Pragmatic and Observational Research

Dovepress

open access to scientific and medical research

Open Access Full Text Article

REVIEW

UK Electronic Healthcare Records for Research: A Scientometric Analysis of Respiratory, Cardiovascular, and COVID-19 Publications

Georgie M Massen[1], Olivia Blamires[2,*], Megan Grainger[2,*], Max Matta[2,*], Rachel Monica Gyemfuah Twumasi[2,*], Tanvi Joshi[2,*], Alex Laity[2,*], Elena Nakariakova[2,*], Thilaksana Thavaranjan[2,*], Aziz Sheikh[3], Jennifer K Quint[1]

[1]School of Public Health, Imperial College London, London, UK; [2]Faculty of Medicine, Imperial College London, London, UK; [3]Usher Institute, University of Edinburgh, Edinburgh, UK

*These authors contributed equally to this work

Correspondence: Georgie M Massen, Imperial College, Level 9, Sir Michael Uren Hub, 86 Wood Lane, White City, W12 0BZ, London, UK, Tel +44 7447104872, Email g.massen21@imperial.ac.uk

Background: Routinely collected electronic healthcare records (EHRs) document many details of a person's health, including demographics, preventive services, symptoms, tests, disease diagnoses and prescriptions. Although not collected for research purposes, these data provide a wealth of information which can be incorporated into epidemiological investigations, and records can be analysed to understand a range of important health questions. We aimed to understand the use of routinely collected health data in epidemiological studies relating to three of the most common chronic respiratory conditions, namely: asthma, chronic obstructive pulmonary disease (COPD) and interstitial lung disease (ILD). We also characterised studies using EHR data to investigate respiratory diseases more generally, relative to cardiovascular disease and COVID-19, to understand trends in the use of these data.

Methods: We conducted a search of the Scopus database, to identify original research articles (irrespective of date) which used data from one of the following most frequently used UK EHR databases: Clinical Practice Research Datalink (including General Practice Research Database (CPRD's predecessor)), The Health Improvement Network and QResearch, defined through the presence of keywords. These databases were selected as they had been previously included in the works of Vezyridis and Timmons.

Findings: A total of 716 manuscripts were included in the analysis of the three chronic respiratory conditions. The majority investigated either asthma or COPD, whilst only 28 manuscripts investigated ILD. The number of publications has increased for respiratory conditions over the past 10 years (888% increase from 2000 to 2022) but not as much as for cardiovascular diseases (1105%). These data have been used to investigate comorbidities, off-target effects of medication, as well as assessing disease incidence and prevalence. Most papers published across all three domains were in journals with an impact factor less than 10.

Plain Language Summary: When people go to healthcare services such as the GP or hospital, details of the encounter are recorded in electronic systems known as electronic healthcare records. Information which is recorded can include symptoms, diagnoses, tests performed and ordered and prescriptions. We looked to understand how these records were being used to conduct epidemiological research, specifically in three respiratory conditions (asthma, chronic obstructive pulmonary disease, and interstitial lung diseases). We analysed information from 716 research papers which investigated one of these three conditions, we also looked more broadly at papers using electronic healthcare records for respiratory, cardiovascular and COVID-19 research. We found that research (published within articles) into these conditions has significantly increased in the past decade, however more research has been published with respect to cardiovascular diseases. We have shown that throughout the COVID-19 pandemic, electronic healthcare records were used extensively to conduct research into this new virus. Research is regularly conducted using electronic healthcare records, to understand diseases as well as treatments, more research is published in cardiovascular diseases than respiratory diseases.

Keywords: electronic health records, COPD, chronic obstructive pulmonary disease, asthma, interstitial lung disease, ILD

Received: 23 April 2024
Accepted: 2 August 2024
Published: 15 August 2024

Introduction

The increasing digitalisation of our society has resulted in an explosion in the availability of a myriad of data, sometimes referred to as "Big Data", here we focus on routinely collected healthcare data. This includes a wide range of data, for example, from medical insurance claims to mortality data, specific drug and disease registries, pharmacy databases, and electronic healthcare records (EHRs), with each database coding and storing information differently. Administrative datasets are often used for billing by insurance companies who renumerate the cost of healthcare in countries with privatised healthcare. An example of this would be the MarketScan Database from the United States of America which brings together data from multiple sources, and includes extensive information, such as diagnoses, medications, dental care, benefit plans and productivity to name a few.[1] Specific databases exist which hold the EHRs of sub-national populations, for example, the Veterans Affairs corporate Data Warehouse which contains the EHRs of American service personnel.[2] EHRs are increasingly used for research globally, as they offer the potential for large sample sizes, a range of study variables, and the inclusion of more generalisable populations compared with prospective cohort studies and surveys. They are important for better understanding the natural history of disease, estimating healthcare resource utilisation (HRU) and can be used to study the effectiveness and safety of treatments in routine care settings.

