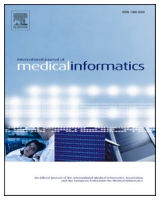
International Journal of Medical Informatics 170 (2023) 104942

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Defining clinical subtypes of adult asthma using electronic health records: Analysis of a large UK primary care database with external validation

Elsie M.F. Horne [a,b,c,*,1], Susannah McLean [a,b], Mohammad A. Alsallakh [a,d,e], Gwyneth A. Davies [a,d], David B. Price [f,g], Aziz Sheikh [a,b], Athanasios Tsanas [a,b]

[a] Asthma UK Centre for Applied Research, Edinburgh, UK
[b] Usher Institute, Edinburgh Medical School, College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, UK
[c] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[d] Population Data Science, Swansea University Medical School, Swansea, UK
[e] Health Data Research UK, Swansea and Edinburgh, UK
[f] Observational and Pragmatic Research Institute (OPRI), Singapore
[g] Centre of Academic Primary Care, Division of Applied Health Sciences, University of Aberdeen, Aberdeen, UK

ARTICLE INFO

Keywords:
Asthma
Electronic health records
Cluster analysis

ABSTRACT

Introduction: Asthma is one of the commonest chronic conditions in the world. Subtypes of asthma have been defined, typically from clinical datasets on small, well-characterised subpopulations of asthma patients. We sought to define asthma subtypes from large longitudinal primary care electronic health records (EHRs) using cluster analysis.

Methods: In this retrospective cohort study, we extracted asthma subpopulations from the Optimum Patient Care Research Database (OPCRD) to robustly train and test algorithms, and externally validated findings in the Secure Anonymised Information Linkage (SAIL) Databank. In both databases, we identified adults with an asthma diagnosis code recorded in the three years prior to an index date. Train and test datasets were selected from OPCRD using an index date of Jan 1, 2016. Two internal validation datasets were selected from OPCRD using index dates of Jan 1, 2017 and 2018. Three external validation datasets were selected from SAIL using index dates of Jan 1, 2016, 2017 and 2018. Each dataset comprised 50,000 randomly selected non-overlapping patients. Subtypes were defined by applying multiple correspondence analysis and k-means cluster analysis to the train dataset, and were validated in the internal and external validation datasets.

Results: We defined six asthma subtypes with clear clinical interpretability: low inhaled corticosteroid (ICS) use and low healthcare utilisation (30% of patients); low-to-medium ICS use (36%); low-to-medium ICS use and comorbidities (12%); varied ICS use and comorbid chronic obstructive pulmonary disease (4%); high (10%) and very high ICS use (7%). The subtypes were replicated with high accuracy in internal (91–92%) and external (84–86%) datasets.

Conclusion: Asthma subtypes derived and validated in large independent EHR databases were primarily defined by level of ICS use, level of healthcare use, and presence of comorbidities. This has important clinical implications towards defining asthma subtypes, facilitating patient stratification, and developing more personalised monitoring and treatment strategies.

Abbreviations: EHR, Electronic Health Record; OPCRD, Optimum Patient Care Research Database; SAIL, Secure Anonymised Information Linkage; ICS, Inhaled Corticosteroid; GINA, Global Initiative for Asthma; MCA, Multiple correspondence analysis; BMI, Body mass index; RF, Random forest; *t*-SNE, *t*-distributed stochastic neighbour embedding; CCI, Charlson Comorbidity Index; COPD, Chronic Obstructive Pulmonary Disease; SABA, Short-Acting Beta-Agonist.

* Corresponding author at: Usher Institute, Old Medical School, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK.
E-mail address: Elsie.Horne@bristol.ac.uk (E.M.F. Horne).
[1] Present address: Population Health Sciences, University of Bristol, Oakfield House, Bristol BS8 2BN, UK.

https://doi.org/10.1016/j.ijmedinf.2022.104942
Received 24 April 2022; Received in revised form 13 November 2022; Accepted 28 November 2022
Available online 7 December 2022

1. Introduction

Asthma is a chronic condition characterised by reversible airway obstruction and has an estimated global prevalence of 262 million [1]. People diagnosed with asthma exhibit diverse symptoms, likely arising from different underlying mechanisms [2]. The Global Initiative for Asthma (GINA) guidelines provide a characterisation of asthma symptom severity as "mild", "moderate" or "severe" based on the current level of treatment required to control symptoms [3]. However, the 2022 update to these guidelines presents limitations of this characterisation: e.g. that it can only be applied retrospectively, that these terms may be used informally without full assessment, and that the term "mild asthma" may encourage complacency [3]. Of 155 deaths included in the National Review of Asthma Deaths, 58% had occurred in people who were being treated for "mild" or "moderate" asthma [4].

More recently, the task of asthma categorisation has been tackled using data analytics methods. "Cluster analysis" is an increasingly popular method for identifying homogeneous patient subgroups based on patterns in clinical characteristics, both in asthma [5,6] and other disease areas [7–9]. Our recent review identified 63 studies that applied cluster analysis with the aim of defining asthma subtypes, and highlighted several practical and methodological limitations that were common across most of the studies reviewed (in particular, small sample sizes, inappropriate feature encoding, lack of stability testing, and lack of formal validation exercises) [10]. The subtypes that have been consistently identified across these studies are primarily defined by the age of disease onset, the presence of allergies and level of eosinophilic inflammation [6].

The aim of this study was to: (1) develop and validate a data analytics framework for identifying asthma subtypes among adults using electronic health records (EHRs), and (2) gain new insights into this heterogeneous patient population. The framework was designed to be generalisable for use with large datasets comprising categorical features, which are key characteristics of datasets derived from longitudinal, coded EHRs.

2. Methods

2.1. Study design and population

The primary datasets were obtained from the Optimum Patient Care Research Database (OPCRD; https://opcrd.co.uk/). The OPCRD dataset comprises medical records of >13.7 million patients from over 800 general practices across the UK (approximately 10 % of the total UK population), drawn from all UK clinical systems (EMIS, TPP SystmOne, InPS Vision, Microtest Evolution). It benefits from a long retrospective period (median time in the database is 7 years, going back to birth for summary diagnostic data in many cases), and contains linked patient-completed respiratory questionnaires. Respiratory-related outcome measures within the OPCRD have been validated using patient reported outcomes [11].

The findings derived from the OPCRD datasets were validated using datasets obtained from the Secure Anonymised Information Linkage (SAIL; https://saildatabank.com) Databank [12–16]. Within the SAIL Databank, the Welsh Longitudinal General Practice dataset comprises the primary care EHR data on >4 million patients from over 76 % of general practices in Wales [17].

The NHS Health Research Authority has approved OPCRD for clinical research purposes (REC reference: 20/EM/0148). Approval for the analysis of OPCRD data in the current study was granted by the Anonymous Data Ethics Protocols and Transparency committee (ADEPT0619). The Secure Anonymised Information Linkage (SAIL) Databank independent Information Governance Review Panel approved the study as part of the Wales Asthma Observatory project (0317). No additional research ethics approval was need as only anonymised data were used.

For each of OPCRD and SAIL, three cohorts were derived corresponding to three index dates (Jan 1, 2016, 2017, and 2018). This was to assess the generalisability for the clusters to future time-points. Inclusion criteria were as follows: (1) the patient had a Read code corresponding to asthma in the three years prior to the index date; and (2) the patient was registered at a general practice located in England for the OPCRD cohorts and Wales for the SAIL cohorts. The exclusion criteria are described in Appendix A.1.1.

