Identifying COPD in routinely collected electronic health records: a systematic scoping review

Shanya Sivakumaran [1], Mohammad A. Alsallakh[1], Ronan A. Lyons[1], Jennifer K. Quint [2] and Gwyneth A. Davies[1]

[1]Population Data Science, Swansea University Medical School, Swansea, UK. [2]National Heart and Lung Institute, Imperial College London, London, UK.

Corresponding author: Shanya Sivakumaran (shanya.sivakumaran@swansea.ac.uk)

Shareable abstract (@ERSpublications)
Inconsistency in methods of identifying COPD in electronic health records and the lack of clinically important variables in healthcare databases widely used for research are persisting constraints in harnessing the potential of EHRs worldwide https://bit.ly/2TtqNKV

Cite this article as: Sivakumaran S, Alsallakh MA, Lyons RA, *et al*. Identifying COPD in routinely collected electronic health records: a systematic scoping review. *ERJ Open Res* 2021; 7: 00167-2021 [DOI: 10.1183/23120541.00167-2021].

This article has supplementary material available from openres.ersjournals.com

Received: 10 March 2021
Accepted: 24 June 2021

Abstract

Although routinely collected electronic health records (EHRs) are widely used to examine outcomes related to COPD, consensus regarding the identification of cases from electronic healthcare databases is lacking. We systematically examine and summarise approaches from the recent literature.

MEDLINE *via* EBSCOhost was searched for COPD-related studies using EHRs published from January 1, 2018 to November 30, 2019. Data were extracted relating to the case definition of COPD and determination of COPD severity and phenotypes.

From 185 eligible studies, we found widespread variation in the definitions used to identify people with COPD in terms of code sets used (with 20 different code sets in use based on the ICD-10 classification alone) and requirement of additional criteria (relating to age (n=139), medication (n=31), multiplicity of events (n=21), spirometry (n=19) and smoking status (n=9)). Only seven studies used a case definition which had been validated against a reference standard in the same dataset. Various proxies of disease severity were used since spirometry results and patient-reported outcomes were not often available.

To enable the research community to draw reliable insights from EHRs and aid comparability between studies, clear reporting and greater consistency of the definitions used to identify COPD and related outcome measures is key.

Introduction

COPD is a common, chronic condition characterised by persistent respiratory symptoms and irreversible expiratory airflow limitation, usually caused by chronic exposure to inhaled noxious substances. In clinical practice, COPD can be diagnosed in patients suspected to have the condition by use of spirometry, which also aids in the assessment of disease severity. Patients with COPD can also be grouped by their pattern of exacerbations and symptom burden, and this phenotyping guides treatment decisions [1]. Guidelines produced by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) divide patients into "GOLD groups" based on these characteristics when delineating treatment strategies.

Research using routinely collected data from electronic health records (EHRs) and administrative databases to study COPD has seen an upsurge in recent years, as the wealth of data accumulated as a by-product of routine clinical care has made large, diverse populations accessible to researchers. However, this data has not been generated for the purpose of research, and important information such as spirometry results and patient-reported outcome measures are not often accessible in the data sources. Alternative measures are thus frequently used to identify individuals with COPD, as well as determine disease severity and phenotypes, though the extent to which the definitions used have been assessed for validity is unclear.

Although the need to focus on the accuracy of case definitions has been emphasised [2], there is still significant heterogeneity in the definitions used to identify common conditions in routinely collected data [3–5]. In this systematic scoping review, we sought to summarise the range of methods used to identify COPD, its severity and phenotypes in EHRs, and determine what proportion of case definitions in use have been validated against reference standards.

Methods

We conducted a systematic scoping review [6] to answer our research questions: 1) how were individuals with COPD identified in EHRs in the recent literature, 2) how many methods of case identification had been validated against reference standards, 3) how were COPD severity and phenotypes defined and 4) what important data are missing from the data sources used in these studies?

Search strategy and eligibility criteria

A broad search strategy was developed to gather studies which used EHRs to identify individuals with COPD (supplementary table S1). MEDLINE *via* EBSCOhost was searched January 15, 2020 for articles published between January 1, 2018 and November 30, 2019. Our search was limited to those written in the English language. There were no limitations as to study design.