EHRs specifically are a collection of data relating to individuals enrolled within a medical center. These data are collected during routine appointments and collectively form a patient's health record, detailing appointments as well as diagnoses, reported symptoms, tests and referrals to other care services.[3] It is also possible to view prescribing data, including type of medication and the date these were prescribed, some also contain or can be linked with dispensing records.[4] It is also possible to use these data to understand which medications are commonly prescribed and potential associations between the effects of a medication and a defined outcome.

EHRs are a fruitful data source for epidemiological research, they can be used to understand a plethora of aspects of disease burden and are frequently used in epidemiological research, to estimate the incidence and prevalence of disease, conduct cohort and case control studies and to better understand comorbidities, reasons for hospitalisation and cause of death.[1,5–8] The MarketScan databases are a family of administrative claims databases that contain data on inpatient and outpatient claims, outpatient prescription claims, clinical utilization records, and healthcare expenditures. The three main databases available for use are each composed of a convenience sample for one of the following patient populations: (1) patients with employer-based health insurance from contributing employers, (2) Medicare beneficiaries who possess supplemental insurance paid by their employers, and (3) patients with Medicaid in one of eleven participating states. Eleven supplemental databases are available, which are utilized to overcome the limited clinical data available in the core MarketScan databases. There are several limitations to this database, primarily related to the fact that individuals or their family members within two of the core databases mandatorily possess some form of employer-based health insurance, which prevents the dataset from being nationally representative. Nonetheless, this database provides detailed and rigorously maintained claims data to identify healthcare utilization patterns among this cohort of patients. The longitudinal form of these data allows for studies using extended follow up to be conducted, and some clinical trials now link to EHRs to understand long term drug effects, disease outcomes or the outcome of complex interventions, given most trials are only able to "follow up" participants during a predefined, often short period of time in which the outcome is measured. As data are recorded using clinical coding systems, it is possible to investigate the occurrence of diseases across multiple EHRs therefore gaining greater insight into disease prevalence as well as differences in the care provided.[9]

As EHR data are routinely collected in practice, using these data are convenient as it does not require significant additional burden to collate data; therefore, study costs are significantly less. The precision of these data, especially recorded in primary care mean that it is possible to study a wide range of epidemiological questions, this is not the case in clinical trials where only specific predetermined data can be collected which in turn can limit potential analyses; another benefit of this is the opportunity to perform adjusted analyses which can provide greater insight into the effect of an exposure relative to a person's overall health. As these data are longitudinal, it is possible to perform long-term follow up

which would not be possible in other studies. It has been previously shown that data provided by the Clinical Practice Research Datalink are representative of age, sex, deprivation and UK geography,[5] however it is also important to acknowledge some potential biases, information regarding prisoners and members of the armed forces, people paying to receive private health care as well as the homeless population.[10] When using EHR data it is also important to acknowledge the potential biases caused by missing data and carefully consider the impact this would have on analyses and interpretation.

Given that it has previously been documented that the use of EHRs for research has significantly increased since 2010, we looked to explore this further in the context of respiratory disease research.[1] Here, we analysed the use of the three most used UK EHRs in the fields of cardiovascular disease, respiratory disease and more specifically COVID-19. It is known that respiratory data (such as incidence and prevalence) are often lacking, and it has been previously shown that respiratory research is severely underfunded, as a result we aimed to explore the impact of lack of funding on the output of epidemiological research in comparison with cardiovascular disease research (which receives a larger amount of funding) and COVID-19 research given the changing attitude towards data studies and data access during the pandemic.[11,12] We then focused on three of the most common chronic long term respiratory conditions, namely asthma, chronic obstructive pulmonary disease (COPD) and, interstitial lung disease (ILD).

Methods

On the 10th of August 2023, we conducted searches of the Scopus database for six topics (asthma, COPD, ILD, respiratory disease, cardiovascular disease and, COVID-19). We applied an advanced search of titles, abstracts, and keywords for a combination of search terms relating to EHR databases and one of the six research areas.

Inclusion Criteria

We utilised the databases filtering system to select only studies on humans and published before the 1st of January 2023. We only included original research articles.

Data Extraction

Information regarding the publications was extracted directly from Scopus. Six data files were stored in Excel (Version 2302) and used for analysis of each of the three respiratory conditions as well as the broader fields of cardiovascular, respiratory and COVID-19 research. We used the following variables from the extracted data: authors, year of publication, publisher, author keywords, citations, contact details.

