purposes. *J Eval Clin Pract* 17.6 (2010), 1140–1146.
- [446] Lea NC, Nicholls J, Dobbs C, Sethi N, Cunningham J, et al. Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research. *JMIR Medical Informatics* 4.2 (2016), e22.
- [447] Stockdale J, Cassell J, and Ford E. Giving something back: A systematic review and ethical enquiry of public opinions on the use of patient data for research in the United Kingdom and the Republic of Ireland. *Wellcome Open Res* 3 (2018), 6.
- [448] Corburn J. Urban planning and health disparities: Implications for research and practice. *Plan Pract Res* 20.2 (2005), 111–126.
- [449] Barton H, Grant M, Mitcham C, and Tsourou C. Healthy urban planning in European cities. *Health Promotion International* 24.Supplement 1 (2009), i91–i99.
- [450] Smith JR, Noble MJ, Musgrave S, Murdoch J, Price GM, et al. The at-risk registers in severe asthma (ARRISA) study: a cluster-randomised controlled trial examining effectiveness and costs in primary care. *Thorax* 67.12 (2012), 1052–1060.
- [451] Smith JR, Musgrave S, Payerne E, Noble M, Sims EJ, 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* 19.1 (2018).

- [452] Wyatt JC and Wright P. Design should help use of patients' data. *Lancet* 352.9137 (1998), 1375-1378.
- [453] Moskowitz A, McSparron J, Stone DJ, and Celi LA. Preparing a new generation of clinicians for the era of big data. *Harv Med Stud Rev* 2.1 (2015), 24.
- [454] Friedman C, Rindfleisch TC, and Corn M. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *J Biomed Inform* 46.5 (2013), 765-773.
- [455] Kukafka R, Ancker JS, Chan C, Chelico J, Khan S, et al. Redesigning electronic health record systems to support public health. *J Biomed Inform* 40.4 (2007), 398-409.
- [456] Chalmers DJ, Deakyne SJ, Payan ML, Torok MR, Kahn MG, et al. Feasibility of Integrating Research Data Collection into Routine Clinical Practice Using the Electronic Health Record. *J Urol* 192.4 (2014), 1215-1220.
- [457] Jones KH, Laurie G, Stevens L, Dobbs C, Ford DV, et al. The other side of the coin: Harm due to the non-use of health-related data. *International Journal of Medical Informatics* 97 (2017), 43-51.
- [458] Asthma Audit Development Project. URL: <https://www.rcplondon.ac.uk/projects/asthma-audit-development-project>.
- [459] Asthma Audit Feasibility Study Team. National Asthma Audit Feasibility Study: Phase 1 Full Report. Royal Collage of Physicians, 2017.
- [460] Coiera E. The Forgetting Health System. URL: <https://coiera.com/2015/10/07/the-forgetting-health-system/>.
- [461] Learning Health System for Asthma (NCT00287391). 2017. URL: <https://clinicaltrials.gov/ct2/show/NCT03000491>.
- [462] Use of Routine Data in Research. URL: <https://www.healthandcareresearch.gov.wales/use-of-routine-data-in-research/>.
- [463] Sacristán JA and Dilla T. No big data without small data: learning health care systems begin and end with the individual patient. *J Eval Clin Pract* 21.6 (2015), 1014-1017.
- [464] Howard S, Lang A, Patel M, Sharples S, and Shaw D. Electronic monitoring of adherence to inhaled medication in asthma. *Curr Respir Med Rev* 10.1 (2014), 50-63.
- [465] Ryan D, Blakey J, Chisholm A, Price D, Thomas M, et al. Use of electronic medical records and biomarkers to manage risk and resource efficiencies. *Eur Clin Respir J* 4.1 (2017), 1293386.
- [466] van der Schee MP, Paff T, Brinkman P, Alderen WMC van, Haarman EG, et al. Breathomics in Lung Disease. *Chest* 147.1 (2015), 224-231.
- [467] Bos LD, Sterk PJ, and Fowler SJ. Breathomics in the setting of asthma and chronic obstructive pulmonary disease. *J Allergy Clin Immunol* 138.4 (2016), 970-976.
- [468] Turner SW, Ayres JG, Macfarlane TV, Mehta A, Mehta G, et al. A methodology to establish a database to study gene environment interactions for childhood asthma. *BMC Med Res Methodol* 10 (2010), 107.
- [469] Ram S, Zhang W, Williams M, and Pengetnze Y. Predicting Asthma-Related Emergency Department Visits Using Big Data. *IEEE J Biomed Health Inform* 19.4 (2015), 1216-1223.
- [470] Dai H, Lee BR, and Hao J. Predicting Asthma Prevalence by Linking Social Media Data and Traditional Surveys. *Ann Am Acad Pol Soc Sci* 669.1 (2016), 75-92.

- [471] Mavragani A, Sampri A, Sypsa K, and Tsagarakis KP Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era. *JMIR Public Health Surveill* 4.1 (2018), e24.
- [472] Bian J, Topaloglu U, and Yu F. Towards large-scale twitter mining for drug-related adverse events. *Proceedings of the 2012 international workshop on Smart health and well-being - SHB '12*. ACM Press, 2012.
- [473] Wongsuphasawat K, Guerra Gómez JA, Plaisant C, Wang TD, Taieb-Maimon M, et al. Life-Flow: visualizing an overview of event sequences. *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 2011, 1747-1756.
- [474] Monroe M, Lan R, Lee H, Plaisant C, and Shneiderman B. Temporal event sequence simplification. *IEEE Trans Vis Comput Graph* 19.12 (2013), 2227-2236.
- [475] Welsh Results Reports Service (WRRS). URL: <http://www.wales.nhs.uk/nwis/page/53509/>.
- [476] Scott PJ, Cornet R, McCowan C, Peek N, Fraccaro P, et al. Informatics for Health 2017: Advancing both science and practice. *Journal of Innovation in Health Informatics* 24.1 (2017), 1.

Appendices

Appendix A

Chapter 2 Appendix

A.1 Additional results for Chapter 2

Table A.1.1: Search query used in the systematic scoping review.

Search ID	Query
#1	Search ("humans"[mh] AND English[lang] AND ("2012/11/19"[PDat] : "2015/11/18"[PDat]) AND ("loাত্রfull text"[sb] AND (Asthma[mh] OR "Anti-Asthmatic Agents"[mh]) AND (Asthmatic[tiab] OR Asthmatics[tiab]) NOT "Comment" [pt] NOT "Editorial"[pt] NOT "Letter" [pt] NOT "review"[pt] NOT "Meta-Analysis" [pt] NOT "clinical trial"[pt] NOT "Randomized [pt] NOT "Clinical Trial, Phase I" [pt] NOT "Clinical Trial, Phase II" [pt] NOT "Clinical Trial, Phase III" [pt] NOT "Clinical Trial, Phase IV" [pt] NOT "Controlled Clinical Trial" [pt] NOT "Clinical Topic" [Mesh] NOT "double-blind" [All] NOT "placebo-controlled" [All] NOT "case reports" [pt] NOT "pilot study" [All] NOT "pilot projects" [Mesh] NOT "Prospective Studies" [Mesh])
#2	"GPRD" OR "CPRD" OR "Clinical Practice Research Datalink" OR "General Practice Research Database" OR "SAIL databank" OR "Secure Anonymised Information Linkage Databank" OR "Episode Statistics" OR ("HES" AND "England") OR "Mediplus" OR "DIN-LINK" OR "QRResearch" OR "RIRL" OR "Research in Real Life" OR "Paediatric Intensive Care Audit Network" OR "Scottish Drug Misuse Database" OR "Prescribing Information System" OR "Maternity and Neonatal Linked Database" OR "Office for National Statistics" OR ("ONS" AND ("UK" OR "LWales" OR "PEDW" OR "Primary Care dataset" OR "Primary Care GP dataset" OR "Maternity and Neonatal Linked Database" OR "Prescribing Information System" OR "Scottish Drug Misuse Database" OR "PEDW" OR "Primary Care dataset" OR "Primary Care GP dataset" OR "Maternity and Neonatal Linked Database" OR "Prescribing Information System" OR "Scottish Drug Misuse Database" OR "Scottish Morbidity Records" OR "Scottish morbidity" OR "SMR01" OR "SMR00" OR "Outpatient Attendance dataset" OR "SMR01" OR "General and Day Case dataset" OR "Department of Health Victoria Australia" OR "Clalit Health Service computerized databases" OR "National Health Insurance Research Database" OR "NHIS" OR "Portuguese Anti-Doping authority database" OR "Children's Hospital Srebrnjak Database" OR "CHSD" OR "Practice Team Information" OR "Norwegian Prescription Database" OR "National Health Insurance Claims Database" OR "Longitudinal Health Insurance Database" OR "LHID" OR "Medical Birth Registry" OR "Medical Birth Register" OR "Statistics Norway" OR "National Patient Register" OR "Medco Health Solutions administrative database" OR "Discharge Abstract Database" OR "Ontario Asthma Database" OR "Ontario COPD Database" OR "Ontario Hypertension Database" OR "Ontario Diabetes Database" OR "Surveillance, Epidemiology and End Results" OR "SEER" OR "National Board of Health and Welfare and Statistics" OR "Prescribed Drug Register" OR "Optimum Patient Care Research Database" OR "OPCRD" OR "Hospital Discharge Register" OR "Cause of Death Register" OR "Register of Population and Public Health" OR "British Thoracic Society Difficult Asthma Registry" OR "InterAction Database" OR "IADB" OR "Total Population Register" OR "Multi-Generation Register" OR "Prescribed Drug Register" OR "National Patient Register" OR "NPR" OR "Statistics Denmark" OR "Odense Pharmaco-Epidemiological Database" OR "Register of Medicinal Product Statistics" OR "RMPS" OR "Register of Medicinal Product Statistics" OR "RMPS" OR "National Hospital Register" OR "Hospital In-Patient Enquiry" OR "HIPE" OR "Utrecht General Practitioner Research Network" OR "Christelijke Medische Vereniging" OR "MigMed2" OR "Hospital Discharge Registers" OR "Ambulatory Care Classification System" OR "ACCS" OR "Physician Claims Database" OR "Medical Services Plan" OR "Abstracts Database" OR "Régie de l'Assurance Maladie du Québec" OR "RAMQ" OR "MED-ECHO" OR "Fichier des événements démographiques" OR "The Health Improvement Network" OR "PharMetrics" OR "National Inpatient Sample" OR "Mutuelle Générale de l'Education Nationale" OR "INSS Unified Benefit System"
#3	("Premier" [All] OR "Solucient" [All] OR "Cerner" [All] OR "Ingenix" [All] OR "LabRx" [All] OR "IHCIS" [All] OR "marketscan" [All] OR "market scan" [All] OR "Medstat" [All] OR "Thomson" [All] OR "pharmetrics" [All] OR "healthcore" [All] OR "united healthcare" [All] OR "UnitedHealthcare" [All] OR "UHC" [All] OR "Research Database" [All] OR "Group Health" [All] OR "HCUP" [All] OR "Health Cost" [All] AND "Utilization Project" [All]) OR ("Health Care Cost" [All] AND "Utilization Project" [All]) OR "MEPS" [All] OR "Medical Expenditure Panel Survey" [All] OR "NAMCS" [All] OR "Hospital Ambulatory Medical Care Survey" [All] OR "National Ambulatory Medical Care Survey" [All] OR "NHIS" [All] OR "National Health Interview Survey" [All] OR "Kaiser" [All] OR "Kaiser-Permanente" [All] OR "Kaiser Permanente" [All] OR "HMO Research" [All] OR "Health Maintenance Organization" [All] OR "HMO" [All] OR "Cleveland Clinic" [All] OR "Lovelace" [All] OR "Department of Defense" [All] OR "Henry Ford" [All] OR "i3 Drug Safety" [All] OR "i3" [All] OR "Aetna" [All] OR "Humana" [All] OR "Wellpoint" [All] OR "IMS" [All] OR "Intercontinental" [All] OR "Services" [All] OR "IMS Health" [All] OR "Geisinger" [All] OR "GE Healthcare" [All] OR "MQIC" [All] OR "PHARMO" [All] OR "Institute for Drug Outcome Research" [All] OR "Pilgrim" [All] OR "Sound" [All] OR "Regenstrief" [All] OR "Saskatchewan" [All] OR "Tayside" [All] OR "MEMO" [All] OR "Veterans Affairs" [All] OR "Partners Healthcare" [All] OR "Mayo Clinic" [All] OR "Fleming" [All] OR "Epidemiology" [All] OR "Indiana Health Information Exchange" [All] OR "Indiana Health" [All] OR "Intermountain" [All] OR "blue cross" [All] OR "health partners" [All] OR "health plan" [All] OR "services" [All] OR "Nationwide Inpatient Sample" [All] OR "National Inpatient Sample" [All] OR "medicaid" [All] OR "medicare" [All] OR "MediPlus" [All] OR "Outcome Assessment" [All] OR "tiab" OR (RAMQ [tiab] OR Cigna [tiab] OR (british columbia [tiab] AND (health [tiab] OR (data [tiab] OR (database [tiab] OR (population [tiab]))) OR (CIHI [All Fields] OR (manitoba [tiab] OR (center for health policy [all fields] OR (population [tiab] OR (health insurance [tiab]))) OR (ontario [tiab] AND ((population [tiab] OR (OHIP [tiab] OR (registered persons database [tiab] OR (ICES [All Fields] OR (Institute for Clinical Evaluative Sciences [All Fields]))) OR (Alberta [tiab] AND (health [tiab] OR (data [tiab] OR (database [tiab] OR (population [tiab] OR (pop [tiab] OR (Alberta Health and Wellness [All Fields]))) OR "ICD-9-CM" [All Fields] OR "ICD-10-CM" [All Fields] OR "ICD-9" [All] OR "ICD-10" [All] OR "international statistical classification" [All] OR "classification of diseases" [All] OR "Database Management Systems" [Mesh] OR "Medical Records Systems, Computerized" [Mesh] OR "CPT" [All] OR "Current procedural terminology" [Mesh] OR "OPCS-4" OR "Read code" OR "SNOMED-CT" OR "J45" OR "H33" OR "insurance database" [All] OR "insurance databases" [All] OR "health insurance claim" OR "health insurance" OR "claim data" OR "claims data" OR ("claims" [tw] AND "administrative" [tw]) OR "Insurance Claim Review" [mh] OR (medical OR pharmacy) AND claim) OR (medical OR pharmacy) AND "Insurance Claim Reporting" [mh] OR "routine data" OR "routine health data" OR "routine clinical data" OR "routine electronic data" OR "routinely collected data" OR "routinely-collected health data" OR "drug surveillance" [All] OR "pharmacy data" OR "dispensing data" OR "administrative data" OR "administrative health data" OR "health administration" OR "data" [tw] AND "administrative" [tw]) OR "database analysis" OR "register" OR "registry" OR "Databases, Factual" [Mesh] OR "Databases as topic" [Mesh] OR "Data Warehouse" [All] OR "Record Linkage" [Mesh] OR "record-linkage" OR "record linkage")
#4	#1 AND (#2 OR #3)

Table A.1.2: Charting table showing the data extracted from the reviewed articles.

Variable	Notes
General	
Title/Year	
Country	
Study design	
Routine data sources used	We extracted the routine datasets from the articles. The variables that are measured or derived from these datasets are listed in the table below.
Algorithms and case definitions	
Asthma	Includes asthma severity, comorbidity and validity reporting. Includes study between the study-sites and the study-sites selection which can be used for validation. Also includes validity reporting. Also includes validity reporting.
Asthma severity	
Asthma control	
Asthma exacerbation	
Clarity of reporting routine data-related methods	
<i>Title and abstract</i>	
RECORD 1.1: Types or names of routine data sources used are mentioned	
RECORD 1.2: Geographical regions covered by the routine data sources used are mentioned	
RECORD 1.2: Study-time frame is mentioned	
RECORD 1.3: Record linkage is mentioned (if used)	
<i>Methods</i>	
RECORD 6.1: Selection process of study population is mentioned in detail; clinical codes for asthma case definitions are reported	Clinical codes for asthma case definitions are reported in the supplementary material.
RECORD 6.2: Validation for case definitions	
RECORD 6.3 and 12.3: Record-linkage, if used, is sufficiently explained	
RECORD 7.1: List of codes used in study variables	
RECORD 12.1: Authors explained their level of access to database population	
RECORD 12.2: Data cleaning is explained	
<i>Results</i>	
RECORD 13.1: Details of study population selection	
<i>Discussion</i>	
RECORD 19.1: Implications of using routine data for asthma research (e.g. misclassification bias, unmeasured confounding, missing data, and changing eligibility over time)	

Variable	Notes
RECORD 22.1: Information on how to access study protocol, raw data, and programming code is mentioned	

Table A.1.3: Geographical distribution of the reviewed studies.

Country	Number of studies
US	52
Taiwan	20
Canada	12
Sweden	4
Denmark	4
UK	3
Republic of Korea	2
Israel	2
France	2
Finland	2
Europe	2
USA, UK	1
Spain	1
Singapore	1
Portugal	1
Netherlands	1
Korea	1
Italy	1
Iran	1
Australia	1

Table A.1.4: Study designs of the reviewed studies.

Study design	Number of studies
cohort study, retrospective, using routine database(s)	62
cross-sectional / prevalence study	27
nested case-control	5
cohort study, prospective, using routine database(s)	5
validation study	3
time series analysis	3
population based cross-sectional ecological study	2
cohort study, retrospective, linked to self-reported data	1
cohort study, retrospective, linked to medical charts	1
cohort study, retrospective, linked to death registry	1
case-crossover study	1
case-control study	1
case-control	1

Table A.1.5: Types of EHR-derived data sources used in the reviewed articles.

Type	Number of studies
health insurance claim	72
medical records or medical administrative data	39
dispensing	13
mortality with causes of death	2
public health surveillance database	1
medical birth register	1
health insurance claim + medical records	1
drug adverse effect surveillance	1
disease register	1

Table A.1.6: Algorithms used to identify asthma patients.

Label	Algorithm
asthma	<p>asthma encounter (position = unspecified) ≥ 1 IP (position = unspecified) ≥ 1 Rx ≥ 1 IP (position = unspecified) ≥ 1 OR OP (position = unspecified) ≥ 1 OR ED (position = unspecified) ≥ 1 asthma encounter (position = 1) ≥ 1 IP (position = unspecified) ≥ 1 OR OP (position = unspecified) ≥ 2 IP (position = unspecified) ≥ 1 OR OP (position = unspecified) ≥ 2 IP (position = unspecified) ≥ 1 OR OP (position = unspecified) ≥ 1 IP (position = 1) ≥ 1 ED (position ≤ 3) ≥ 1 asthma encounter (position = unspecified) ≥ 1 OR Rx ≥ 1 asthma encounter (position = unspecified) ≥ 1 AND Rx ≥ 2 within 12 months SABA ≥ 1 AND (ICS, inhaled anticholinergics, Theo, LTRA, OCS, Combo) ≥ 2 OR LABA-ICS ≥ 1 Rx ≥ 1 within 12 months Rx > 1 OR omalizumab ≥ 1 within 12 months OP (position = unspecified) ≥ 3 OR IP (position = unspecified) ≥ 1 OP (position = unspecified) ≥ 2 OR IP (position = unspecified) ≥ 2 OP (position = unspecified) ≥ 2 OR IP (position = 1) ≥ 1 within 12 months OP (position ≤ 2) ≥ 2 OR IP (position = unspecified) ≥ 1 OR ED (position = unspecified) ≥ 1 OP (position ≤ 2) ≥ 2 OR ED (position = 1) ≥ 1 OR IP (position = 1) ≥ 1 IP OR ED (position = 1 or second to a respiratory diagnosis) ≥ 1 IP (position ≥ 1) ≥ 1 OR OP (position ≥ 1) ≥ 2 within 2 years IP (position ≥ 1) ≥ 1 OR OP (position ≥ 1) ≥ 2 IP (position ≥ 1) ≥ 1 OR OP (position ≥ 1) ≥ 1</p>

ED = emergency department visit; GP = general practitioner visit; ICS = inhaled corticosteroid; IP = inpatient hospitalisation; LABA = long-acting beta agonist; oral corticosteroids; OP = outpatient visit; RSV = respiratory syncytial virus; SABA = short-acting beta-2 agonists

Table A.1.6: Algorithms used to identify asthma patients.(cont'd)

Label	Algorithm
	IP (position = unspecified) \geq 1 OR OP (position = unspecified) \geq 3 within 36 months
	IP (position = unspecified) \geq 1 OR OP (position = unspecified) \geq 2 within 12 months
	IP (position = unspecified) \geq 1 OR OP (position = unspecified) \geq 2 OR ED (position = unspecified) \geq 2 OR with
	IP (position = unspecified) \geq 1 OR OP (position = unspecified) \geq 1 OR ED (position = unspecified) \geq 1
	IP (position = unspecified) \geq 1 OR OP (position = unspecified) \geq 1 within 1 year
	IP (position = unspecified) \geq 1 OR OP (position = unspecified) \geq 1 OR Rx \geq 4 within 12 months
	IP (position = unspecified) \geq 1 OR (OP (position = unspecified) + ED (position = unspecified)) \geq 2
	IP (position = unspecified) \geq 1 AND Rx \geq 1
	IP (position = 1) OR OP (position = 1) ever
	IP (position = 1) \geq 1 OR IP (position = 2 or 3, following pneumonia/influenza, respiratory failure, RSV/bronchi
	IP (position \leq 2) \geq 1
	GP (position \leq 2) \geq 1 OR IP (position \leq 2) \geq 1 OR ED (position \leq 2) \geq 1 OR asthma urgent care visit (position
	ED (position = any) \geq 1 OR wheeze \geq 1
	ED (position = 1) \geq 1
	ED (position = 1 to 11) \geq 1
	based on ICS
	based on asthma medications
	asthma encounter (position = unspecified) \geq 2 within 12 months
	asthma encounter (position = unspecified) \geq 2 ever
	asthma encounter (position = unspecified) \geq 2
	asthma encounter (position = unspecified) \geq 1 within 12 months
	asthma encounter (position = unspecified) \geq 1 OR Rx within 6 months
	asthma encounter (position = unspecified) \geq 1 OR Rx \geq 2 ever
	asthma encounter (position = unspecified) \geq 1 OR Rx \geq 1 ever
	asthma encounter (position = unspecified) \geq 1 OR ICS \geq 1 within 12 months
	asthma encounter (position = unspecified) \geq 1 OR asthma medications \geq 2
	asthma encounter (position = unspecified) \geq 1 ever
	asthma encounter (position = unspecified) \geq 1 AND Rx \geq 2 within 24 months
	asthma encounter (position = unspecified) \geq 1 AND Rx \geq 2
	asthma encounter (position = unspecified) \geq 1 AND Rx \geq 1 within 12 months
	asthma encounter (position = unspecified) \geq 1 AND Rx \geq 1
	asthma encounter (position = unspecified) \geq 1 AND current Rx \geq 2
	asthma encounter (position = unspecified) \geq 1 AND current Rx \geq 1
	asthma encounter (position = unspecified) \geq 1 AND Rx \geq 2
	asthma encounter (position = unspecified) \geq 1 AND Rx \geq 1

ED = emergency department visit; GP = general practitioner visit; ICS = inhaled corticosteroid; IP = inpatient hospitalisation; LABA = long-acting beta agonist; oral corticosteroids; OP = outpatient visit; RSV = respiratory syncytial virus; SABA = short-acting beta-2 agonists

Table A.1.6: Algorithms used to identify asthma patients.(cont'd)

Label	Algorithm
current asthma	(asthma encounter (position = 1) \geq 1 OR asthma encounter (position \geq 1) \geq 4) AND (asthma prescriptions \geq 2 within 5 years asthma encounter (position = unspecified) \geq 1
current GP-reported asthma	asthma encounter (position = unspecified) \geq 1 within 12 months
current treated asthma	asthma encounter (position = unspecified) \geq 1 AND Rx \geq 1 within 12 months
treated asthma	Rx $>$ = 3 within 12 months
persistent asthma	Rx \geq 4 OR IP \geq 1 OR ED (position = 1) \geq 1 OR (OP (position = any) \geq 1 AND Rx \geq 2) within 12 months Rx \geq 4 OR IP \geq 1 OR ED (position = 1) \geq 1 OR (OP (position = any) \geq 1 AND Rx \geq 2) within 24 months Rx \geq 4 OR IP \geq 1 OR ED (position = 1) \geq 1 OR (OP (position = any) \geq 1 AND Rx \geq 2) within 12 months Rx \geq 4 OR IP \geq 1 OR ED (position = 1) \geq 1 OR (OP (position = any) \geq 1 AND Rx \geq 2) IP (position = unspecified) \geq 1 OR ED (position = unspecified) \geq 1 OR OCS \geq 3) within 12 months

ED = emergency department visit; GP = general practitioner visit; ICS = inhaled corticosteroid; IP = inpatient hospitalisation; LABA = long-acting beta agonist; oral corticosteroids; OP = outpatient visit; RSV = respiratory syncytial virus; SABA = short-acting beta-2 agonists

Table A.1.7: Approaches used in identifying asthma patients.

Criteria base on	Diagnostic label used	Number of studies
	'asthma'	68
Asthma diagnostic/management codes	persistent asthma	1
	acute asthma	1
	current asthma	1
	current GP-reported and diagnosed asthma	1
Asthma diagnostic/management codes AND asthma prescription codes	'asthma'	11
	current treated asthma	1
	persistent asthma	2
Asthma prescription codes	asthma	22
	treated asthma	1
	persistent asthma	4

Table A.1.8: Age restriction approaches used in asthma patient identification.

Age limits	Studies	Number of studies
Minimum age limits		
6 months	A1	1
2 years	A2-A7	6
3 years	A8-A10	3
5 years	A11-A13	3
Maximum age limits		
44 years	A14	1
55 years	A15	1
60 years	A16	1
64 years	A17	1

Table A.1.9: Co-morbidities and conditions based on which asthma patients were excluded.

Condition	Number of studies
COPD	11
Cystic fibrosis	13
Pulmonary embolism	3
Bronchiectasis	4
Pulmonary hypertension	4
Congestive heart failure	3
Emphysema	3
Chronic bronchitis	2
Immunodeficiency	2
Churg Strauss syndrome	1
Wegener syndrome	1
Sarcoidosis	1
Smoker over age of 60	1
Pneumonia	1
Anti-cholinergic prescription as a proxy of COPD	1
Chronic respiratory failure	1
Achondroplasia	1
Bronchopulmonary dysplasia	1
Respiratory cancer	1
Active or past tobacco use	1
Primary ciliary dyskinesia	1
Tracheomalacia	1
Bronchiolitis/RSV infection	2
Pneumoconiosis	1
Other lung diseases due to external agents	1
Psychosis	1
“Perinatal respiratory condition”	1
Tracheostomy	1
Gastrostomy	1

Table A.1.10: Algorithms used to ascertain asthma severity using EHR data.

Variable	Algorithm	Interval (months)	Appears in
Mild asthma	either 500 mg/day of ICS monotherapy (in beclomethasone chlorofluorocarbon equivalents) OR 250 mg/day of ICS + additional controller AND either ≤ 3 SABA doses per week on average (each = 2 salbutamol 100mg puffs) OR both 4–10 doses of SABA per week on average AND no moderate to severe asthma exacerbation (defined as asthma ED visit OR asthma hospitalisation OR short-course OCS)	12	[A18]
Moderate asthma	NOT mild asthma NOR severe asthma as defined in the same study	12	[A18]
Severe asthma	> 1000 mg/day of ICS AND one of > 3 SABA per week on average OR ≥ 1 moderate to severe asthma exacerbation OR both lower doses of ICS with >10 SABA doses per week on average AND 1 moderate to severe asthma exacerbation > 6 albuterol refills per year	12	[A2]
	GINA step 4 or higher	unclear	[A19]

ED = emergency department; GINA = Global Initiative for Asthma; HEDIS = Healthcare Effectiveness Data and Information Set; ICS = inhaled corticosteroid; corticosteroids; OP = outpatient; SABA = short-acting β_2 agonists.

Table A.1.10: Algorithms used to ascertain asthma severity using EHR data. (cont'd).

Variable	Algorithm	Interval (months)	Appears in
	continuous treatment with ICS (at least 800 mg budesonide daily or equivalent [500 mg fluticasone]) and (LABA)	12	[A20]
	presence of persistent asthma according to the HEDIS criteria associated with readmission	12	[A4]
	OR		
	presence of complex chronic condition within the prior year associated with readmission		
	based on number of ICS, LABA, and OCS prescriptions	24	[A21]
	based on number of asthma prescriptions (including OCS)	12	[A22]
	based on asthma hospitalisation, asthma Ed visits, outpatient visits for asthma exacerbation, number of SABA dispensings, number of OCS dispensings	12	[A23]
	based on number of asthma hospitalisations, asthma ED visits, SABA prescriptions, OCS prescriptions, and asthma exacerbations over 6 months	6	[A13]
	based on acute OCS course, mean daily SABA dose, number of asthma consultations with no acute OCS	12	[A16]
	ICS (>800 mg budesonide daily) AND second controller	12	[A14]
	OR		
	ICS–LABA		
	OR		
	omalizumab		
	According to GINA 2006 classification of severity	unclear	[A24]
	Based on OCS prescriptions	unclear	[A25]
	Number of OP over variable follow–up periods	variable	[A26]
'More severe asthma'	≥ 2 SABA prescriptions within 90 days of ICS prescriptions	3	[A27]

ED = emergency department; GINA = Global Initiative for Asthma; HEDIS = Healthcare Effectiveness Data and Information Set; ICS = inhaled corticosteroid; corticosteroids; OP = outpatient; SABA = short-acting β_2 agonists.

Table A.1.10: Algorithms used to ascertain asthma severity using EHR data. (cont'd).

Variable	Algorithm	Interval (months)	Appears in
	HEDIS criteria for persistent asthma: ≥ 1 asthma hospitalisation OR ≥ 1 asthma ED visit OR ≥ 4 asthma prescriptions OR both ≥ 4 asthma outpatient visits AND ≥ 2 asthma prescriptions ≥ 1 asthma hospitalisations or ED visits	24 12	[A6] [A28]
Low-risk asthma	no asthma ED visits AND no asthma hospitalisations AND < 15 β-agonist canisters dispensed AND no OCS dispensed	12	[A9]
Moderate-risk asthma	no asthma ED visits AND no asthma hospitalisations AND only one of: ≥ 15 β-agonist canisters dispensed OR ≥ 1 OCS dispensings	12	[A9]
High-risk asthma	≥ 1 asthma ED visits OR ≥ 1 asthma hospitalisations OR both: ≥ 15 β-agonist canisters dispensed AND ≥ 1 OCS dispensings	12	[A9]

ED = emergency department; GINA = Global Initiative for Asthma; HEDIS = Healthcare Effectiveness Data and Information Set; ICS = inhaled corticosteroid; corticosteroids; OP = outpatient; SABA = short-acting β₂ agonists.

Table A.1.11: Algorithms used to ascertain asthma exacerbation using EHR data.