To reduce computation time, samples of 50,000 non-overlapping patients were randomly selected from the eligible patients for each index date and data source (except for the 2016 OPCRD cohort, from which 100,000 patients were randomly selected and split into train and test sets). The purpose of each of these samples is outlined in Fig. 1. Two independent EHR databases and three separate datasets (corresponding to three time-periods) from each database were used to assess the extent to which findings were biased to a certain database or time-period.

2.2. Feature derivation

Forty-five categorical features were derived from the primary care EHR data (Table 1). These features were chosen based on previous studies of asthma patients using primary care EHRs [18,19], clinical relevance to asthma subtypes [6], and the availability of data across both the OPCRD and SAIL Databank. We chose to include a wide range of features that were linked to asthma both directly (e.g. medications for asthma) and indirectly (e.g. presence of comorbidities). Multiple correspondence analysis (MCA) was used to detect underlying structures in the dataset and represent them in fewer dimensions prior to the application of k-means cluster analysis (Table 2).

2.3. Identifying subtypes

Table 2 describes the methods that were applied to the 2016 OPCRD data to identify the subtypes. The data were split randomly into non-overlapping train and test datasets, each of 50,000 samples (Fig. 1).

2.4. Validation

To validate the subtypes internally, the RF was used to predict cluster labels for the 2017 and 2018 OPCRD datasets. The ground truth cluster labels were derived in these datasets by applying MCA and k-means cluster analysis using the number of MCA-derived features and number of clusters determined from the train dataset. The subtypes were validated externally using three datasets derived from SAIL Databank (Fig. 1).

For both the internal and external validations, comparisons between the RF-assigned subtypes and the cluster labels determined using the cluster analysis methodology were reported as confusion matrices, balanced accuracies and Jaccard similarity coefficients. Bar plots of the features with the 10 highest RF feature importance values for each subtype were used to compare characteristics of the clusters across the 2016 OPCRD and 2016 SAIL datasets.

2.5. Interpretation

The subtypes were interpreted through a combination of summary statistics, RF feature importance values and scatter plots of the data projected to two-dimensional space using *t*-distributed stochastic

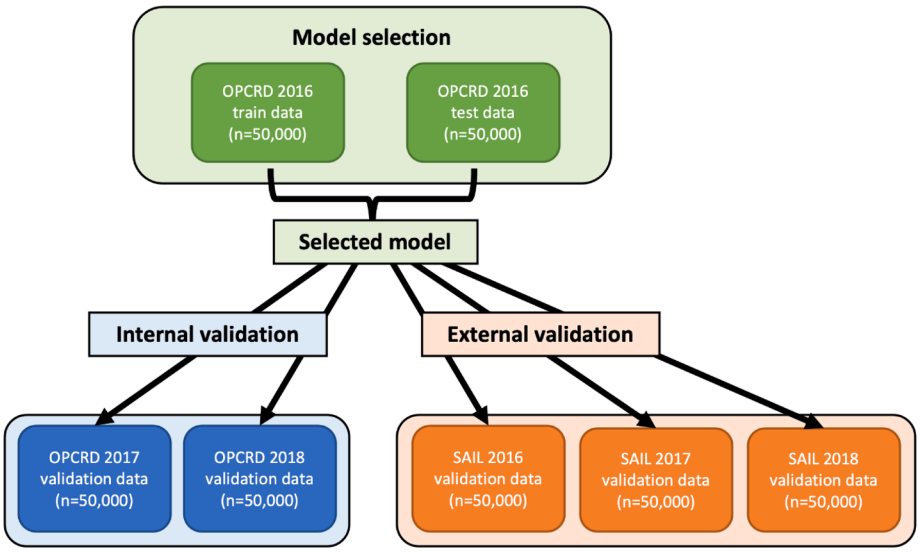


Fig. 1. Illustration of the model selection and validation framework. Each square represents a non-overlapping sample of 50,000 patients. The OPCRD 2016 train data is the main 'train dataset' used in the study. The selected model was chosen based on performance in the OPCRD 2016 test data.

neighbour embedding (*t*-SNE; see Section A.1.2) [25].

2.6. Reporting guidelines and availability of code

This study was reported in line with the STROBE and RECORD statements [26,27]. The completed checklist is provided as Appendix B. Analysis code is publicly available at https://github.com/elsie-h/asthma-subtypes. All statistical analysis was performed with R (version 4.0.3).

2.7. Role of the funders

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

3. Results

3.1. Study design and population

The flow of eligible patients into the study is illustrated in Fig. A.2.

3.2. Feature derivation

Summary statistics for all features derived from OPCRD and SAIL stratified by year are given in Table A.2.

3.3. Identifying subtypes

Based on the scree plot (Fig. A.3), MCA solutions with 4–14 dimensions were used as input for the *k*-means cluster analysis. For all MCA solutions, the average silhouette width was greatest for the *2*-cluster solution (Fig. A.5). In all cases this was driven by one large, well-defined cluster, while the other cluster was poorly defined, but small. See Fig. A.6 for silhouettes plots for all stable cluster solutions for the *12*-dimensional MCA solution. For the *4*- and *5*-dimension MCA solutions, the *5*-cluster solution had the next greatest average silhouette width, and for *6*- to *14*-dimension MCA solutions, the *6*-cluster solution had the next greatest average silhouette width. These solutions were taken forward to the classification step, along with the *6*-dimension *7*-cluster solution that was favoured by the gap statistic (Fig. A.7).

The *6*-cluster solution derived from the *12*-dimension MCA solution had the highest mean Jaccard coefficient when compared with the labels assigned by the RF model, so was selected as the final solution (Fig. A.9).

3.4. Interpretation

The datasets were partitioned into groups according to the cluster labels as predicted by the RF, we refer to these groups as "subtypes". Table 3 provides summary statistics for key features across the subtypes. Table A.4 summarises all 45 categorical features across the subtypes. Fig. 2 (plot A) illustrates the six subtypes in two-dimensional space projected by *t*-SNE. Fig. 2 (plot B) illustrates that the feature that best separates the clusters is the number of inhaled corticosteroid (ICS) prescriptions. The number of asthma reviews (plot C), a COPD diagnosis (plot D) and the Charlson Comorbidity Index (CCI; plot E) also separate the clusters, albeit to a lesser extent than ICS. Additional *t*-SNE plots are provided in Fig. A.13.

3.5. Validation

The subtypes were replicated with 91–92 % balanced accuracy in the internal validation (datasets extracted from OPCRD; Fig. 3) and 84–86 % in the external validation (datasets extracted from SAIL; Fig. 4).

4. Discussion

4.1. Summary

We defined the following six subtypes of adult asthma from primary care EHR data (percentage of patients in the 2016 OPCRD test dataset assigned to each subtype in parentheses): "low ICS use and low healthcare utilisation" (30 %); "low-to-medium ICS use" (36 %); "low-to-medium ICS use and comorbidities" (12 %); "varied ICS use and co-morbid COPD" (4 %); "high ICS use" (10 %); and "very high ICS" (7 %)

Table 1
Description of the input features.