Statistical Analysis

We analysed the number of publications per year per respiratory disease as well as the total number of publications; this analysis was also carried out using the broader cardiovascular, respiratory and COVID-19 EHR manuscripts (only relevant for comparison from 2020 onwards). We described the five journals most commonly publishing research, respective to the three respiratory diseases, including data on the Journal Impact Factor (JIF) and publisher; this was also completed for the respiratory, cardiovascular and, COVID-19 manuscripts. All analyses were conducted individually by two people (GMM, OB, MG, MM, RT, TJ, AL, EN, TT) and then compared with one another to ensure that the results were valid. Through analysis of contact details, we determined the contributing countries. We analysed author lists to determine the people who contributed most to research within the fields of asthma, COPD and, ILD, this was also visualised using VosViewer.[13] We explored the number of citations per disease, identifying the most cited publications as well as how many publications were yet to be cited. We used the author keywords to analyse keywords, we produced counts per keyword per disease and split these keywords into conditions, medications and "other". We removed keywords which were specific to the disease eg, in the analysis of COPD condition keywords we removed "COPD" and "Chronic Obstructive Pulmonary Disease".

Results

Of the three broad disease areas (cardiovascular, respiratory and COVID-19), most publications were in the field of cardiovascular disease research (Figure 1A). The number of publications per year has steadily increased since 2010, with both cardiovascular and respiratory research following similar trends (Figure 1B). However, since the emergence of COVID-19, EHRs have been used extensively to study COVID-19; a total of 289 papers were published between the beginning of 2020 and the end of 2022 which used EHR to research COVID-19; comparatively 368 respiratory papers were published and 565 cardiovascular manuscripts in the same time period.

Three hundred and eighty-nine manuscripts were identified which studied asthma, 229 manuscripts were found that studied COPD and 28 investigated ILD (Figure 1C). Overall, a steady increase in the number of publications using routinely collected health records for respiratory research was observed since 2010. A significant increase in the number of publications for both asthma and COPD studies were identified from 2010 onwards, demonstrating an increased use of EHRs for research (Figure 1D).

Of the top five journals publishing cardiovascular EHR research; two journals (108 manuscripts) were within the field of pharmacology whilst two of the journals were general medical research journals (127 manuscripts) (Table 1A). Comparatively, of the five journals most commonly publishing respiratory research, one journal focused on pharmacology (32 manuscripts) (Table 1B). The COVID-19 research using EHRs was published in a broad range of journals including public health specific and general medical journals (Table 1C). Of the five journals, which most published asthma manuscripts using EHRs, two of the journals were general respiratory journals, whilst one focused on pharmacoepidemiology and the other two journals were general medical journals (Table 1D). However, of the COPD