Variable	Study	Algorithm	OCS				IP	ED
			alone	+ OP	+ IP, ED, OP or GP	+ IP or ED		
Exacerba- tion	[A29]	≥ 1 OCS prescription for < 21 days OR ≥ 4 asthma GP visits per year OR ≥ 5 SABA prescriptions per year	< 21 days					
	[A14]	≥ 1 OCS prescription OR Hospitalisation or ED visit for asthma, status asthmaticus, pneumonia, dyspnoea, or respiratory insufficiency	≥ 1			p	p	
	[A31]	asthma hospitalisation OR asthma ED visit OR OCS pharmacy claim	p			p	p	
	[A19]	OCS prescription within 7 days of any asthma encounter (which may include hospitalisation, ED, outpatient, or GP visit, ascertained with the ICD–9 code 493 as a primary diagnosis or as a secondary diagnosis provided the primary diagnosis is another respiratory condition) Variation: asthma encounter = asthma hospitalisation or ED visit only			within 7 days	within 7 days		

p = present; a = absent; OCS = oral corticosteroids; AE = asthma exacerbation; SABA = short-acting β 2 agonists; ED = emergency department; ICD = International Classification of Diseases; OP = outpatient; GP = general practitioner.

Table A.1.1.1: Algorithms used to ascertain asthma exacerbation using EHR data. (cont'd)

Variable	Study	Algorithm	OCS				IP	ED
			alone	+ OP	+ IP, ED, OP or GP	+ IP or ED		
	[A21]	OCS with asthma as indication OR asthma ED visit OR asthma hospitalisation	indica- tion is asthma				p	p
	[A22]	OCS prescription OR number of asthma GP visits OR hospitalisation for asthma (as a primary diagnosis; variation: as a primary or secondary diagnosis)	p				p	
	[A32]	Occurrence, after 3 months from previous asthma hospitalisation, if any, of: OCS short-course OR asthma ED visit (ICD-9-CM = 493) OR asthma hospitalisation (ICD-9-CM = 493)	p				p	p
	[A33]	Primary hospital discharge diagnosis of asthma exacerbation					p	

p = present; a = absent; OCS = oral corticosteroids; AE = asthma exacerbation; SABA = short-acting β 2 agonists; ED = emergency department; ICD = International Classification of Diseases, 9th Revision, Clinical Modification; OP = outpatient; GP = general practitioner.

Table A.1.11: Algorithms used to ascertain asthma exacerbation using EHR data. (cont'd)

Variable	Study	Algorithm	OCS				IP	ED
			alone	+ OP	+ IP, ED, OP or GP	+ IP or ED		
	[A34]	ED visit with primary diagnosis of asthma OR outpatient visit with diagnosis of asthma exacerbation OR diagnosis of asthma with OCS prescription (< 14-day supply) within 5 days OR hospitalisation with diagnosis of asthma (primary) or asthma exacerbation (any position)		< 14-day supply; within 5 days			p	p
	[A35]	OCS use OR asthma ED visit OR asthma hospitalisation	p				p	p
	[A23]	outpatient visit with primary diagnosis of asthma (ICD-9-CM = 493) and OCS dispensing within 5 days OR asthma ED visit (ICD-9-CM = 493.xx) OR asthma hospitalisation (ICD-9-CM = 493.xx)		within 5 days			p	p

p = present; a = absent; OCS = oral corticosteroids; AE = asthma exacerbation; SABA = short-acting β 2 agonists; ED = emergency department; ICD = International Classification of Diseases, 9th Revision, Clinical Modification; OP = outpatient; GP = general practitioner.

Table A.1.1.1: Algorithms used to ascertain asthma exacerbation using EHR data. (cont'd)

Variable	Study	Algorithm	OCS				IP	ED
			alone	+ OP	+ IP, ED, OP or GP	+ IP or ED		
	[A36]	ED visit with any asthma diagnosis OR hospitalisation with primary diagnosis asthma OR OCS with asthma claim within 7 days			within 7 days		p	p
	[A37]	one-off OCS prescription (short-course)	p					
	[A38]	OCS within 7 days of an encounter with diagnosis of exacerbation or uncontrolled asthma			P			
	[A39]	≥ 1 asthma ED visits OR ≥ 1 asthma hospitalisations OR OCS prescriptions	p				p	p
	[A40]	asthma ED visit (ICD-9-CM = 493) AND/OR asthma hospitalisation (ICD-9-CM = 493)					p	p
	[A8]	Encounter with asthma exacerbation code						
	[A16]	acute OCS OR unscheduled asthma hospitalisation OR ED visit	p				p	p

p = present; a = absent; OCS = oral corticosteroids; AE = asthma exacerbation; SABA = short-acting β_2 agonists; ED = emergency department; ICD = International Classification of Diseases, 9th revision, Clinical Modification; OP = outpatient; GP = general practitioner.

Table A.1.11: Algorithms used to ascertain asthma exacerbation using EHR data. (cont'd)

Variable	Study	Algorithm	OCS				IP	ED
			alone	+ OP	+ IP, ED, OP or GP	+ IP or ED		
	[A17]	new occurrence (after ≥ 8 -day wash-up period) of: Both Asthma outpatient visit (with a code for acute exacerbation, status asthmaticus, acute asthma attack, uncontrolled asthma, asthmatic bronchitis) AND OCS dispensing within 7 days OR Asthma ED visit or hospitalisation (asthma diagnosis position = 1 OR position = 2 following a primary respiratory diagnosis)		p			p	p
	[A41]	Asthma hospitalisation OR Asthma ED visit OR Asthma OP visit with OCS prescription		p			p	p
	[A24]	Based on rescue medications						
Moderate-to-severe exacerbation	[A18]	OCS short-course OR asthma ED visit OR asthma hospitalisation	p				p	p

p = present; a = absent; OCS = oral corticosteroids; AE = asthma exacerbation; SABA = short-acting β_2 agonists; ED = emergency department; ICD = International Classification of Diseases; OP = outpatient; GP = general practitioner.

Table A.1.11: Algorithms used to ascertain asthma exacerbation using EHR data. (cont'd)

Variable	Study	Algorithm	OCS				IP	ED
			alone	+ OP	+ IP, ED, OP or GP	+ IP or ED		
	[A13]	OCS within 7 days of asthma outpatient visit OR Asthma ED visit		within 7 days			p	
Moderate exacerbation	[A42]	≥ 1 ED visits for asthma AND no hospitalisation for asthma				a	p	
Severe exacerbation	[A42]	≥ 1 hospitalisation for asthma as a primary or admission diagnosis					p	

p = present; a = absent; OCS = oral corticosteroids; AE = asthma exacerbation; SABA = short-acting β 2 agonists; ED = emergency department; ICD = International Classification of Diseases; OP = outpatient; GP = general practitioner.

Table A.1.12: Algorithms used to assess asthma control using EHR data

Variable	Algorithm	Interval	App
Low control/ uncontrolled asthma	≥ 600 doses (1 dose = 1 puff) of SABA in the recent year OR ≥ 1 exacerbation in the recent year, defined as: ≥ 1 hospitalisation or ED visit associated with ICD–10 code for asthma, status asthmaticus, pneumonia, dyspnoea, or respiratory insufficiency OR ≥ 1 OCS prescription	12 months	[A1]
	≥ 1 hospitalisation or ED visit OR dispensing of OCS for ≥ 3 days	12 months	[A3]
	≥ 1 ED or OP visit for asthma OR ≥ 1 antibiotic prescriptions	unclear	[A4]
	≥ 1 moderate to severe asthma exacerbation AND > 3 and 10 SABA doses per week on average for mild and moderate/severe asthma, respectively	12 months	[A1]
	≥ 2 acute care contact within 1 month OR ≥ 3 reliever inhaler uses per week OR severe exacerbation requiring ICU/intubation in the last 3 months OR asthma hospitalisation in the last 3 months	1-3 months	[A2]

LRTI = lower respiratory tract infection; SABA = short-acting β agonists; OCS = oral corticosteroids; GP = general practitioner; ED = emergency department; ICU = intensive care unit

Table A.1.12: Algorithms used to assess asthma control using EHR data (cont'd)

Variable	Algorithm	Interval	App
	at the assessment date > 2 asthma drug classes OR ≥ 1 SABA OR in 12 months ≥ 1 OCS OR ≥ 6 SABA OR ≥ 1 asthma ED visits OR ≥ 1 asthma hospitalisations	12 months	[A3]
Low-risk asthma control	Absence of all the following: hospitalisation, ED, and unscheduled outpatient visits for asthma (ascertained by any asthma or LRTI codes) GP consultation for LRTI requiring antibiotics acute course of OCS	12 months	[A1]
	based on number of OCS prescriptions per year	12 months	[A4]
Impairment-domain asthma control	based on number of β -agonists prescriptions per year	12 months	[A4]
	> 2 salbutamol puffs per day (> 200 μ g in the UK and > 180 μ g in the US)	12 months	[A1]

LRTI = lower respiratory tract infection; SABA = short-acting β agonists; OCS = oral corticosteroids; GP = general practitioner; ED = emergency department; IC

Table A.1.12: Algorithms used to assess asthma control using EHR data (cont'd)

Variable	Algorithm	Interval	App
Overall asthma control	based on impairment–domain and risk–domain asthma control algorithms used by the same study	12 months	[A1

LRTI = lower respiratory tract infection; SABA = short-acting β agonists; OCS = oral corticosteroids; GP = general practitioner; ED = emergency department; ICD

References for Chapter 2 appendix


- [A1] Parikh K, Davis AB, and Pavuluri P. Do we need this blood culture? *Hosp Pediatr* 4.2 (2014), 78-84.
- [A2] Wu CL, Andrews AL, Teufel RJ, and Basco WT. Demographic predictors of leukotriene antagonist monotherapy among children with persistent asthma. *J. Pediatr.* 164.4 (2014), 827-831.e1.
- [A3] Kaiser SV, Bakel LA, Okumura MJ, Auerbach AD, Rosenthal J, et al. Risk Factors for Prolonged Length of Stay or Complications During Pediatric Respiratory Hospitalizations. *Hosp Pediatr* 5.9 (2015), 461-73.
- [A4] Kenyon CC, Rubin DM, Zorc JJ, Mohamad Z, Faerber JA, et al. Childhood Asthma Hospital Discharge Medication Fills and Risk of Subsequent Readmission. *J. Pediatr.* 166.5 (2015), 1121-7.
- [A5] Parikh K, Hall M, Mittal V, Montalbano A, Mussman GM, et al. Establishing benchmarks for the hospitalized care of children with asthma, bronchiolitis, and pneumonia. *Pediatrics* 134.3 (2014), 555-62.
- [A6] Capo-Ramos DE, Duran C, Simon AE, Akinbami LJ, and Schoendorf KC. Preventive asthma medication discontinuation among children enrolled in fee-for-service Medicaid. *J Asthma* 51.6 (2014), 618-26.
- [A7] Lachance L, Benedict MB, Doctor LJ, Gilmore LA, Kelly C, et al. Asthma coalition effects on vulnerable sub groups of children: the most frequent users of health care and the youngest. *J Asthma* 51.5 (2014), 474-9.
- [A8] Bhattacharjee R, Choi BH, Gozal D, and Mokhlesi B. Association of adenotonsillectomy with asthma outcomes in children: a longitudinal database analysis. *PLoS Med.* 11.11 (2014), e1001753.
- [A9] Chang J, Freed GL, Prosser LA, Patel I, Erickson SR, et al. Comparisons of health care utilization outcomes in children with asthma enrolled in private insurance plans versus medicaid. *J Pediatr Health Care* 28.1 (2014), 71-9.

- [A10] Liu X, Olsen J, Pedersen LH, Agerbo E, Yuan W, et al. Antidepressant use during pregnancy and asthma in the offspring. *Pediatrics* 135.4 (2015), e911-7.
- [A11] Jena AB, Ho O, Goldman DP, and Karaca-Mandic P. The Impact of the US Food and Drug Administration Chlorofluorocarbon Ban on Out-of-pocket Costs and Use of Albuterol Inhalers Among Individuals With Asthma. *JAMA Intern Med* 175.7 (2015), 1171-9.
- [A12] Malhotra K, Baltrus P, Zhang S, McRoy L, Immergluck LC, et al. Geographic and racial variation in asthma prevalence and emergency department use among Medicaid-enrolled children in 14 southern states. *J Asthma* 51.9 (2014), 913-21.
- [A13] Adimadhyam S, Schumock GT, Walton S, Joo M, McKell J, et al. Risk of arrhythmias associated with ipratropium bromide in children, adolescents, and young adults with asthma: a nested case-control study. *Pharmacotherapy* 34.4 (2014), 315-23.
- [A14] Bülow A von, Kriegbaum M, Backer V, and Porsbjerg C. The prevalence of severe asthma and low asthma control among Danish adults. *J Allergy Clin Immunol Pract* 2.6 (2014), 759-67.
- [A15] Hasegawa K, Tsugawa Y, Brown DFM, and Camargo CA. A population-based study of adults who frequently visit the emergency department for acute asthma. California and Florida, 2009-2010. *Ann Am Thorac Soc* 11.2 (2014), 158-66.
- [A16] Martin RJ, Price D, Roche N, Israel E, Aalderen WMC van, et al. Cost-effectiveness of initiating extrafine- or standard size-particle inhaled corticosteroid for asthma in two health-care systems: a retrospective matched cohort study. *NPJ Prim Care Respir Med* 24 (2014), 14081.
- [A17] Zeiger RS, Schatz M, Li Q, Chen W, Khatri DB, et al. High blood eosinophil count is a risk factor for future asthma exacerbations in adult persistent asthma. *J Allergy Clin Immunol Pract* 2.6 (2014), 741-50.
- [A18] Blais L, Kettani FZ, and Forget A. Associations of maternal asthma severity and control with pregnancy complications. *J Asthma* 51.4 (2014), 391-8.
- [A19] Schatz M, Meckley LM, Kim M, Stockwell BT, and Castro M. Asthma exacerbation rates in adults are unchanged over a 5-year period despite high-intensity therapy. *J Allergy Clin Immunol Pract* 2.5 (2014), 570-4.e1.
- [A20] Nordlund B, Melén E, Schultz ES, Grönlund H, Hedlin G, et al. Prevalence of severe childhood asthma according to the WHO. *Respir Med* 108.8 (2014), 1234-7.
- [A21] Ismaila A, Corriveau D, Vaillancourt J, Parsons D, Stanford R, et al. Impact of adherence to treatment with fluticasone propionate/salmeterol in asthma patients. *Curr Med Res Opin* 30.7 (2014), 1417-25.
- [A22] Laforest L, Licaj I, Devouassoux G, Chatte G, Martin J, et al. Asthma drug ratios and exacerbations: claims data from universal health coverage systems. *Eur. Respir. J.* 43.5 (2014), 1378-86.
- [A23] Dilokthornsakul P, Chaikyakunapruk N, Schumock GT, and Lee TA. Calendar time-specific propensity score analysis for observational data: a case study estimating the effectiveness of inhaled long-acting beta-agonist on asthma exacerbations. *Pharmacoepidemiol Drug Saf* 23.2 (2014), 152-64.
- [A24] Tan NC, Nadkarni NV, Lye WK, Sankari U, et al. Ten-year longitudinal study of factors influencing nocturnal asthma symptoms among Asian patients in primary care. *NPJ Prim Care Respir Med* 25 (2015), 15064.
- [A25] Garne E, Hansen AV, Morris J, Zaupper L, Addor MC, et al. Use of asthma medication during pregnancy and risk of specific congenital anomalies: A European case-malformed control study. *J Allergy Clin Immunol* 136 (6 2015), 1496-502.e1-7.
- [A26] Jian ZH, Huang JY, Lin FCF, Nfor ON, Jhang KM, et al. The use of corticosteroids in patients with COPD or asthma does not decrease lung squamous cell carcinoma. *BMC Pulm Med* 15 (2015), 154.
- [A27] Rust G, Zhang S, Holloway K, and Tyler-Hill Y. Timing of emergency department visits for childhood asthma after initial inhaled corticosteroid use. *Popul Health Manag* 18.1 (2015), 54-60.
- [A28] Fung V, Graetz I, Galbraith A, Hamity C, Huang J, et al. Financial barriers to care among low-income children with asthma: health care reform implications. *JAMA Pediatr* 168.7 (2014), 649-56.
- [A29] Confino-Cohen R, Brufman I, Goldberg A, and Feldman BS. Vitamin D, asthma prevalence and asthma exacerbations: a large adult population-based study. *Allergy* 69.12 (2014), 1673-80.
- [A30] Fuhlbrigge A, Peden D, Apter AJ, Boushey HA, Camargo CA, et al. Asthma outcomes: Exacerbations. *J Allergy Clin Immunol* 129.3 (2012), S34-S48.
- [A31] Tunceli O, Williams SA, Kern DM, Elhefni H, Pethick N, et al. Comparative effectiveness of budesonide-formoterol combination and fluticasone-salmeterol combination for asthma management: a United States retrospective database analysis. *J Allergy Clin Immunol Pract* 2.6 (2014), 719-26.
- [A32] Tan CC, McDowell KM, Fenchel M, Szczesniak R, and Kerckmar CM. Spirometry use in children hospitalized with asthma. *Pediatr. Pulmonol.* 49.5 (2014), 451-7.
- [A33] Nanchal R, Kumar G, Majumdar T, Taneja A, Patel J, et al. Utilization of mechanical ventilation for asthma exacerbations: analysis of a national database. *Respir Care* 59.5 (2014), 644-53.
- [A34] Sumino K, O'Brian K, Bartle B, Au DH, Castro M, et al. Coexisting chronic conditions associated with mortality and morbidity in adult patients with asthma. *J Asthma* 51.3 (2014), 306-14.
- [A35] Li L, Vollmer WM, Butler MG, Wu P, Kharbanda EO, et al. A comparison of confounding adjustment methods for assessment of asthma controller medication effectiveness. *Am. J. Epidemiol.* 179.5 (2014), 648-59.
- [A36] Hagiwara M, Delea TE, and Stanford RH. Health-care utilization and costs with fluticasone propionate and fluticasone propionate/salmeterol in asthma patients at risk for exacerbations. *Allergy Asthma Proc* 35.1 (2014), 54-62.

- [A37] Ali AK, Hartzema AG, Winterstein AG, Segal R, Lu X, et al. Application of multicategory exposure marginal structural models to investigate the association between long-acting beta-agonists and prescribing of oral corticosteroids for asthma exacerbations in the Clinical Practice Research Datalink. *Value Health* 18.2 (2015), 260-70.
- [A38] Schatz M, Zeiger RS, Yang SJ, Chen W, Crawford W, et al. Change in asthma control over time: predictors and outcomes. *J Allergy Clin Immunol Pract* 2.1 (), 59-64.
- [A39] Wu AC, Li L, Fung V, Kharbanda EO, Larkin EK, et al. Use of leukotriene receptor antagonists are associated with a similar risk of asthma exacerbations as inhaled corticosteroids. *J Allergy Clin Immunol Pract* 2.5 (), 607-13.
- [A40] Tse SM, Charland SL, Stanek E, Herrera V, Goldfarb S, et al. Statin use in asthmatics on inhaled corticosteroids is associated with decreased risk of emergency department visits. *Curr Med Res Opin* 30.4 (2014), 685-93.
- [A41] Kim S, Kim J, Park SY, Um HY, Kim K, et al. Effect of pregnancy in asthma on health care use and perinatal outcomes. *J Allergy Clin Immunol* 136 (5 2015), 1215-23.e1-6.
- [A42] Blais L, Kettani FZ, Forget A, Beauchesne MF, and Lemièrè C. Asthma exacerbations during the first trimester of pregnancy and congenital malformations: revisiting the association in a large representative cohort. *Thorax* 70.7 (2015), 647-52.
- [A43] Keast SL, Thompson D, Farmer K, Smith M, Nesser N, et al. Impact of a prior authorization policy for montelukast on clinical outcomes for asthma and allergic rhinitis among children and adolescents in a state Medicaid program. *J Manag Care Spec Pharm* 20.6 (2014), 612-21.
- [A44] Sullivan PW, Campbell JD, Ghushchyan VH, and Globe G. Outcomes before and after treatment escalation to Global Initiative for Asthma steps 4 and 5 in severe asthma. *Ann. Allergy Asthma Immunol.* 114.6 (2015), 462-9.

A.2 Published paper related to Chapter 2

Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review

Mohammad A Al Sallakh, MD ^{1,a}, Eleftheria Vasileiou, MPH^{2,a}, Sarah E Rodgers, PhD^{1,b}, Ronan A Lyons, MD^{1,b}, Aziz Sheikh, MD^{2,a,b} and Gwyneth A Davies, MD^{1,a}

¹Swansea University Medical School, Singleton Park, Swansea, SA2 8PP, UK

²Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland, UK

^aAsthma UK Centre for Applied Research

^bThe Farr Institute of Health Informatics Research

Short Title

Defining and assessing asthma using EHR data

Correspondence

Mohammad A Al Sallakh, MD, MSc

Data Science Building, Swansea University, Singleton Park, Swansea, SA2 8PP, United Kingdom

Phone: +

Email: M.A.AISallakh@swansea.ac.uk

This is an author-submitted, peer-reviewed version of an article that has been accepted for publication in the European Respiratory Journal, prior to copy-editing, formatting and typesetting. This version of the article may not be duplicated or reproduced without prior permission from the copyright owner, the European Respiratory Society. The publisher is not responsible or liable for any errors or omissions in this version of the article or in any version derived from it by any other parties. The final, copy-edited, published article, which is the version of record, is available online (<https://doi.org/10.1183/13993003.00204-2017>) from the European Respiratory Journal without a subscription 18 months after the date of issue publication.

Abstract

There is currently no consensus on approaches to defining asthma or assessing asthma outcomes using electronic health record (EHR)-derived data. We explored these approaches in the recent literature, and examined the clarity of reporting.

We systematically searched for asthma-related articles published between 1-1-2014 and 31-12-2015, extracted the algorithms used to identify asthma patients and assess severity, control and exacerbations, and examined how the validity of these outcomes was justified.

From 113 eligible articles, we found significant heterogeneity in the algorithms used to define asthma (n=66 different algorithms), severity (n=18), control (n=9), and exacerbations (n=24). For the majority of algorithms (n=106), validity was not justified. In the remaining cases, approaches ranged from using algorithms validated in the same databases, to using non-validated algorithms that were based on clinical judgement or clinical guidelines. The implementation of these algorithms was sub-optimally described overall.

Although EHR-derived data are now widely used to study asthma, the approaches being used are significantly varied and are often underdescribed, rendering it difficult to assess the validity of studies and compare their findings. Given the substantial growth in this body of literature, it is crucial that scientific consensus is reached on the underlying definitions and algorithms.

Keywords: Algorithms; asthma; electronic health records; quality of reporting; reproducibility.

Introduction

Asthma is in clinical practice a diagnosis based on the patient history, examination and objective tests [1]. It is however increasingly considered to represent a heterogeneous group of disorders with different phenotypes and endotypes [2]. The clinical definitions of asthma and its key outcomes, including disease severity, control, and attacks/exacerbations have been the subject of vigorous debate [3–8].

Particular challenges arise in the context of epidemiologic studies where validated operational definitions are needed [9, 10]. These studies are, increasingly, being undertaken using electronic health record (EHR)-derived data, which adds a further layer of complexity as the use of valid and reliable approaches is essential in order to ensure the reproducibility of research findings [11].

In order to assess current approaches, we systematically interrogated the recent EHR-based asthma literature. Our specific objectives were to: i) describe the different methods of defining asthma and assessing disease severity, control and exacerbations in EHR-based studies; ii) investigate whether authors reported on the validity of those methods; and iii) assess their reporting practices.

Methods

We conducted a systematic scoping review based on Arksey and O'Malley's five-stage framework, including identifying the research question, identifying relevant studies, study selection, data charting and collating, summarising and reporting the results [12]. The research questions were: (1) How were asthma and its key outcomes defined using EHR data in the recent literature? (2) How did authors report on the validity of their EHR-based algorithms? (3) How clearly were the EHR-related methods reported?

Eligibility criteria and search strategy

We searched PubMed using a broad query (Table E1) to retrieve asthma studies that used EHR-derived data and were published between January 1, 2014 and December 31, 2015. The search query was iteratively improved by adding many variations and equivalents of the keywords "EHR" and "routinely collected data" as well as named data sources found in the literature. Only articles written in English were included.

Study selection

We excluded non-relevant articles by reviewing titles and abstracts, referring to the full-text when needed. We included only articles where asthma was a main finding. For the purpose of this review, we limited the concept of EHR-derived data to coded, objective, individual-level data that were generated as a by-product of routine health care.

Data extraction and synthesis

From each of the eligible articles, we extracted and summarised information from the full text and online supplements, including basic bibliography, setting (country) and design; names and types of EHR-derived data sources used; algorithms to identify asthma patients, assess disease severity, control, exacerbation; and how authors reported on algorithm validity. In this context, we referred to 'validation' as any attempt to assess the algorithm's concurrent or construct validity. We used the RECORD Statement's 13-items checklist to assess the clarity of reporting of other EHR-related aspects such as clinical code lists used in the algorithms, and the implications of using EHR data in asthma research. The RECORD Statement is a recently introduced extension to the STROBE Statement which helps improve the reporting of observational studies conducted using routinely collected data [13]. Table E2 describes the data extraction and charting tool. Article screening and data extraction were performed independently by two authors (MAS and EV) with a third author arbitrating (GAD).

Role of the funding sources

The funding sources had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

Results

We included 113 articles in the review. [Figure 1](#) shows the study selection process. Most studies were conducted in the United States (US), Taiwan, and Canada ([Table E3](#)), and employed longitudinal designs ([Table E4](#)). The most commonly used data types were health insurance claims followed by medical record repositories and dispensing databases ([Table E5](#)).

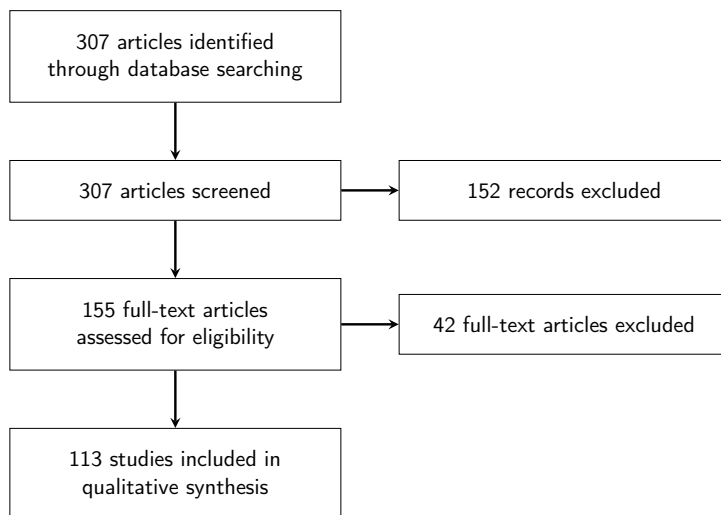


Figure 1: Flowchart for study selection in this scoping review.

Defining asthma

We identified 66 different algorithms to define asthma under seven diagnostic labels ([Table E6](#)).

‘*Persistent asthma*’ was defined over 12 and 24 months using the US Healthcare Effectiveness Data and Information Set (HEDIS) criteria [[14](#)], which involved assessing for any of the following asthma-related events: (1) emergency department (ED) visit, (2) hospitalisation, (3) outpatient visit and two asthma prescriptions, or (4) four asthma prescriptions [[15–18](#)]; by HEDIS criteria except “four asthma prescriptions” [[19](#)]; and by any asthma encounter (hospitalisation or ED visit) or using oral corticosteroids (OCS) for three or more days [[20](#)].

‘*Current asthma*’ was defined by any asthma encounter in the last three years [[21](#)].

‘*Current general practitioner (GP)-reported and diagnosed asthma*’ was defined as any asthma encounter in the last 12 months, and ‘*current GP-reported, diagnosed and treated asthma*’ as the same plus any asthma prescription in the same period [[22](#)].

Patients with treated asthma were otherwise required to have at least three dispensing events of asthma treatments in three different quarters of the year [[23](#)].

‘*Acute asthma*’ was defined using any asthma diagnosis codes in ED or inpatient data [[24](#)].

In the remaining studies, the label ‘*asthma*’ was defined using various algorithms, some of which were similar to those of the aforementioned more specific labels.

The intervals over which asthma diagnostic/management and prescription codes were queried were specified in 31 and 8 studies, respectively. The positions of diagnostic codes in the encounter (i.e. primary or secondary) were specified in 37 studies.

We identified five approaches in these algorithms: requiring diagnostic/management events, prescription events, or both (Table E7). In addition, to exclude likely non-asthma patients, some studies applied additional non-asthma criteria to restrict the study population based on age (Table E8) and/or comorbidities (Table E9).

Assessing asthma severity

Eighteen studies used 20 different algorithms to assess asthma severity (Table E10), as binary (i.e. severe vs. non-severe asthma) [15, 23, 25–38] or ordinal variables (mild, moderate, and severe asthma [39]; or low, moderate, and high-risk asthma [40]). The algorithms were based on one or more of the following asthma-related variables: number and/or dosage of prescriptions—namely SABA, inhaled corticosteroids (ICS), OCS, and leukotriene receptor antagonist (LTRA)—and number of hospitalisations, ED and outpatient visits. Almost all algorithms (17) used prescriptions (either alone or with other variables), while one algorithm was based only on hospitalisations and ED visits [36]. The intervals over which asthma severity was assessed were three [29], six [38], 12 [15, 23, 28, 30–32, 34, 36, 37, 39, 40], 24 months [33, 35], or unclear [26, 27].

Assessing asthma control

Nine studies assessed asthma control using 11 algorithms, in 9 of which the interval was 12 months, in one 1-3 months, and in the remaining study this was unclear (Table E12). Uncontrolled asthma was defined by a minimum number/dose of SABA prescriptions [30, 31, 39, 41, 42]; any or short-course OCS prescriptions [30, 31, 41–44]; any hospitalisation or ED visit with either diagnosis of asthma [27, 30, 31, 41–43, 45] or — in already diagnosed asthma patients — diagnosis of status asthmaticus, pneumonia, dyspnoea, or respiratory insufficiency [30]; unscheduled outpatient visits for asthma or lower respiratory tract infections (LRTI) [31]; and GP consultations for LRTI requiring antibiotics in asthma patients [31]. Asthma impairment was defined based on the required SABA use, namely an average of more than two salbutamol puffs per day [31]. One study assessed asthma control based on number of OCS and SABA prescriptions per year (without giving any further details about the actual algorithm) [41].

Defining exacerbations

Twenty-four studies defined exacerbations using EHR-derived data (Table E11), as a dichotomous variable (absent vs. present) [16, 17, 23, 27, 30–32, 35, 37–39, 42–44, 46–54], or stratified into absent, moderate and severe [55]. Oral corticosteroid prescriptions were used as a marker for exacerbations in 17 studies, either alone [23, 30, 31, 35, 39, 42, 47, 48, 53] or with a concurrent asthma encounter (e.g., a GP, outpatient, or ED visit, or hospitalisation within five or seven days) [16, 17, 32, 37, 38, 46, 52, 54]. In one study, exacerbations were defined by a minimum of six short-acting beta-2 agonist (SABA) prescriptions per year [47]. Other definitions included an outpatient code of ‘asthma exacerbation’ [52], asthma hospitalisation [23, 30, 32, 35, 37, 39, 43,

Table 1: Practices of reporting or justifying the validity of algorithms to define and assess asthma using EHR-derived data.