Feature	Description	Categories	Read codes[5]
Sex	As recorded in primary care record	F; M;	–
Index age	Age in years on index date split into quintiles	18–33; 34–45; 46–56; 57–68; 69–100;[1]	–
Smoking status	Smoking status recorded	Current; ex; non;[2]	smoking.csv
Age of diagnosis	Age in years at recording of first asthma diagnosis code in primary care record	<18; ≥18; [2]	asthma_diagnosis.csv
Body mass index (BMI)	Most recent BMI code recorded in the two-year pre-index period (kg/m^2)	<20; 20–24; 25–29; 30–39; ≥40; missing; [2]	BMI.csv
Blood eosinophil count	Most recent blood eosinophil count recorded in the two-year pre-index period (cells per microlitre)	≤400; >400; missing;[2]	blood_eosinophils.csv
Emergency events	Number of asthma-related A&E attendances or hospital admission codes recorded in the pre-index year (codes recorded within 14 days of each other treated as one occurrence)	0; ≥1; [1]	emergency.csv
Exacerbations	Number of asthma exacerbation codes recorded in the pre-index year (codes recorded within 14 days of each other treated as one occurrence)	0; ≥1; [1]	exacerbation.csv
Asthma reviews	Number of distinct dates on which asthma review codes were recorded in the two-year pre-index period	0; 1; 2; ≥3; [a]	asthma_review.csv
Upper respiratory tract infection (URTI)	Number of codes recorded in pre-index year (codes recorded within 14 days of each other treated as one occurrence)	0; 1; ≥2; [1]	infections.csv
Lower respiratory tract infection (LRTI; includes influenza)			
Influenza vaccination	Influenza vaccination code recorded in pre-index year	No; yes;	influenza_vaccination.csv
Percent of predicted peak expiratory flow (ppPEF)	Most recent ppPEF recorded in the two-year pre-index period	<60 %; 60–80 %; >80 %; not valid; missing; [2,3]	PEFR.csv
Percent of predicted forced expiratory volume in 1 second (ppFEV1)	Most recent ppFEV1 recorded in the two-year pre-index period	<60 %; 60–80 %; 80–100 %; >100 %; not valid; missing; [2,3]	spirometry.csv
Asthma action plan	Asthma action plan code recorded in two-year pre-index period	No; yes;	asthma_plan.csv
Peak expiratory flow (PEF) home monitoring	PEF home monitoring code recorded in two-year pre-index period	No; yes;	PEFR_home.csv
Royal College of Physicians Three Questions (RCP3Q)	Most recent RCP3Q score recorded in the two-year pre-index period	0; 1; 2; 3; missing;	RCP3Q.csv
Anaphylaxis	Anaphylaxis code ever recorded (active if "adrenaline pen prescribed" code recorded in the two-year pre-index period)	Absent; ever; active;	allergic_conditions.csv & allergy_prescriptions.csv
Angioedema or urticaria	Angioedema or urticaria code ever recorded	Absent; ever;	
Allergic conjunctivitis	Allergic conjunctivitis code ever recorded (active if prescription for "medication for eye allergy" also recorded in pre-index year)	Absent; ever; active;	
Eczema	Eczema code ever recorded (active if prescription for "topical preparation for eczema" also recorded in pre-index year)	Absent; ever; active;	
Rhinitis	Rhinitis code ever recorded (active if prescription for "medication for nasal allergy" also recorded in pre-index year)	Absent; ever; active;	
Drug allergy	Drug allergy code ever recorded	Absent; ever;	
Food allergy	Food allergy code ever recorded	Absent; ever;	
Other allergy	Other allergy code ever recorded	Absent; ever;	
Oral antihistamine	Active if oral antihistamine prescription code recorded in pre-index year	Absent; ever;	
Anxiety	Anxiety code ever recorded (active if prescription of "medication for anxiety" also recorded in pre-index year)	Absent; ever; active;	comorbidities.csv & comorbidity_prescriptions.csv
Chronic cardiac disease (CCD)	CCD code ever recorded (active if prescription of "medication for anxiety" also recorded in pre-index year)	Absent; ever;	
Chronic obstructive pulmonary disease (COPD)	COPD code ever recorded	Absent; ever;	
Depression	Depression code ever recorded (active if prescription of "medication for depression" also recorded in pre-index year)	Absent; ever; active;	
Diabetes	Diabetes code ever recorded	Absent; ever;	
Gastro-oesophageal reflux disease (GORD)	GORD code ever recorded (active if prescription of "medication for GORD" also recorded in pre-index year)	Absent; ever; active;	
Nasal polyps	Nasal polyps code ever recorded	Absent; ever;	
Beta-blocker	Number of beta-blocker prescription codes in the pre-index year	0; ≥1; [1]	comedications.csv
Non-steroidal anti-inflammatory drug (NSAID)	Number of NSAID prescription codes in the pre-index year	0; ≥1; [1]	
Paracetamol	Number of paracetamol prescription codes in pre-index year	0; ≥1; [1]	
Statin	Number of statin prescription codes in pre-index year	0; ≥1; [1]	
Inhaled corticosteroids (ICS)	Number of ICS prescription codes in pre-index year	0; 1–4; 5–8; ≥9; [1]	asthma_prescriptions.csv
Short-acting beta-agonist (SABA)	Number of SABA prescription codes in pre-index year	0; 1–4; 5–8; ≥9; [1]	

(*continued on next page*)

Table 1 (*continued*)

Feature	Description	Categories	Read codes[5]
Long-acting beta-agonist (LABA)	Number of LABA prescription codes in pre-index year	0; 1–4; 5–8; ≥9; [1]	
Short-acting muscarinic-agonist (SAMA)	Number of SAMA prescription codes in pre-index year	0; ≥1; [1]	
Long-acting muscarinic-agonist (LAMA)	Number of LAMA prescription codes in pre-index year	0; ≥1; [1]	
Leukotriene receptor antagonist (LTRA)	Number of LTRA prescription codes in pre-index year	0; ≥1; [1]	
Methylxanthine	Number of methylxanthine prescription codes in pre-index year	0; ≥1; [1]	
Oral corticosteroids (OCS)	Number of OCS prescription codes recorded on the same date as an asthma-related code in pre-index year	0; 1; 2; ≥3; [1]	asthma_prescriptions.csv, asthma_qof.csv, asthma_review.csv, asthma_plan.csv, emergency.csv, exacerbation.csv, RCP3Q.csv
Charlson comorbidity index (CCI)[4]	Number of CCI categories with a code recorded ever	0–1; 2; ≥3; [1]	CCI.csv

[1] Categorisation based on feature distribution.
[2] Established categorisation.
[3] Samples for which data had been recorded, but were not valid were categorised as "not valid", while samples in which no data had been recorded were categorised as "missing". This was to distinguish between samples for which the data had been recorded, but were not usable (e.g. the patient's ppPEF had been measured but incorrectly recorded), and samples for which no data had been recorded (e.g. the patient's ppPEF had not been measured).
[4] Used to summarise subtypes but not used as an input feature for multiple correspondence analysis.
[5] Files can be found in https://github.com/elsie-h/asthma-subtypes/tree/main/read_codes.

use.

The "low ICS use and low healthcare utilisation" subtype was labelled as such because 89 % and 84 % of patients had received no ICS or short-acting beta-agonist (SABA) prescriptions respectively in the pre-index year, and 87 % had no record of an asthma review in the two-year pre-index period. Patients assigned to the "low-to-medium ICS use" subtype had a median of 1 (IQR 3) ICS and 1 (IQR 2) SABA prescriptions in the pre-index year, 93 % had records of ≥ 1 asthma reviews in the two-year pre-index period, and 11 % had CCI > 2 (the lowest across the six subtypes). The numbers of ICS and SABA prescriptions had a similar distribution in the "low-to-medium ICS use and comorbidities", which had the highest prevalence of comorbidity across the six subtypes (68 % had CCI > 2). The variation in the number of ICS and SABA prescriptions was greatest in the "varied ICS use and comorbid COPD", and 98 % of patients in this subtype had a record of COPD. The "high" and "very high ICS use" subtypes were principally defined by the number of ICS prescriptions in the pre-index year: 98 % of patients in the "high ICS use" had 5–8 ICS prescriptions in the pre-index year, and 99 % of patients in the "very high ICS use" had ≥ 9. These subtypes had varied SABA use, and the highest number of exacerbations across the six subtypes (14 % and 15 % respectively).