EHRs included routinely collected, individual level data from administrative databases, disease registries, electronic health records and any other electronic databases that were generated as a by-product of routine healthcare. Studies using solely survey or trial data were excluded, along with studies not reporting original data. We included studies identifying a study population of individuals with COPD or using COPD as a primary outcome, but not those where COPD was just contained in a list of covariates.

Study selection and data extraction

Articles that did not fit the above eligibility criteria were excluded. Screening was initially conducted using titles and abstracts, and full-texts were accessed when necessary. Information extracted from articles deemed eligible for inclusion related to core study details, definitions of COPD diagnosis, severity and phenotypes, and quality appraisal (supplementary table S2).

Article screening and data extraction were performed independently by two authors (S.S. and M.A.A.) for 20% of studies. S.S. then completed screening and extraction, with discussion with the wider study group when necessary.

Results

Our search strategy identified 1226 articles for screening (supplementary table S1), of which 189 met our eligibility criteria. We were able to access the full text for 185 of these, which are included in our review. Most studies were conducted in North America, Taiwan and the UK (supplementary table S3). We included studies with a range of designs, including retrospective cohort, self-controlled case series, quasi-experimental, nested case–control, case crossover and descriptive/exploratory studies.

Identifying COPD

Studies often identified individuals with COPD using clinical codes. The most frequently used coding scheme was the International Statistical Classification of Diseases and Related Health Problems (ICD) [7], either alone or in conjunction with other coding schemes (supplementary table S4). However, studies using the same coding scheme did not always use the same list of codes ("code set") from within the scheme – 57 studies incorporating ICD-10-based codes used 20 different code sets to detect COPD (supplementary table S5). Some studies did not report the specific code set used for case identification.

In order to achieve greater accuracy of case definitions, many studies used additional inclusion and exclusion criteria in their definition of COPD. 139 (75%) used age as a criterion, with the lower age limit varying from 18 to 66 (supplementary table S6). Some studies (21; 11%) required multiple COPD-related event or claim codes. Some (25; 14%) gave more weight to inpatient care codes, requiring multiple COPD-related codes if arising from primary care or outpatient care, but only one if arising from inpatient care. 19 (10%) mandated presence of a spirometry code but results of spirometry were not always taken into account. Nine (5%) specified ever-smoking as a criterion (supplementary table S7). A COPD-related medication code was required by 31 out of 171 (18%) studies (not including studies whose aim was to investigate COPD medications, since these would have automatically required presence of the medication). The specific medication requirements and reporting of this varied by study, from requiring "the prescription of at least one bronchodilator" [8] to mandating a greater frequency of medication use, with "COPD medication use at least twice per year" [9] (where COPD medications were "long-acting

muscarinic antagonist, long-acting beta-2 agonist (LABA), inhaled corticosteroid (ICS), ICS plus LABA, short-active muscarinic antagonist (SAMA), short-acting beta-2 agonist (SABA), SAMA plus SABA, methylxanthines, systemic corticosteroids and systemic beta agonists") [9]. One study stated only that patients were required to have been prescribed a "respiratory medication" but did not elaborate further [10].

With regard to exclusion criteria, 25 (14%) studies excluded those with a previous asthma diagnosis, some excluded additional comorbid respiratory conditions and a few excluded individuals using specific medications, such as leukotriene receptor antagonists which are mostly used in people with asthma.

COPD severity

Spirometry was utilised by 25 (14%) studies in their assessment of COPD severity, 8 (4%) as a binary measure, 17 (11%) as an ordinal measure. Most studies did not assess disease severity in any form, specifying that this was because they lacked the clinical data necessary. Proxies of severity were sometimes used, ranging from chronic medication use (n=16; 9%) and measures relating to exacerbations (n=16; 9%), to serum bicarbonate levels (n=1; 0.5%) [11], to algorithms purporting to represent "complexity" (n=1; 0.5%) [12].