Algorithm validity was justified by	Number of algorithms				
	Identifying asthma patients	Assessing severity	Assessing control	Defining exacerbation	Total per category
Validation of the same algorithm in the same database	14	1	1	1	17
Validation of the same algorithm in different database(s)	2	6	3	2	13
Validation of other diseases' algorithms in the same database	2	0	0	0	2
Validation of other diseases' algorithms in different database(s)	1	0	0	0	1
Being consistent with similar studies in the same database	1	0	1	0	2
Being consistent with similar studies in different database(s)	1	0	0	1	2
Validation or concordance analysis in the same study	4	0	0	0	4
Being based on nationally developed algorithms	3	0	0	2	5
Relying on the validity of database coding	5	0	0	0	5
Being based on clinical guidelines	0	3	0	0	3
Not justified	76	8	4	18	106

44, 46, 48, 50, 51, 53–55], asthma ED visit [16, 30–32, 35, 37, 38, 43, 44, 46, 48, 51–54], or hospitalisation with diagnosis of status asthmaticus, or — in already diagnosed asthma patients — diagnosis of pneumonia, dyspnoea, or respiratory insufficiency [30].

Clarity of reporting

Overall, the reporting of methodological aspects of using EHR-derived data was suboptimal. The majority of studies presented no information on the algorithms' validity. Among studies that reported on the validity, we identified 10 practices of reporting or justifying on the validity of algorithms (Table 1): (1) performing validation or concordance analysis in the same study against other measures based on different data sources (e.g., medical record review or patient-reported measures); (2) referring to previous validation of similar algorithms in the same or (3) different databases; (4) referring to previous validation of similar algorithms for different diseases in the same or (5) different database (6); using algorithms 'consistent' with previous studies in the same or (7) different databases; (8) using nationally developed algorithms; (9) using algorithms based on clinical guidelines; (10) and relying on previous validation of the database content. Some studies did not provide clear algorithms for asthma severity or control, but only referred to their components [23, 35, 37, 38, 41].

Of the 113 reviewed studies, 40 studies used record-linkage, of which 17 mentioned it in the abstract, and 28 provided at least some explanation in the full text. The geographical region, time frame of data, and types or names of the data sources were mentioned in 83, 91, and 104 abstracts, respectively. Eighty-three studies reported their extent of access to the data sources. The intervals over which the algorithms were applied were often not reported. One hundred and eleven studies touched on the implications of using EHR data to study asthma. Of these, 64 and 63 studies discussed the risk of misclassification bias and unmeasured confounding, respectively. Six studies acknowledged the possible changes over time in data quality and coding practices and the entailing changes in case definition eligibility and accuracy. Five studies explained their data cleansing procedures. Finally, no study shared the programming codes of data preparation and analysis.

Discussion

Statement of principal findings

This systematic analysis of the contemporaneous asthma literature has found evidence of considerable international activity in using EHR-derived data to study a variety of asthma populations and outcomes. Importantly, we also found wide variations in the approaches used with limited attention being paid to the validity of the underlying algorithms used and suboptimal reporting of studies. This poses a major challenge to the interpretation and reproducibility of this important, emerging body of research inquiry.

Strengths and limitations

To our knowledge, this is the first systematic exercise to investigate the quality of reporting on EHR-based studies, especially the validity of measures, in the context of asthma. In undertaking this work, we used robust approaches which involved two people independently selecting studies and undertaking data extraction. The findings may also apply to other chronic diseases. This review had no geographic limits, but it was confined to assessing the recent literature. Examining the most recent asthma literature is most likely to provide meaningful insights on current practices. A limitation is that we did not systematically check whether the references provided to support the claimed validity of algorithms in question actually provided sufficient evidence of validity. For example, differences might exist between the algorithms used in a given study and those previously validated.

Interpretation in the light of previous studies

Although EHR-derived data are convenient resources for research, they are originally collected for other purposes, and usually suffer from missing or incorrect data and potential biases [56–58]. In addition, EHR systems usually fail to capture complete and accurate clinical information at the point of care due to design limitations and inefficient use of these systems by clinicians to document clinical data [59, 60].

These issues impose challenges on their use to assess a complex and heterogeneous condition such as asthma. For example, asthma diagnosis codes, which are commonly used solely for patient identification, may be recorded after a trial or wrong diagnosis, and do not capture undiagnosed patients [61]. In addition, many EHR-derived databases often lack important variables, such as lung function, indication of dispensed medications, adherence to treatment, and lifestyle, which are vital for identifying and assessing asthma patients. These challenges are however not insurmountable. In this review, we found several techniques intended to improve algorithm accuracy such as age limitation, comorbidity exclusion, and diagnosis position restriction.

Ideally, algorithms should be validated in the databases in which they are used. However, this was often not the case. Instead, using algorithms with only reasonable face validity based on clinical guidelines or clinical judgement is a very common practice in EHR-based studies

[62, 63]. These approaches assume that clinical codes in the database accurately represent the patient's actual health care events [62].

Under-reporting on implementation details and methods' validity compromises transparency and reproducibility, a crucial issue in medical research. It has been previously found that in EHR-based studies, full lists of clinical codes were often not reported [64]. A recent, large-scale reproducibility exercise identified similar challenges due to suboptimal reporting of EHR-based studies, particularly sharing code lists and algorithms [65].

The significant methodological heterogeneity we found in EHR-based asthma assessment algorithms reflects, in addition to the content differences between the databases used, the lack of consensus on the clinical definitions in the first place despite continuous standardisation efforts [5, 6, 66, 67]. The focus of our work was to examine asthma definitions and their validity specifically in the context of EHR, but this highlights the fundamental need to reach consensus on clinical asthma definitions and the appropriate validation of asthma diagnosis. For example, there is still an active debate on whether lung function is essential to establish asthma diagnosis [7, 8]. A recent study also found significant variation in algorithms to assess asthma severity from health insurance data [68]. Unjustified inter-study variation in the operational definitions of the same clinical concepts creates challenges for comparability, meta-analysis and evidence synthesis. These issues have been raised for asthma [69] and other allergic conditions such as peanut allergy [70, 71] and anaphylaxis [72], where wide variations in findings were potentially attributed to inconsistent case definitions.

Implications for policy, practice and research

This review sheds light on the opportunities offered by the increasingly ubiquitous EHRs, but also highlights considerable heterogeneity and suboptimal reporting of EHR-based asthma assessment algorithms and the implications of these practices on comparability and reproducibility of studies.

Developing reliable algorithms to assess asthma outcomes using EHR data is a non-trivial challenge. In addition, standardising such algorithms across different populations may be impractical since databases differ in content, validity may not hold across different populations, and no best practice currently exists [68]. Similar challenges arise when comparing asthma epidemiology between multiple populations [73]. These methodologic issues, in addition to suboptimal reporting, should be considered when interpreting and synthesising evidence from geographically dispersed studies.

With the accelerating availability of EHR-derived data and their use to study asthma, we believe that consideration needs to be given to convening an international task force to work on the harmonisation of those algorithms under uniform and consistent clinical labels, while considering the differences between populations and databases. In addition, validation of these algorithms in the respective populations should be given a high priority. Furthermore, to allow more accurate assessment of asthma from EHR data, efforts are needed to improve the capture and coding of asthma-related data at the point of care [74] which requires more efficient EHR systems [59, 60]. In addition, emerging data sources such as patient-generated data and wearables need to be harnessed [75]. Finally, to improve the clarity of reporting on EHR-related methodological aspects, we strongly advocate the adoption of the RECORD Statement as an extension of the STROBE Statement by both authors and journal editors [13]. Optimal reporting should include complete code lists, detailed algorithms and validity assessment. Implications of using EHR-

derived data to study a complex condition such as asthma should be clearly communicated to enable judgement of internal and external validity.

In summary, we have found that there is considerable international interest in exploiting EHR-derived data to study asthma, but that there are considerable variations in the approaches used. These variations are compounded by sub-optimal reporting of methods, which makes it difficult to assess the reproducibility of research. Given the substantial investments taking place in EHRs globally, this body of work is likely to grow significantly in the coming years. It is therefore important that the asthma-interested research community works to place it on a solid footing in order to ensure the quality and reproducibility of this work.

Authors' contributions

MAS, SER and GAD, AS, and RAL developed the concept and methods. MAS conducted the literature search. MAS and EV independently reviewed the studies with GAD arbitrating. All authors contributed to the development of methods, interpretation of findings, and manuscript writing, and critically reviewed and approved the final manuscript.

Conflict of Interest Statement

Aziz Sheikh reports grants from Asthma UK during the conduct of the study.

Support statement: This work was funded by Health and Care Research Wales and Abertawe Bro Morgannwg University Health Board. It was carried out with the support of the Asthma UK Centre for Applied Research [AUK-AC-2012-01]. We also acknowledge the support from The Farr Institute of Health Informatics Research. The Farr Institute is supported by a 10-funder consortium: Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the Medical Research Council, the National Institute of Health Research, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates), the Wellcome Trust, (MRC Grant Nos: CIPHER MR/K006525/1, Scotland MR/K007017/1).

References

- [1] Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention (2015 update). 2015.
- [2] Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med* 2012;**18**: 716–725.
- [3] Hargreave FE and Nair P. The definition and diagnosis of asthma. *Clin Exp Allergy* 2009;**39**: 1652–1658.
- [4] . A plea to abandon asthma as a disease concept. *Lancet* 2006;**368**: 705.
- [5] Bousquet J, Mantzouranis E, Cruz AA, A1 t-Khaled N, Baena-Cagnani CE, Bleecker ER, et al. Uniform definition of asthma severity, control, and exacerbations: document presented for the World Health Organization Consultation on Severe Asthma. *J Allergy Clin Immunol* 2010;**126**: 926–938.
- [6] Reddel HK, Bateman ED, Becker A, Boulet LP, Cruz AA, Drazen JM, et al. A summary of the new GINA strategy: a roadmap to asthma control. *Eur Respir J* 2015;**46**: 622–639.
- [7] Akker ILv d., Zeijden H van der, and Verheij TJ. Is spirometry essential in diagnosing asthma? Yes. *Br J Gen Pract* 2016;**66**: 484–484.
- [8] Levy ML. Is spirometry essential in diagnosing asthma? No. *Br J Gen Pract* 2016;**66**: 485–485.
- [9] Toelle BG, Peat JK, Salome CM, Mellis CM, and Woolcock AJ. Toward a Definition of Asthma for Epidemiology. *Am Rev Respir Dis* 1992;**146**: 633–637.
- [10] Pekkanen J and Pearce N. Defining asthma in epidemiological studies. *Eur Respir J* 1999;**14**: 951–957.
- [11] Ioannidis JP. Why Most Published Research Findings Are False. *PLoS Medicine* 2005;**2**:e124.
- [12] Arksey H and O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005;**8**: 19–32.
- [13] Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 2015;**12**:e1001885.
- [14] Use of appropriate medications for people with asthma. *HEDIS 2003*. Vol. 2. Washington, DC: National Committee for Quality Assurance, 2003, 25–28.
- [15] Wu CL, Andrews AL, Teufel RJ, and Basco WT. Demographic predictors of leukotriene antagonist monotherapy among children with persistent asthma. *J. Pediatr.* 2014;**164**:827–831.e1.
- [16] Zeiger RS, Schatz M, Li Q, Chen W, Khatry DB, Gossage D, et al. High blood eosinophil count is a risk factor for future asthma exacerbations in adult persistent asthma. *J Allergy Clin Immunol Pract* 2014;**2**: 741–50.
- [17] Schatz M, Zeiger RS, Yang SJ, Chen W, Crawford W, Sajjan S, et al. Change in asthma control over time: predictors and outcomes. *J Allergy Clin Immunol Pract* ;**2**: 59–64.

- [18] Jena AB, Ho O, Goldman DP, and Karaca-Mandic P. The Impact of the US Food and Drug Administration Chlorofluorocarbon Ban on Out-of-pocket Costs and Use of Albuterol Inhalers Among Individuals With Asthma. *JAMA Intern Med* 2015;**175**: 1171–9.
- [19] McRoy L, Weech-Maldonado R, and Kilgore M. The relationship between direct to consumer advertising (DTCA) and asthma-related emergency department use among Medicaid-enrolled children. *J Asthma* 2014;**51**: 922–6.
- [20] Wu AC, Butler MG, Li L, Fung V, Kharbanda EO, Larkin EK, et al. Primary adherence to controller medications for asthma is poor. *Ann Am Thorac Soc* 2015;**12**: 161–6.
- [21] Tomasallo CD, Hanrahan LP, Tandias A, Chang TS, Cowan KJ, and Guilbert TW. Estimating Wisconsin asthma prevalence using clinical electronic health records and public health data. *Am J Public Health* 2014;**104**:e65–73.
- [22] Mukherjee M, Gupta R, Farr A, Heaven M, Stoddart A, Nwaru BI, 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* 2014;**4**:e006647.
- [23] Laforest L, Licaj I, Devouassoux G, Chatte G, Martin J, and Ganse EV. Asthma drug ratios and exacerbations: claims data from universal health coverage systems. *Eur. Respir. J.* 2014;**43**: 1378–86.
- [24] Lemke LD, Lamerato LE, Xu X, Booza JC, Reiners JJ, Iii DMR, et al. Geospatial relationships of air pollution and acute asthma events across the Detroit-Windsor international border: study design and preliminary results. *J Expo Sci Environ Epidemiol* 2014;**24**: 346–57.
- [25] Jian ZH, Huang JY, Lin FCF, Nfor ON, Jhang KM, Ku WY, et al. The use of corticosteroids in patients with COPD or asthma does not decrease lung squamous cell carcinoma. *BMC Pulm Med* 2015;**15**: 154.
- [26] Garne E, Hansen AV, Morris J, Zaupper L, Addor MC, Barisic I, et al. Use of asthma medication during pregnancy and risk of specific congenital anomalies: A European case-malformed control study. *J Allergy Clin Immunol* 2015;**136**:1496–502.e1-7.
- [27] Tan NC, Nadkarni NV, Lye WK, Sankari U, et al. Ten-year longitudinal study of factors influencing nocturnal asthma symptoms among Asian patients in primary care. *NPJ Prim Care Respir Med* 2015;**25**: 15064.
- [28] Kenyon CC, Rubin DM, Zorc JJ, Mohamad Z, Faerber JA, and Feudtner C. Childhood Asthma Hospital Discharge Medication Fills and Risk of Subsequent Readmission. *J. Pediatr.* 2015;**166**: 1121–7.
- [29] Rust G, Zhang S, Holloway K, and Tyler-Hill Y. Timing of emergency department visits for childhood asthma after initial inhaled corticosteroid use. *Popul Health Manag* 2015;**18**: 54–60.
- [30] Bülow A von, Kriegbaum M, Backer V, and Porsbjerg C. The prevalence of severe asthma and low asthma control among Danish adults. *J Allergy Clin Immunol Pract* 2014;**2**: 759–67.
- [31] Martin RJ, Price D, Roche N, Israel E, Aalderen WMC van, Grigg J, et al. Cost-effectiveness of initiating extrafine- or standard size-particle inhaled corticosteroid for asthma in two health-care systems: a retrospective matched cohort study. *NPJ Prim Care Respir Med* 2014;**24**: 14081.
- [32] Schatz M, Meckley LM, Kim M, Stockwell BT, and Castro M. Asthma exacerbation rates in adults are unchanged over a 5-year period despite high-intensity therapy. *J Allergy Clin Immunol Pract* 2014;**2**:570–4.e1.
- [33] Capo-Ramos DE, Duran C, Simon AE, Akinbami LJ, and Schoendorf KC. Preventive asthma medication discontinuation among children enrolled in fee-for-service Medicaid. *J Asthma* 2014;**51**: 618–26.
- [34] Nordlund B, Melén E, Schultz ES, Grönlund H, Hedlin G, and Kull I. Prevalence of severe childhood asthma according to the WHO. *Respir Med* 2014;**108**: 1234–7.
- [35] Ismaila A, Corriveau D, Vaillancourt J, Parsons D, Stanford R, Su Z, et al. Impact of adherence to treatment with fluticasone propionate/salmeterol in asthma patients. *Curr Med Res Opin* 2014;**30**: 1417–25.
- [36] Fung V, Graetz I, Galbraith A, Hamity C, Huang J, Vollmer WM, et al. Financial barriers to care among low-income children with asthma: health care reform implications. *JAMA Pediatr* 2014;**168**: 649–56.
- [37] Dilokthornsakul P, Chaiyakunapruk N, Schumock GT, and Lee TA. Calendar time-specific propensity score analysis for observational data: a case study estimating the effectiveness of inhaled long-acting beta-agonist on asthma exacerbations. *Pharmacoepidemiol Drug Saf* 2014;**23**: 152–64.
- [38] Adimadhyam S, Schumock GT, Walton S, Joo M, McKell J, and Lee TA. Risk of arrhythmias associated with ipratropium bromide in children, adolescents, and young adults with asthma: a nested case-control study. *Pharmacotherapy* 2014;**34**: 315–23.
- [39] Blais L, Kettani FZ, and Forget A. Associations of maternal asthma severity and control with pregnancy complications. *J Asthma* 2014;**51**: 391–8.
- [40] Chang J, Freed GL, Prosser LA, Patel I, Erickson SR, Bagozzi RP, et al. Comparisons of health care utilization outcomes in children with asthma enrolled in private insurance plans versus Medicaid. *J Pediatr Health Care* 2014;**28**: 71–9.
- [41] Sullivan PW, Campbell JD, Ghushchyan VH, and Globe G. Outcomes before and after treatment escalation to Global Initiative for Asthma steps 4 and 5 in severe asthma. *Ann. Allergy Asthma Immunol.* 2015;**114**: 462–9.
- [42] Ali AK, Hartzema AG, Winterstein AG, Segal R, Lu X, and Hendeles L. Application of multicategory exposure marginal structural models to investigate the association between long-acting beta-agonists and prescribing of oral corticosteroids for asthma exacerbations in the Clinical Practice Research Datalink. *Value Health* 2015;**18**: 260–70.

- [43] Wu AC, Li L, Fung V, Kharbanda EO, Larkin EK, Vollmer WM, et al. Use of leukotriene receptor antagonists are associated with a similar risk of asthma exacerbations as inhaled corticosteroids. *J Allergy Clin Immunol Pract* ;2: 607–13.
- [44] Tan CC, McDowell KM, Fenchel M, Szczesniak R, and Kerckmar CM. Spirometry use in children hospitalized with asthma. *Pediatr Pulmonol*. 2014;**49**: 451–7.
- [45] Keast SL, Thompson D, Farmer K, Smith M, Nesser N, and Harrison D. Impact of a prior authorization policy for montelukast on clinical outcomes for asthma and allergic rhinitis among children and adolescents in a state Medicaid program. *J Manag Care Spec Pharm* 2014;**20**: 612–21.
- [46] Kim S, Kim J, Park SY, Um HY, Kim K, Kim Y, et al. Effect of pregnancy in asthma on health care use and perinatal outcomes. *J Allergy Clin Immunol* 2015;**136**:1215–23.e1-6.
- [47] Confino-Cohen R, Brufman I, Goldberg A, and Feldman BS. Vitamin D, asthma prevalence and asthma exacerbations: a large adult population-based study. *Allergy* 2014;**69**: 1673–80.
- [48] Tunceli O, Williams SA, Kern DM, Elhefni H, Pethick N, Wessman C, et al. Comparative effectiveness of budesonide-formoterol combination and fluticasone-salmeterol combination for asthma management: a United States retrospective database analysis. *J Allergy Clin Immunol Pract* 2014;**2**: 719–26.
- [49] Bhattacharjee R, Choi BH, Gozal D, and Mokhlesi B. Association of adenotonsillectomy with asthma outcomes in children: a longitudinal database analysis. *PLoS Med*. 2014;**11**:e1001753.
- [50] Nanchal R, Kumar G, Majumdar T, Taneja A, Patel J, Dagar G, et al. Utilization of mechanical ventilation for asthma exacerbations: analysis of a national database. *Respir Care* 2014;**59**: 644–53.
- [51] Tse SM, Charland SL, Stanek E, Herrera V, Goldfarb S, Litonjua AA, et al. Statin use in asthmatics on inhaled corticosteroids is associated with decreased risk of emergency department visits. *Curr Med Res Opin* 2014;**30**: 685–93.
- [52] Sumino K, O'Brian K, Bartle B, Au DH, Castro M, and Lee TA. Coexisting chronic conditions associated with mortality and morbidity in adult patients with asthma. *J Asthma* 2014;**51**: 306–14.
- [53] Li L, Vollmer WM, Butler MG, Wu P, Kharbanda EO, and Wu AC. A comparison of confounding adjustment methods for assessment of asthma controller medication effectiveness. *Am. J. Epidemiol*. 2014;**179**: 648–59.
- [54] Hagiwara M, Delea TE, and Stanford RH. Health-care utilization and costs with fluticasone propionate and fluticasone propionate/salmeterol in asthma patients at risk for exacerbations. *Allergy Asthma Proc* 2014;**35**: 54–62.
- [55] Blais L, Kettani FZ, Forget A, Beauchesne MF, and Lemièrre C. Asthma exacerbations during the first trimester of pregnancy and congenital malformations: revisiting the association in a large representative cohort. *Thorax* 2015;**70**: 647–52.
- [56] Schneeweiss S and Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005;**58**: 323–337.
- [57] Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. 2015;.
- [58] Hemkens LG, Contopoulos-Ioannidis DG, and Ioannidis JPA. Routinely collected data and comparative effectiveness evidence: promises and limitations. *Can Med Assoc J* 2016;.
- [59] Sheikh A, Cornford T, Barber N, Avery A, Takian A, Lichtner V, et al. Implementation and adoption of nationwide electronic health records in secondary care in England: final qualitative results from prospective national evaluation in "early adopter" hospitals. *BMJ* 2011;**343**: d6054–d6054.
- [60] Frenkel LD. Electronic health records-Applications for the allergist/immunologist: All that glitters is not gold. *Allergy Asthma Proc* 2016;**37**: 273–278.
- [61] Huzel L, Roos LL, Anthonisen NR, and Manfreda J. Diagnosing asthma: the fit between survey and administrative database. *Can Respir J* 2002;**9**: 407–412.
- [62] Manuel DG, Rosella LC, and Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ* 2010;**341**: c4226.
- [63] Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;**21**: 221–230.
- [64] Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PLoS One* 2014;**9**:e99825.
- [65] Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, and Bartels DB. Transparency and Reproducibility of Observational Cohort Studies Using Large Healthcare Databases. *Clin Pharmacol Ther* 2016;**99**: 325–332.
- [66] Bel EH, Sousa A, Fleming L, Bush A, Chung KF, Versnel J, et al. Diagnosis and definition of severe refractory asthma: an international consensus statement from the Innovative Medicine Initiative (IMI). *Thorax* 2011;**66**: 910–917.
- [67] Chung KF, Wenzel SE, Brozek JL, Bush A, Castro M, Sterk PJ, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. *Eur Respir J* 2014;**43**: 343–373.
- [68] Jacob C, Haas JS, Bechtel B, Kardos P, and Braun S. Assessing asthma severity based on claims data: a systematic review. *Eur J Health Econ* 2016;.

- [69] Ford ES. The epidemiology of obesity and asthma. *J Allergy Clin Immunol: In Practice* 2005;**115**:897–909, quiz 910.
- [70] Kotz D, Simpson CR, and Sheikh A. Incidence, prevalence, and trends of general practitioner–recorded diagnosis of peanut allergy in England, 2001 to 2005. *J Allergy Clin Immunol* 2011;**127**:623–630.e1.
- [71] Custovic A and Nicolaou N. Peanut allergy: overestimated in epidemiology or underdiagnosed in primary care? *J Allergy Clin Immunol: In Practice* 2011;**127**: 631–632.
- [72] Panesar S, Javad S, Silva Dd, Nwaru B, Hickstein L, Muraro A, et al. The epidemiology of anaphylaxis in Europe: a systematic review. *Allergy* 2013;**68**: 1353–1361.
- [73] Nwaru BI, Mukherjee M, Gupta RP, Farr A, Heaven M, Stoddart A, et al. Challenges of harmonising data from UK national health surveys: a case study of attempts to estimate the UK prevalence of asthma. *J R Soc Med* 2015;:.
- [74] Mukherjee M, Wyatt JC, Simpson CR, and Sheikh A. Usage of allergy codes in primary care electronic health records: a national evaluation in Scotland. *Allergy* 2016;:.
- [75] Howie L, Hirsch B, Locklear T, and Abernethy AP. Assessing The Value Of Patient-Generated Data To Comparative Effectiveness Research. *Health Aff (Millwood)* 2014;**33**: 1220–1228.

Appendix B

Chapter 3 Appendix

B.1 Making sense of patient-reported currently treated asthma using routinely collected data

Mohammad Al Sallakh,¹ Sarah Rodgers,¹ Ronan Lyons,¹ Aziz Sheikh,² Gwyneth Davies.¹ P148 Making sense of patient-reported currently treated asthma using routinely collected data. *Thorax* 2016;71:A163-A164.

¹Medical School, Swansea University, Swansea, UK; ²Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK

An abstract presented at British Thoracic Society Winter Meeting 2016.

Background: Currently treated asthma (CTA) is commonly assessed in epidemiological studies and is typically self-reported.

Aims: To investigate how patient understanding of this label compared with objective measures from routinely collected data.

Methods: We obtained the valid CTA responses of individuals aged 16+ from the Welsh Health Survey 2014, who also had linked records in the GP dataset of the Secure Anonymised Information Linkage databank and complete GP registrations between 2009-2014. We queried their recent prescriptions and whether they had ever asthma diagnosis. We examined the concordance between self-reported CTA and each of 'ever prescriptions', 'ever diagnosis', and 'prescriptions in varying backward intervals from mid-2014', with the latter repeated by adding 'ever diagnosis'.

Results: Of 4,291 eligible people, 10.2% self-reported CTA, of these 11.2% and 22.4% had no prescriptions in the past 12 months and no recorded asthma diagnosis ever. For concordance between self-reported CTA and each of 'ever prescrip-

tions' and 'ever diagnosis', Cohen's kappa was 0.42 and 0.68. For concordance between self-reported CTA and 'prescriptions in backward intervals', kappa was 0.76 for the 12-month interval but peaked to 0.77 at 9-months. After adding 'ever diagnosis', the kappa became 0.78 for the 12-month measure (which represents the treated asthma criteria of the Quality of Outcomes Framework, QOF), and peaked to 0.79 at 18-months.

Conclusions: In Wales, self-reported CTA agreed well with the QOF treated asthma criteria, but slightly better with 'any prescriptions in last 18 months and ever diagnosis'. However, the concordance remains suboptimal, demonstrating that objective measures from routinely collected data are preferred over self-reported CTA.

Funding: Health and Care Research Wales and ABMU Health Board. Supported by Asthma UK Centre for Applied Research (AUK-AC-2012-01) and the Farr Institute @ CIPHER.

B.2 Read Codes sets used to define the observed variables in the latent class analysis

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3.

Read code	Description
Asthma Diagnosis Codes	
173A	Exercise induced asthma
102..	Asthma confirmed
663V.	Asthma severity
663V0	Occasional asthma
663V1	Mild asthma
663V2	Moderate asthma
663V3	Severe asthma
9Q21.	Patient in asthma study
H3120	Chronic asthmatic bronchitis
H33%%	Asthma
Asthma GP Visits	
173A.	Exercise induced asthma
173c.	Occupational asthma
173d.	Work aggravated asthma
178	Asthma trigger
1780.	Aspirin induced asthma
1781.	Asthma trigger - pollen
1782.	Asthma trigger - tobacco smoke
1783.	Asthma trigger - warm air
1784.	Asthma trigger - emotion
1785.	Asthma trigger - damp
1786.	Asthma trigger - animals
1787.	Asthma trigger - seasonal
1788.	Asthma trigger - cold air
1789.	Asthma trigger - respiratory infection
178A.	Asthma trigger - airborne dust
178B.	Asthma trigger - exercise
102..	Asthma confirmed
388t	Royal College of Physicians asthma assessment
388t.	Royal College of Physicians asthma assessment
38DL.	Asthma control test
38DV.	Mini asthma quality of life questionnaire
38QM.	Childhood Asthma Control Test
661M1	Asthma self-management plan agreed
661N1	Asthma self-management plan review
663N.	Asthma disturbing sleep
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
663P0	Asthma limits activities 1 to 2 times per month
663P1	Asthma limits activities 1 to 2 times per week
663P2	Asthma limits activities most days
663Q.	Asthma not limiting activities
663U.	Asthma management plan given
663V.	Asthma severity
663V0	Occasional asthma
663V1	Mild asthma
663V2	Moderate asthma
663V3	Severe asthma
663W.	Asthma prophylactic medication used
663d.	Emergency asthma admission since last appointment

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β_2 agonists; OCS = oral corticosteroids; SABA = short-acting β_2 agonists.

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3. (cont'd).

Read code	Description
663e.	Asthma restricts exercise
663e0	Asthma sometimes restricts exercise
663e1	Asthma severely restricts exercise
663f.	Asthma never restricts exercise
663h.	Asthma - currently dormant
663j.	Asthma - currently active
663m.	Asthma accident and emergency attendance since last visit
663n.	Asthma treatment compliance satisfactory
663p.	Asthma treatment compliance unsatisfactory
663q.	Asthma daytime symptoms
663r.	Asthma causes night symptoms 1 to 2 times per month
663s.	Asthma never causes daytime symptoms
663t.	Asthma causes daytime symptoms 1 to 2 times per month
663u.	Asthma causes daytime symptoms 1 to 2 times per week
663v.	Asthma causes daytime symptoms most days
663w.	Asthma limits walking up hills or stairs
663x.	Asthma limits walking on the flat
663y.	Number of asthma exacerbations in past year
66Y5.	Change in asthma management plan
66Y9.	Step up change in asthma management plan
66YA.	Step down change in asthma management plan
66YC.	Absent from work or school due to asthma
66YE.	Asthma monitoring due
66YJ.	Asthma annual review
66YK.	Asthma follow-up
66YP.	Asthma night-time symptoms
66YQ.	Asthma monitoring by nurse
66YR.	Asthma monitoring by doctor
66YZ.	Does not have asthma management plan
66Yp.	Asthma review using Royal College of Physicians three questions
66Yq.	Asthma causes night time symptoms 1 to 2 times per week
66Yr.	Asthma causes symptoms most nights
66Ys.	Asthma never causes night symptoms
66Yu.	Number of days absent from school due to asthma in past 6 months
679J.	Health education - asthma
679J0	Health education - asthma self management
679J1	Health education - structured asthma discussion
679J2	Health education - structured patient focused asthma discussion
8791.	Further asthma - drug prevent.
8793.	Asthma control step 0
8794.	Asthma control step 1
8795.	Asthma control step 2
8796.	Asthma control step 3
8797.	Asthma control step 4
8798.	Asthma control step 5
8B3j.	Asthma medication review
8CMA0	Patient has a written asthma personal action plan
8CRO.	Asthma clinical management plan
8H2P.	Emergency admission, asthma
8HTT.	Referral to asthma clinic
9N1d.	Seen in asthma clinic
9N1d0	Seen in school asthma clinic
9NI8.	Asthma outreach clinic
9NNX.	Under care of asthma specialist nurse
9OJ..	Asthma monitoring admin.
9OJ1.	Attends asthma monitoring
9OJ2.	Refuses asthma monitoring
9OJ3.	Asthma monitor offer default
9OJ4.	Asthma monitor 1st letter
9OJ5.	Asthma monitor 2nd letter
9OJ6.	Asthma monitor 3rd letter
9OJ7.	Asthma monitor verbal invite
9OJ8.	Asthma monitor phone invite
9OJ9.	Asthma monitoring deleted
9OJA.	Asthma monitoring check done
9OJB.	Asthma monitoring invitation SMS (short message service) text message
9OJC.	Asthma monitoring invitation email

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β_2 agonists; OCS = oral corticosteroids; SABA = short-acting β_2 agonists.