These subtypes generalised well at two future time-points (Fig. 3), and in an additional EHR database from a different UK nation (the SAIL Databank; Fig. 4). The characteristics of the subtypes were consistent across OPCRD and SAIL (Figs. A.14-A.19). This finding was reassuring, given the differences in the distribution of certain features in the datasets extracted from OPCRD and SAIL (e.g., there were typically more missing data, lower prevalence of comorbidities and fewer prescriptions for asthma medications in the datasets extracted from OPCRD compared to SAIL; Table A.2).

4.2. Findings in context

Asthma subtypes that have been consistently defined in previous studies were characterised by the age of disease onset, the presence of allergy and level of eosinophilic inflammation [5,6]. It was unsurprising that none of the subtypes identified in this study were characterised by these features, given the nature of primary care EHR data. Arguably the best approximation for age of asthma onset considered herein was the first date on which an asthma Read code was recorded, which does not necessarily correspond to true disease onset [28]. A full discussion of the limitations of EHR data for deriving features relating to age of disease onset, presence of allergy and the level of eosinophilic inflammation is given in Appendix A.3.1.

As the subtypes identified in this study were largely defined by the number of prescriptions for asthma medication, it is tempting to align these subtypes with the GINA treatment-based categorisation of severity [3]. We refrain from doing this for three reasons. First, our datasets did not include information on asthma medication dose. While the number of prescriptions could be used as a proxy for the dose, this is not validated. It would therefore be challenging to reliably infer the GINA category from a practical perspective. Second, EHRs do not accurately capture control of asthma symptoms. While SABA prescriptions, exacerbations, and the Royal College of Physicians three questions score may capture control to some extent, this gives an incomplete view. If we did consider these proxies to sufficiently capture asthma control, a substantial proportion of patients in each subtype would be considered to have uncontrolled symptoms, in whom the GINA characterisation should not be assessed. Third, the 2022 update of the GINA guidelines raises limitations of this severity characterisation, as described in our Introduction. In particular, the guidelines state that "the term 'mild asthma' should generally be avoided in clinical practice" and "for population-level observational studies, if clinical details are not available, describe the prescribed (or dispensed) treatment, without imputing severity" [3].

The emerging "treatable traits" model of care for airways diseases proposes that treatment for airways diseases goes beyond diagnostic labels such as asthma or COPD, taking a more complete view of the patient that considers pulmonary, extrapulmonary and lifestyle/behavioural traits [2,29]. While subtypes could help to facilitate the treatable traits model of care, it is clear from the limitations discussed in the following subsection and in Appendix A.3.1 that primary care EHRs would not provide a sufficient data source for deriving such subtypes. Future studies could extend the analytical framework proposed in this study to incorporate additional linked datasets.

Table 2
Statistical methods used to identify subtypes.

Method	Description
Multiple correspondence analysis (MCA)	MCA was applied to project the 45 categorical features to a lower-dimensional continuous feature space [20]. All 45 features were assigned equal weight in the MCA.To ensure that the projection was stable, four subsets, each of size 10,000 samples, were randomly selected from the training dataset without replacement. MCA was carried out on the training dataset, and independently on each of the four subsets. The scree plots [21] of the MCA results from the full dataset and each of the subsets were plotted and stability was subjectively assessed by the degree of overlap between the plots. Using the scree plot corresponding to MCA on the full dataset, a region was identified of acceptable dimensions to retain. Fig. A.1 (A) shows an illustration of this region, highlighted in yellow. The MCA scores were computed for up to the maximum number of dimensions in the identified region (the right-most yellow column in Fig. A.1 (B)). The correlation ratios for each of the original 45 features and each of the MCA scores were plotted. The distributions of the resulting scores were visualised using frequency polygons.
k-means	*K*-means cluster analysis was applied to the computed MCA scores. This was repeated for a range of MCA scores, as illustrated in Fig. A.1 (B). The stability of solutions for $k = 2, \ldots, 10$ were assessed using the framework outlined by Hennig et al. [22] This framework was implemented using the "clusterboot" function of the "fpc" R package[1], drawing 100 subsets each of size 10,000. Cluster solutions were considered stable if the average Jaccard coefficient was >0.85 (this cut-off is proposed by Hennig in the documentation for the "clusterboot" function of the "fpc" R package). The gap statistic, proposed by Tibshirani et al., [23] were calculated for all *k*. To speed up computation, the gap statistic was calculated for subsamples of size 10,000 (selected randomly without replacement). This was repeated over four different subsamples to assess whether findings were consistent across subsamples. The silhouette widths were calculated for all *k* for which the cluster solution was stable. The gap statistic and silhouette widths were used to select the best value of *k* (for each number of MCA dimensions) using the following rules of thumb: Based on the gap statistic, the best value of k was selected using the criterion outlined by Tibshirani et al. [23] If the solution corresponding to *k* was not stable based on the Jaccard coefficients calculated in the stability framework, the next best value of k was selected. Based on the silhouette widths, the value of k with the greatest average silhouette width was selected. Plots of the silhouette widths were also inspected to give a visual impression of cluster validity. This resulted in a value of *k* being selected for each number of MCA dimensions. Each of these solutions were taken forward to the classification step. If the Gap statistic and silhouette widths indicated different values of *k*, multiple solutions were taken forward.
Random forest (RF)	We trained a RF presenting the 45 categorical features as inputs and using the cluster labels as the output. We developed different RF models, using the cluster labels computed from the different possible cluster solutions as a function of the MCA features fed into k-means. The RF hyper-parameters (number of trees and number of features over which to optimize each split) were determined using the "tuneRF" function of the "randomForest" R package. [24] Each of the RF models were used to replicate the cluster labels in the test dataset. The ground truth cluster labels were derived in the test dataset by applying the clustering process described above, using the hyper-parameters determined from the train dataset. Jaccard coefficients were calculated between each of the replicated clusters and the ground truth cluster labels in the test dataset. The performance of the model was summarized by the average Jaccard coefficient. The final model (i.e. the selected dimensionality and number of clusters) was the model that maximised the average Jaccard coefficient between the replicated clusters and the true clusters. From here on, the outputs of the final model are referred to as subtypes.

[1] https://cran.r-project.org/web/packages/fpc/index.html.

4.3. Strengths

We used an analytical framework that comprised MCA to transform the categorical EHR-derived features into a more compact form, *k*-means to identify the clusters, and RF prediction models to investigate cluster label replication. This analytical framework addressed many limitations of previous asthma subtyping studies, identified in our recent review [10]. The use of EHR databases as the data source allowed for a large sample size (we used a training dataset of 50,000 patients from OPCRD, compared to the median sample size of 195 in the reviewed studies) [10]. We tested and validated our findings in a further three internal and three external datasets, each comprising 50,000 patients.