COPD phenotypes

Coexisting asthma was a phenotype examined by 15 (8%) studies and was generally identified by presence of a previous asthma diagnosis code, but the specific code sets or identification methods were not always reported. 12 (6%) studies compared those with high *versus* low blood eosinophil counts or concentrations, although the thresholds used to determine high and low differed by study (supplementary table S8) [13–15]. Three studies performed sensitivity analyses to examine the effect of using differing thresholds [16–18]. Eight (4%) studies categorised individuals by their GOLD groups (*i.e.* taking into account both exacerbation history and symptom burden), and 18 (10%) examined individuals by exacerbation history. Again, there was variation in the code sets or algorithms used to identify an exacerbation, and the thresholds for high *versus* low exacerbators.

Validation of case definitions

Eight (4%) studies in our review compared their definition of COPD against a reference standard and provided sufficient information that a measure of validity could be calculated, although this may not have been their primary purpose. Two of these studies identified themselves as "validation studies" and went on to report measures of validity [19, 20].

Of the remaining 177 studies, a further 7 (4%) used a definition of COPD that had been previously validated against a reference standard in the same database used for their research. Additional studies referred to their case definition being "based on" validated definitions, but used code sets different to those validated [21–23], or did not report the codes they used [18, 24, 25].

Some studies conducted analyses to justify the validity of their findings in different ways, such as performing sensitivity analyses using different definitions of COPD [26].

Reporting

Only one study referred to the REporting of studies Conducted using Observational Routinely collected Data (RECORD) guidance [27]. 15 (8%) studies stated that they used a particular coding scheme to identify people with COPD but did not report the code set used. 107 (58%) studies did not report whether they could access data related to (one of) smoking or spirometry. 44 (24%) reported that smoking information was not available within their data source (supplementary table S7), and 60 (32%) reported that spirometry events were unavailable.

Discussion

Principal findings

Electronic databases of routinely collected health data are used internationally to advance knowledge about "real world" COPD by the research community. This systematic scoping review has demonstrated significant variability in the methods used by researchers to identify individuals with COPD and describe disease severity and phenotypes using routine data.

Only a limited number of studies used definitions that had been validated against reference standards in the same database used for their study. Some studies referred to previous validation studies, but as they did not report the code list they had used, it was not clear whether they used the same validated case definition. The RECORD guidance [27] advocates for provision of a "complete list of codes and algorithms used to

classify exposures, outcomes, confounders, and effect modifiers" in order to enhance research transparency.

Interpretation and implications

Datasets generated as a by-product of routine clinical care are increasingly important and useful in research, given the size, heterogeneity and unselected nature of the populations they provide access to. However, there are pitfalls to their use, and among them is the use of case definitions with unknown validity. Limitations in the reporting of code sets used by researchers further hinder comparability and reproducibility.

For EHRs to provide meaningful insights, the case definitions used must be able to accurately detect individuals with the condition in question. One way of ensuring this is to use definitions validated against a reference standard. However, a definition validated in one database may not be transferable for use in others. One validated definition for COPD in the UK's Clinical Practice Research Datalink [28] specifies being a current or ex-smoker as an inclusion criterion, given that COPD is uncommon in never-smokers in the UK. However, in countries with a higher contribution of alternative risk factors to the development of COPD [29], necessitating being an ever-smoker would likely reduce the sensitivity of this definition, as will happen in all geographies where smoking prevalence is falling. Additionally, different research questions may necessitate different case definitions – if investigators wanted to prioritise specificity over sensitivity, a more restrictive definition would be used, and vice versa – but clarity in the rationale would be useful to readers.

In addition to case definitions, the availability and accuracy of disease severity and phenotyping measures is imperative for studies to be able to adequately adjust for potential confounding. However, many administrative databases do not contain clinical detail at this level, so attempts at adjusting for severity in analyses often use proxy measures, the choice of which may be determined by data availability and not be validated against "true" disease severity. More often, no attempt at adjusting for disease severity is made, leading to the potential of unmeasured confounding influencing results. Facilitating inclusion of clinically important variables (such as common investigation results and patient-reported outcome measures) into electronic health databases commonly used for research would be a useful and important intervention in improving research outputs. Through the lens of COPD, inclusion of spirometry results and UK Medical Research Council (MRC) dyspnoea scale scores would play a significant role in advancing research in the field. Similarly, inclusion of echocardiogram results and New York Heart Association (NYHA) class would likely be helpful for cardiovascular disease research. However, even when databases do hold such information, there may be a high rate of missing data (*e.g.* 65% of patients in one study had no spirometry recorded) [30]. This reflects real world patterns, and levels of missingness are likely to vary geographically due to historic differences in clinical practice, or national incentive schemes.