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3. (cont'd).

Read code	Description
90JZ.	Asthma monitoring admin.NOS
9Q21.	Patient in asthma study
9hA..	Exception reporting: asthma quality indicators
9hA1.	Excepted from asthma quality indicators: Patient unsuitable
SLF7.	Antiasthmatic poisoning
SLF7z	Antiasthmatic poisoning NOS
COPD Diagnosis Codes	
H3...	Chronic obstructive pulmonary disease
H31	Chronic bronchitis
H32	Emphysema
H36	Mild chronic obstructive pulmonary disease
H37	Moderate chronic obstructive pulmonary disease
H38	Severe chronic obstructive pulmonary disease
H39	Very severe chronic obstructive pulmonary disease
H3A	End stage chronic obstructive airways disease
H3y	Other specified chronic obstructive airways disease
H3z	Chronic obstructive airways disease NOS
H4640	Chronic emphysema due to chemical fumes
H4641	Obliterative bronchiolitis due to chemical fumes
H5832	Eosinophilic bronchitis
Hyu30	[X]Other emphysema
Hyu31	[X]Other specified chronic obstructive pulmonary disease
H3101	Smokers' cough
H31y0	Chronic tracheitis
COPD GP Visits	
66YL.	Chronic obstructive pulmonary disease follow-up
66YS.	Chronic obstructive pulmonary disease monitoring by nurse
66YT.	Chronic obstructive pulmonary disease monitoring by doctor
66YB	Chronic obstructive pulmonary disease monitoring
66YM	Chronic obstructive pulmonary disease annual review
90i	Chronic obstructive pulmonary disease monitoring administration
SABA Inhalers	
c11	SALBUTAMOL [ORAL PREPARATIONS]
c12	SALBUTAMOL [PARENTERAL PREPARATIONS]
c13	SALBUTAMOL [INHALATION PREPARATIONS]
c14	TERBUTALINE SULPHATE [RESPIRATORY USE]
c15	FENOTEROL HYDROBROMIDE
c1E	SALBUTAMOL [INHALATION PREPARATIONS 2]
OCS	
fe6	PREDNISOLONE [ENDOCRINE]
fe61.	PREDNISOLONE 1mg tablets
fe62.	PREDNISOLONE 5mg tablets
fe64.	*DELTA-PHORICOL 5mg tablets
fe65.	DELTACORTRIL ENTERIC 2.5mg tablets
fe66.	DELTACORTRIL ENTERIC 5mg tablets
fe67.	*DELTALONE 1mg tablets
fe68.	*DELTALONE 5mg tablets
fe69.	*DELTASTAB 1mg tablets
fe6a.	*DELTASTAB 5mg tablets
fe6c.	*PRECORTISYL 1mg tablets
fe6d.	*PRECORTISYL 5mg tablets
fe6e.	PRECORTISYL FORTE 25mg tablets
fe6f.	*PREDNESOL 5mg tablets
fe6g.	*SINTISONE 5mg tablets
fe6h.	PREDNISOLONE 2.5mg e/c tablets
fe6i.	PREDNISOLONE 5mg e/c tablets
fe6j.	PREDNISOLONE 5mg soluble tablets
fe6k.	PREDNISOLONE 50mg tablets
fe6l.	DILACORT 5mg gastro-resistant tablets

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β 2 agonists; OCS = oral corticosteroids; SABA = short-acting β 2 agonists.

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3. (cont'd).

Read code	Description
fe6m.	DILACORT 2.5mg gastro-resistant tablets
fe6n.	PEVANTI 2.5mg tablets
fe6o.	PEVANTI 25mg tablets
fe6p.	PEVANTI 5mg tablets
fe6q.	PEVANTI 10mg tablets
fe6r.	PEVANTI 20mg tablets
fe6s.	PREDNISOLONE 20mg tablets
fe6t.	PREDNISOLONE 10mg tablets
fe6v.	*PREDNISOLONE 2.5mg tablets
fe6w.	*PREDNISOLONE 2.5mg tablets
fe6z.	PREDNISOLONE 25mg tablets
fe63	*CODELSOL 32mg/2mL injection
fe6b	DELTA TAB 25mg/1mL injection
fe6u	PREDNISOLONE 32mg/2mL injection
fe6y	PREDNISOLONE 125mg/5mL injection
LABA Inhalers	
c19..	SALMETEROL XINAFOATE
c191.	SALMETEROL 25microgram inhaler
c192.	*SEREVENT 25microgram inhaler
c193.	SEREVENT 50microgram diskhaler
c194.	SEREVENT 50micrograms disk refill
c195.	SALMETEROL 50micrograms disks+disk inhaler
c196.	SALMETEROL 50micrograms disk refill
c197.	SALMETEROL 50micrograms breath-actuated dry powder inhaler
c198.	SEREVENT 50micrograms Accuhaler
c199.	SEREVENT 25micrograms Evohaler
c19z.	SALMETEROL 25micrograms CFC-free inhaler
c1B..	BAMBUTEROL HYDROCHLORIDE
c1B1.	BAMBEC 10mg tablets
c1B2.	BAMBEC 20mg tablets
c1B3.	BAMBUTEROL HYDROCHLORIDE 10mg tablets
c1B4.	BAMBUTEROL HYDROCHLORIDE 20mg tablets
c1C..	FORMOTEROL
c1C1.	FORMOTEROL FUMARATE 12micrograms inhalation capsules+inhaler
c1C2.	FORADIL 12micrograms inhalation capsules+inhaler
c1C3.	FORMOTEROL FUMARATE DIHYDRATE 6micrograms breath-act dry powder inhaler
c1C4.	FORMOTEROL FUMARATE DIHYDRATE 12micrograms breath-act dry powder inhaler
c1C5.	OXIS 6micrograms Turbohaler
c1C6.	OXIS 12micrograms Turbohaler
c1C7.	ATIMOS MODULITE 12micrograms metered dose inhaler
c1C8	FORMOTEROL EASYHALER 12micrograms breath-act dry powder inhaler
c1Cy	FORMOTEROL FUMARATE DIHYDRATE 12micrograms breath-act dry powder inhaler
c1Cz.	FORMOTEROL FUMARATE DIHYDRATE 12micrograms metered dose inhaler
c1a..	TULOBUTEROL HYDROCHLORIDE
c1a1.	*TULOBUTEROL 2mg tablets
c1a2.	*BRELOMAX 2mg tablets
c1a3.	*RESPACAL 2mg tablets
c1a4.	TULOBUTEROL 1mg/5mL sugar free liquid
c1a5.	RESPACAL 1mg/5mL sugar free liquid
ICS-LABA Combination Inhalers	
c1D	SALMETEROL+FLUTICASONE PROPIONATE
c1c	FLUTICASONE PROPIONATE+FORMOTEROL FUMARATE
c67	BUDESONIDE+FORMOTEROL
c6A	BECLOMETASONE+FORMOTEROL
c6B	FLUTICASONE+VILANTEROL
ICS Inhalers	
c6...	CORTICOSTEROIDS [RESPIRATORY USE]

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β_2 agonists; OCS = oral corticosteroids; SABA = short-acting β_2 agonists.

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3. (cont'd).

Read code	Description
c61..	BECLOMETASONE DIPROPIONATE [RESPIRATORY USE]
c611.	BECLOFORTE 250microgram inhaler
c612.	BECOTIDE-50 50microgram inhaler
c613.	BECOTIDE 100micrograms rotacaps
c614.	BECOTIDE 200micrograms rotacaps
c615.	*BECOTIDE rotahaler device
c616.	BECOTIDE 50micrograms/mL nebuliser solution
c617.	BECOTIDE-100 100microgram inhaler
c618.	*VOLUMATIC spacer device
c619.	BECODISK 100micrograms diskhaler 14x8
c61A.	BECLOMETASONE DIPROPIONATE 400micrograms disks+disk inhaler
c61B.	BECLOMETASONE DIPROPIONATE 400micrograms disk refill
c61C.	BECLOMETHASONE DIPROPIONATE 250micrograms inhaler+spacer device
c61E.	BECLOMETASONE DIPROPIONATE 250micrograms breath-actuated aerosol inhaler
c61F.	BECLOMETASONE DIPROPIONATE 100micrograms breath-actuated aerosol inhaler
c61G.	*FILAIR 50micrograms inhaler
c61H.	*FILAIR 100micrograms inhaler
c61J.	FILAIR FORTE 250micrograms inhaler
c61K.	BECLAZONE 50micrograms inhaler
c61L.	BECLAZONE 100micrograms inhaler
c61M.	BECLAZONE 250micrograms inhaler
c61N.	BECLAZONE 50 EASI-BREATHE inhaler
c61O.	BECLAZONE 100 EASI-BREATHE inhaler
c61P.	BECLAZONE 250 EASI-BREATHE inhaler
c61Q.	BECLOFORTE INTEGRA 250micrograms inhaler+compact spacer
c61R.	BECLOFORTE INTEGRA 250micrograms refill
c61S.	BECLOMETHASONE DIPROPIONATE 250micrograms inhaler+compact spacer
c61T.	BECLOMETHASONE DIPROPIONATE 250micrograms compact spacer refill
c61U.	BECLOMETHASONE rotahaler device
c61V.	BECLOMETHASONE DIPROPIONATE 50micrograms vortex metered dose inhaler
c61W.	*BDP 50micrograms Spacehaler
c61X.	BECLOMETHASONE DIPROPIONATE 100micrograms vortex metered dose inhaler
c61Y.	*BDP 100micrograms Spacehaler
c61Z.	BECLOMETHASONE DIPROPIONATE 250micrograms vortex metered dose inhaler
c61a.	BECODISK 200micrograms diskhaler 14x8
c61b.	BECOTIDE 400micrograms rotacaps
c61c.	BECODISK 100micrograms disk refill 14x8
c61d.	BECODISK 200micrograms disk refill 14x8
c61e.	BECODISK 400micrograms diskhaler 7x8
c61f.	BECODISK 400micrograms disk refill 7x8
c61g.	BECLOFORTE VM 250micrograms inhaler+volumatic
c61h.	BECLOMETASONE DIPROPIONATE 400micrograms inhalation capsules
c61i.	BECOTIDE-200 200microgram inhaler
c61j.	*AEROBEC 50microgram Autohaler
c61k.	AEROBEC FORTE 250micrograms Autohaler
c61l.	AEROBEC 100microgram Autohaler
c61m.	BECLOFORTE DISKHALER 400micrograms 14x8
c61n.	BECLOFORTE DISKS 400micrograms disk refill 14x8
c61p.	BECLOMETASONE DIPROPIONATE 100micrograms disks+disk inhaler
c61q.	BECLOMETASONE DIPROPIONATE 200micrograms disks+disk inhaler
c61r.	BECLOMETASONE DIPROPIONATE 100micrograms disk refill
c61s.	BECLOMETASONE DIPROPIONATE 200micrograms disk refill
c61u.	BECLOMETASONE DIPROPIONATE 200micrograms inhaler
c61v.	BECLOMETASONE DIPROPIONATE 50micrograms inhaler
c61w.	BECLOMETASONE DIPROPIONATE 100micrograms inhalation capsules
c61x.	BECLOMETASONE DIPROPIONATE 200micrograms inhalation capsules
c61y.	BECLOMETHASONE DIPROPIONATE 50micrograms/mL nebuliser solution
c61z.	BECLOMETASONE DIPROPIONATE 100micrograms inhaler
c62..	BECLOMETASONE COMPOUNDS
c621.	*VENTIDE inhaler
c622.	*VENTIDE Rotacaps
c623.	*VENTIDE paediatric Rotacaps

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β_2 agonists; OCS = oral corticosteroids; SABA = short-acting β_2 agonists.

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3. (cont'd).

Read code	Description
c624.	*VENTIDE Rotahaler device
c63..	*BETAMETHASONE VALERATE
c631.	*BEXTASOL 100microgram inhaler
c63z.	BETAMETHASONE 100micrograms inhaler
c64..	BUDESONIDE [RESPIRATORY USE]
c641.	PULMICORT 200micrograms inhaler 200dose
c642.	PULMICORT 200micrograms refill 100dose
c643.	PULMICORT 200micrograms refill 200dose
c644.	PULMICORT LS 50micrograms inhaler
c645.	PULMICORT LS 50micrograms refill
c646.	*NEBUHALER spacer device
c647.	PULMICORT 200microgram inhaler 100dose
c649.	PULMICORT 400microgram Turbohaler 50dose
c64A.	BUDESONIDE 200micrograms refill cannister
c64B.	BUDESONIDE 50micrograms spacer inhaler
c64C.	PULMICORT 200micrograms spacer inhaler
c64D.	PULMICORT LS 50micrograms spacer inhaler
c64E.	PULMICORT 200micrograms inhaler with NebuChamber
c64F.	BUDESONIDE 200micrograms/dose dry powder cartridge refill
c64G.	NOVOLIZER BUDESONIDE 200micrograms/dose dry powder cartridge refill
c64H.	EASYHALER BUDESONIDE 100micrograms breath-actuated dry powder inhaler
c64I.	EASYHALER BUDESONIDE 200micrograms breath-actuated dry powder inhaler
c64J.	EASYHALER BUDESONIDE 400micrograms breath-actuated dry powder inhaler
c64K.	PULMICORT 100micrograms CFC-free inhaler
c64a.	PULMICORT 500micrograms Respules 2mL unit
c64b.	PULMICORT 1mg Respules 2mL unit
c64c.	PULMICORT 100microgram Turbohaler 200dose
c64d.	BUDESONIDE 100micrograms breath-actuated dry powder inhaler
c64e.	BUDESONIDE 50micrograms refill cannister
c64g.	BUDESONIDE 200micrograms breath-actuated dry powder inhaler
c64h.	BUDESONIDE 400micrograms breath-actuated dry powder inhaler
c64i.	BUDESONIDE 500micrograms/2mL nebuliser solution
c64j.	BUDESONIDE 1mg/2mL nebuliser solution
c64k.	*BUDESONIDE 200 Cyclocaps
c64l.	*BUDESONIDE 400 Cyclocaps
c64m.	BUDESONIDE 200micrograms inhalation capsules
c64n.	BUDESONIDE 400micrograms inhalation capsules
c64o.	BUDESONIDE 200micrograms inhaler with spacer device
c64p.	NOVOLIZER BUDESONIDE 200micrograms/dose dry powder cartridge and refillable inhaler device
c64u.	BUDESONIDE 200micrograms/dose dry powder cartridge and refillable inhaler device
c64v.	BUDESONIDE 200micrograms inhaler
c64x.	*BUDESONIDE refill 200dose
c64y.	BUDESONIDE 50micrograms inhaler
c64z.	BUDESONIDE 200micrograms spacer inhaler
c65..	FLUTICASONE PROPIONATE [RESPIRATORY USE]
c651.	FLIXOTIDE 50micrograms diskhaler
c652.	FLIXOTIDE 100micrograms diskhaler
c653.	FLIXOTIDE 250micrograms diskhaler
c654.	FLUTICASONE PROPIONATE 50micrograms disks+disk inhaler
c655.	FLUTICASONE PROPIONATE 100micrograms disks+disk inhaler
c656.	FLUTICASONE PROPIONATE 250micrograms disks+disk inhaler
c657.	FLIXOTIDE 50micrograms disk refill
c658.	FLIXOTIDE 100micrograms disk refill
c659.	FLIXOTIDE 250micrograms disk refill
c65A.	FLUTICASONE PROPIONATE 50micrograms disk refill
c65B.	FLUTICASONE PROPIONATE 100micrograms disk refill
c65C.	FLUTICASONE PROPIONATE 250micrograms disk refill
c65D.	FLIXOTIDE 25micrograms inhaler
c65E.	FLIXOTIDE 50micrograms inhaler
c65F.	FLIXOTIDE 125micrograms inhaler
c65G.	FLUTICASONE PROPIONATE 25micrograms inhaler
c65H.	FLUTICASONE PROPIONATE 50micrograms inhaler
c65I.	FLUTICASONE PROPIONATE 125micrograms inhaler
c65K.	FLIXOTIDE 250micrograms inhaler
c65L.	FLIXOTIDE 500micrograms diskhaler

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β 2 agonists; OCS = oral corticosteroids; SABA = short-acting β 2 agonists.

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3. (cont'd).

Read code	Description
c65M.	FLIXOTIDE 500micrograms disk refill
c65N.	FLUTICASON PROPIONATE 500micrograms disks+disk inhaler
c65O.	FLUTICASON PROPIONATE 500micrograms disk refill
c65P.	FLUTICASON PROPIONATE 50micrograms breath-actuated dry powder inhaler
c65Q.	FLUTICASON PROPIONATE 100micrograms breath-actuated dry powder inhaler
c65R.	FLUTICASON PROPIONATE 250micrograms breath-actuated dry powder inhaler
c65S.	FLUTICASON PROPIONATE 500micrograms breath-actuated dry powder inhaler
c65T.	FLIXOTIDE 50micrograms Accuhaler
c65U.	FLIXOTIDE 100micrograms Accuhaler
c65V.	FLIXOTIDE 250micrograms Accuhaler
c65W.	FLIXOTIDE 500micrograms Accuhaler
c65X.	FLUTICASON PROPIONATE 0.5mg/2mL nebulisation units
c65Y.	FLUTICASON PROPIONATE 2mg/2mL nebulisation units
c65Z.	FLIXOTIDE 0.5mg/2mL Nebules
c65a.	FLIXOTIDE 2mg/2mL Nebules
c65b.	FLUTICASON PROPIONATE 125micrograms CFC-free inhaler
c65c.	FLUTICASON PROPIONATE 250micrograms CFC-free inhaler
c65d.	FLIXOTIDE 125micrograms Evohaler
c65e.	FLIXOTIDE 250micrograms Evohaler
c65f.	FLUTICASON PROPIONATE 50micrograms CFC-free inhaler
c65g.	FLIXOTIDE 50micrograms Evohaler
c66..	BECLOMETASONE DIPROPIONATE [RESPIRATORY USE 2]
c661.	*BDP 250micrograms Spacehaler
c662.	BECOTIDE 50 EASI-BREATHE inhaler
c663.	BECOTIDE 100 EASI-BREATHE inhaler
c664.	BECLOFORTE EASI-BREATHE 250micrograms inhaler
c665.	QVAR 50 inhaler
c666.	QVAR 100 inhaler
c667.	QVAR 50 Autohaler
c668.	QVAR 100 Autohaler
c669.	*BECLAZONE 200 inhaler
c66A.	BECLOMETASONE DIPROPIONATE 50micrograms breath-actuated dry powder inhaler
c66B.	BECLOMETASONE DIPROPIONATE 100micrograms breath-actuated dry powder inhaler
c66C.	BECLOMETASONE DIPROPIONATE 250micrograms breath-actuated dry powder inhaler
c66D.	ASMABEC 50micrograms Clickhaler
c66E.	ASMABEC 100micrograms Clickhaler
c66F.	ASMABEC 250micrograms Clickhaler
c66G.	BECLOMETASONE DIPROPIONATE 400micrograms breath-actuated dry powder inhaler
c66H.	BECLOMETASONE DIPROPIONATE 200micrograms breath-actuated dry powder inhaler
c66I.	PULVINAL BECLOMETHASONE DIPROPIONATE 100micrograms breath-actuated dry powder inhaler
c66J.	PULVINAL BECLOMETHASONE DIPROPIONATE 200micrograms breath-actuated dry powder inhaler
c66K.	PULVINAL BECLOMETHASONE DIPROPIONATE 400micrograms breath-actuated dry powder inhaler
c66L.	*BECLOMETASONE 100 cyclocaps
c66M.	*BECLOMETASONE 200 cyclocaps
c66N.	*BECLOMETASONE 400 cyclocaps
c66P.	BECODISK 100micrograms diskhaler 15x8
c66Q.	BECODISK 200micrograms diskhaler 15x8
c66R.	BECODISK 400micrograms diskhaler 15x8
c66S.	BECODISK 100micrograms disk refill 15x8
c66T.	BECODISK 200micrograms disk refill 15x8
c66U.	BECODISK 400micrograms disk refill 15x8
c66V.	BECLOMETASONE DIPROPIONATE 50micrograms CFC-free inhaler
c66W.	BECLOMETASONE DIPROPIONATE 100micrograms CFC-free inhaler
c66X.	BECLOMETASONE DIPROPIONATE 50micrograms CFC-free breath-actuated aerosol inhaler

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β 2 agonists; OCS = oral corticosteroids; SABA = short-acting β 2 agonists.

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3. (cont'd).

Read code	Description
c66Y.	BECLOMETASONE DIPROPIONATE 100micrograms CFC-free breath-actuated aerosol inhaler
c66Z.	QVAR EASI-BREATHE 50micrograms CFC-free breath-actuated dry powder inhaler
c66a.	QVAR EASI-BREATHE 100micrograms CFC-free breath-actuated dry powder inhaler
c66c.	CLENIL MODULITE 50micrograms CFC-free inhaler
c66d.	CLENIL MODULITE 100micrograms CFC-free inhaler
c66e.	CLENIL MODULITE 200micrograms CFC-free inhaler
c66f.	CLENIL MODULITE 250micrograms CFC-free inhaler
c66g.	BECLOMETASONE DIPROPIONATE 200micrograms CFC-free inhaler
c66h.	BECLOMETASONE DIPROPIONATE 250micrograms CFC-free inhaler
c68..	MOMETASONE [RESPIRATORY USE]
c681.	MOMETASONE FUROATE 200micrograms breath-actuated dry powder inhaler
c682.	MOMETASONE FUROATE 400micrograms breath-actuated dry powder inhaler
c683.	ASMANEX TWISTHALER 200micrograms breath-actuated dry powder inhaler
c684.	ASMANEX TWISTHALER 400micrograms breath-actuated dry powder inhaler
c69..	CICLESONIDE
c691.	ALVESCO 160micrograms inhaler
c692.	ALVESCO 80micrograms inhaler
c69y.	CICLESONIDE 80micrograms inhaler
c69z.	CICLESONIDE 160micrograms inhaler
COPD-specific Prescriptions	
8BMW.	Issue of chronic obstructive pulmonary disease rescue pack
81610	Chronic obstructive pulmonary disease rescue pack not indicated
81EZ.	Chronic obstructive pulmonary disease rescue pack declined
a46..	GLYCOPYRRONIUM BROMIDE [ANTISPASMODIC]
a46z.	*GLYCOPYRRONIUM 2mg tablets
c1b..	INDACATEROL
c1b1.	ONBREZ BREEZHALER 150micrograms inhalation capsules+inhaler
c1b2.	INDACATEROL 150micrograms inhalation capsules+inhaler
c1b3.	ONBREZ BREEZHALER 300micrograms inhalation capsules+inhaler
c1b4.	INDACATEROL 300micrograms inhalation capsules+inhaler
c1d..	OLODATEROL
c1d1.	STRIVERDI RESPIMAT 2.5micrograms inhaler
c1d2.	OLODATEROL 2.5micrograms inhaler
c1e..	INDACATEROL+GLYCOPYRRONIUM
c1e2.	INDACATEROL+GLYCOPYRRONIUM 85mcg/43mcg inh powder caps+inh
c3...	ANTICHOLINERGIC BRONCHODILATORS
c31..	IPRATROPIUM BROMIDE [1]
c311.	*ATROVENT 20micrograms inhaler
c312.	ATROVENT 500microgram/2mL nebuliser solution
c313.	ATROVENT FORTE 40microgram inhaler
c314.	ATROVENT 250microgram/1mL nebuliser solution
c315.	ATROVENT 20micrograms Autohaler
c316.	STERI-NEB IPRATROPIUM 250micrograms/1mL nebulisation units
c317.	STERI-NEB IPRATROPIUM 500micrograms/2mL nebulisation units
c318.	ATROVENT 40micrograms Aerocaps refill pack
c319.	ATROVENT 40micrograms Aerocaps+Aerohaler device
c31A.	IPRATROPIUM BROMIDE 40mcg inhalation capsules
c31B.	IPRATROPIUM BROMIDE 40mcg inhalation capsules+inhaler device
c31C.	RESPONTIN 250micrograms/1mL Nebules
c31D.	RESPONTIN 500micrograms/2mL Nebules
c31E.	TROPIOVENT 250micrograms/1mL Steripoules
c31F.	TROPIOVENT 500micrograms/2mL Steripoules
c31G.	ATROVENT 20micrograms CFC-free inhaler
c31t.	IPRATROPIUM BROMIDE 20micrograms CFC-free inhaler
c31u.	IPRATROPIUM 20micrograms breath-actuated aerosol inhaler
c31v.	IPRATROPIUM 250micrograms/1mL nebuliser solution
c31w.	IPRATROPIUM 500micrograms/2mL nebuliser solution
c31x.	IPRATROPIUM 20micrograms inhaler
c31y.	IPRATROPIUM 250micrograms/mL nebuliser solution
c31z.	IPRATROPIUM 40microgram inhaler
c32..	OXITROPIUM BROMIDE

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β_2 agonists; OCS = oral corticosteroids; SABA = short-acting β_2 agonists.

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3. (cont'd).

Read code	Description
c321.	OXITROPIUM 100micrograms/dose inhaler 200dose
c322.	OXIVENT 100micrograms/dose inhaler 200dose
c323.	OXIVENT 100micrograms Autohaler
c324.	OXITROPIUM 100micrograms breath-actuated aerosol inhaler
c33..	TIOTROPIUM
c331.	TIOTROPIUM 18micrograms inhalation capsules
c332.	TIOTROPIUM 18micrograms capsules with inhaler device
c333.	TIOTROPIUM 2.5micrograms inhalation cartridges with inhaler device
c33x.	SPIRIVA RESPIMAT 2.5micrograms inhalation cartridges with Respimat inhaler device
c33y.	SPIRIVA COMBOPACK 18micrograms capsules with HandiHaler inhaler device
c33z.	SPIRIVA 18micrograms inhalation capsules
c34..	ACLIDINIUM
c341.	EKLIRA GENUAIR 322micrograms/dose dry powder inhaler
c342.	ACLIDINIUM 322micrograms/dose dry powder inhaler
c35..	UMECLIDINIUM
c351.	INCRUSE ELLIPTA 55micrograms/dose dry powder inhaler
c352.	UMECLIDINIUM 55micrograms/dose dry powder inhaler
c51J	UMECLIDINIUM+VILANTEROL 55mcg/22mcg dry powder inhaler
c51L.	ACLIDINIUM+FORMOTEROL FUMARATE DIHYD 340mcg/12mcg pdr inh
c51N	TIOTROPIUM+OLODATEROL 2.5micrograms/2.5micrograms inhaler
cl1..	ROFLUMILAST
cl11.	DAXAS 500micrograms tablets
cl1z.	ROFLUMILAST 500micrograms tablets
pc3..	OXYGEN CYLINDERS
pc31.	OXYGEN BP 1360litres cylinder
pc32.	OXYGEN GAS cylinder AD
pc33.	OXYGEN GAS cylinder AF 1360L
pc34.	OXYGEN GAS cylinder C
pc35.	OXYGEN GAS cylinder D
pc36.	OXYGEN GAS cylinder E
pc37.	OXYGEN GAS cylinder F 1360L
pc38.	OXYGEN GAS cylinder G
pc39.	OXYGEN GAS cylinder J
pc3A.	OXYGEN GAS cylinder PD
pc3B.	OXYGEN GAS cylinder SD
pc3C.	OXYGEN gas cylinder DD
pc3D.	OXYGEN gas cylinder HD
pc3E.	OXYGEN gas cylinder RD
pc3F.	OXYGEN gas cylinder DF
pc3G.	OXYGEN gas cylinder HX
pc3H.	OXYGEN GAS cylinder FC
Current Smoking	
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
137C.	Keeps trying to stop smoking
137G.	Trying to give up smoking
137H.	Pipe smoker
137J.	Cigar smoker
137M.	Rolls own cigarettes
137P.	Cigarette smoker
137R.	Current smoker
137V.	Smoking reduced
137X.	Cigarette consumption
137Y.	Cigar consumption
137Z.	Tobacco consumption NOS
137a.	Pipe tobacco consumption
137b.	Ready to stop smoking
137c.	Thinking about stopping smoking
137d.	Not interested in stopping smoking
137e.	Smoking restarted
137f.	Reason for restarting smoking
137h.	Minutes from waking to first tobacco consumption

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β 2 agonists; OCS = oral corticosteroids; SABA = short-acting β 2 agonists.

Table B.2.1: Read Codes sets used to define the observed variables in the latent class modelling in Chapter 3. (cont'd).

Read code	Description
137m.	Failed attempt to stop smoking
137o.	Waterpipe tobacco consumption
Ex-smoking	
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
137K.	Stopped smoking
137N.	Ex pipe smoker
137O.	Ex cigar smoker
137S.	Ex smoker
137T.	Date ceased smoking
137j.	Ex-cigarette smoker
137l.	Ex roll-up cigarette smoker

COPD = chronic obstructive pulmonary disease; ICS = inhaled corticosteroids; LABA = long-acting β 2 agonists; OCS = oral corticosteroids; SABA = short-acting β 2 agonists.

B.3 Item-response probabilities for the competing latent class models

The following diagrams shows item-response probabilities in the classes of each of the competing models. The names and levels of the observed variables are shown on the left-most column in each diagram. Model diagnostics are shown below each diagram and includes Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), likelihood ratio chi-squared statistic (Gsq), Pearson's Chi-square goodness of fit statistic (Chisq), maximum log-likelihood value (l_{lik}), number of iterations needed (numiter), number of individuals included in the modelling (N), and number of estimated parameters (i.e. the number of degrees of freedom used; npar).