4.4. Limitations

Identifying a clinical cohort from EHR data is typically a trade-off between sensitivity and specificity. This is especially true for conditions such as asthma, for which diagnosis is challenging, error-prone, and can take months or years [30,31]. Our cohort comprised people with an asthma Read code in the previous three years, based on a previously validated method (although some alterations were made due to differences in coding systems) [32]. This method prioritises sensitivity (avoids erroneously excluding people with asthma), but as a result may include some people without asthma. The positive predictive value from the validation study on which this method was based was 86 % [32]. As in the validation study [32], we did not exclude people with COPD to avoid erroneously excluding people with co-existing asthma and COPD [33].

Difficulties in representing clinical features are limitations of using EHR data in research. We took various steps to mitigate biases associated with these limitations. To avoid misclassification bias, we used validated EHR phenotyping algorithms where available, and otherwise derived our own algorithms with input from clinical collaborators. We have provided our codelists in a public GitHub repository. As omitting patients with missing data could introduce selection bias (data are typically “missing not at random” in primary care EHRs [34]), we instead included this information as a feature category.

Despite steps to mitigate the limitations of EHR data, there remains the challenge that EHR data alone give an incomplete view of a patient’s condition. E.g., the features related to medication use were derived from counts of Read codes, and did not incorporate dosing instructions, whether the prescription was dispensed, the patient’s adherence to their medication nor their inhaler technique, as such information was unavailable in one or both databases. Additionally, information relating to signs and symptoms experienced by the patient are not reliably coded in EHRs. Finally, insights derived from EHRs are intrinsically biased according to diagnostic labels and clinician/patient behaviours during the period of data collection.

Table 3
Summary statistics for subtypes across key features.

Subtype	% of dataset[1]	Age[2]	ICS[2]	SABA[2]	Asthma review[3]	RCP3Q ≥ 1[4]	Exacerbation[5]	COPD[6]	CCI > 2[7]	Obese[8]
1. Low medication use and low healthcare utilisation	30 %	44 (26)	0 (0)	0 (0)	13 %	11 %	3 %	5 %	16 %	35 %
2. Low-to-medium medication use	36 %	45 (23)	1 (3)	1 (2)	93 %	58 %	7 %	3 %	11 %	31 %
3. Low-to-medium medication use and comorbidities	12 %	69 (16)	2 (4)	1 (3)	83 %	51 %	8 %	12 %	68 %	46 %
4. Varied medication use and comorbid COPD	4 %	70 (17)	7 (8)	5 (9)	73 %	63 %	9 %	98 %	52 %	33 %
5. High medication use	10 %	56 (24)	6 (2)	4 (5)	88 %	61 %	14 %	9 %	27 %	40 %
6. Very high medication use	7 %	60 (24)	12 (3)	8 (10)	89 %	69 %	15 %	13 %	37 %	43 %

[1] OPCRD 2016 test dataset.
[2] Median (interquartile rage) for Age and the number of inhaled corticosteroids (ICS) and short-acting beta-agonist (SABA) prescriptions recorded in pre-index year.
[3] Percentage of patients with one of more asthma reviews in the two-year pre-index period.
[4] Percent of patients with a Royal College of Physicians three questions (RCP3Q) score ≥ 1 (missing RCP3Q score grouped with RCP3Q score = 0 for percent calculation).
[5] Percentage of patients with an exacerbation recorded in the two-year pre-index period.
[6] Percentage of patients with a recording of chronic obstructive pulmonary disease (COPD) and chronic cardiac disease (CCD) ever.
[7] Percentage of patients with Charlson Comorbidity Index (CCI) > 2. Note that CCI was not included as a feature in the identification of the subtypes but is used here to succinctly summarise comorbidity across the clusters. See Table A.1 for the full list of comorbidity features used to identify subtypes.
[8] Percentage of patients with body mass index (BMI) ≥ 30 kg/m^2 (patients with missing BMI excluded from percentage calculation).

4.5. Clinical implications

This study provides evidence that subtypes of asthma are identifiable from primary care EHRs, and that they generalise across time-points and EHR databases. The identified subtypes have clear clinical interpretations, but there remains heterogeneity within the subtypes that requires further information to unpick. The clinical value of the subtypes at this stage is that they point to the information required to enable clinically meaningful actions, as illustrated by the following two examples.

The "low ICS use and low healthcare utilisation" is characterised by lack of data, due to the patient having little interaction with primary care. For some patients, this may reflect well-controlled asthma or a misdiagnosis of asthma. However, 3 % of patients in this subtype had an exacerbation recorded in the pre-index year, and 3 % had five or more SABA prescriptions recorded in the pre-index year, suggesting uncontrolled asthma in a small number of patients. This subtype could also be indicative of wider issues such as the fragmentation of healthcare records, with separate records for out of hours care, accident and emergency, and hospital care. Although the data collected when patients visit these different services should be captured in the primary care EHR, the frequency and accuracy of this record transfer is variable [19,35]. Therefore, the identification of this subtype has the potential to flag barriers to care, misdiagnoses, and data transfer issues.

At the other end of the spectrum, the "very high ICS use" subtype could flag a different range of issues. Patients assigned this subtype with low SABA use could be considered for stepping down treatment, in line with the GINA strategy of finding the patient's minimum effective level of treatment to minimise potential side-effects of medication and reduce medication costs [36,37]. High SABA use in this subtype (a proxy for poor symptom control) suggests that the patient may benefit from a medication review to assess inhaler technique, consider alterative controller options, and/or refer to a specialist asthma clinic. This medication pattern may indicate that the patient's asthma diagnosis requires re-evaluation, as dysfunctional breathing could be a symptom of another undiagnosed condition [38]. Consequently, identifying this subtype could facilitate better prioritisation and targeting of services to those with the greatest need and therefore inform service planning.

Using these subtypes to summarise asthma populations could help with management and resource planning at the practice level, and could be useful for understanding regional differences in the asthma population. E.g., although the same subtypes were identified in the OPCRD and SAIL datasets (samples from England and Wales respectively), in SAIL a higher proportion were assigned to the "very high ICS use" subtype (19.0 % compared to 7.4 % in OPCRD, Fig. A.19).

4.6. Conclusions

We have developed and validated a generalisable data-driven framework to define stable, reproducible, and clinically meaningful subtypes of adult asthma from datasets derived from two independent primary care EHR databases. This has important clinical implications towards assigning asthma subtypes, facilitating patient stratification, and developing more personalised monitoring and treatment regimens.

Summary table

What is already known on this topic?

- Asthma is an umbrella term for an unknown number of disease subtypes.
- Clinically relevant subtypes of asthma can be identified through the application of cluster analysis, and previous studies reported on small cohorts.

What this study added to our knowledge?

- This is the first study to demonstrate the application of cluster analysis to datasets derived from large longitudinal electronic health record (EHR) databases to identify subtypes of asthma.