Strengths and limitations

This is the first review, to our knowledge, to systematically examine methods of identifying individuals with COPD in routinely collected EHRs. This approach has allowed us to objectively demonstrate the variability in research practice in the field. We applied broad inclusion criteria ensuring representation across the research field but confined our review to recent literature in order to ensure relevancy. We did not include studies where COPD was only relevant due to being contained in a list of covariates (as is often done, for example as part of the Charlson Comorbidity Index) [31], since an accurate case definition for this purpose holds less importance. However, this means that our review does not fully encompass the whole spectrum of the use of "COPD" in electronic health data research.

Conclusions

Although the interrogation of routinely collected EHRs is now commonplace in investigating important research questions related to COPD, and provides huge value when used carefully, persistent limitations constrain the quality of this research. The lack of clinically important variables in widely used databases limits researchers' ability to adjust for confounders such as disease severity. Variation in methods to identify COPD and define outcome measures restricts comparability between studies. With the contribution of EHRs to COPD research continuing to increase internationally, ensuring greater consistency of case definitions and optimisation of reporting is key to enhancing the reliability of research outputs.

Provenance: Submitted article, peer reviewed.

Ethics: Ethical review was not required since this study was a review of previously published work.

Conflict of interest: S. Sivakumaran has nothing to disclose. M.A. Alsallakh has nothing to disclose. R.A. Lyons reports payment to Swansea University to support research using population health datasets from Health Data Research UK (HDRUK), outside the submitted work. J.K. Quint has nothing to disclose. G.A. Davies has nothing to disclose.

Support statement: This study was funded by Swansea University Medical School with the support of BREATHE – The Health Data Research Hub for Respiratory Health (MC_PC_19004), which is funded through the UK Research and Innovation Industrial Strategy Challenge Fund and delivered through Health Data Research UK. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Singh D, Agusti A, Anzueto A, *et al.* Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease: the GOLD science committee report 2019. *Eur Respir J* 2019; 53: 1900164.
2 Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ* 2010; 341: c4226.
3 Crossfield SSR, Lai LYH, Kingsbury SR, *et al.* Variation in methods, results and reporting in electronic health record-based studies evaluating routine care in gout: a systematic review. *PLoS One* 2019; 14: e0224272.
4 al Sallakh MA, Vasileiou E, Rodgers SE, *et al.* Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. *Eur Respir J* 2017; 49: 1700204.
5 Rubbo B, Fitzpatrick NK, Denaxas S, *et al.* Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: a systematic review and recommendations. *Int J Cardiol* 2015; 187: 705–711.
6 Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Social Res Methodol Theory Practice* 2005; 8: 19–32.
7 World Health Organisation. International Statistical Classification of Diseases and related health Problems, 10th Revision. 2010 Edition; 2011. https://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf Date last accessed 20 January 2021.
8 Lai CC, Wu CH, Wang YH, *et al.* The association between COPD and outcomes of patients with advanced chronic kidney disease. *Int J Chron Obstruct Pulmon Dis* 2018; 13: 2899–2905.
9 Park SC, Kim YS, Kang YA, *et al.* Hemoglobin and mortality in patients with COPD: a nationwide population-based cohort study. *Int J Chron Obstruct Pulmon Dis* 2018; 13: 1599–1605.
10 Huang HH, Chen SJ, Chao TF, *et al.* Influenza vaccination and risk of respiratory failure in patients with chronic obstructive pulmonary disease: a nationwide population-based case-cohort study. *J Microbiol Immunol Infect* 2019; 52: 22–29.
11 Vazquez Guillamet R, Ursu O, Iwamoto G, *et al.* Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records. *Health Informatics J* 2018; 24: 394–409.
12 Bishwakarma R, Zhang W, Li YL, *et al.* Metformin use and health care utilization in patients with coexisting chronic obstructive pulmonary disease and diabetes mellitus. *Int J Chron Obstruct Pulmon Dis* 2018; 13: 793–800.
13 Choi J, Oh JY, Lee YS, *et al.* The association between blood eosinophil percent and bacterial infection in acute exacerbation of chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis* 2019; 14: 953–959.
14 Whittaker HR, Müllerova H, Jarvis D, *et al.* Inhaled corticosteroids, blood eosinophils, and FEV1 decline in patients with COPD in a large UK primary health care setting. *Int J Chron Obstruct Pulmon Dis* 2019; 14: 1063–1073.
15 Hamad GA, Cheung W, Crooks MG, *et al.* Eosinophils in COPD: how many swallows make a summer? *Eur Respir J* 2018; 51: 1702177.
16 Li Q, Larivée P, Courteau J, *et al.* Greater eosinophil counts at first COPD hospitalization are associated with more readmissions and fewer deaths. *Int J Chron Obstruct Pulmon Dis* 2019; 14: 331–341.
17 Müllerová H, Hahn B, Simard EP, *et al.* Exacerbations and health care resource use among patients with COPD in relation to blood eosinophil counts. *Int J COPD* 2019; 14: 683–692.
18 Müllerová H, Meeraus WH, Galkin DV, *et al.* Clinical burden of illness among patients with severe eosinophilic COPD. *Int J COPD* 2019; 14: 741–755.
19 Ho TW, Ruan SY, Huang CT, *et al.* Validity of ICD9-CM codes to diagnose chronic obstructive pulmonary disease from National Health Insurance claim data in Taiwan. *Int J Chron Obstruct Pulmon Dis* 2018; 13: 3055–3063.
20 Su VYF, Yang KY, Yang YH, *et al.* Use of ICS/LABA combinations or LAMA is associated with a lower risk of acute exacerbation in patients with coexistent COPD and asthma. *J Allergy Clin Immunol* 2018; 6: 1927–1935.e3.
21 Lawson CA, Mamas MA, Jones PW, *et al.* Association of medication intensity and stages of airflow limitation with the risk of hospitalization or death in patients with heart failure and chronic obstructive pulmonary disease. *JAMA Netw Open* 2018; 1: e185489.