#classes: 1



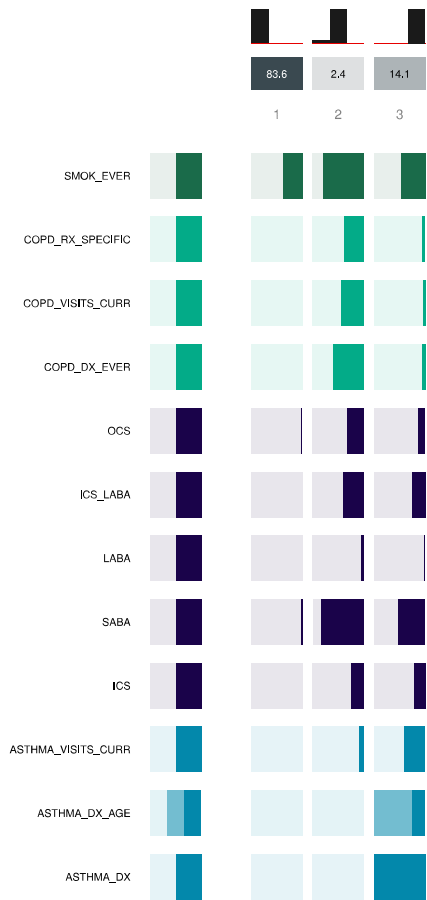
AIC = 297943.83, BIC = 298058.48, Gsq = 109287.92, Chisq = 139774016756.94, llik = -148958.91, numiter = 1/3000, N = 50000/50000, npar = 13

#classes: 2



AIC = 219636.16, BIC = 219874.3, Gsq = 30952.25, Chisq = 7159142.63, llik = -109791.08, numiter = 19/3000, N = 50000/50000, npar = 27

#classes: 3



AIC = 207915.8, BIC = 208277.41, Gsq = 19203.89, Chisq = 3695583.03, Ilik = -103916.9, numiter = 258/3000, N = 50000/50000, npar = 41

#classes: 4



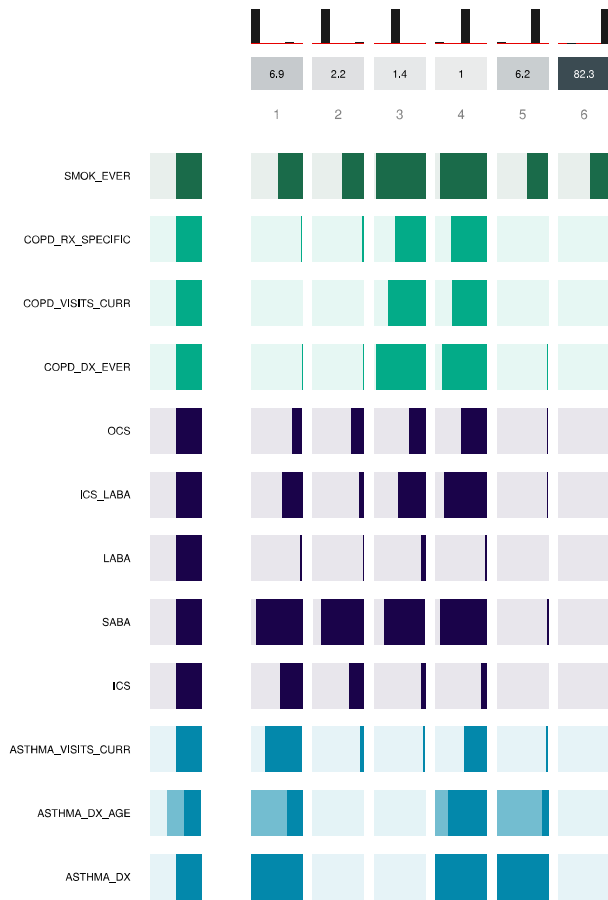
AIC = 197295.36, BIC = 197780.45, Gsq = 8555.45, Chisq = 66993.38, llik = -98592.68, numiter = 53/3000, N = 50000/50000, npar = 55

#classes: 5



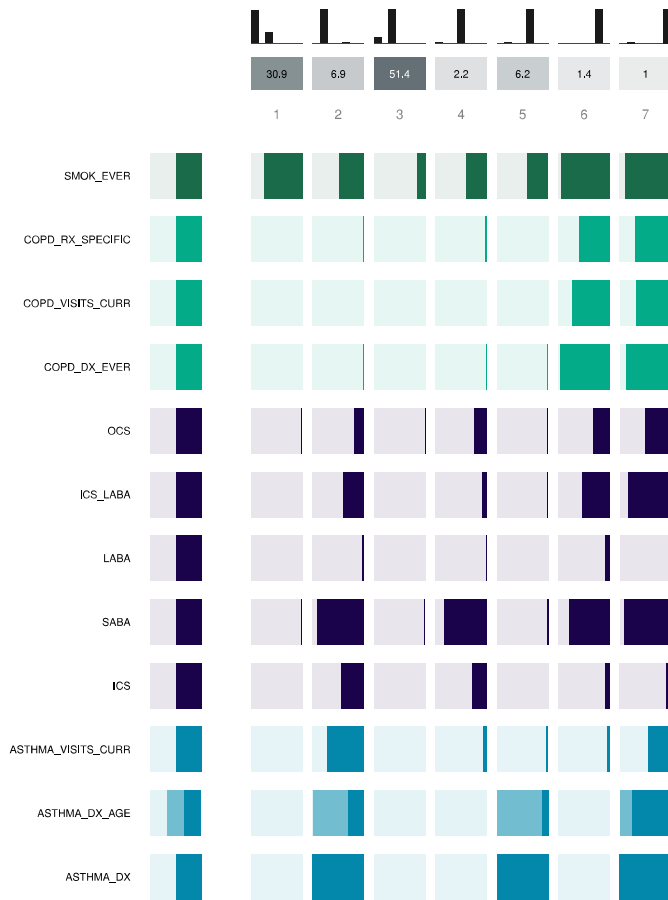
AIC = 195593.31, BIC = 196201.88, Gsq = 6825.4, Chisq = 24435.3, llik = -97727.66, numiter = 76/3000, N = 50000/50000, npar = 69

#classes: 6



AIC = 192656.15, BIC = 193388.19, Gsq = 3860.24, Chisq = 12060.83, llik = -96245.08, numiter = 202/3000, N = 50000/50000, npar = 83

#classes: 7



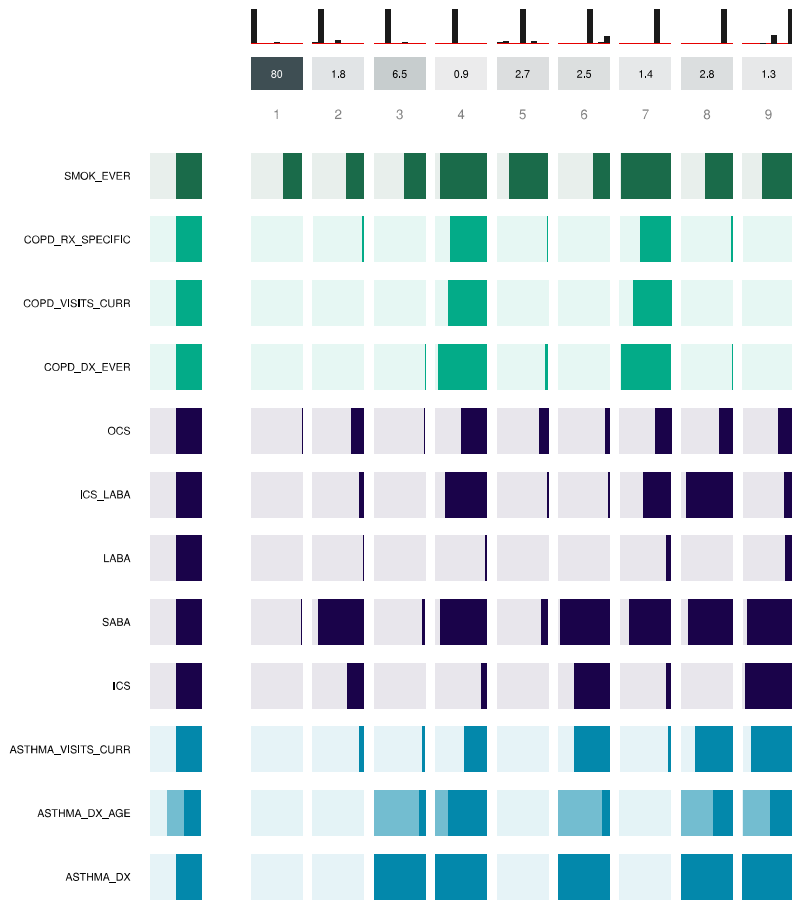
AIC = 192592.52, BIC = 193448.03, Gsq = 3768.61, Chisq = 11973.55, llik = -96199.26, numiter = 3000/3000, N = 50000/50000, npar = 97

#classes: 8



AIC = 190965.89, BIC = 191944.88, Gsq = 2113.98, Chisq = 9427.48, llik = -95371.94, numiter = 1836/3000, N = 50000/50000, npar = 111

#classes: 9



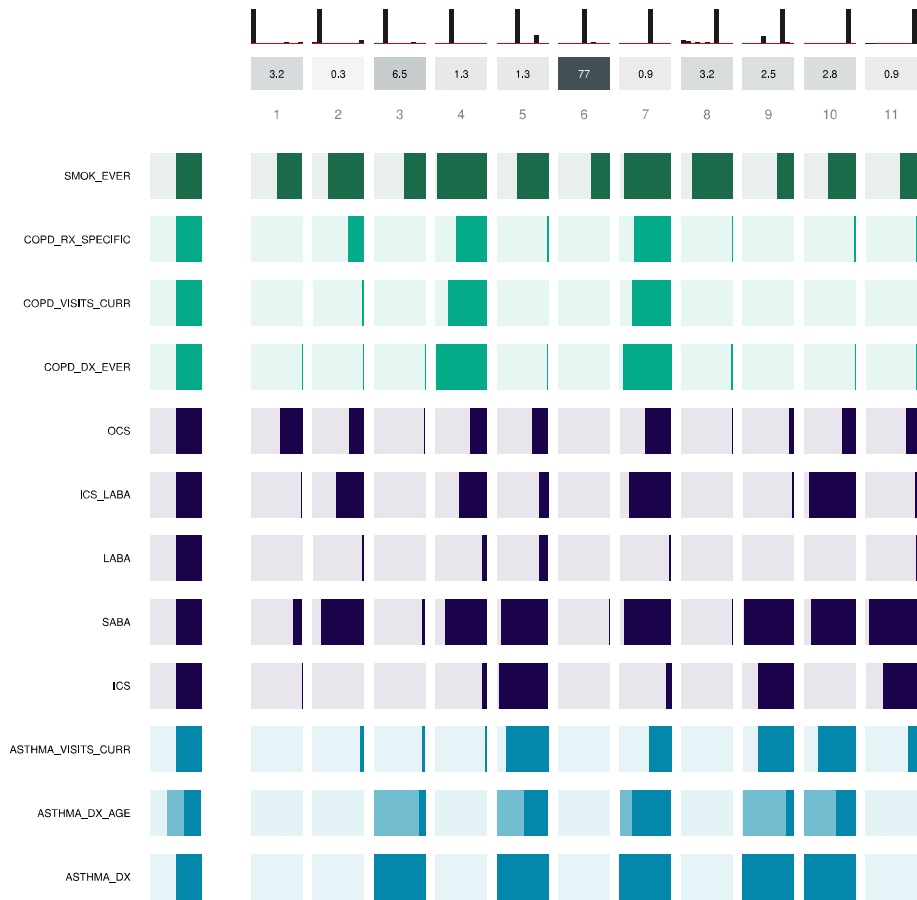
AIC = 190829.52, BIC = 191931.99, Gsq = 1949.61, Chisq = 7716.2, llik = -95289.76, numiter = 3000/3000, N = 50000/50000, npar = 125

#classes: 10



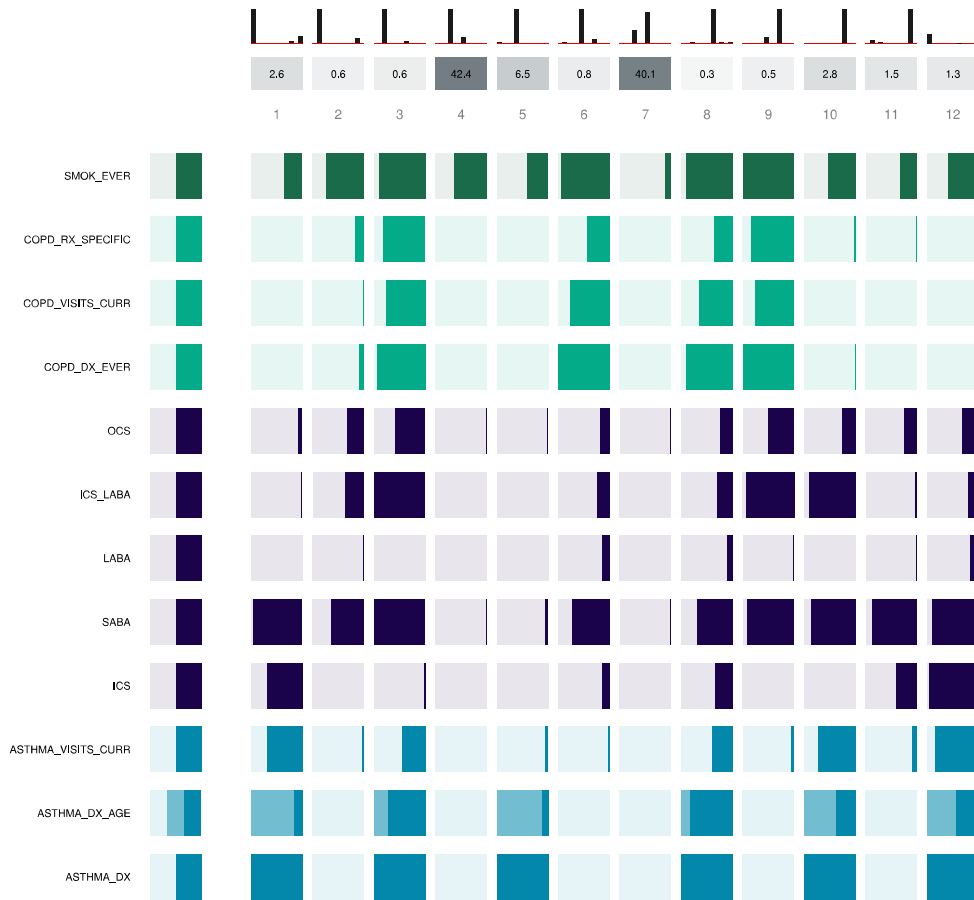
AIC = 190899.72, BIC = 192125.67, Gsq = 1991.81, Chisq = 8809.87, llik = -95310.86, numiter = 3000/3000, N = 50000/50000, npar = 139

#classes: 11



AIC = 190744.87, BIC = 192094.3, Gsq = 1808.96, Chisq = 7230.18, llik = -95219.44, numiter = 3000/3000, N = 50000/50000, npar = 153

#classes: 12



AIC = 190461.81, BIC = 191934.71, Gsq = 1497.9, Chisq = 3916.75, llik = -95063.9, numiter = 3000/3000, N = 50000/50000, npar = 167

B.4 Related study protocol: identifying patients with asthma-COPD overlap syndrome using latent class analysis of electronic health record data

The following study protocol is related to the work presented in [Chapter 3](#) and is focused on identifying people with asthma-COPD overlap syndrome using latent class analysis of electronic health record data. It has been published in *npj Primary Care Respiratory Medicine* (DOI: 10.1038/s41533-018-0088-4, <https://www.nature.com/articles/s41533-018-0088-4>).

Identifying patients with asthma-chronic obstructive pulmonary disease overlap syndrome using latent class analysis of electronic health record data: a study protocol

Mohammad A Al Sallakh, MD^{B1,a}, Sarah E Rodgers, PhD^{1,b}, Ronan A Lyons, MD^{1,b}, Aziz Sheikh, MD^{2,a,b} and Gwyneth A Davies, MD^{1,a}

¹Swansea University Medical School, Singleton Park, Swansea, SA2 8PP, UK

²Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh

^aAsthma UK Centre for Applied Research

^bThe Farr Institute of Health Informatics Research

Correspondence

Mohammad A Al Sallakh, MD, MSc

Data Science Building, Swansea University, Singleton Park, Swansea, SA2 8PP, United Kingdom

Phone: 

Email: M.A.Alsallakh@swansea.ac.uk

Introduction

Asthma and chronic obstructive pulmonary disease (COPD) are two common different clinical diagnoses with overlapping clinical features. Global Initiative for Asthma (GINA) defined asthma based on variable respiratory symptoms and expiratory airflow limitation.¹ On the other hand, the Global Initiative for Chronic Obstructive Lung Disease (GOLD) defined COPD based on persistent respiratory symptoms and airflow limitations.² While asthma affects people from the early school age, COPD mainly affects those aged over 40 years with a smoking history. Clinically, the differentiation between the two diseases and identifying their overlap in those older people can be challenging.¹ Co-existence of clinical features of both conditions along with persistent airflow limitation has been recently recognized by a joint committee publication between GOLD and GINA as the asthma–COPD overlap syndrome (ACOS).³

However, there are currently no universally agreed consensus clinical definitions for the diagnosis of asthma,^{4–9} COPD,^{10,11} and ACOS.^{12–15} Subsequently, the prevalence of these three conditions is highly dependent on the different available case definitions and data sources.^{16–20}

In studies conducted using electronic health records (EHR), identifying patient groups is further complicated by the limitations of these data, such as missing data and coding errors.^{21–23} Despite the lack of consensus clinical definitions, we expect EHR data of people with “ACOS” to be systematically different from those with “asthma only” or “COPD only”. Case definitions aiming to differentiate between those patient groups based solely on clinical knowledge or face validity may be inaccurate, and validating them with traditional methods, e.g., review of full patient records, is time-consuming and labour-intensive. Clustering methods overcome these challenges by automatically identifying subgroups in the population that best explain the patterns in high-dimensional EHR data, without an *a priori* hypothesis about those subgroups and their labels.²⁴ Latent class analysis (LCA) is such a

method that can probabilistically identify patients with asthma and/or COPD using the available recorded data.

Aims

We plan to develop an LCA model to identify and characterise patients with asthma, COPD and ACOS in Wales. Based on this LCA model, we will derive a classification algorithm, and compare its performance with commonly used objective and self-reported case definitions for asthma and COPD.

Methods

We will use primary care data on asthma and COPD recorded in or before 2014 for a sample of the Welsh population to find, using LCA, clinically meaningful classes (i.e. clusters) related to the two conditions in that year. We will follow the STROBE²⁵ and RECORD Statements²⁶ in reporting the full study.

Data sources

We will use the following two de-identified datasets from the Secure Anonymised Information

Linkage (SAIL) Databank in Wales:^{27,28}

- The Welsh Demographic Service (WDS) which contains demographic and administrative information for the National Health Services (NHS) patient in Wales.
- The General Practitioner (GP) dataset which contains primary care events, such as diagnoses, clinical findings, prescriptions codified in Read codes by general practitioners.

At the time of writing of this protocol, the most recent extract of the GP dataset was in March 2017, covering about 80% of GP surgeries in Wales.

Patient population

The study sample will be randomly selected from the total population of Wales within the SAIL Databank in 2014. The sampling will be stratified by general practices to improve their representativeness. We will determine the sample size based on the computational capacity in the SAIL Databank which will be available for this study. The sampling frame will include all individuals who were aged at least 40 years on 1-1-2014.

Latent class modelling

LCA is a finite mixture modelling method that aims to divide a sample into classes or clusters related to a set of observed variables.^{24,29} LCA assumes that the patterns in these observed variables can be explained by, in addition to measurement errors, a hidden categorical variable that divides the sample into a pre-defined number of distinct classes.

In our study, we will construct observed variables from asthma- and COPD-related events recorded in the GP Dataset. The construction of observed variables will be based on their usefulness, from a clinical perspective, for identifying and distinguishing between patients with asthma and/or COPD. These variables will include diagnosis, GP visits, and prescriptions related to asthma and COPD, as well as history of allergy (including atopic eczema/dermatitis, food allergy, allergic rhinitis, and anaphylaxis) and smoking history (see [Table 1](#)). GP visits and prescriptions will be queried during 2014, while the other events will be queried in or any time before 2014.

Model parameters include proportions of the latent classes, and probabilities of observing the levels of observed variables in each latent class, a.k.a item–response probabilities.

Parameters are estimated by the expectation–maximisation (EM) algorithm, which iteratively searches for maximum–likelihood parameter values for which the data are more likely to be observed.³⁰ Based on observed characteristics, each individual is assigned

membership probability in each latent class,²⁹ and is finally assigned to the latent class of maximum membership probability.³¹

We will begin the modelling for two latent classes and will then iteratively increase the numbers of latent classes. Model selection will be based on model diagnostics and interpretability.

We will look for a model for which the Bayesian Information Criterion (BIC)^{32,33} is ideally minimum, or becomes ‘stabilised’, indicating no significant improvement in information gain beyond a certain number of classes. In addition, the selected model should be clinically relevant; we will use the estimated item–response probabilities to assign labels consistent with “asthma”, “COPD”, “both” (ACOS), and “none” to the latent classes. We will use class shares as prevalence estimates for these clinical labels among the age groups of 40 and over in 2014.

LCA modelling will be performed using the R package *poLCA* (version 1.4.1, 2014).³⁴

Derivation of a classification algorithm

Based on the LCA model, we will derive a classification algorithm to identify patients with asthma, COPD and ACOS according to their characteristics. To do so, we will perform recursive partitioning³⁵ using the assigned latent classes as labels and the aforementioned observed variables as predictors. We will use the R package *rpart* (version 4.1-11, 2017)³⁶ for this purpose.

Comparison with other case definitions

We will compare the LCA model and the derived classification algorithm with other objective and self-reported measures. As objective measures, we will use definitions used in the Quality of Outcomes Framework (QOF) 2014–2015 indicators for ‘treated asthma’ (AST001) and ‘COPD’ (COPD001).³⁷ From the Welsh Health Survey (WHS) 2014,³⁸ we will use self-

reported responses on current treatment of ‘asthma’, ‘emphysema’, and ‘spells of bronchitis that have lasted over 3 years’, with any of the latter two representing currently-treated COPD. We will treat invalid and missing responses as negative responses. We will perform the comparisons in the group of the WHS 2014 participants who were aged 40 years or over on 1-1-2014, and whose responses were successfully linked to the SAIL Databank. We will calculate diagnostic accuracy measures of the LCA model and the classification algorithm against each of the above case definitions and vice versa.

Ethics, timeline and dissemination

We obtained an approval to use the SAIL Databank from the Information Governance Review Panel. NHS Research Ethics Committee approval for this study is not required because we will only use anonymised data. The data extraction and statistical analysis will be performed between March and May 2018. The full paper will be submitted for publication in a respiratory care-related peer-reviewed journal in due course.

Discussion

While the interest in ACOS is growing, there is no consensus definition for this emerging and debated concept,³⁹ leading to wide variations in prevalence and impaired comparability between studies. With the increasing use of EHR data to study asthma and COPD, it is important to develop operational definitions for ACOS based on such data. In this study, we will perform LCA on recorded events of diagnosis, prescriptions, and healthcare utilisation for asthma and COPD in routinely collected primary care data. By including observed variables for asthma and COPD in the same model, we will be able to identify patients with either or both conditions (i.e. ACOS).

An inherent limitation of routinely collected EHR data is the lack of vital pieces of information that are often used to make diagnoses at the point of care. Unlike diagnosis and prescriptions which are generally well coded, important diagnostic tests such as lung

function and peripheral eosinophil count are often poorly and inconsistently recorded in primary care datasets. These missing data would have been potentially useful for improving the accuracy of our model. However, it is often difficult to assess data missingness in event-based databases. The GP Dataset in the SAIL Databank is a long-format dataset, in which each row contains a dated code representing a single primary care event. The presence of a code usually indicates that the corresponding event occurred. However, when a code is absent, it is often impossible to ascertain whether the event did not occur or whether it was simply not recorded or coded. This is a particular challenge for events that are known to be poorly recorded. Therefore, since the quality of observed variables is essential in LCA, we will only include variables that are thought to be of reasonable quality in the SAIL Databank. In interpreting the results, we will consider the limitations of EHR-derived data such as the possibility of missing or incorrect codes and the changes in coding practices over time.

LCA itself has limitations. The construction of observed variables, model selection and interpretation involves a level of subjectivity. The model's interpretation and usefulness depends largely on the choice and structure of observed variables. In our LCA modelling, the clinical meaning of the latent classes will be based on surrogate variables, such as diagnosis, GP visits, and prescriptions, rather than on more direct disease markers such as clinical and laboratory findings. Nevertheless, we hypothesise that LCA of these surrogate variables can reasonably distinguish between patients with asthma, COPD, and ACOS. This will also provide an opportunity to assess how clustering based on these surrogate variables will perform compared with that based on disease markers.⁴⁰⁻⁴⁷ Comparing our LCA model and classification algorithm against other objective and self-reported measures will provide useful information about their validity and performance.

Contributions: All authors contributed to refinement of the study protocol and manuscript writing and critically reviewed and approved the final manuscript. Mohammad A Al Sallakh developed the statistical methods.

Funding and Support: This work is funded by Health and Care Research Wales and Abertawe Bro Morgannwg University Health Board. It is carried out with the support of the Asthma UK Centre for Applied Research [AUK-AC-2012-01]. We also acknowledge the support from The Farr Institute of Health Informatics Research. The Farr Institute is supported by a 10-funder consortium: Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the Medical Research Council, the National Institute of Health Research, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates), the Wellcome Trust, (MRC Grant Nos: CIPHER MR/K006525/1, Scotland MR/K007017/1).

Conflict of Interest: AS is the Editor-in-Chief of npj PCRM. The remaining authors declare no conflict of interest.

References

1. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention (2017 update) (2017).
2. Global Initiative for Chronic Obstructive Lung Disease. Global strategy for prevention, diagnosis and management of COPD (2018).
3. Global Initiative for Asthma and Global Initiative for Chronic Obstructive Lung Disease. Diagnosis of diseases of chronic airflow limitation: asthma, COPD and asthma–COPD overlap syndrome (ACOS). (2015).
4. Hargreave FE & Nair P. The definition and diagnosis of asthma. *Clin Exp Allergy* 39, 1652–1658 (2009).
5. A plea to abandon asthma as a disease concept. *Lancet* 368, 705 (2006).
6. Bousquet J *et al.* Uniform definition of asthma severity, control, and exacerbations: document presented for the World Health Organization Consultation on Severe Asthma. *J Allergy Clin Immunol* 126, 926–938 (2010).
7. Reddel HK *et al.* A summary of the new GINA strategy: a roadmap to asthma control. *Eur Respir J* 46, 622–639 (2015).
8. van den Akker IL, van der Zeijden H & Verheij TJ. Is spirometry essential in diagnosing asthma? Yes. *Br J Gen Pract* 66, 484–484 (2016).
9. Levy ML. Is spirometry essential in diagnosing asthma? No. *Br J Gen Pract* 66, 485–485 (2016).
10. Brusasco V. Spirometric definition of COPD: exercise in futility or factual debate? *Thorax* 67, 569–570 (2012).
11. Vestbo J. COPD: Definition and Phenotypes. *Clin Chest Med* 35, 1–6 (2014).
12. Bateman ED, Reddel HK, van Zyl-Smit RN & Agusti A. The asthma–COPD overlap syndrome: towards a revised taxonomy of chronic airways diseases? *Lancet Respir Med* 3, 719–728 (2015).
13. Bujarski S, Parulekar AD, Sharafkhaneh A & Hanania NA. The asthma COPD overlap syndrome (ACOS). *Curr Allergy Asthma Rep* 15, 509 (2015).
14. McDonald VM & Gibson PG. To define is to limit: perspectives on asthma–COPD overlap syndrome and personalised medicine. *Eur Respir J* 49, 1700336 (2017).
15. Miravittles M. Diagnosis of asthma–COPD overlap: the five commandments. *Eur Respir J* 49, 1700506 (2017).
16. Ford ES. The epidemiology of obesity and asthma. *J Allergy Clin Immunol Pract* 115, 897–909, quiz 910 (2005).
17. 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 (2016).

18. Bonten TN *et al.* Defining asthma–COPD overlap syndrome: a population-based study. *Eur Respir J* 49, 1602008 (2017).
19. Halbert R, Isonaka S, George D & Iqbal A. Interpreting COPD Prevalence Estimates. *Chest* 123, 1684–1692 (2003).
20. Viegi G *et al.* Definition, epidemiology and natural history of COPD. *Eur Respir J* 30, 993–1013 (2007).
21. Schneeweiss S & Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 58, 323–337 (2005).
22. Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. *Public Health Research & Practice* 25 (2015).
23. Al Sallakh MA *et al.* Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. *Eur Respir J* 49 (2017).
24. Howard R, Rattray M, Prosperi M & Custovic A. Distinguishing Asthma Phenotypes Using Machine Learning Approaches. *Curr Allergy Asthma Rep* 15, 38 (2015).
25. von Elm E *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 61, 344–349 (2008).
26. Benchimol EI *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 12, e1001885 (2015).
27. Lyons RA *et al.* The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 9, 3 (2009).
28. Ford DV *et al.* The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 9, 157 (2009).
29. Linda M. Collins STL. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. 285 Seiten. ISBN: 0470228393 (John Wiley & Sons Inc, 2010).
30. Dempster AP, Laird NM & Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol*, 1–38 (1977).
31. McLachlan G & Peel D. Finite Mixture Models 1st ed. ISBN: 9780471006268 (Wiley-Interscience, 2000).
32. Schwarz G. Estimating the dimension of a model. *Ann Stat* 6, 461–464 (1978).
33. Nylund KL, Asparouhov T & Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Struct Equ Modeling* 14, 535–569 (2007).
34. Linzer DA & Lewis JB. polCA: An R package for polytomous variable latent class analysis. *J Stat Softw* 42, 1–29 (2011).
35. Strobl C, Malley J & Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* 14, 323–348 (2009).

36. Therneau TM & Atkinson EJ. An Introduction to Recursive Partitioning Using the RPART Routines (2015).
37. General Medical Services Contract: Quality and Outcomes Framework Statistics for Wales, 2014-15. Report. 2015.
38. Welsh Assembly Government. Welsh Health Survey 2014: Health status, illnesses, and other conditions (2015).
39. Rodrigo GJ, Neffen H & Plaza V. Asthma-chronic obstructive pulmonary disease overlap syndrome: a controversial concept. *Curr Opin Allergy Clin Immunol* 17, 36–41 (2017).
40. Haldar P *et al.* Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 178, 218–224 (2008).
41. Moore WC *et al.* Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* 181, 315–323 (2010).
42. Garcia-Aymerich J *et al.* Phenotyping asthma, rhinitis and eczema in MeDALL population-based birth cohorts: an allergic comorbidity cluster. *Allergy* (2015).
43. Weatherall M *et al.* Distinct clinical phenotypes of airways disease defined by cluster analysis. *Eur Respir J* 34, 812–818 (2009).
44. Mäkikyrö EMS, Jaakkola MS & Jaakkola JJK. Subtypes of asthma based on asthma control and severity: a latent class analysis. *Respir Res* 18 (2017).
45. Weinmayr G *et al.* Asthma phenotypes identified by latent class analysis in the ISAAC phase II Spain study. *Clin Exp Allergy* 43, 223–232 (2013).
46. Burgel PR *et al.* Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J* 36, 531–539 (2010).
47. Ghebre MA *et al.* Biological clustering supports both Dutch and British hypotheses of asthma and chronic obstructive pulmonary disease. *J Allergy Clin Immunol* 135, 63–72.e10 (2015).

Appendix C

Chapter 4 Appendix

C.1 SAIL Databank Datasets used in the Observatory

Table C.1.1: Data fields of the datasets that were used for the development of the Wales Asthma Observatory. These metadata are from the SAIL Databank in 2018.