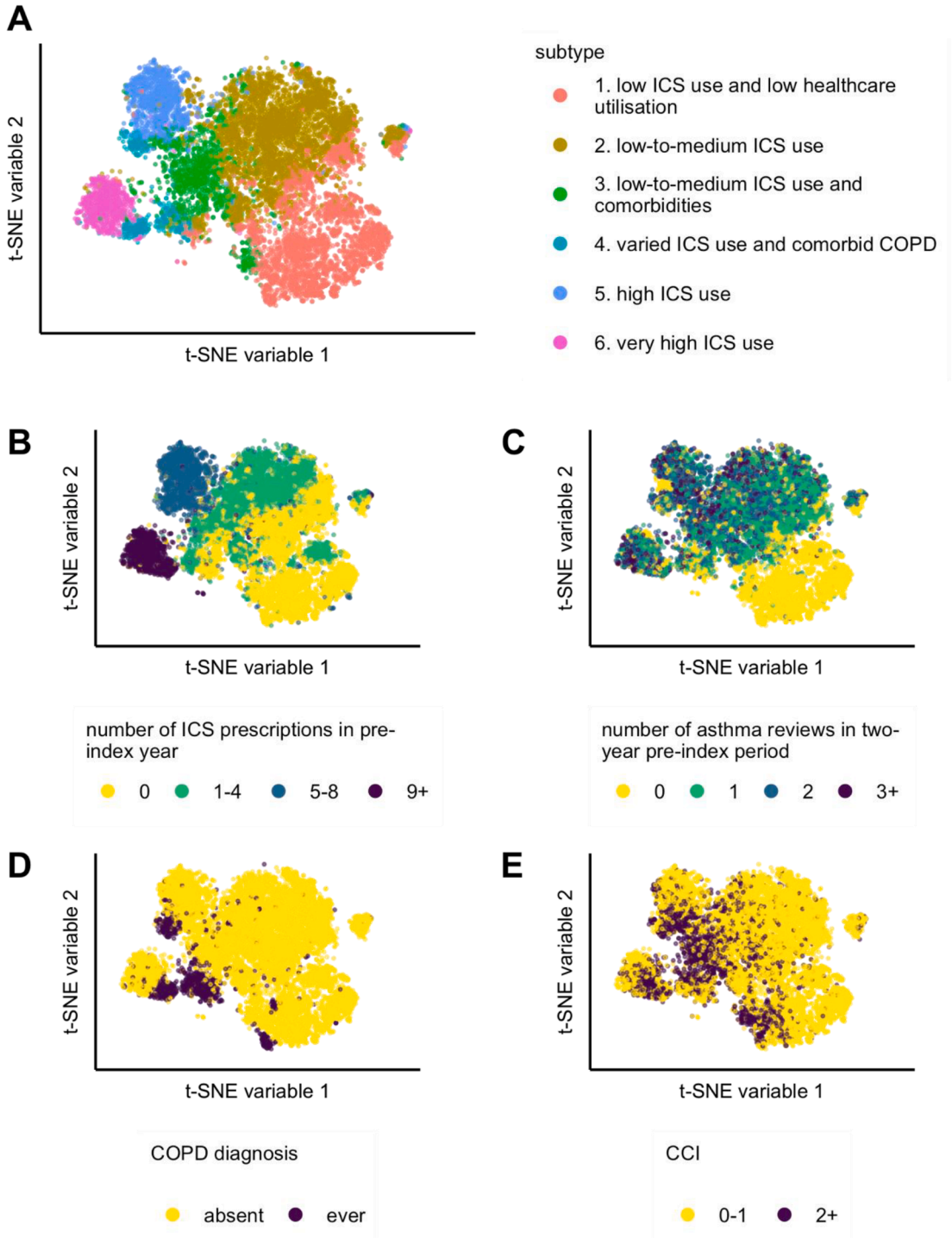


Fig. 2. *t*-SNE plots labelled by subtype (A), and EHR-derived features (B – E). CCI: Charlson comorbidity index; COPD: chronic obstructive pulmonary disease; ICS: inhaled corticosteroids. Note that CCI was not included as a feature in the identification of the subtypes but is used here to succinctly summarise comorbidity across the clusters. See Table A.1 for the full list of comorbidity features used to identify subtypes.

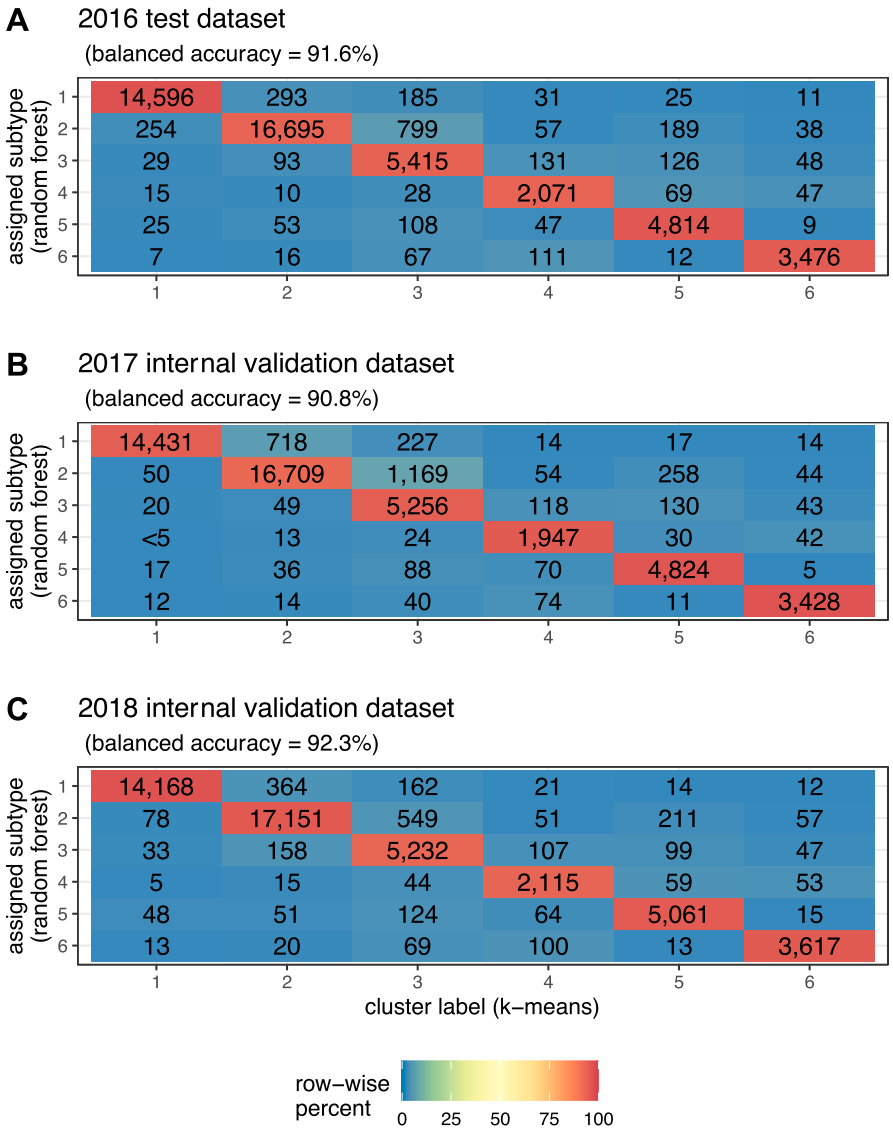


Fig. 3. Confusion matrices for OPCRD datasets. Subtype labels: (1) low ICS use and low healthcare utilisation; (2) low-to-medium ICS use; (3) low-to-medium ICS use and comorbidities; (4) varied ICS use and comorbid COPD; (5) high ICS use; (6) very high ICS use.