22 To T, Zhu J, Gray N, *et al.* Asthma and chronic obstructive pulmonary disease overlap in women: incidence and risk factors. *Ann Am Thorac Soc* 2018; 15: 1304–1310.

23 Chalmers JD, Poole C, Webster S, *et al.* Assessing the healthcare resource use associated with inappropriate prescribing of inhaled corticosteroids for people with chronic obstructive pulmonary disease (COPD) in GOLD groups A or B: an observational study using the Clinical Practice Research Datalink (CPRD). *Respir Res* 2018; 19: 63.

24 Oshagbemi OA, Keene SJ, Driessen JHM, *et al.* Trends in moderate and severe exacerbations among COPD patients in the UK from 2005 to 2013. *Respir Med* 2018; 144: 1–6.

25 Landis S, Suruki R, Maskell J, *et al.* Demographic and clinical characteristics of COPD patients at different blood eosinophil levels in the UK clinical practice research datalink. *J Chronic Obstruct Pulmon Dis* 2018; 15: 177–184.

26 Mcguire K, Aviña-Zubieta JA, Esdaile JM, *et al.* Risk of incident chronic obstructive pulmonary disease in rheumatoid arthritis: a population-based cohort study. *Arthritis Care Res (Hoboken)* 2019; 71: 602–610.

27 Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 2015; 12: e1001885.

28 Quint JK, Müllerova H, DiSantostefano RL, *et al.* Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ Open* 2014; 4: e005540.

29 Raherison C, Girodet PO. Epidemiology of COPD. *Eur Respir Rev* 2009; 18: 213–221.

30 Ohar JA, Loh CH, Lenoir KM, *et al.* A comprehensive care plan that reduces readmissions after acute exacerbations of COPD. *Respir Med* 2018; 141: 20–25.

31 Charlson M, Szatrowski TP, Peterson J, *et al.* Validation of a combined comorbidity index. *J Clin Epidemiol* 1994; 47: 1245–1251.