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
Welsh Demographic Service (WDS) - AR_PERS (Administrative Register - Persons)						
pers_id_e	Encrypted Person Id	integer	0	0	5218464	An encrypted unique identifier for the Demographic Service.
alf_e	Encrypted Anonymised Linking Field	bigint	0	0	5218464	The Anonymised Linking Field, within the database, is derived from the person's name. The encryption occurs in NWIS and is not supplied in the data extract the person is linked to.
wob	Week of Birth	date	0	0	6880	The date of the Monday that occurs closest to this data item is limited, however, to the first of the month.
dod	Date of Death	date	4394400	84.21	10079	The date of death for the individual.
gndr_cd	Gender code (also known as sex)	character	0	0	3	This is the sex (gender) of person.
avail_from_dt	Available from date	date	0	0	1	Date when the data made available.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
Welsh Demographic Service (WDS) - AR_PERS_ADD (Administrative Register - Addresses)						
pers_id_e	Encrypted Person Id	integer	0	0	5240178	An encrypted unique identifier for the Demographic Service.
ralf_e	Encrypted Residential Anonymous Linking Field	bigint	1242454	8.79	1332673	The encrypted Residential Anonymous Linking Field derived from the persons postal code, described in the paper "Residential Anonymisation: a novel information infrastructure for a secure environment and individuals health care".
ralf_sts_cd	Residential Anonymous Linking Field_status_code	character	0	0	3	The status code generated when the Residential Anonymous Linking Field is created.
uprn_gas_match_cd	UPRN quality match code	character	134288	0.95	1260	The UPRN (Unique Property Reference Number) quality match code.
lsoa_cd	Local Super Output Area code	character	1228010	8.69	1896	The Local Super Output Area code.
row_sts	Row status	character	0	0	2	Row status code.
from_dt	From date	date	0	0	28413	From date.
to_dt	To date	date	0	0	10428	To date.
avail_from_dt	Available from date	date	0	0	1	Date when the data made available.
Welsh Demographic Service (WDS) - AR_PERS_GP (Administrative Register - GP registration)						
pers_id_e	Encrypted Person Id	integer	0	0	5240179	An encrypted unique identifier for the Demographic Service.
prac_cd_e	Encrypted GP practice code	integer	177360	1.25	663	The encrypted GP practice code.
row_sts	Row status	character	0	0	2	Row status code.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
from_dt	From date	date	0	0	30723	From date.
to_dt	To date	date	0	0	12754	To date.
avail_from_dt	Available from date	date	0	0	1	Date when the data made available.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
Welsh Longitudinal General Practice (WLGP)						
alf_e	Encrypted Anonymised Linking Field	bigint	3462255	0.15	4016617	The Anonymised Linking Field, v database, is derived from the p encryption occurs in NWIS and supplied in the data extract the
alf_sts_cd	ALF status code	character	0	0	5	Status code assigned when deriv
alf_mtch_pct	ALF match percentage	decimal	2367210983	99.68	90859	Match percentage assigned whe Field.
prac_cd_e	Encrypted GP practice code	integer	0	0	340	The encrypted GP practice code
local_num_e	NA	bigint	0	0	274452	NA
gndr_cd	Gender code (also known as sex)	character	0	0	3	This is the sex (gender) of perso
wob	Week of Birth	date	0	0	6857	The date of the Monday that occ to this data item is limited, how events.
lsoa_cd	Local Super Output Area code	character	18212565	0.77	6606	The Local Super Output Area of
reg_cat_cd	Registration category code	character	0	0	36	This denotes the registration sta
event_dt	Event date	date	0	0	52003	The date the event occurred.
event_yr	Event year	smallint	0	0	273	The year the event occurred.
event_cd_vrs	Event code version	character	0	0	2	This denotes the coding classific collated from different GP pract are a variety of coding classifica use Read Code version 2.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
event_cd	Event code	character	0	0	267925	This code documents clinical info. In the majority of practices this will be a description denoting a sign, symptom, or procedure.
event_val	Event value	decimal	1048003348	44.13	118407	This value is associated to the Event code. It could be blood pressure or the number of tablets.
episode	Episode	character	0	0	6	This denotes the type of episode.
sequence	Sequence	integer	0	0	108853875	This numeric sequence denotes the order of the event. For example if a blood pressure is taken within an appointment.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
Patient Episode Database for Wales (PEDW) - SPELL						
prov_unit_cd	Provider unit code	character	0	0	662	This is the organisation code of the provider which identifies the health care provider responsible for the treatment of the patient.
spell_num_e	Encrypted spell number	integer	<5	0	18555146	A number (alphanumeric) to provide a unique provider spell for a health care episode.
gndr_cd	Gender code (also known as sex)	character	0	0	7	This is the sex (gender) of person.
res_dha_cd	District health authority of residence code	character	30526	0.16	1276	The District Health Authority in which the patient resides.
admis_yr	Admission year	character	72	0	104	The year of the beginning of a hospital admission.
admis_dt	Admission date	date	72	0	10242	This is the beginning of a hospital admission when a consultant has assumed responsibility for the patient. This may be before the patient has been admitted, completed and the patient is treated.
fin_admis_yr	Financial admission year	character	72	0	102	The financial year of the beginning of a hospital admission.
admis_mthd_cd	Admission method code	character	974	0.01	21	This is the method of admission to hospital.
admis_source_cd	Admission source code	character	8550	0.04	49	This is the source of admission to hospital.
intended_management_cd	Intended management code	character	1930	0.01	8	The intended pattern of bed use for the patient when made to admit, and only applies to inpatient admissions. This categorization describes whether the patient is admitted as a day case or overnight. Occasionally the patients treated as day cases are admitted overnight. Therefore another code is used to describe what actually happens.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
disch_yr	Discharge year	character	21984	0.11	31	The year of discharge from the Hospital for a patient dies or is discharged from a hospital bed(s) within a single hospital provider or consultant episode of care and spell.
disch_dt	Discharge date	date	21983	0.11	9481	Date of discharge from the Hospital for a patient dies or is discharged from a hospital bed(s) within a single hospital provider or consultant episode of care and spell.
fin_disch_yr	Financial discharge year	character	21984	0.11	30	The financial year of the end of a patient's hospital discharge.
disch_mthd_cd	Discharge method code	character	0	0	8	This is the method of discharge for a patient.
disch_destination_cd	Discharge destination code	character	30548	0.16	75	The classification of where a patient is discharged, spell, or a note that the patient died.
dur_elect_wait	Duration of elective wait	decimal	9064942	46.83	3271	This is the waiting time from the time a patient is referred where the treatment actually takes place.
pat_class_cd	Patient classification code	character	5833	0.03	9	A coded classification of Patients' Hospital Provider Spell.
spell_dur	Spell duration	integer	22093	0.11	3163	The period of time in days between the start and discharge date of the provider spell.
admis_spec_cd	Admission speciality code	character	0	0	145	This is the specialty under which a patient is admitted, either be the same as the specialty function or a different specialty function. Note that both specialty function should be based on the specialty function.
disch_spec_cd	Discharge speciality code	character	0	0	146	This is the specialty under which a patient is discharged, either be the same as the specialty function or a different specialty function. Note that both specialty function should be based on the specialty function.
ua_cd	Unitary Authority code	character	62105	0.32	196	The unitary authority in which the patient is treated.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
res_ward_cd	Ward code of residence	character	86052	0.44	320	The electoral ward in which the
reg_gp_cd_e	Encrypted registered GP code	integer	22083	0.11	42463	The encrypted unique GP Practic
reg_gp_prac_cd_e	Encrypted registered GP practice code	integer	1239686	6.4	10192	The encrypted unique GP Practic
ref_cd_e	Encrypted referrer code	integer	897630	4.64	47174	This is the nationally recognized may be a General Medical Pract (GDP), Consultant or Independ GDP, Consultant or Independen
ref_org_cd	Referring organisation code	character	1054418	5.45	11670	The code of the referring organisi
admis_dec_dt	Admission decision date	date	4919073	25.41	10071	This is the date upon which the c
local_health_grp_cd	Local health group code	integer	19357775	100	0	The local health group associate
curr_prov_unit_cd	Current provider unit code	character	0	0	333	The current organisation code of identifies the health care provic treatment of the patient. To ena years, a current field can be use version/name.
curr_res_dha_cd	Current district health authority of residence code	character	199	0	312	The current District Health Auth longitudinal analysis over a pe map previously used codes to c
curr_lo- cal_health_grp_cd	Current local health group code	integer	19357775	100	0	The current local health group a enable longitudinal analysis ov used to map previously used co
curr_ua_cd	Current Unitary Authority code	character	62105	0.32	196	The current unitary authority in longitudinal analysis over a pe map previously used codes to c

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
curr_res_ward_cd	Current ward residence code	character	86052	0.44	320	The current electoral ward in which the patient is living at the time of the longitudinal analysis over a period of time. The map previously used codes to categorise patients into wards.
pat_id_e	Encrypted patient identifier	integer	0	0	3923115	An encrypted unique Patient Identifier.
alf_e	Encrypted Anonymised Linking Field	bigint	1026066	5.3	3228910	The Anonymised Linking Field, used in the database, is derived from the patient's unique identifier. The encryption occurs in NWIS and is not visible in the data extract the Observatory uses.
case_rec_num_e	Encrypted casenote record number	integer	3211	0.02	4964329	This is the case record number. It is unique to the health care provider.
alf_sts_cd	ALF status code	character	0	0	5	Status code assigned when deriving the ALF.
alf_mtch_pct	ALF match percentage	decimal	19057851	98.45	36332	Match percentage assigned when deriving the ALF.
hrg_localpayment_cd	HRG local payment code	varchar	11598773	59.92	1918	The Healthcare Resource Group code for local payment.
hrg_localpayment_desc	HRG local payment description	varchar	0	0	3079	The Healthcare Resource Group code for local payment spell.
hrg_referencecost_cd	HRG reference cost code	varchar	11601219	59.93	4439	The Healthcare Resource Group code for reference cost.
hrg_referencecost_desc	HRG reference cost description	varchar	0	0	6152	The Healthcare Resource Group code for reference cost overall spell.
hrg_v31_cd	HRG version 31 code	varchar	5887305	30.41	572	Healthcare Resource Group code for version 31.
hrg_v31_desc	HRG version 31 description	varchar	0	0	572	Healthcare Resource Group code for version 31 codes (v3.1).
hrg_v35_cd	HRG version 35 code	varchar	5886058	30.41	610	Healthcare Resource Group code for version 35.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
hrg_v35_desc	HRG version 31 description	varchar	0	0	609	Healthcare Resource Group codes (v3.5)
lsoa_cd_2001	Local Super Output Area code 2001	character	92527	0.48	34261	The Local Super Output Area of
avail_from_dt	Available from date	date	0	0	1	Date when the data made availa
Patient Episode Database for Wales (PEDW) - EPISODE						
prov_unit_cd	Provider unit code	character	0	0	663	This is the organisation code of t identifies the health care provid treatment of the patient.
spell_num_e	Encrypted spell number	integer	<5	0	18555146	A number (alphanumeric) to pro provider spell for a health care
epi_num	Episode number	character	0	0	99	A number used to identify episod each consultant episode in a ho
epi_str_yr	Episode start year	character	573	0	101	The year of the start of a stay, an period of time.
epi_str_dt	Episode start date	date	573	0	10084	This is the start date of a stay, an period of time.
epi_end_yr	Episode end year	character	0	0	29	The year of the end of a stay, an period of time.
epi_end_dt	Episode end date	date	0	0	9477	This is the end of a stay, an episod time.
prov_site_cd	Provider site code	character	0	0	1949	This is the organisation code of t identifies the health care provid treatment of the patient.
age_epi_str_yr	Age at start of episode	integer	3198	0.01	196	The age of patient at start date o
age_epi_str_under1	Age at start of episode for under 1s	integer	21632323	99.3	182	Is the age of patient at start date episode.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
con_spec_main_cd	Consultant main speciality code	character	7507	0.03	162	A unique code identifying each M Colleges. Specialties are division body systems (dermatology), ag medicine), clinical function (rhe combinations of these factors. O Colleges and Faculties should b and Specialist Medical Practice 2003 and European Primary an 1998.
con_spec_cd_of_treat	Consultant treatment speciality code	character	0	0	148	This is the specialty under which either be the same as the speci specialty or a different specialty specialty function. Note that bo specialty function should be bas
epi_dur	Episode duration	decimal	645	0	2879	The period of time in days betwe the End Date of Consultant Epi in days between the Start Date period for unfinished episodes.
diag_cd_123	Diagnosis code (3 digits)	varchar	689333	3.16	2415	The first 3 digits of the Diagnost
diag_cd_4	Diagnosis code (4th digit)	varchar	693337	3.18	13	The 4th digit of the Diagnostic c
diag_cd_1234	Diagnosis code (4 digits)	varchar	693337	3.18	10963	The first 4 digits of the Diagnost
oper_cd_123	Operation code (3 digits)	varchar	9834137	45.14	1449	The first 3 digits of the Procedur
oper_cd_4	Operation code (4th digits)	varchar	9847548	45.21	17	The 4th digit of the Procedure c
oper_cd	Operation code	varchar	9847548	45.21	7869	The full Procedure code availabl

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
curr_prov_unit_cd	Current provider unit code	varchar	0	0	333	The current organisation code of the provider identifies the health care provider for the treatment of the patient. To enable comparison over years, a current field can be used with version/name.
curr_prov_site_cd	Current provider site code	varchar	0	0	1291	The current organisation code of the provider identifies the health care provider for the treatment of the patient. To enable comparison over years, a current field can be used with version/name.
fin_epi_end_yr	Financial episode end year	character	0	0	28	The financial year of the end of a spell.
hsw_first_epi_in_spell	HSW first episode in spell	integer	21784142	100	0	Flag to show if an episode is first in a spell.
site_cd_of_treat	Site code of treatment	character	3348076	15.37	1885	
gmc_con_cd_e	Encrypted GMC consultant code	integer	519033	2.38	36175	Nationally agreed form for consultant registration. General Medical Council (GMC) Registration Number will be used.
ua_cd	Unitary Authority code	character	70916	0.33	196	The unitary authority in which the patient lives.
hrg_localpayment_cd	HRG local payment code	varchar	12752672	58.54	4441	The Healthcare Resource Group code for the local payment.
hrg_localpayment_desc	HRG local payment description	varchar	0	0	6155	The Healthcare Resource Group description for the local payment.
hrg_referencecost_cd	HRG reference cost code	varchar	12702016	58.31	1920	The Healthcare Resource Group code for the reference cost.
hrg_referencecost_desc	HRG reference cost description	varchar	0	0	3083	The Healthcare Resource Group description for the reference cost.
hrg_v31_cd	HRG version 31 code	varchar	6604546	30.32	572	Healthcare Resource Group code for version 31.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
hrg_v31_desc	HRG version 31 description	varchar	0	0	572	Healthcare Resource Group codes (v3.1)
hrg_v35_cd	HRG version 35 code	varchar	6603156	30.31	610	Healthcare Resource Group codes (v3.5)
hrg_v35_desc	HRG version 31 description	varchar	0	0	609	Healthcare Resource Group codes (v3.5)
avail_from_dt	Available from date	date	0	0	1	Date when the data made available
Patient Episode Database for Wales (PEDW) - DIAG						
prov_unit_cd	Provider unit code	character	0	0	642	This is the organisation code of the provider which identifies the health care provider for the treatment of the patient.
spell_num_e	Encrypted spell number	integer	<5	0	18129197	A number (alphanumeric) to provide a unique provider spell for a health care episode.
epi_num	Episode number	character	0	0	99	A number used to identify each consultant episode in a hospital.
diag_num	Diagnosis number	integer	0	0	14	A number used to identify the primary ICD Diagnosis. 1 relates to the primary ICD Diagnosis and 13 relates to the secondary ICD diagnostic codes.
diag_cd_123	Diagnosis code (3 digits)	character	0	0	2668	The first 3 digits of the Diagnostic code.
diag_cd_4	Diagnosis code (4th digit)	character	81534	0.11	33	The 4th digit of the Diagnostic code.
diag_cd_56	Diagnosis code (5th and 6th digits)	character	65323041	91.52	262	The 5th and 6th digit of the Diagnostic code.
diag_cd_1234	Diagnosis code (4 digits)	character	0	0	15420	The first 4 digits of the Diagnostic code.
diag_cd	Diagnosis code	character	0	0	40086	The full Diagnostic code available.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
avail_from_dt	Available from date	date	0	0	1	Date when the data made available
Emergency Department Data Set (EDDS)						
record_id	Record ID	character	8506326	100	0	This is the field to identify the type
prov_unit_cd	Provider unit code	character	0	0	14	This is the organisation code of the provider that identifies the health care provider responsible for the treatment of the patient.
prov_site_cd	Provider site code	character	0	0	23	This is the organisation code of the provider that identifies the health care provider responsible for the treatment of the patient.
admin_arr_dt	Administrative arrival date	date	0	0	3327	Accident and Emergency Attendance. Accident and Emergency reception has arrived and needs to be seen in the Department. Notification could be by themselves, or a person accompanying
admin_arr_tm	Administrative arrival time	time	0	0	1440	Accident and Emergency Attendance. Accident and Emergency reception has arrived and needs to be seen in the Department. Notification could be by themselves, or a person accompanying
alf_e	Encrypted Anonymised Linking Field	bigint	272963	3.21	2537941	The Anonymised Linking Field, within the database, is derived from the patient's unique identifier. The encryption occurs in NWIS and is supplied in the data extract the
nhs_no_ind	NA	character	636636	7.48	30	NA
prac_cd_e	Encrypted GP practice code	integer	189972	2.23	12840	The encrypted GP practice code
dob_year	Year of Date of Birth	character	1168	0.01	128	The year of the Date of birth of patient

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
age	Age at start of admission	smallint	1168	0.01	143	The age of patient at start date of admission.
sex	Sex (also known as gender code)	character	414	0	7	This is the sex (gender) of person.
lsoa_cd	Local Super Output Area code	character	211271	2.48	35994	The Local Super Output Area of residence.
ref_cd_e	NA	bigint	472888	5.56	23123	NA
arrival_mode	Mode of arrival	character	101884	1.2	18	The principal means by which a patient arrives at the Emergency Department.
amb_incid_no_e	Encrypted ambulance incident number	integer	7357724	86.5	1059762	When a patient arrives by ambulance, this is the incident number allocated by the Ambulance Service.
site_cd_of_treat	Site code of treatment	character	0	0	51	The site code of the treatment.
health_event_dt	Health event date	date	1784422	20.98	4266	Date of the incident / acute medical admission to the Emergency Department Attendance Group.
health_event_tm	Health event time	time	2909289	34.2	1440	This is the time of the incident / acute medical admission to the Emergency Department Attendance Group.
attend_group	Attendance group	character	218653	2.57	25	A general reason for an Accident and Emergency Attendance.
attend_category	Attendance Category	character	6751	0.08	6	Accident and Emergency Attendance Category. This is the category of patient is making a first or follow-up attendance to the Emergency Department.
diag_cd_1	Diagnosis Type code 1	character	943809	11.1	756	A broad list of diagnosis types, with the most common being the Emergency Department Attendance Category.
diag_cd_2	Diagnosis Type code 2	character	6706626	78.84	670	A broad list of diagnosis types, with the most common being the Emergency Department Attendance Category.
diag_cd_3	Diagnosis Type code 3	character	6987207	82.14	459	A broad list of diagnosis types, with the most common being the Emergency Department Attendance Category.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
diag_cd_4	Diagnosis Type code 4	character	7022431	82.56	270	The 4th digit of the Diagnostic code
diag_cd_5	Diagnosis Type code 5	character	7027918	82.62	151	A broad list of diagnosis types, with Emergency Department Attending
diag_cd_6	Diagnosis Type code 6	character	7029211	82.64	98	A broad list of diagnosis types, with Emergency Department Attending
anat_area_cd_1	Anatomical area code 1	character	1125400	13.23	39	A list of parts of the human body
anat_area_cd_2	Anatomical area code 2	character	6884335	80.93	40	A list of parts of the human body
anat_area_cd_3	Anatomical area code 3	character	7005530	82.36	39	A list of parts of the human body
anat_area_cd_4	Anatomical area code 4	character	7024604	82.58	39	A list of parts of the human body
anat_area_cd_5	Anatomical area code 5	character	7028269	82.62	36	A list of parts of the human body
anat_area_cd_6	Anatomical area code 6	character	7029132	82.63	34	A list of parts of the human body
side_cd_1	Anatomical side code 1	character	3067296	36.06	9	An indication of the side of the human body
side_cd_2	Anatomical side code 2	character	8149792	95.81	9	An indication of the side of the human body
side_cd_3	Anatomical side code 3	character	8269796	97.22	10	An indication of the side of the human body
side_cd_4	Anatomical side code 4	character	8290114	97.46	9	An indication of the side of the human body
side_cd_5	Anatomical side code 5	character	8294677	97.51	8	An indication of the side of the human body
side_cd_6	Anatomical side code 6	character	8295877	97.53	6	An indication of the side of the human body
treat_cd_1	Treatment code 1	character	1383221	16.26	17	A broad list of types of treatment received by patient as a result of an Accident

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
treat_cd_2	Treatment code 2	character	6609855	77.71	17	A broad list of types of treatment patient as a result of an Accident
treat_cd_3	Treatment code 3	character	7008801	82.4	17	A broad list of types of treatment patient as a result of an Accident
treat_cd_4	Treatment code 4	character	7164431	84.22	17	A broad list of types of treatment patient as a result of an Accident
treat_cd_5	Treatment code 5	character	7222299	84.91	17	A broad list of types of treatment patient as a result of an Accident
treat_cd_6	Treatment code 6	character	7238663	85.1	17	A broad list of types of treatment patient as a result of an Accident
invest_cd_1	Investigation code 1	character	1561708	18.36	14	A broad list of types of investigation diagnosis during and Accident
invest_cd_2	Investigation code 2	character	6873257	80.8	11	A broad list of types of investigation diagnosis during and Accident
invest_cd_3	Investigation code 3	character	7161646	84.19	11	A broad list of types of investigation diagnosis during and Accident
invest_cd_4	Investigation code 4	character	7373525	86.68	11	A broad list of types of investigation diagnosis during and Accident
invest_cd_5	Investigation code 5	character	7501448	88.19	11	A broad list of types of investigation diagnosis during and Accident
invest_cd_6	Investigation code 6	character	7518976	88.39	10	A broad list of types of investigation diagnosis during and Accident
admin_end_dt	Administrative end date	date	61624	0.72	3329	This is the date that the patients
admin_end_tm	Administrative end time	time	82466	0.97	1440	This is the time that the patients
add_details	Additional incident details	varchar	8506326	100	0	A record of any additional details Emergency Department Attend the dataset.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
discharge	Outcome of attendance	character	161927	1.9	27	This records the outcome of the Attendance.
location_type	Location place type	character	338871	3.98	69	The type of place where the phys the Accident and Emergency D
road_user	Road user	character	767921	9.03	15	This is the nature of the patients
presenting_complaint	Presenting complaint	varchar	8506326	100	0	This is the presenting complaint and Emergency Department.
mech_of_inj	Mechanism of injury	character	564890	6.64	180	The mechanics of how the physio
activity	Activity at time of injury	character	666342	7.83	18	What the patient was doing at th picture of how the physical or c
sport	Sporting activity	character	1088994	12.8	37	The sport in which the patient w the Accident and Emergency D
crn_pseud_e	Encrypted Pseudonimised Case Record Number	integer	0	0	3221753	This is the case record number. health care provider.
alf_sts_cd	ALF status code	character	0	0	5	Status code assigned when deriv
alf_mtch_pct	ALF match percentage	decimal	8399224	98.74	27110	Match percentage assigned whe Field.
alcohol_ind	Alcohol indicator	character	405798	4.77	29	In the clinical opinion of the Em alcohol in the presenting patien
avail_from_dt	Available from date	date	0	0	1	Date when the data made availa
batch_num	Batch number	smallint	0	0	1	The batch number is a sequentia SAIL.
triage_cat	Triage category	character	371878	4.37	16	The triage category is assigned t by medical or nursing staff in a

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
clinical_ref_dt	Clinical referral date	date	9377699	17.93	9864	The Clinical Referral Date (CRD) is the start of a period of waiting either for an episode of treatment such as elective surgery or for a consultant appointment. It is used for booking patients, waiting lists used for booking patients, and other circumstances. It is not used to calculate waiting list dates.
waiting_list_dt	Waiting list date	date	24015264	45.92	8635	The Waiting List Date is set initially as the date of the first appointment. It is used to calculate waiting list dates. It is used to calculate waiting list dates for Health Board/Trust performance targets. It is not used to calculate waiting list dates for booking or to order inpatient or outpatient services or for surgery. There are a number of factors that can lead to a patient's reinstatement to a waiting list for a different consultant. These include rescheduling an appointment, a patient's decision to be reinstated to a waiting list for a different consultant, or a patient's decision to remain with a consultant for a different consultant.
priority_type_cd	Priority type new patients	character	31340596	59.92	5	This is the priority of a request for a consultant appointment, that is, where the patient is to be provided by a Consultant. For a Follow Up Attendance, Priority Type must be set to 'Follow Up'.
source_of_ref_cd	Source of referral outpatients code	character	326558	0.62	35	This is a classification which is used to identify the source of an Outpatient Episode or Outpatient Appointment.
con_spec_main_cd	Consultant main speciality code	character	41308	0.08	163	A unique code identifying each Medical Speciality. Specialities are divisional codes for body systems (dermatology), age groups (paediatrics), clinical function (rheumatology), and combinations of these factors. Codes are used by Medical Colleges and Faculties should be used to identify the Medical College and Specialist Medical Practice. Codes are used in the 2003 and European Primary Care Survey 1998.
con_spec_cd_of_treat	Consultant treatment speciality code	character	19168	0.04	186	This is the specialty under which a patient is treated. It can either be the same as the specialty code or a different specialty code. Note that both specialty and specialty function should be based on the specialty function.

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
local_spec_cd	Local sub specialty	character	8737323	16.71	358	This is a locally or nationally defined specialty boundaries.
clinic_purpose_cd	Clinic Purpose code	character	7355159	14.06	46859	This is the function of an outpatient national classification of functional function titles must be decided shown, such as: a. Obstetric and
gmc_con_cd_e	Encrypted GMC consultant code	integer	225378	0.43	24479	Nationally agreed form for consultant General Medical Council (GMC) Consultant or locum Consultant Registration Number will be used
att_id_e	Encrypted attendance identifier	integer	0	0	43721493	A sequential number or time of day appointment to be uniquely identified
admin_cat_cd	Administrative category code	character	1859	0	17	This is to indicate whether the patient etc.
loc_type_cd	Location type code	character	1464101	2.8	26	This is a classification of location
site_cd_of_treat	Site code of treatment	character	21452	0.04	1928	
med_staff_type_cd	Medical staff type seeing patient code	character	1531441	2.93	11	A classification of the type of medical Outpatient attendance.
attend_dt	Attendance date	date	0	0	5230	This is the date of an attendance
first_attend_cd	First Attendance Category	character	8262	0.02	5	The first attendance is the start of attendance in a series with the following a referral.
attend_cd	Attendance code	character	61143	0.12	13	This indicates whether a person patient did not attend, it also includes given.
outcome_cd	Outcome of attendance	character	3388267	6.48	14	This records the outcome of the

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
last_dna_cancel_dt	Last DNA or patient cancelled date	date	48882640	93.46	7572	This derived item should only be used for treatment
oper_sts_flg	Operation status flag	character	1704072	3.26	11	Operation status should be used for knowledge regarding the operation
oper_cd_123	Operation code (3 digits)	character	49048246	93.78	943	The first 3 digits of the Procedure code
oper_cd_4	Operation code (4th digit)	character	49150721	93.97	34	The 4th digit of the Procedure code
lsoa_cd	Local Super Output Area code	character	231313	0.44	25100	The Local Super Output Area of the patient
alf_sts_cd	ALF status code	character	0	0	5	Status code assigned when derived from ALF
alf_mtch_pct	ALF match percentage	decimal	52135801	99.68	33811	Match percentage assigned when derived from ALF Field.
avail_from_dt	Available from date	date	0	0	1	Date when the data made available
Annual District Death Extract (ADDE)						
alf_e		bigint	35345	5.12	655547	
alf_sts_cd		character	0	0	2	
death_annual-record_ind_cd		character	37857	5.48	1	
count_death		integer	0	0	1	
death_dt		date	14	0	7949	
death_dt_valid		varchar	0	0	2	

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
death_reg_dt		timestamp	0	0	5706	
death_reg_dt_valid		varchar	0	0	1	
neonatal_ind_flg		integer	0	0	3	
dec_urbanrural_cd		character	<5	0	8	
dec_stats_curr_census_lsoa_cd		varchar	18	0	1909	
dec_stats_curr_census_la_cd		varchar	18	0	22	
dec_stats_curr_census_la_previous_cd		varchar	18	0	22	
dec_stats_curr_census_country_cd		varchar	18	0	1	
dec_stats_curr_census_health_org_cd		varchar	18	0	7	
dec_stats_curr_census_health_org_previous_cd		varchar	18	0	7	
dec_stats_prev_census_lsoa_cd		varchar	76	0.01	1896	
dec_stats_prev_census_la_cd		varchar	76	0.01	22	
dec_stats_prev_census_la_prev_cd		varchar	76	0.01	7	
dec_stats_prev_census_country_cd		varchar	76	0.01	1	

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
dec_stats_prev_census_health_org_cd		varchar	76	0.01	22	
dec_stats_prev_census_health_org_prev_cd		character	76	0.01	7	
dec_sha_cd		varchar	0	0	1	
dec_health_org_cd		varchar	0	0	7	
dec_sex_cd		character	0	0	2	
deathcause_diag_underlying_cd		varchar	2160	0.31	4205	
deathcause_diag_1_cd		varchar	<5	0	2997	
deathcause_diag_2_cd		varchar	167660	24.27	4714	
deathcause_diag_3_cd		varchar	416684	60.31	4292	
deathcause_diag_4_cd		varchar	567539	82.15	3623	
deathcause_diag_5_cd		varchar	640374	92.69	2683	
deathcause_diag_6_cd		varchar	671192	97.15	1879	
deathcause_diag_7_cd		varchar	683353	98.91	1245	
deathcause_diag_8_cd		varchar	687856	99.56	750	
death_communal_establishment_cd_e		bigint	0	0	2663	
death_urbanrural_cd		character	145133	21.01	8	

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
deathstats_curr_census_lsoa_cd		varchar	145135	21.01	4215	
deathstats_curr_census_la_cd		varchar	145135	21.01	340	
deathstats_curr_census_la_prev_cd		varchar	145270	21.03	333	
deathstats_curr_census_country_cd		varchar	145135	21.01	2	
deathstats_curr_census_health_org_cd		varchar	145135	21.01	157	
deathstats_curr_census_health_org_prev_cd		character	165362	23.93	7	
deathstats_prev_census_lsoa_cd		varchar	145137	21.01	4198	
deathstats_prev_census_la_cd		varchar	145137	21.01	340	
deathstats_prev_census_la_prev_cd		varchar	145137	21.01	157	
deathstats_prev_census_country_cd		varchar	145137	21.01	2	
deathstats_prev_census_health_org_cd		varchar	145137	21.01	340	
deathstats_prev_census_health_org_prev_cd		varchar	165362	23.93	7	
death_ccg_cd		varchar	670674	97.07	209	
death_countyanddistrict_cd		varchar	145156	21.01	172	

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
death_health_org_cd		varchar	145156	21.01	157	
death_nhs_establishment_ind_cd		varchar	160663	23.25	2	
death_establishment_type_cd		varchar	160663	23.25	49	
death_postcode_imputation_ind_cd		varchar	690891	100	1	
dec_age		varchar	0	0	204	
dec_age_unit_cd		varchar	0	0	4	
dec_occ_type_cd		varchar	6	0	5	
dec_or_mother_socioeconomic_class_cd		varchar	447238	64.73	49	
dec_husband_or_father_socioeconomic_class_cd		varchar	578935	83.8	48	
dec_or_mother_occ_class_cd		varchar	447220	64.73	944	
dec_husband_or_father_occ_class_cd		varchar	578930	83.79	904	
dec_or_mother_retired_ind_cd		varchar	618493	89.52	1	
dec_husband_or_father_retired_ind_cd		varchar	663323	96.01	1	
dec_birthcountry_cd		varchar	21	0	330	

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
dec_governmentoffice_region_cd		varchar	0	0	1	
dec_countyanddistrict_cd		varchar	0	0	22	
dec_county_cd		varchar	0	0	2	
dec_countydistrict_cd		varchar	0	0	22	
dec_ccg_cd		varchar	690892	100	0	
dec_lsoa_cd		varchar	0	0	1909	
dec_ward_cd		varchar	0	0	148	
deathcause_diag_sec_cause_cd		varchar	665526	96.33	658	
deathcause_rowpos_1_cd		varchar	95743	13.86	8	
deathcause_rowpos_2_cd		varchar	253788	36.73	10	
deathcause_rowpos_3_cd		varchar	479168	69.35	10	
deathcause_rowpos_4_cd		varchar	609554	88.23	10	
deathcause_rowpos_5_cd		varchar	663883	96.09	8	
deathcause_rowpos_6_cd		varchar	682433	98.78	9	

Table C.1.1: SAIL Databank Datasets used in the Observatory (cont'd).