- Patient clusters based on information in primary care EHRs were characterised by level of healthcare utilisation, level of inhaled corticosteroid use and comorbidity.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: David Price has advisory board membership with AstraZeneca, Boehringer Ingelheim, Chiesi, Mylan, Novartis, Regeneron Pharmaceuticals, Sanofi Genzyme, Thermofisher; consultancy agreements with Airway Vista Secretariat, AstraZeneca, Boehringer Ingelheim, Chiesi, EPG Communication Holdings ltd, FIECON ltd, Fieldwork International, GlaxoSmithKline, Mylan, Mundipharma, Novartis, OM Pharma SA, PeerVoice, Phadia AB, Spirosure Inc, Strategic North Limited, Synapse Research Management Partners S.L., Talos Health Solutions, Theravance and WebMD Global LLC; grants and unrestricted funding for investigator-initiated studies (conducted through Observational and Pragmatic Research Institute Pte ltd) from AstraZeneca, Boehringer Ingelheim, Chiesi, Mylan, Novartis, Regeneron Pharmaceuticals, Respiratory Effectiveness Group, Sanofi Genzyme, Theravance and UK National Health Service; payment for lectures/speaking engagements from AstraZeneca, Boehringer Ingelheim, Chiesi, Cipla, GlaxoSmithKline, Kyorin, Mylan, Mundipharma, Novartis, Regeneron Pharmaceuticals and Sanofi Genzyme; payment for travel/accommodation/meeting expenses from AstraZeneca, Boehringer Ingelheim, Mundipharma, Mylan, Novartis, Thermofisher; stock/stock options from AKL Research and Development ltd which produces phytopharmaceuticals; owns 74 % of the social enterprise Optimum Patient Care ltd (Australia and UK) and 92.61 % of Observational and Pragmatic Research Institute Pte ltd (Singapore); 5 % shareholding in Timestamp which develops adherence monitoring technology; is peer reviewer for grant committees of the UK Efficacy and Mechanism Evaluation programme, and Health Technology Assessment; and was an expert witness for GlaxoSmithKline. All other authors have no conflict of interest to declare.

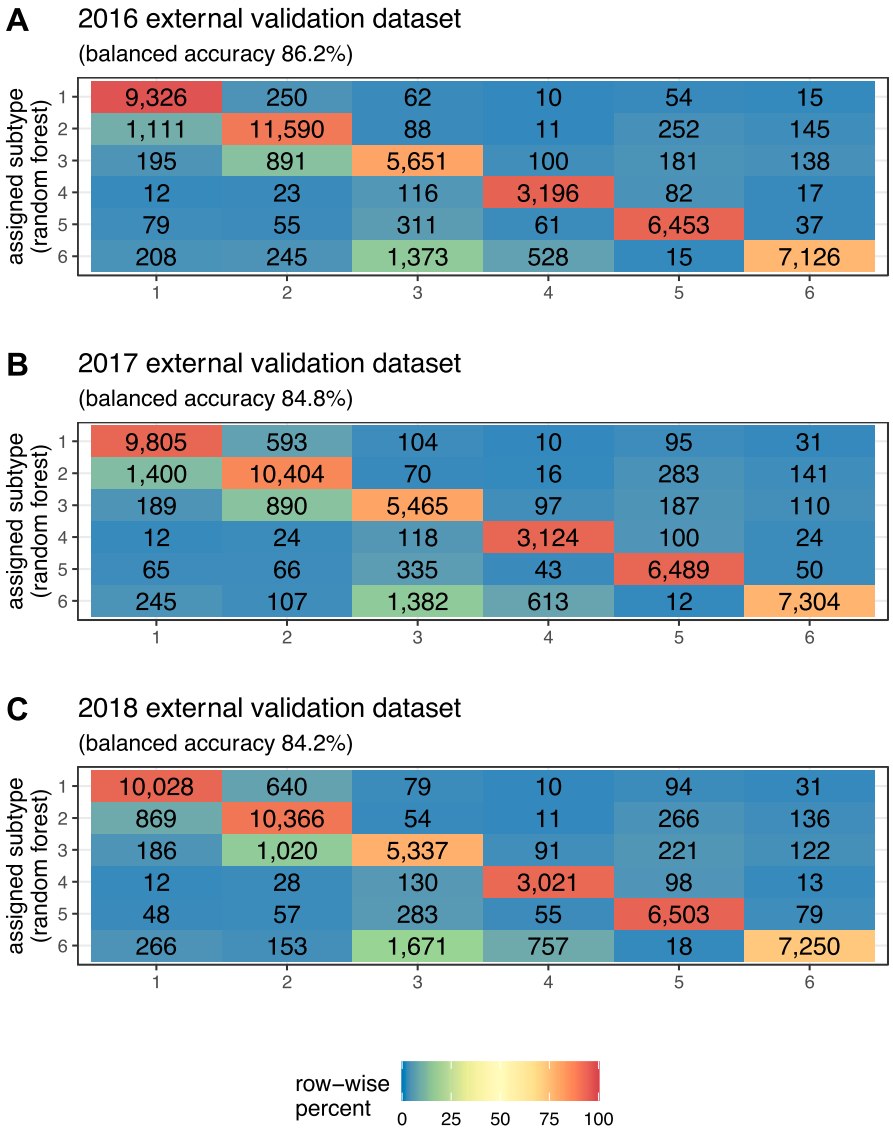


Fig. 4. Confusion plots for SAIL datasets. Subtype labels: (1) low ICS use and low healthcare utilisation; (2) low-to-medium ICS use; (3) low-to-medium ICS use and comorbidity; (4) varied ICS use and comorbid COPD; (5) high ICS use; (6) very high ICS use.

Acknowledgments

EMFH was supported by a Medical Research Council PhD Studentship (eHERC/Farr). This work is carried out with the support of the Asthma UK Centre for Applied Research [AUK-AC-2012-01] and Health Data Research UK which receives its funding from HDR UK Ltd (HDR-5012) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and the Wellcome Trust. The funders had no role in the study and the decision to submit this work to be considered for publication.

This Project is based in part/wholly on Data from the Optimum Patient Care Research Database (opcrd.co.uk) obtained under licence from Optimum Patient Care Limited and its execution is approved by recognised experts affiliated to the Respiratory Effectiveness Group. However, the interpretation and conclusion contained in this report are those of the author/s alone.

This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers who make anonymised data available for research. SAIL is not responsible for the interpretation of these data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2022.104942.

References

[1] GBD 2019 Disease and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396:1204–22. 10.1016/S0140-6736(20)30925-9.

[2] I.D. Pavord, R. Beasley, A. Agusti, G.P. Anderson, E. Bel, G. Brusselle, et al., After asthma: redefining airways diseases, Lancet 391 (2017) 10118, https://doi.org/10.1016/S0140-6736(17)30879-6.

[3] Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention, 2022.

[4] Royal College of Physicians. Why asthma still kills: The National Review od Asthma Deaths (NRAD). vol. 33. 2012. 10.1055/s-0032-1326964.

[5] M.E. Kuruvilla, F.E.H. Lee, G.B. Lee, Understanding asthma phenotypes, endotypes, and mechanisms of disease, Clin. Rev. Allergy Immunol. (2018) 2, https://doi.org/10.1007/s12016-018-8712-1.

[6] R. Kaur, G. Chupp, Phenotypes and endotypes of adult asthma: Moving toward precision medicine, J. Allergy Clin. Immunol. 144 (2019) 1–12, https://doi.org/10.1016/j.jaci.2019.05.031.

[7] E. Ahlqvist, P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, et al., Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables, Lancet Diabetes Endocrinol. 6 (2018) 361–369, https://doi.org/10.1016/S2213-8587(18)30051-2.

[8] R.W. Grant, J. McCloskey, M. Hatfield, C. Uratsu, J.D. Ralston, E. Bayliss, et al., Use of latent class analysis and k-means clustering to identify complex patient profiles, JAMA Netw. Open 3 (2020) 1–13, https://doi.org/10.1001/jamanetworkopen.2020.29068.

