npj | Primary Care Respiratory Medicine

www.nature.com/npjpcrm

PROTOCOL OPEN

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[1,2], Sarah E Rodgers[1,3], Ronan A Lyons[1,3], Aziz Sheikh[2,3,4] and Gwyneth A Davies[1,2]

Asthma and chronic obstructive pulmonary disease (COPD) are two common different clinical diagnoses with overlapping clinical features. Both conditions have been increasingly studied using electronic health records (EHR). Asthma-COPD overlap syndrome (ACOS) is an emerging concept where clinical features from both conditions co-exist, and for which, however, there is no consensus definition. Nonetheless, we expect EHR data of people with ACOS to be systematically different from those with "asthma only" or "COPD only". We aim to develop a latent class model to understand the overlap between asthma and COPD in EHR data. From the Secure Anonymised Information Linkage (SAIL) databank, we will use routinely collected primary care data recorded in or before 2014 in Wales for people who aged 40 years or more on 1st Jan 2014. Based on this latent class model, we will train a classification algorithm and compare its performance with commonly used objective and self-reported case definitions for asthma and COPD. The resulting classification algorithm is intended to be used to identify people with ACOS, 'asthma only', and 'COPD only' in primary care datasets.

npj Primary Care Respiratory Medicine (2018)28:22 ; doi:10.1038/s41533-018-0088-4

BACKGROUND

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 limitation.[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 recognised 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 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 explains 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 STROBE[25] and RECORD Statements[26] in reporting the full study.

Data sources

We will use the following two deidentified datasets from the Secure Anonymised Information Linkage (SAIL) Databank in Wales:[27,28]

[1]Swansea University Medical School, Singleton Park, Swansea SA2 8PP, UK; [2]Asthma UK Centre for Applied Research, Edinburgh and Swansea, UK; [3]The Farr Institute of Health Informatics Research, Edinburgh and Swansea, UK and [4]Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK
Correspondence: Mohammad A Al Sallakh (M.A.Alsallakh@swansea.ac.uk)

Received: 11 October 2017 Revised: 23 May 2018 Accepted: 23 May 2018
Published online: 20 June 2018

- The Welsh Demographic Service (WDS) which contains demographic and administrative information for the National Health Services (NHS) patients in Wales.
- The General Practitioner (GP) dataset which contains primary care events, such as diagnoses, clinical findings, and 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 1st January 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 will 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 will be 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 partitioning[35] 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 1st January 2014, and whose responses where 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.

Table 1. Observed variables that will be used in the latent class model

Variable	Time interval for calculation	Categories
Asthma related		
Asthma diagnosis codes	Ever	0, 1+
Age at asthma first diagnosis codes (if any)	–	<40, ≥40, no diagnosis
Asthma GP visits codes	Last year	0, 1+
COPD related		
COPD diagnosis codes	Ever	0, 1+
COPD GP visits codes	Last year	0, 1+
COPD-specific prescriptions codes*	Last year	0, 1+
Prescriptions		
ICS codes	Last year	0, 1+
SABA codes	Last year	0, 1+
LABA codes	Last year	0, 1+
ICS+LABA codes	Last year	0, 1+
OCS codes	Last year	0, 1+
LTRA codes	Last year	0, 1+
Others		
Allergy history**	Ever	No, yes
Smoking history	Ever	No, yes
Gender	–	Male, female

Abbreviations: *COPD* = chronic obstructive pulmonary disease, *ICS* = inhaled corticosteroids, *GP* = general practitioner, *LTRA* = leukotriene receptor antagonists, *LABA* = long-acting $\beta2$ agonists, *OCS* = oral corticosteroids, *SABA* = short-acting $\beta2$ 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.

**Allergy includes atopic eczema/dermatitis, food allergy, allergic rhinitis, and anaphylaxis.

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 the LCA model and the classification algorithm against other objective and self-reported measures will provide useful information about their validity and performance.

ACKNOWLEDGEMENTS

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).

AUTHOR CONTRIBUTIONS

All authors contributed to refinement of the study protocol and manuscript writing and critically reviewed and approved the final manuscript. M.A.A.S developed the statistical methods.

ADDITIONAL INFORMATION

Competing interests: The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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, F. E. & Nair, P. The definition and diagnosis of asthma. *Clin. Exp. Allergy* **39**, 1652–1658 (2009).
5. The Lancet. A plea to abandon asthma as a disease concept. *Lancet* **368**, 705 (2006). https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(06)69257-X.pdf.
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, H. K. 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, I. L., van der Zeijden, H. & Verheij, T. J. Is spirometry essential in diagnosing asthma? Yes. *Br. J. Gen. Pract.* **66**, 484–484 (2016).
9. Levy, M. L. 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, E. D., Reddel, H. K., van Zyl-Smit, R. N.& 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, A. D., Sharafkhaneh, A. & Hanania, N. A. The asthma COPD overlap syndrome (ACOS). *Curr. Allergy Asthma Rep.* **15**, 509 (2015).
14. McDonald, V. M. & Gibson, P. G. To define is to limit: perspectives on asthma–COPD overlap syndrome and personalised medicine. *Eur. Respir. J.* **49**, 1700336 (2017).
15. Miravitlles, M. Diagnosis of asthma–COPD overlap: the five commandments. *Eur. Respir. J.* **49**, 1700506 (2017).
16. Ford, E. S. The epidemiology of obesity and asthma. *J. Allergy Clin. Immunol. Pract.* **115**, 897–909 (2005). quiz 910.
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**, 113 (2016).
18. Bonten, T. N. 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. *Pub. Health Res. Prac.* **25**, e2541540 (2015).
23. Al Sallakh, M.A. et al. Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. *Eur. Respir. J.* **49**, 1700204 (2017).
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, E. I. et al. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLOS Med.*.**12**, e1001885 (2015).
27. Lyons, R. A. et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform. Decis. Mak.* **9**, 3 (2009).
28. Ford, D. V. et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv. Res* **9**, 157 (2009).
29. Collins, L.M. & Lanza S.T. Latent class and latent transition analysis: with applications in the social, behavioral, and health sciences (John Wiley & Sons Inc, Hoboken, NJ, 2010). ISBN: 0470228393..
30. Dempster, A.P., Laird, N.M., & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Stat Methodol* **39**, 1–38 (1977).
31. McLachlan G. & Peel D. *Finite Mixture Models* 1st ed (Wiley-Interscience, New York, NY, 2000). ISBN: 9780471006268.
32. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).

33. Nylund, K. L., Asparouhov, T. & Muthén, B. O. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Struct. Equ. Model.* **14**, 535–569 (2007).
34. Linzer, D. A. & Lewis, J. B. 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 T. M. & Atkinson E. J. 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, G. J., 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, W. C. 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* **70**, 973–984 (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, E.M.S., Jaakkola, M.S., & Jaakkola, J.J.K. Subtypes of asthma based on asthma control and severity: a latent class analysis. *Respir. Res.* **18**, 24 (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, P. R. et al. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur. Respir. J.* **36**, 531–539 (2010).
47. Ghebre, M. A. et al. Biological clustering supports both Dutch and British hypotheses of asthma and chronic obstructive pulmonary disease. *J. Allergy Clin. Immunol.* **135**, 63–72 (2015). e10.