Column Name	Friendly Name	Type	NULL Count	NULL %	Distinct Values	Description
deathcause_row- pos_7_cd		varchar	688336	99.63	9	
deathcause_row- pos_8_cd		varchar	690079	99.88	7	
avail_from_dt		date	0	0	1	

Table C.1.2: Frequency of events and number of patients in calendar year for each of the SAIL datasets used in the Wales Asthma Observatory.

Calendar Year	GP events		Hospital spells		Hospital episodes		A&E visits		Outpatient visits	
	events	unique patients	events	unique patients	events	unique patients	events	unique patients	events	unique patients
1990	7,260,835	1,023,620	27	26	55					
1991	11,039,966	1,213,249	2,531	1,984	2,765					
1992	14,975,143	1,387,088	3,639	2,638	4,122					
1993	21,219,312	1,622,826	4,724	3,590	5,257					
1994	24,780,191	1,761,958	4,514	3,442	5,346					
1995	29,693,724	1,816,797	194,590	135,965	542,548					
1996	33,553,308	1,903,954	330,591	214,305	733,582					
1997	36,941,292	1,948,478	448,121	268,530	759,288					
1998	40,879,926	1,979,208	696,920	423,332	766,851					
1999	45,368,448	2,036,630	825,021	478,138	893,548					
2000	52,882,234	2,186,095	882,144	500,038	958,027					
2001	60,298,298	2,174,307	891,733	502,539	971,566					<5
2002	70,072,481	2,224,281	848,668	468,889	927,433					508
2003	87,285,461	2,311,618	847,108	459,863	934,746					199,417
2004	109,398,271	2,412,617	853,757	464,346	951,083					2,509,030
2005	120,102,299	2,431,003	845,085	464,324	952,725					3,495,586
2006	128,606,434	2,424,984	874,025	472,479	986,353					3,643,120
2007	136,172,653	2,429,926	897,308	482,080	1,009,833					3,740,792
2008	140,725,972	2,404,258	923,730	491,836	1,039,673		31	26		4,014,820
2009	147,750,672	2,442,412	959,062	505,742	1,103,224		502,848	370,390		4,238,246
2010	150,579,349	2,432,561	932,304	484,631	1,085,917		790,630	545,100		4,074,717
2011	155,580,366	2,423,954	956,771	492,803	1,120,353		875,508	588,193		4,250,334
2012	161,045,727	2,426,940	964,655	498,914	1,140,630		986,641	645,852		4,225,830
2013	162,633,799	2,458,714	968,383	491,903	1,144,771		981,242	641,478		4,131,658
2014	160,806,382	2,466,769	987,251	494,742	1,164,864		974,479	634,024		4,168,053
2015	157,859,221	2,423,260	1,002,137	500,288	1,180,826		967,562	631,174		4,303,931
2016	162,516,240	2,406,713	1,027,955	511,280	1,213,104		983,261	638,625		4,492,596
2017	164,649,368	2,379,126	1,027,525	511,917	1,223,682		1,001,592	646,859		4,429,663
2018	86,638,008	2,051,394	684,796	374,779	811,446		601,854	429,276		2,619,541

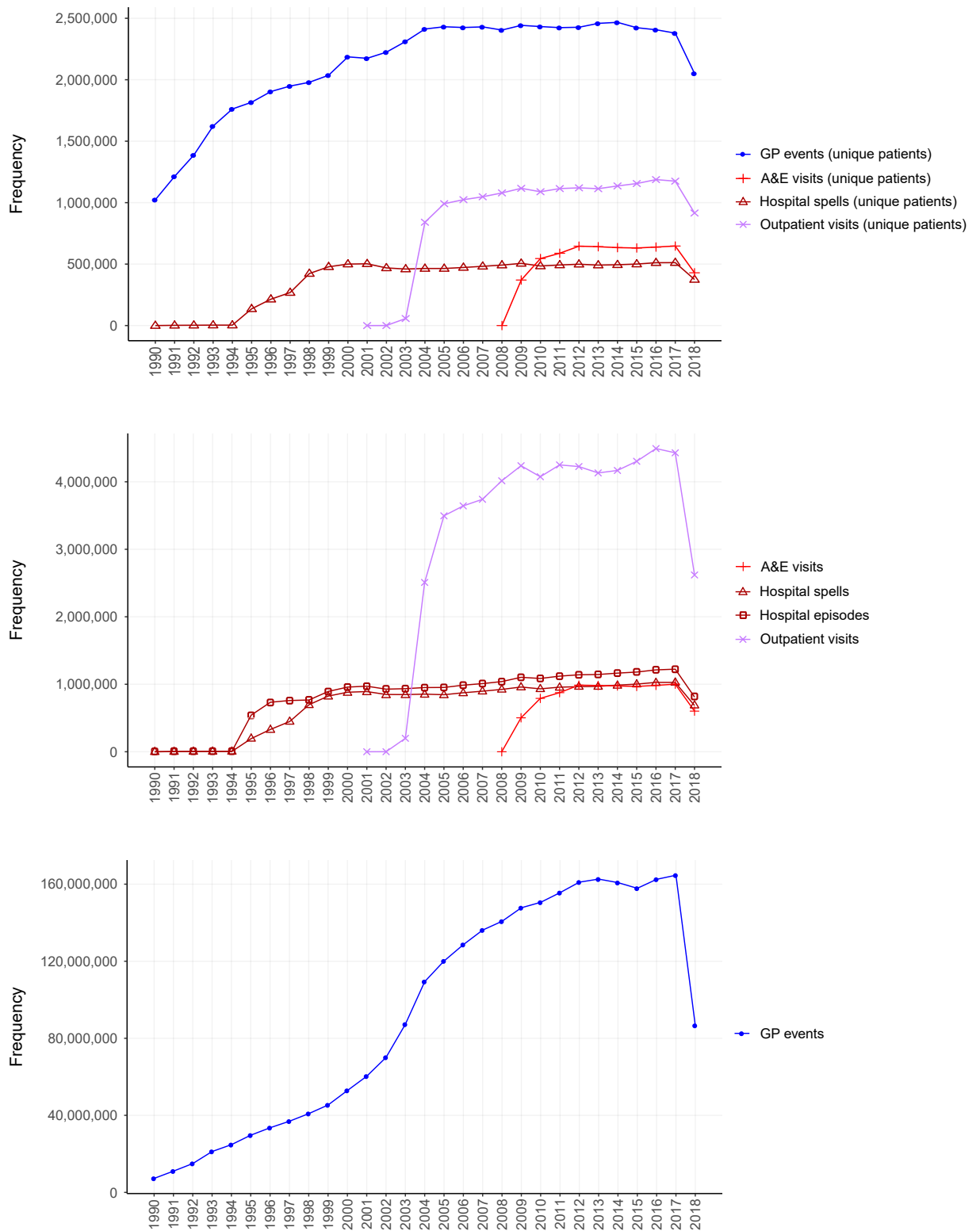


Figure C.1.1: Frequency of events and number of patients in calendar years for each of the SAIL datasets used in the Wales Asthma Observatory.

C.2 SAIL IGRP approval letter



10 April 2015

Dr Gwyneth Davies

College of Medicine
ILS1
Swansea University
Singleton Park,
Swansea, SA2 8PP

Dear Gwyneth

Re: Wales Asthma Observatory

Your proposal to use the SAIL databank has been assessed by the SAIL Collaboration Review System (CRS). The CRS consists of the SAIL Management Team and the Information Governance Review Panel (IGRP). The membership of the IGRP is comprised of senior representatives from:

- British Medical Association (BMA)
- National Research Ethics Service (NRES)
- Public Health Wales
- NHS Wales Informatics Service (NWIS)
- Involving People

After careful consideration the proposal has been given **approval** to commence with analysis.

The project has been given a SAIL project number of 0317.

Creation of project specific data view

Work will now commence on the creation of the project specific data view. The analyst working on this will be Mohammad Al Sallakh and they will be in contact with you to confirm your data specification.

Publication statement

All publications must acknowledge the use of SAIL data.

Yours sincerely

Cynthia McNerney

Information Governance Coordinator

SAIL DATABANK

Institute of Life Science 2
College of Medicine
Swansea University
Singleton Park
Swansea
SA2 8PP

SAILDatabank@swansea.ac.uk

www.SAILDatabank.com

Figure C.2.1: Approval letter by the SAIL Databank IGRP panel for using SAIL data in the development of the Wales Asthma Observatory.

C.3 A tool for automatic characterisation of cohorts using primary care data.

The following abstract has been presented at the Informatics for Health 2017 congress, 24-26 April 2017 in Manchester, UK [476, page 164].

Abstract no. 594 A tool to improve the efficiency and reproducibility of research using electronic health record databases

Mohammad Al Sallakh and Gwyneth Davies, Swansea University Medical School, Swansea

Sarah Rodgers, Farr Institute, CIPHER, Swansea

Ronan Lyons, Farr Institute, CIPHER, Swansea

Aziz Sheikh, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh

Introduction: Interrogation of routine electronic health record (EHR) databases often involves repetitive programming tasks, such as manually constructing and modifying complex database queries, requiring significant time from an experienced data analyst. The objective was to develop a tool to automate the selection and characterisation of cohorts from primary care databases to be used by data analysts and researchers.

Methods: We identified a set of common elementary approaches to query clinical variables from the primary care database of the Secure Anonymised Information Linkage databank. We then designed an easy-to-use web-based user interface to allow using combinations of these approaches as 'building blocks' for querying more complex variables. We created an R programme to automatically generate and execute the corresponding Structured Query Language (SQL) queries.

Results: The developed prototype allows researchers to query clinical information from primary care databases based on the following elementary variable types: (1) count of events of interest (e.g. asthma prescriptions) or their distinct dates (2) the code or date of the earliest or latest event of interest (e.g. type of the earliest smoking cessation prescription) (3) the code or date of the event of maximum or minimum value (e.g., maximum BMI recording ever) and (4) count of events of interest having complex temporal constraints with other events (e.g.,

count of asthma doctor visits with oral steroid prescriptions within one week). Researchers may choose fixed, dynamic, or individualised query intervals. Algorithms are saved on a web server as versioned and shareable objects. The prototype integrates with a Read Codes dictionary and a shareable codeset repository allowing researchers to keep a record of codes used for reporting transparency.

Discussion: The developed prototype provides a scalable, versatile solution for the implementation of complex cohort selection and characterisation algorithms using primary care databases. The automatic generation of SQL queries reduces human errors and should enable rapid and scalable implementation of these algorithms, which has the potential to improve research efficiency and reproducibility. In addition, the graphical user interface allows researchers with no programming skills to interrogate the data. The tool is under active development to improve the functionality and usability, and we look forward to testing it in other databases and assessing its suitability in different research contexts. We plan to make this tool available under an open source licence.

GP-ACT A tool for automatic characterisation of cohorts using primary care data

Mohammad Al Sallakh^{1*} Sarah Rodgers¹ Ronan Lyons¹ Aziz Sheikh² Gwyneth Davies¹
 * 594803@swansea.ac.uk 01792 60 2349



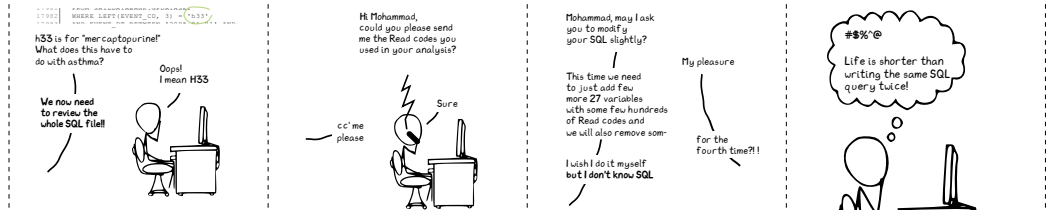
Interrogation of EHR databases for research is often an unstreamlined, poorly documented process that hinders transparency and reproducibility.

Manual writing of programming scripts increases the risk of unnoticed human errors.

Inefficient sharing/reuse of programming scripts and clinical codes, wastes time and effort.

Non-technical members of the research team cannot directly interrogate the data.

Analysts spend significant time on repetitive work, which could be otherwise automated.

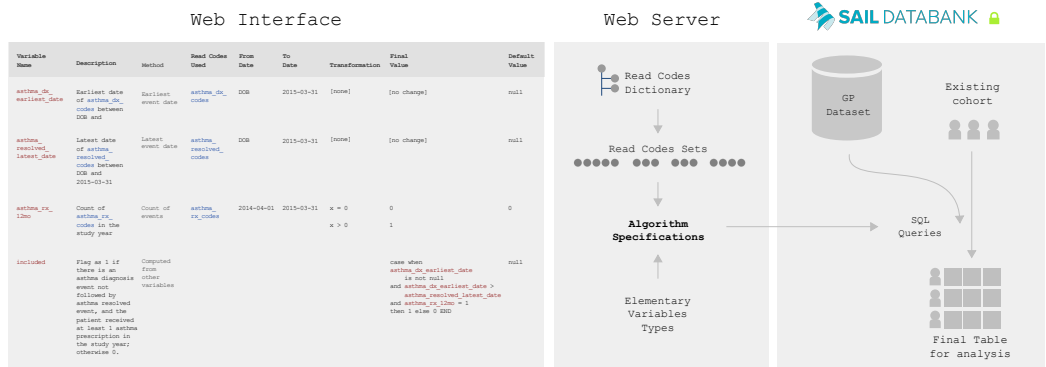


-- So, what is GP-ACT? --

It is a platform that allows researchers and analysts to query complex clinical information from primary care databases based on common elementary variable types. Users can maintain and share clinical codes and phenotyping algorithms as versioned and cite-able objects on a web server, and can run automatically-generated SQL queries against the database.

Currently supported variable types:

- (1) Count of events of interest or their distinct dates
- (2) Code or date of the earliest or latest event of interest
- (3) Code or date of the event of maximum or minimum value
- (4) Count of events of interest having temporal constraints with other events



-- How does it help? --

GP-ACT provides a scalable, versatile solution for implementing cohort characterisation algorithms.

Automatic SQL queries generation -> less human errors, rapid implementation, improved research efficiency, and direct involvement of non-technical researchers

Sharable code sets and algorithms -> better research transparency and reproducibility



Figure C.3.1: A poster presented at the Informatics for Health 2017 congress in Manchester, United Kingdom.

C.4 Read codes used in the assessment of data quality in Chapter 4

Table C.4.1: Asthma diagnosis Read codes.

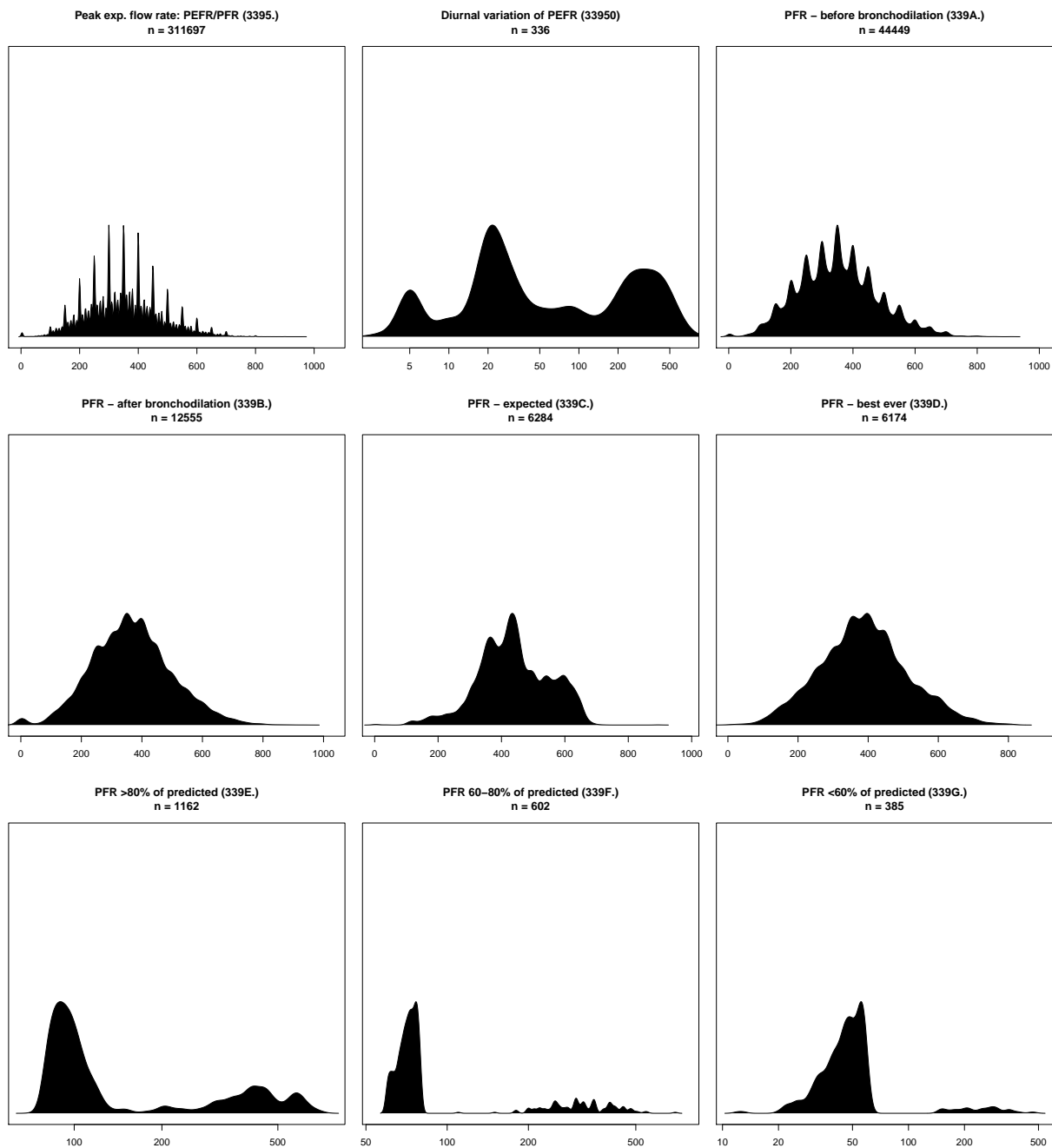
Read code	Description
<i>Asthma diagnosis codes</i>	
173A.	Exercise induced asthma
H3120	Chronic asthmatic bronchitis
H33..	Asthma
H330.	Extrinsic (atopic) asthma
H3300	Extrinsic asthma without status asthmaticus
H3301	Extrinsic asthma with status asthmaticus
H330z	Extrinsic asthma NOS
H331.	Intrinsic asthma
H3310	Intrinsic asthma without status asthmaticus
H3311	Intrinsic asthma with status asthmaticus
H331z	Intrinsic asthma NOS
H332.	Mixed asthma
H333.	Acute exacerbation of 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
<i>Asthma resolved codes</i>	
21262	Asthma resolved
212G.	Asthma resolved

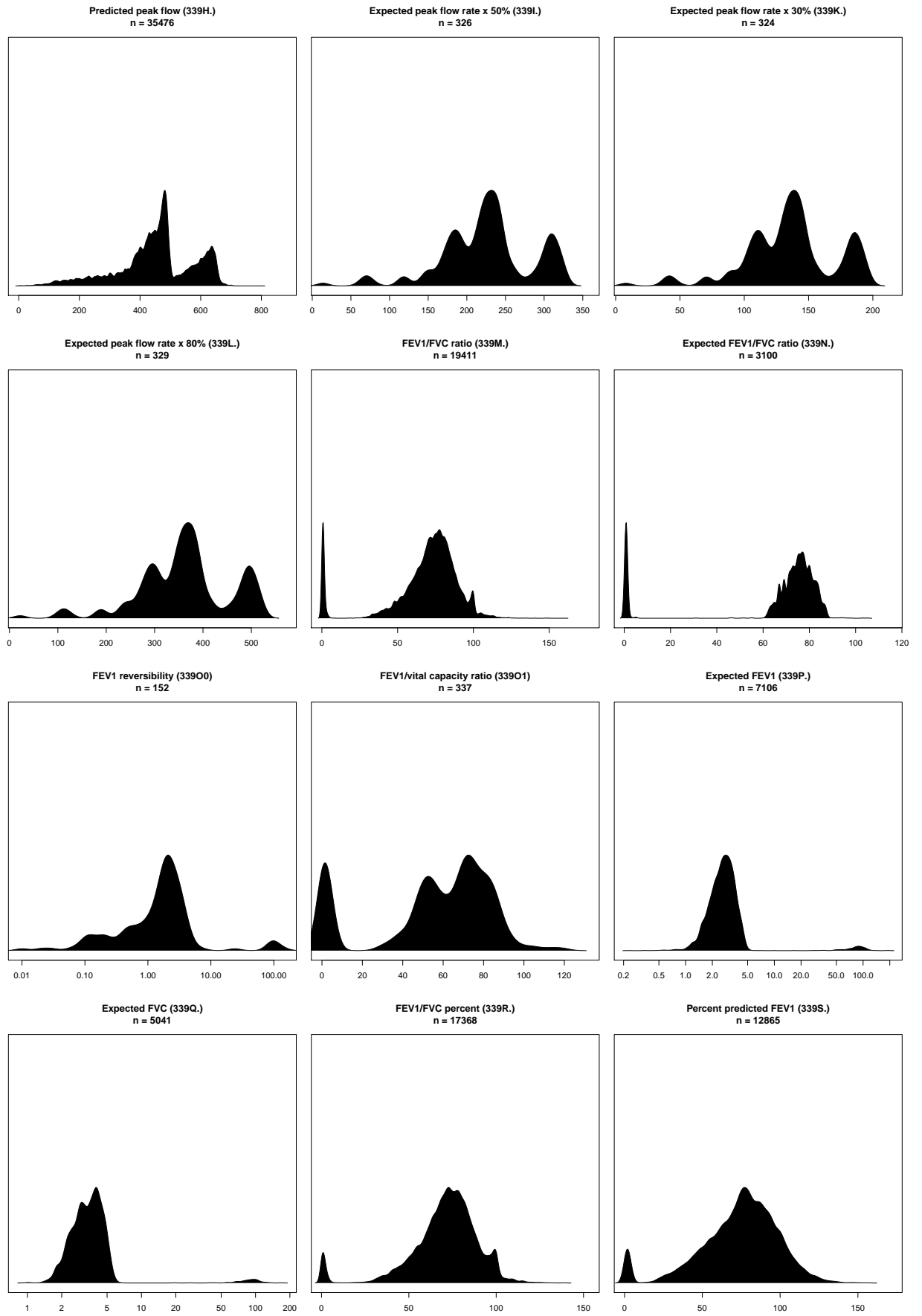
Table C.4.3: Asthma-related event groups chosen for the coding quality analysis.

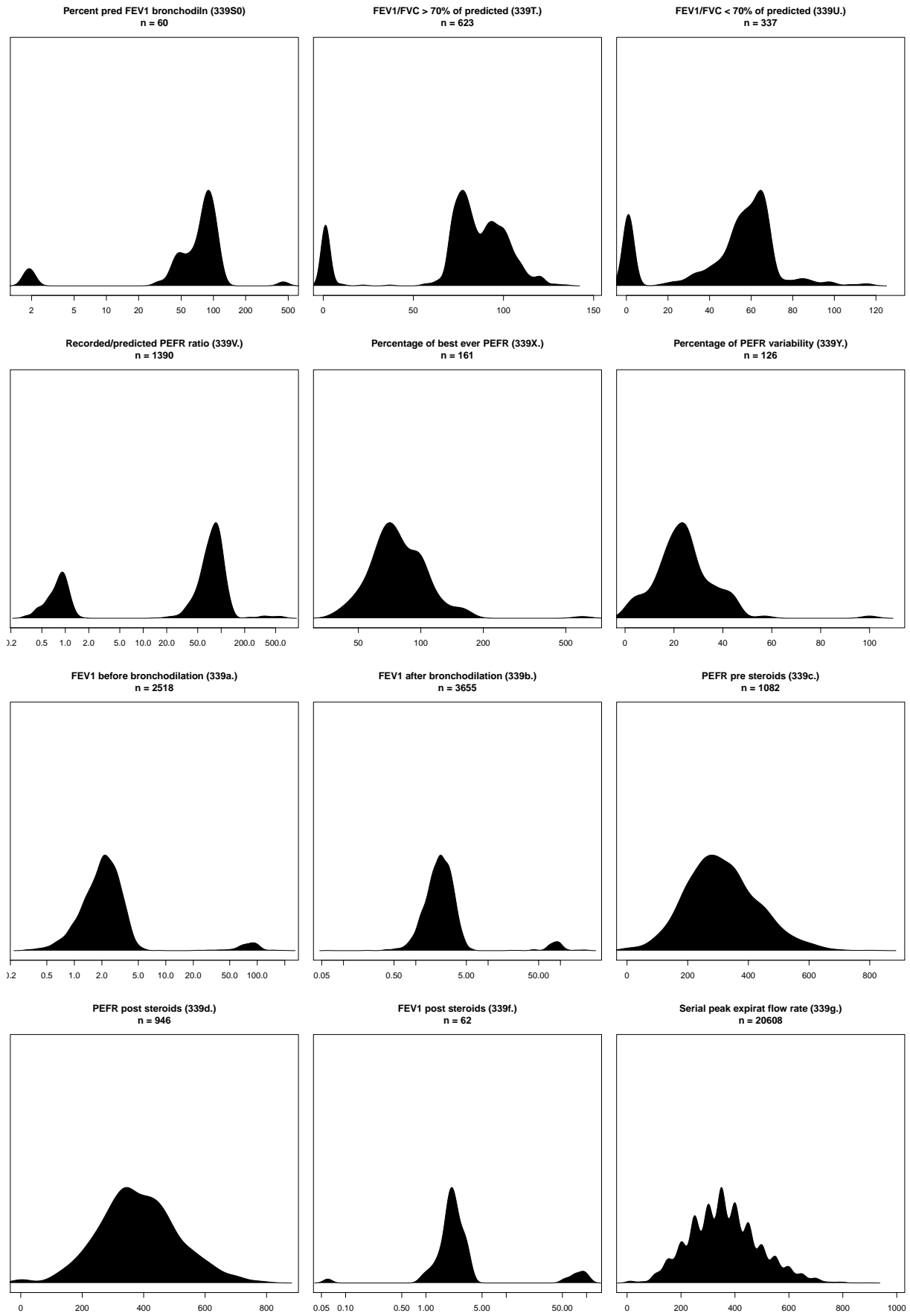
Read code	Description
Asthma triggers	
178..	Asthma trigger
1780.	Aspirin induced asthma
1781.	Asthma trigger - pollen
1782.	Asthma trigger - tobacco smoke
1783.	Asthma trigger - warm air
1784.	Asthma trigger - emotion
1785.	Asthma trigger - damp
1786.	Asthma trigger - animals
1787.	Asthma trigger - seasonal
1788.	Asthma trigger - cold air
1789.	Asthma trigger - respiratory infections
178A.	Asthma trigger - airborne dust
178B.	Asthma trigger - exercise
Asthma severity	
663V1	Mild asthma
663V2	Moderator asthma
663V3	Severe asthma
Asthma control steps	
8793.	Asthma control step 0
8793.	Asthma control step 1
8793.	Asthma control step 2
8793.	Asthma control step 3
8793.	Asthma control step 4
8793.	Asthma control step 5
Spirometry	
33G1.	Spirometry reversibility positive
33H1.	Positive reversibility test to salbutamol
33I1.	Positive reversibility test to ipratropium bromide
33J1.	Positive reversibility test to a combination of salbutamol and ipratropium bromide
33K1.	Positive reversibility test to corticosteroids
663J.	Airways obstruction reversible
745D4	Post bronchodilator spirometry
8HRC.	Referral for spirometry
Serum eosinophil count	
42K..	Eosinophil count
42K1.	Eosinophil count normal
42K3.	Eosinophil count raised
42KZ.	Eosinophil count NOS (not otherwise specified)
42b9.	Percentage eosinophil
Serum total IgE	
43J7.	IgE
43Jw.	Total immunoglobulin E level

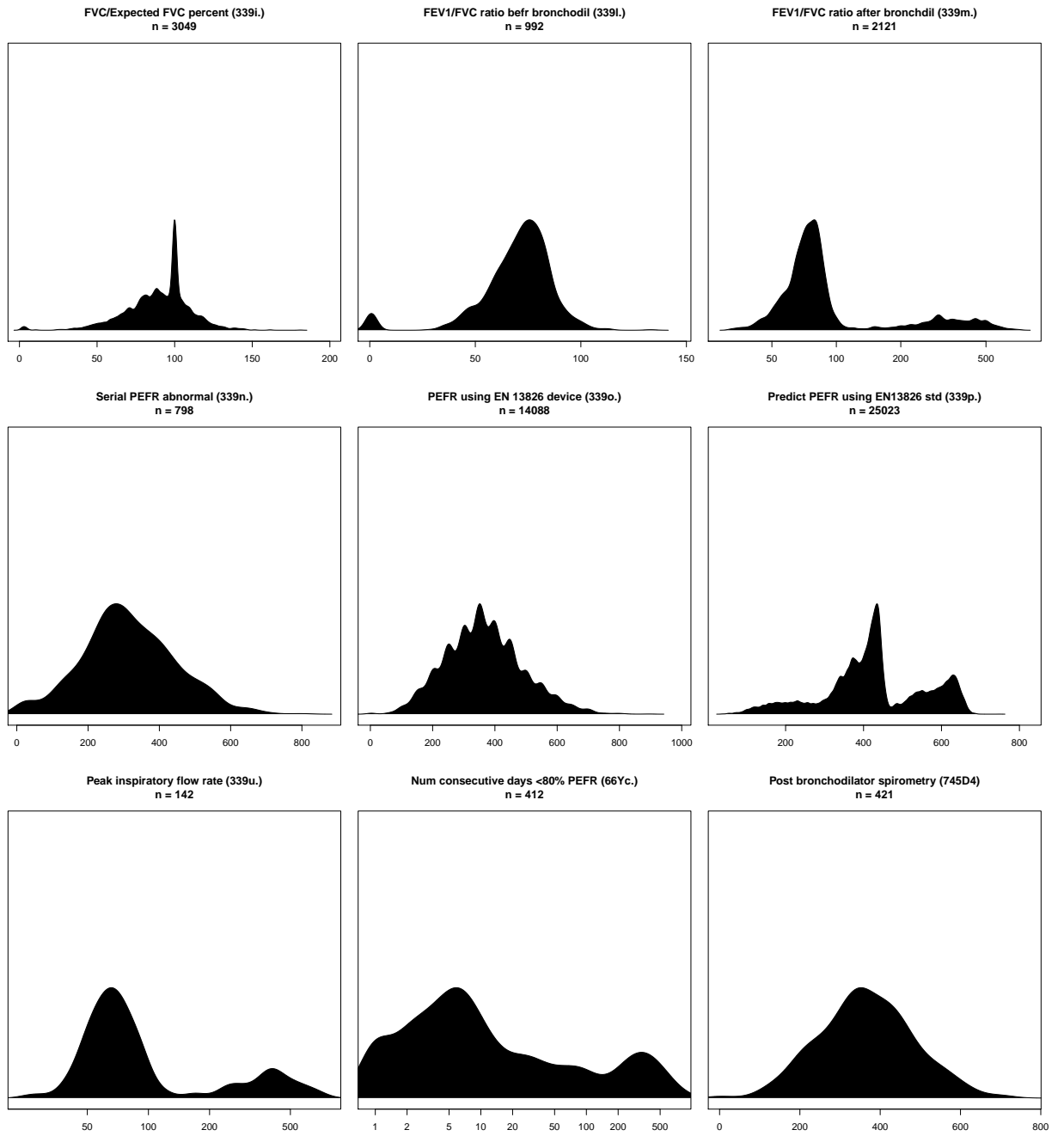
C.5 Density distributions of lung function test values

Figure C.5.1: Beanplots showing density distributions for lung function event values.