[9] K.E. Nnoaham, K.F. Cann, Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population? BMC Public Health 20 (2020) 1–10, https://doi.org/10.1186/s12889-020-08930-z.

[10] E. Horne, H. Tibble, A. Sheikh, A. Tsanas, Challenges of clustering multimodal clinical data: a review of applications in asthma subtyping, JMIR Med. Informatics 8 (2020) e16452.

[11] G. Colice, A. Chisholm, A.L. Dima, H.K. Reddel, A. Burden, R.J. Martin, et al., Performance of database-derived severe exacerbations and asthma control measures in asthma: responsiveness and predictive utility in a UK primary care database with linked questionnaire data, Pragmatic Obs. Res. 9 (2018) 29–42, https://doi.org/10.2147/por.s151615.

[12] R.A. Lyons, K.H. Jones, G. John, C.J. Brooks, J.-P. Verplancke, D.V. Ford, et al., The SAIL databank: linking multiple health and social care datasets, BMC Med. Inf. Decis. Making 9 (2009) 3, https://doi.org/10.1186/1472-6947-9-3.

[13] D.V. Ford, K.H. Jones, J.-P. Verplancke, R.A. Lyons, G. John, G. Brown, et al., The SAIL Databank: building a national architecture for e-health research and evaluation, BMC Health Serv. Res. 9 (2009) 157, https://doi.org/10.1186/1472-6963-9-157.

[14] K.H. Jones, D.V. Ford, C. Jones, R. Dsilva, S. Thompson, C.J. Brooks, 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, https://doi.org/10.1016/j.jbi.2014.01.003.

[15] S.E. Rodgers, R.A. Lyons, R. Dsilva, K.H. Jones, C.J. Brooks, D.V. Ford, et al., Residential Anonymous Linking Fields (RALFs): a novel information infrastructure to study the interaction between the environment and individuals' health, J. Public Health (Oxf.) 31 (2009) 582–588, https://doi.org/10.1093/pubmed/fdp041.

[16] S.E. Rodgers, J.C. Demmler, R. Dsilva, R.A. Lyons, Protecting health data privacy while using residence-based environment and demographic data, Health Place 18 (2012) 209–217, https://doi.org/10.1016/j.healthplace.2011.09.006.

[17] D. Thayer, A. Rees, J. Kennedy, H. Collins, D. Harris, J. Halcox, et al., Measuring follow-up time in routinely-collected health datasets: Challenges and solutions, PLoS One 15 (2020) 1–11, https://doi.org/10.1371/journal.pone.0228545.

[18] J.D. Blakey, D.B. Price, E. Pizzichini, T.A. Popov, B.D. Dimitrov, D.S. Postma, et al., Identifying Risk of Future Asthma Attacks Using UK Medical Record Data: A Respiratory Effectiveness Group Initiative, J. Allergy Clin. Immunol. Pr. 5 (2017) 1015–1024.e8, https://doi.org/10.1016/j.jaip.2016.11.007.

[19] D. Ryan, J. Blakey, A. Chisholm, D. Price, M. Thomas, B. Ställberg, et al., Use of electronic medical records and biomarkers to manage risk and resource efficiencies, Eur. Clin. Respir. J. 4 (2017) 1293386, https://doi.org/10.1080/20018525.2017.1293386.

[20] F. Husson, S. Lê, J. Pagès, Exploratory Multivariate Analysis by Example Using R, CRC Press, 2017.

[21] R.B. Cattell, The Scree Test For The Number Of Factors, Multivariate Behav. Res. 1 (1966) 245–276, https://doi.org/10.1207/s15327906mbr0102_10.

[22] C. Hennig, Cluster-wise assessment of cluster stability, Comput. Stat. Data Anal. 52 (2007) 258–271, https://doi.org/10.1016/j.csda.2006.11.025.

[23] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Soc. Ser. B Stat Methodol. 63 (2001) 411–423, https://doi.org/10.1111/1467-9868.00293.

[24] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (2002) 18–22.

[25] L. van der Maaten, Accelerating t-SNE using Tree-Based Algorithms. vol. 15. MIT Press; 2014.

[26] E. von Elm, D.G. Altman, M. Egger, S.J. Pocock, P.C. Gotzsche, J. P. Vandenbroucke, Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies, BMJ 335 (2007) 806–808, https://doi.org/10.1136/bmj.39335.541782.AD.

[27] E.I. Benchimol, L. Smeeth, A. Guttmann, K. Harron, D. Moher, I. Petersen, et al., The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement, PLoS Med. 12 (2015) e1001885.

[28] F. Nissen, I.J. Douglas, H. Müllerová, N. Pearce, C.I. Bloom, L. Smeeth, et al., Clinical profile of predefined asthma phenotypes in a large cohort of UK primary care patients (Clinical Practice Research Datalink), J. Asthma Allergy 12 (2019) 7–19, https://doi.org/10.2147/JAA.S182013.

[29] A. Agusti, N. Barnes, A.A. Cruz, P.G. Gibson, L.G. Heaney, H. Inoue, et al., Moving towards a Treatable Traits model of care for the management of obstructive airways diseases, Respir. Med. 187 (2021), 106572, https://doi.org/10.1016/j.rmed.2021.106572.

[30] B. Marklund, A. Tunsäter, C. Bengtsson, How often is the diagnosis bronchial asthma correct? Fam. Pract. 16 (1999) 112–116, https://doi.org/10.1093/fampra/16.2.112.

[31] A. Akindele, L. Daines, D. Cavers, H. Pinnock, A. Sheikh, Qualitative study of practices and challenges when making a diagnosis of asthma in primary care, Npj Prim. Care Respir. Med. (2019) 29, https://doi.org/10.1038/s41533-019-0140-z.

[32] F. Nissen, D.R. Morales, H. Mullerova, L. Smeeth, I.J. Douglas, J.K. Quint, Validation of asthma recording in the Clinical Practice Research Datalink (CPRD), BMJ Open 7 (2017) 1–8, https://doi.org/10.1136/bmjopen-2017-017474.

[33] J.B. Soriano, K. Davis, B. Coleman, G. Visick, D. Mannino, N.B. Pride, The Proportional Venn Diagram of Obstructive Lung Disease, Chest (2003) 124, https://doi.org/10.1378/chest.124.2.474.

[34] M. Phelan, N.A. Bhavsar, A. Goldstein, Illustrating Informed Presence Bias in Electronic Health Records Data : How Patient Interactions with a Health System Can Impact Inference, Gener Evid Methods to Improv Patient Outcomes (2017) 5, https://doi.org/10.5334/egems.243.

[35] K.J. Rothnie, H. Mullerova, S.L. Thomas, J.S. Chandan, L. Smeeth, J.R. Hurst, et al., Recording of hospitalizations for acute exacerbations of COPD in UK electronic health care records, Clin. Epidemiol. 8 (2016) 771–782, https://doi.org/10.2147/clep.s117867.

[36] C.I. Bloom, L. de Preux, A. Sheikh, J.K. Quint, Health and cost impact of stepping down asthma medication for UK patients, 2001–2017: A population-based observational study, PLoS Med. 17 (2020) 2001–2017, https://doi.org/10.1371/journal.pmed.1003145.

[37] Global Initiative for Asthma. Global initiative for asthma: Asthma management and prevention, 2019. 2019.

[38] L.P. Boulet, Influence of comorbid conditions on asthma, Eur. Respir. J. 33 (2009) 897–906, https://doi.org/10.1183/09031936.00121308.