Appendix D

Chapter 5 Appendix

D.1 Meeting abstract

I presented the following abstract about the findings in [Chapter 5](#) at The Lancet's Public Health Science Conference 2017 in London and the European Respiratory Society International Congress 2017 in Milan.

Socioeconomic deprivation and inequalities in asthma care in Wales

Mohammad A Al Sallakh, Sarah E Rodgers, Ronan A Lyons, Aziz Sheikh, Gwyneth A Davies

Abstract

Background Area-based deprivation indices are widely used to study health inequalities. We explored whether inequality exists for asthma care across socioeconomic deprivation levels in Wales.

[Redacted text block containing the main body of the abstract]

Appendix E

Clinical codes

These codes were used in asthma case definitions and outcome variables in [Chapter 4](#) and [Chapter 5](#).

Code	Description
Asthma GP Visits (Read codes)	
173c.	Occupational asthma
173d.	Work aggravated asthma
178..	Asthma trigger
1780.	Aspirin induced asthma
1781.	Asthma trigger - pollen
1782.	Asthma trigger - tobacco smoke
1783.	Asthma trigger - warm air
1784.	Asthma trigger - emotion
1785.	Asthma trigger - damp
1786.	Asthma trigger - animals
1787.	Asthma trigger - seasonal
1788.	Asthma trigger - cold air
1789.	Asthma trigger - respiratory infection
178A.	Asthma trigger - airborne dust
178B.	Asthma trigger - exercise
102..	Asthma confirmed
388t.	Royal College of Physicians asthma assessment
38DL.	Asthma control test
38DV.	Mini asthma quality of life questionnaire
38QM.	Childhood Asthma Control Test
661M1	Asthma self-management plan agreed
661N1	Asthma self-management plan review
663d.	Emergency asthma admission since last appointment
663e.	Asthma restricts exercise
663e0	Asthma sometimes restricts exercise
663e1	Asthma severely restricts exercise
663f.	Asthma never restricts exercise
663h.	Asthma - currently dormant
663j.	Asthma - currently active
663m.	Asthma accident and emergency attendance since last visit
663n.	Asthma treatment compliance satisfactory
663N.	Asthma disturbing sleep
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 treatment compliance unsatisfactory
663P.	Asthma limiting activities
663P0	Asthma limits activities 1 to 2 times per month
663P1	Asthma limits activities 1 to 2 times per week
663P2	Asthma limits activities most days

Table E.1: (cont'd).

Code	Description
663q.	Asthma daytime symptoms
663Q.	Asthma not limiting activities
663r.	Asthma causes night symptoms 1 to 2 times per month
663s.	Asthma never causes daytime symptoms
663t.	Asthma causes daytime symptoms 1 to 2 times per month
663u.	Asthma causes daytime symptoms 1 to 2 times per week
663U.	Asthma management plan given
663v.	Asthma causes daytime symptoms most days
663V.	Asthma severity
663V0	Occasional asthma
663V1	Mild asthma
663V2	Moderate asthma
663V3	Severe asthma
663w.	Asthma limits walking up hills or stairs
663W.	Asthma prophylactic medication used
663x.	Asthma limits walking on the flat
663y.	Number of asthma exacerbations in past year
66Y5.	Change in asthma management plan
66Y9.	Step up change in asthma management plan
66YA.	Step down change in asthma management plan
66YC.	Absent from work or school due to asthma
66YE.	Asthma monitoring due
66YJ.	Asthma annual review
66YK.	Asthma follow-up
66Yp.	Asthma review using Royal College of Physicians three questions
66YP.	Asthma night-time symptoms
66YQ.	Asthma monitoring by nurse
66Yq..	Asthma causes night time symptoms 1 to 2 times per week
66Yr.	Asthma causes symptoms most nights
66YR.	Asthma monitoring by doctor
66Ys.	Asthma never causes night symptoms
66Yu.	Number of days absent from school due to asthma in past 6 months
66YZ.	Does not have asthma management plan
679J.	Health education - asthma
679J0	Health education - asthma self management
679J1	Health education - structured asthma discussion
679J2	Health education - structured patient focused asthma discussion
8791.	Further asthma - drug prevent.
8793.	Asthma control step 0
8794.	Asthma control step 1
8795.	Asthma control step 2
8796.	Asthma control step 3
8797.	Asthma control step 4
8798.	Asthma control step 5
8B3j.	Asthma medication review
8CMA0	Patient has a written asthma personal action plan
8CR0.	Asthma clinical management plan
8H2P.	Emergency admission, asthma
8HTT.	Referral to asthma clinic
9hA..	Exception reporting: asthma quality indicators
9hA1.	Excepted from asthma quality indicators: Patient unsuitable
9N1d.	Seen in asthma clinic
9N1d0	Seen in school asthma clinic
9NI8.	Asthma outreach clinic
9NNX.	Under care of asthma specialist nurse
90J..	Asthma monitoring admin.
90J1.	Attends asthma monitoring
90J2.	Refuses asthma monitoring
90J3.	Asthma monitor offer default
90J4.	Asthma monitor 1st letter
90J5.	Asthma monitor 2nd letter
90J6.	Asthma monitor 3rd letter
90J7.	Asthma monitor verbal invite
90J8.	Asthma monitor phone invite
90J9.	Asthma monitoring deleted
90JA.	Asthma monitoring check done
90JB.	Asthma monitoring invitation SMS (short message service) text message
90JC.	Asthma monitoring invitation email
90JZ.	Asthma monitoring admin.NOS

Table E.1: (cont'd).

Code	Description
9Q21.	Patient in asthma study
SLF7.	Antiasthmatic poisoning
SLF7z	Antiasthmatic poisoning NOS
Asthma Reviews	
66YJ.	Asthma annual review
66YK.	Asthma follow-up
66Yp.	Asthma review using Royal College of Physicians three questions
66YQ.	Asthma monitoring by nurse
8B3j.	Asthma medication review
9OJA.	Asthma monitoring check done
Asthma Emergency Department Visits (A&E code)	
14A	Asthma
Asthma Hospitalisations (ICD-10 codes)	
J45	Asthma
J45.0	Predominantly allergic asthma
J45.1	Nonallergic asthma
J45.8	Mixed asthma
J45.9	Asthma, unspecified
J46	Status asthmaticus
Asthma Prescriptions (Read codes)	
SABA	
c11%%	SALBUTAMOL [ORAL PREPARATIONS]
c12%%	SALBUTAMOL [PARENTERAL PREPARATIONS]
c13%%	SALBUTAMOL [INHALATION PREPARATIONS]
c14%%	TERBUTALINE SULPHATE [RESPIRATORY USE]
c15%%	FENOTEROL HYDROBROMIDE
c1E%%	SALBUTAMOL [INHALATION PREPARATIONS 2]
ICS	
c615.	*BECOTIDE rotahaler device
c616.	BECOTIDE 50micrograms/mL nebuliser solution
c617.	BECOTIDE-100 100microgram inhaler
c618.	*VOLUMATIC spacer device
c619.	BECODISK 100micrograms diskhaler 14x8
c61A.	BECLOMETASONE DIPROPIONATE 400micrograms disks+disk inhaler
c61B.	BECLOMETASONE DIPROPIONATE 400micrograms disk refill
c61C.	BECLOMETHASONE DIPROPIONATE 250micrograms inhaler+spacer device
c61E.	BECLOMETASONE DIPROPIONATE 250micrograms breath-actuated aerosol inhaler
c61F.	BECLOMETASONE DIPROPIONATE 100micrograms breath-actuated aerosol inhaler
c61G.	*FILAIR 50micrograms inhaler
c61H.	*FILAIR 100micrograms inhaler
c61J.	FILAIR FORTE 250micrograms inhaler
c61K.	BECLAZONE 50micrograms inhaler
c61L.	BECLAZONE 100micrograms inhaler
c61M.	BECLAZONE 250micrograms inhaler
c61N.	BECLAZONE 50 EASI-BREATHE inhaler
c61O.	BECLAZONE 100 EASI-BREATHE inhaler
c61P.	BECLAZONE 250 EASI-BREATHE inhaler
c61Q.	BECLOFORTE INTEGRA 250micrograms inhaler+compact spacer
c61R.	BECLOFORTE INTEGRA 250micrograms refill
c61S.	BECLOMETHASONE DIPROPIONATE 250micrograms inhaler+compact spacer
c61T.	BECLOMETHASONE DIPROPIONATE 250micrograms compact spacer refill
c61U.	BECLOMETHASONE rotahaler device
c61V.	BECLOMETHASONE DIPROPIONATE 50micrograms vortex metered dose inhaler
c61W.	*BDP 50micrograms Spacehaler

Table E.1: (cont'd).

Code	Description
c61X.	BECLOMETHASONE DIPROPIONATE 100micrograms vortex metered dose inhaler
c61Y.	*BDP 100micrograms Spacehaler
c61Z.	BECLOMETHASONE DIPROPIONATE 250micrograms vortex metered dose inhaler
c61a.	BECODISK 200micrograms diskhaler 14x8
c61b.	BECOTIDE 400micrograms rotacaps
c61c.	BECODISK 100micrograms disk refill 14x8
c61d.	BECODISK 200micrograms disk refill 14x8
c61e.	BECODISK 400micrograms diskhaler 7x8
c61f.	BECODISK 400micrograms disk refill 7x8
c61g.	BECLOFORTE VM 250micrograms inhaler+volumatic
c61h.	BECLOMETASONE DIPROPIONATE 400micrograms inhalation capsules
c61i.	BECOTIDE-200 200microgram inhaler
c61j.	*AEROBEC 50microgram Autohaler
c61k.	AEROBEC FORTE 250micrograms Autohaler
c61l.	AEROBEC 100microgram Autohaler
c61m.	BECLOFORTE DISKHALER 400micrograms 14x8
c61n.	BECLOFORTE DISKS 400micrograms disk refill 14x8
c61p.	BECLOMETASONE DIPROPIONATE 100micrograms disks+disk inhaler
c61q.	BECLOMETASONE DIPROPIONATE 200micrograms disks+disk inhaler
c61r.	BECLOMETASONE DIPROPIONATE 100micrograms disk refill
c61s.	BECLOMETASONE DIPROPIONATE 200micrograms disk refill
c61u.	BECLOMETASONE DIPROPIONATE 200micrograms inhaler
c61v.	BECLOMETASONE DIPROPIONATE 50micrograms inhaler
c61w.	BECLOMETASONE DIPROPIONATE 100micrograms inhalation capsules
c61x.	BECLOMETASONE DIPROPIONATE 200micrograms inhalation capsules
c61y.	BECLOMETHASONE DIPROPIONATE 50micrograms/mL nebuliser solution
c61z.	BECLOMETASONE DIPROPIONATE 100micrograms inhaler
c62..	BECLOMETASONE COMPOUNDS
c621.	*VENTIDE inhaler
c622.	*VENTIDE Rotacaps
c623.	*VENTIDE paediatric Rotacaps
c624.	*VENTIDE Rotahaler device
c63..	*BETAMETHASONE VALERATE
c631.	*BEXTASOL 100microgram inhaler
c63z.	BETAMETHASONE 100micrograms inhaler
c64..	BUDESONIDE [RESPIRATORY USE]
c641.	PULMICORT 200micrograms inhaler 200dose
c642.	PULMICORT 200micrograms refill 100dose
c643.	PULMICORT 200micrograms refill 200dose
c644.	PULMICORT LS 50micrograms inhaler
c645.	PULMICORT LS 50micrograms refill
c646.	*NEBUHALER spacer device
c647.	PULMICORT 200microgram inhaler 100dose
c649.	PULMICORT 400microgram Turbohaler 50dose
c64A.	BUDESONIDE 200micrograms refill cannister
c64B.	BUDESONIDE 50micrograms spacer inhaler
c64C.	PULMICORT 200micrograms spacer inhaler
c64D.	PULMICORT LS 50micrograms spacer inhaler
c64E.	PULMICORT 200micrograms inhaler with NebuChamber
c64F.	BUDESONIDE 200micrograms/dose dry powder cartridge refill
c64G.	NOVOLIZER BUDESONIDE 200micrograms/dose dry powder cartridge refill
c64H.	EASYHALER BUDESONIDE 100micrograms breath-actuated dry powder inhaler
c64I.	EASYHALER BUDESONIDE 200micrograms breath-actuated dry powder inhaler
c64J.	EASYHALER BUDESONIDE 400micrograms breath-actuated dry powder inhaler
c64K.	PULMICORT 100micrograms CFC-free inhaler
c64a.	PULMICORT 500micrograms Respules 2mL unit
c64b.	PULMICORT 1mg Respules 2mL unit
c64c.	PULMICORT 100microgram Turbohaler 200dose
c64d.	BUDESONIDE 100micrograms breath-actuated dry powder inhaler
c64e.	BUDESONIDE 50micrograms refill cannister
c64g.	BUDESONIDE 200micrograms breath-actuated dry powder inhaler
c64h.	BUDESONIDE 400micrograms breath-actuated dry powder inhaler
c64i.	BUDESONIDE 500micrograms/2mL nebuliser solution
c64j.	BUDESONIDE 1mg/2mL nebuliser solution
c64k.	*BUDESONIDE 200 Cyclocaps
c64l.	*BUDESONIDE 400 Cyclocaps
c64m.	BUDESONIDE 200micrograms inhalation capsules

Table E.1: (cont'd).

Code	Description
c64n.	BUDESONIDE 400micrograms inhalation capsules
c64o.	BUDESONIDE 200micrograms inhaler with spacer device
c64p.	NOVOLIZER BUDESONIDE 200micrograms/dose dry powder cartridge and refillable inhaler device
c64u.	BUDESONIDE 200micrograms/dose dry powder cartridge and refillable inhaler device
c64v.	BUDESONIDE 200micrograms inhaler
c64x.	*BUDESONIDE refill 200dose
c64y.	BUDESONIDE 50micrograms inhaler
c64z.	BUDESONIDE 200micrograms spacer inhaler
c65..	FLUTICASONE PROPIONATE [RESPIRATORY USE]
c651.	FLIXOTIDE 50micrograms diskhaler
c652.	FLIXOTIDE 100micrograms diskhaler
c653.	FLIXOTIDE 250micrograms diskhaler
c654.	FLUTICASONE PROPIONATE 50micrograms disks+disk inhaler
c655.	FLUTICASONE PROPIONATE 100micrograms disks+disk inhaler
c656.	FLUTICASONE PROPIONATE 250micrograms disks+disk inhaler
c657.	FLIXOTIDE 50micrograms disk refill
c658.	FLIXOTIDE 100micrograms disk refill
c659.	FLIXOTIDE 250micrograms disk refill
c65A.	FLUTICASONE PROPIONATE 50micrograms disk refill
c65B.	FLUTICASONE PROPIONATE 100micrograms disk refill
c65C.	FLUTICASONE PROPIONATE 250micrograms disk refill
c65D.	FLIXOTIDE 25micrograms inhaler
c65E.	FLIXOTIDE 50micrograms inhaler
c65F.	FLIXOTIDE 125micrograms inhaler
c65G.	FLUTICASONE PROPIONATE 25micrograms inhaler
c65H.	FLUTICASONE PROPIONATE 50micrograms inhaler
c65I.	FLUTICASONE PROPIONATE 125micrograms inhaler
c65K.	FLIXOTIDE 250micrograms inhaler
c65L.	FLIXOTIDE 500micrograms diskhaler
c65M.	FLIXOTIDE 500micrograms disk refill
c65N.	FLUTICASONE PROPIONATE 500micrograms disks+disk inhaler
c65O.	FLUTICASONE PROPIONATE 500micrograms disk refill
c65P.	FLUTICASONE PROPIONATE 50micrograms breath-actuated dry powder inhaler
c65Q.	FLUTICASONE PROPIONATE 100micrograms breath-actuated dry powder inhaler
c65R.	FLUTICASONE PROPIONATE 250micrograms breath-actuated dry powder inhaler
c65S.	FLUTICASONE PROPIONATE 500micrograms breath-actuated dry powder inhaler
c65T.	FLIXOTIDE 50micrograms Accuhaler
c65U.	FLIXOTIDE 100micrograms Accuhaler
c65V.	FLIXOTIDE 250micrograms Accuhaler
c65W.	FLIXOTIDE 500micrograms Accuhaler
c65X.	FLUTICASONE PROPIONATE 0.5mg/2mL nebulisation units
c65Y.	FLUTICASONE PROPIONATE 2mg/2mL nebulisation units
c65Z.	FLIXOTIDE 0.5mg/2mL Nebules
c65a.	FLIXOTIDE 2mg/2mL Nebules
c65b.	FLUTICASONE PROPIONATE 125micrograms CFC-free inhaler
c65c.	FLUTICASONE PROPIONATE 250micrograms CFC-free inhaler
c65d.	FLIXOTIDE 125micrograms Evohaler
c65e.	FLIXOTIDE 250micrograms Evohaler
c65f.	FLUTICASONE PROPIONATE 50micrograms CFC-free inhaler
c65g.	FLIXOTIDE 50micrograms Evohaler
c66..	BECLOMETASONE DIPROPIONATE [RESPIRATORY USE 2]
c661.	*BDP 250micrograms Spacehaler
c662.	BECOTIDE 50 EASI-BREATHE inhaler
c663.	BECOTIDE 100 EASI-BREATHE inhaler
c664.	BECLOFORTE EASI-BREATHE 250micrograms inhaler
c665.	QVAR 50 inhaler
c666.	QVAR 100 inhaler
c667.	QVAR 50 Autohaler
c668.	QVAR 100 Autohaler
c669.	*BECLAZONE 200 inhaler
c66A.	BECLOMETASONE DIPROPIONATE 50micrograms breath-actuated dry powder inhaler

Table E.1: (cont'd).

Code	Description
c66B.	BECLOMETASONE DIPROPIONATE 100micrograms breath-actuated dry powder inhaler
c66C.	BECLOMETASONE DIPROPIONATE 250micrograms breath-actuated dry powder inhaler
c66D.	ASMABEC 50micrograms Clickhaler
c66E.	ASMABEC 100micrograms Clickhaler
c66F.	ASMABEC 250micrograms Clickhaler
c66G.	BECLOMETASONE DIPROPIONATE 400micrograms breath-actuated dry powder inhaler
c66H.	BECLOMETASONE DIPROPIONATE 200micrograms breath-actuated dry powder inhaler
c66I.	PULVINAL BECLOMETHASONE DIPROPIONATE 100micrograms breath-actuated dry powder inhaler
c66J.	PULVINAL BECLOMETHASONE DIPROPIONATE 200micrograms breath-actuated dry powder inhaler
c66K.	PULVINAL BECLOMETHASONE DIPROPIONATE 400micrograms breath-actuated dry powder inhaler
c66L.	*BECLOMETASONE 100 cyclocaps
c66M.	*BECLOMETASONE 200 cyclocaps
c66N.	*BECLOMETASONE 400 cyclocaps
c66P.	BECODISK 100micrograms diskhaler 15x8
c66Q.	BECODISK 200micrograms diskhaler 15x8
c66R.	BECODISK 400micrograms diskhaler 15x8
c66S.	BECODISK 100micrograms disk refill 15x8
c66T.	BECODISK 200micrograms disk refill 15x8
c66U.	BECODISK 400micrograms disk refill 15x8
c66V.	BECLOMETASONE DIPROPIONATE 50micrograms CFC-free inhaler
c66W.	BECLOMETASONE DIPROPIONATE 100micrograms CFC-free inhaler
c66X.	BECLOMETASONE DIPROPIONATE 50micrograms CFC-free breath-actuated aerosol inhaler
c66Y.	BECLOMETASONE DIPROPIONATE 100micrograms CFC-free breath-actuated aerosol inhaler
c66Z.	QVAR EASI-BREATHE 50micrograms CFC-free breath-actuated dry powder inhaler
c66a.	QVAR EASI-BREATHE 100micrograms CFC-free breath-actuated dry powder inhaler
c66c.	CLENIL MODULITE 50micrograms CFC-free inhaler
c66d.	CLENIL MODULITE 100micrograms CFC-free inhaler
c66e.	CLENIL MODULITE 200micrograms CFC-free inhaler
c66f.	CLENIL MODULITE 250micrograms CFC-free inhaler
c66g.	BECLOMETASONE DIPROPIONATE 200micrograms CFC-free inhaler
c66h.	BECLOMETASONE DIPROPIONATE 250micrograms CFC-free inhaler
c68..	MOMETASONE [RESPIRATORY USE]
c681.	MOMETASONE FUROATE 200micrograms breath-actuated dry powder inhaler
c682.	MOMETASONE FUROATE 400micrograms breath-actuated dry powder inhaler
c683.	ASMANEX TWISTHALER 200micrograms breath-actuated dry powder inhaler
c684.	ASMANEX TWISTHALER 400micrograms breath-actuated dry powder inhaler
c69..	CICLESONIDE
c691.	ALVESCO 160micrograms inhaler
c692.	ALVESCO 80micrograms inhaler
c69y.	CICLESONIDE 80micrograms inhaler
c69z.	CICLESONIDE 160micrograms inhaler
ICS-LABA	
c1D..	SALMETEROL+FLUTICASONE PROPIONATE
c1D1.	SERETIDE 100 Accuhaler
c1D2.	SERETIDE 250 Accuhaler
c1D3.	SERETIDE 500 Accuhaler
c1D4.	SERETIDE 50 Evohaler
c1D5.	SERETIDE 125 Evohaler
c1D6.	SERETIDE 250 Evohaler
c1D7.	SIRDUPLA 25micrograms/125micrograms inhaler
c1D8.	SIRDUPLA 25micrograms/250micrograms inhaler
c1Du.	SALMETEROL+FLUTICASONE PROPIONATE 25micrograms/50micrograms CFC-free inhaler
c1Dv.	SALMETEROL+FLUTICASONE PROPIONATE 25micrograms/125micrograms CFC-free inhaler

Table E.1: (cont'd).

Code	Description
c1Dw.	SALMETEROL+FLUTICASONE PROPIONATE 25micrograms/250micrograms CFC-free inhaler
c1Dx.	SALMETEROL+FLUTICASONE PROPIONATE 50micrograms/100micrograms breath-actuated dry powder inhaler
c1Dy.	SALMETEROL+FLUTICASONE PROPIONATE 50micrograms/250micrograms breath-actuated dry powder inhaler
c1Dz.	SALMETEROL+FLUTICASONE PROPIONATE 50micrograms/500micrograms breath-actuated dry powder inhaler
c1c..	FLUTICASONE PROPIONATE+FORMOTEROL FUMARATE
c1c1.	FLUTIFORM 50micrograms/5micrograms inhaler
c1c2.	FLUTIFORM 125micrograms/5micrograms inhaler
c1c3.	FLUTIFORM 250micrograms/10micrograms inhaler
c1cx.	FLUTICASONE PROPIONATE+FORMOTEROL FUMARATE 250mcg/10mcg inh
c1cy.	FLUTICASONE PROPIONATE+FORMOTEROL FUMARATE 125mcg/5mcg inh
c1cz.	FLUTICASONE PROPIONATE+FORMOTEROL FUMARATE 50mcg/5mcg inh
c67..	BUDESONIDE+FORMOTEROL
c671.	SYMBICORT 100/6 Turbohaler
c672.	SYMBICORT 200/6 Turbohaler
c673.	SYMBICORT 400/12 Turbohaler
c674.	DUORESP SPIROMAX 160mcg/4.5mcg breath-act dry powder inhaler
c675.	DUORESP SPIROMAX 320mcg/9mcg breath-act dry powder inhaler
c67x.	BUDESONIDE+FORMOTEROL FUMARATE DIHYDRATE 400micrograms/12micrograms breath-actuated dry powder inhaler
c67y.	BUDESONIDE+FORMOTEROL FUMARATE DIHYDRATE 200micrograms/6micrograms breath-actuated dry powder inhaler
c67z.	BUDESONIDE+FORMOTEROL FUMARATE DIHYDRATE 100micrograms/6micrograms breath-actuated dry powder inhaler
c6A..	BECLOMETASONE+FORMOTEROL
c6A1.	FOSTAIR 100micrograms/6micrograms inhaler
c6A2.	FOSTAIR NEXTHALER 100micrograms/6micrograms powder inhaler
c6Ay.	BECLOMET DIPROP+FORMOTERL FUMARATE DIHYD 100mcg/6mcg pdr inh
c6Az.	BECLOMETASONE DIPROPIONATE+FORMETEROL FUMARATE DIHYDRATE 100micrograms/6micrograms inhaler
c6B..	FLUTICASONE+VILANTEROL
c6B1.	RELVAR ELLIPTA 184micrograms/22micrograms inhaler
c6B2.	FLUTICASONE FUROATE+VILANTEROL 184mcg/22mcg dry pdr inhaler
c6B3.	RELVAR ELLIPTA 92micrograms/22micrograms inhaler
c6B4.	FLUTICASONE FUROATE+VILANTEROL 92mcg/22mcg dry pdr inhaler
Theophyllines	
c41..	AMINOPHYLLINE
c411.	AMINOPHYLLINE 100mg tablets
c412.	AMINOPHYLLINE 250mg/10mL injection
c413.	AMINOPHYLLINE 500mg/2mL injection
c414.	AMINOPHYLLINE 50mg suppositories
c415.	AMINOPHYLLINE 100mg suppositories
c416.	AMINOPHYLLINE 150mg suppositories
c417.	AMINOPHYLLINE 180mg suppositories
c418.	AMINOPHYLLINE 360mg suppositories
c419.	*THEODROX tablets
c41A.	*NORPHYLLIN 100mg tablets
c41B.	NORPHYLLIN SR 225mg m/r tablets
c41C.	NORPHYLLIN SR 350mg m/r tablets
c41a.	PHYLLOCONTIN CONTINUS 225mg m/r tablets
c41b.	PHYLLOCONTIN FORTE 350mg m/r tablets
c41c.	PHYLLOCONTIN PAEDIATRIC 100mg m/r tablets
c41d.	AMINOPHYLLINE 225mg m/r tablets
c41e.	*PECRAM 225mg m/r tablets
c41f.	AMINOPHYLLINE 350mg m/r tablets
c41g.	AMINOPHYLLINE 100mg m/r tablets
c41h.	*AMNIVENT 225mg m/r tablets
c41i.	*AMNIVENT 350mg m/r tablets
c41j.	MIN-I-JET AMINOPHYLLINE 250mg/10mL injection
c41k.	AMINOPHYLLINE 250mg/10mL prefilled syringe
c41m.	AMINOPHYLLINE HYDRATE 225mg m/r tablets
c43..	THEOPHYLLINE
c431.	*BIOPHYLLINE 125mg/5mL syrup
c432.	*NUELIN 125mg tablets

Table E.1: (cont'd).

Code	Description
c433.	*NUELIN 60mg/5mL liquid
c434.	*LASMA 300mg m/r tablets
c435.	NUELIN SA 175mg m/r tablets
c436.	NUELIN SA-250 250mg m/r tablets
c437.	*PRO-VENT 300mg m/r capsules
c438.	SLO-PHYLLIN 60mg m/r capsules
c439.	SLO-PHYLLIN 125mg m/r capsules
c43A.	THEOPHYLLINE 200mg/10mL injection
c43B.	THEOPHYLLINE 10mg/5mL sugar free solution
c43a.	SLO-PHYLLIN 250mg m/r capsules
c43b.	*THEO-DUR 200mg m/r tablets
c43c.	*THEO-DUR 300mg m/r tablets
c43d.	*THEOGRAD 350mg m/r tablets
c43e.	UNIPHYLLIN CONTINUS 400mg m/r tablets
c43f.	UNIPHYLLIN CONTINUS 200mg m/r tablets
c43g.	LABOPHYLLINE 200mg/10mL injection
c43h.	UNIPHYLLIN CONTINUS 300mg m/r tablets
c43i.	*BIOPHYLLINE 350mg m/r tablets
c43j.	*BIOPHYLLINE 500mg m/r tablets
c43k.	THEOPHYLLINE 500mg m/r tablets
c43m.	*THEOPHYLLINE 125mg/5mL syrup
c43n.	*THEOPHYLLINE 125mg tablets
c43o.	*THEOPHYLLINE 60mg/5mL liquid
c43p.	THEOPHYLLINE 175mg m/r tablets
c43q.	THEOPHYLLINE 250mg m/r tablets
c43r.	THEOPHYLLINE 300mg m/r capsules
c43s.	THEOPHYLLINE 60mg m/r capsules
c43t.	THEOPHYLLINE 125mg m/r capsules
c43u.	THEOPHYLLINE 250mg m/r capsules
c43v.	THEOPHYLLINE 200mg m/r tablets
c43w.	THEOPHYLLINE 300mg m/r tablets
c43x.	THEOPHYLLINE 350mg m/r tablets
c43y.	THEOPHYLLINE 400mg m/r tablets
c43z.	*THEOPHYLLINE 200mg tablets
LTRA	
cA...	LEUKOTRIENE RECEPTOR ANTAGONIST
cA1..	MONTELUKAST
cA11.	MONTELUKAST 10mg tablets
cA12.	MONTELUKAST 5mg chewable tablets
cA13.	SINGULAIR 10mg tablets
cA14.	SINGULAIR PAEDIATRIC 5mg chewable tablets
cA15.	SINGULAIR PAEDIATRIC 4mg chewable tablets
cA16.	SINGULAIR PAEDIATRIC 4mg/sachet granules
cA1y.	MONTELUKAST 4mg/sachet granules
cA1z.	MONTELUKAST 4mg chewable tablets
cA2..	ZAFIRLUKAST
cA21.	ZAFIRLUKAST 20mg tablets
cA22.	ACCOLATE 20mg tablets
OCS	
fe61	PREDNISOLONE 1mg tablets
fe62	PREDNISOLONE 5mg tablets
fe66	DELTACORTRIL ENTERIC 5mg tablets
fe6i	PREDNISOLONE 5mg e/c tablets
fe6j	PREDNISOLONE 5mg soluble tablets
fe6k	PREDNISOLONE 50mg tablets
fe6z	PREDNISOLONE 25mg tablets