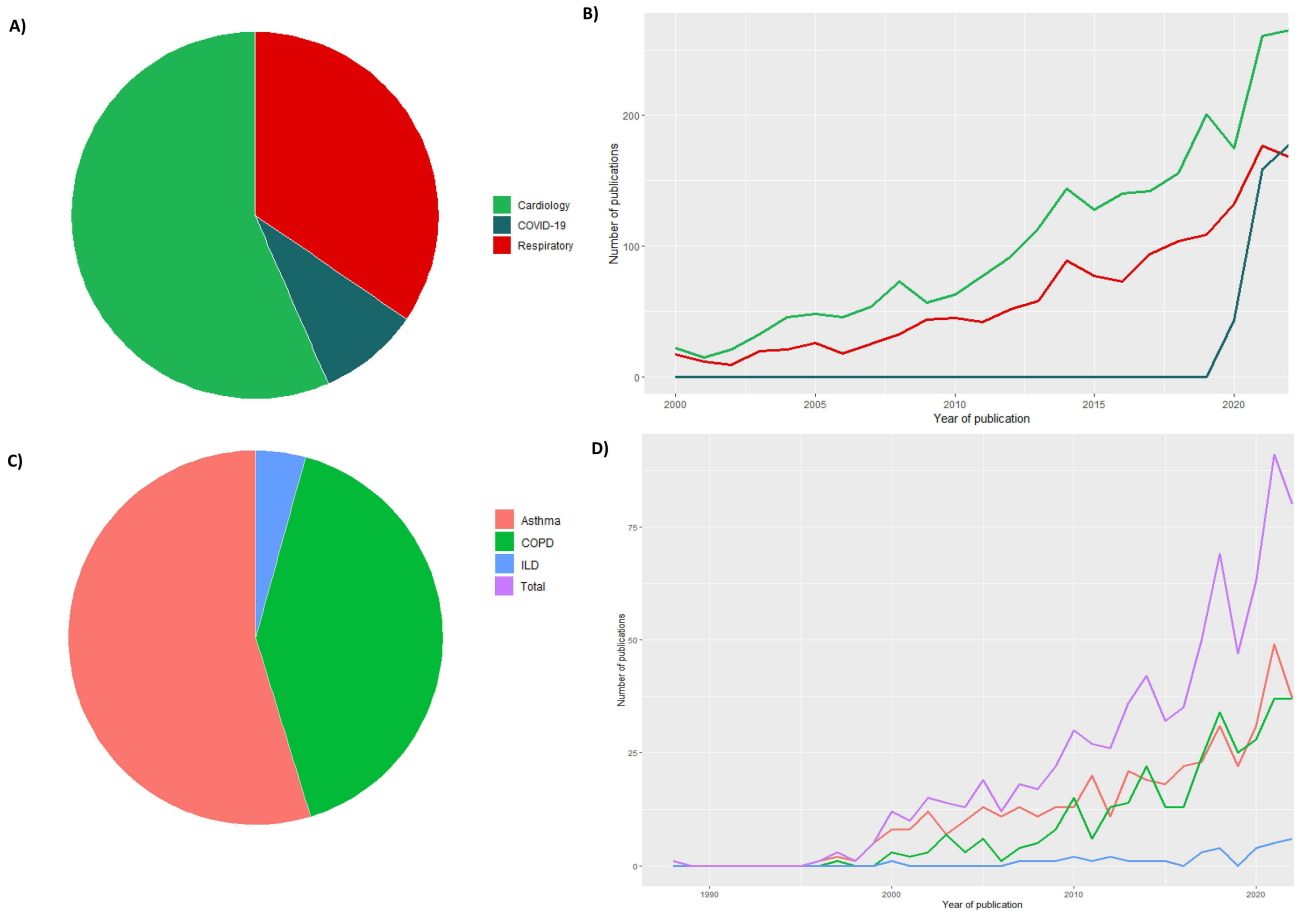

Figure 1 (**A**) Proportion of papers across respiratory, cardiovascular and COVID-19 research. (**B**) Number of papers published per year. (**C**) Proportion of papers across three chronic respiratory conditions. (**D**) Number of asthma, COPD and ILD papers published per year.
Abbreviations: COPD: Chronic Obstructive Pulmonary Disease, ILD: Interstitial Lung Disease.

Table 1 The Top Five Journals Publishing EHR Research

Rank	Journal	No. of Papers	Percentage	Latest Impact Factor	Publisher
A) Cardiovascular					
1	BMJ Open	90	4.53%	2.9	BMJ Publishing Group
2	Pharmacoepidemiology and drug safety	70	3.52%	2.6	John Wiley and Sons Ltd
3	British Journal of Clinical Pharmacology	38	1.91%	3.7	Wiley-Blackwell on behalf of the British Pharmacological Society
4	PLoS ONE	37	1.86%	3.7	Public Library of Science
5	Heart	31	1.56%	5.7	BMJ Publishing Group
B) Respiratory					
1	BMJ Open	53	4.3%	2.9	BMJ Publishing Group
2	British Journal of General Practice	39	3.2%	5.9	Royal College of General Practitioners
3	International Journal of COPD	36	2.9%	5.5	Dove Medical Press Ltd
4	Pharmacoepidemiology and Drug Safety	32	2.6%	2.6	John Wiley and Sons Ltd
5	Chest	20	1.6%	9.6	American College of Chest Physicians
C) COVID-19					
1	BMJ Open	12	4.2%	2.9	BMJ Publishing Group
2	PLoS One	9	3.1%	3.7	Public Library of Science
3	The International Journal of Environmental Research and Public Health	8	2.8%	4.6	MDPI
4	Journal of Medical Internet Research	4	1.4%	7.1	JMIR
5	BMJ Open Quality	4	1.4%	1.4	BMJ Publishing Group
D) Publishers of Asthma manuscripts					
1	BMJ Open	24	6·17%	2.9	BMJ Publishing Group
2	British Journal of General Practice	12	3·08%	5.9	Royal College of General Practitioners
3	Pharmacoepidemiology and Drug Safety	11	2·83%	2.6	John Wiley and Sons Ltd
4	Thorax	10	2·57%	10.0	BMJ Publishing Group
5	European Respiratory Journal	10	2·57%	24.3	European Respiratory Society
E) Publishers of COPD manuscripts					
1	International Journal of COPD	35	11·71%	5.5	Dove Medical Press Ltd
2	COPD: Journal of Chronic Obstructive Pulmonary Disease	13	4·35%	2.2	Taylor and Francis Ltd.
3	BMJ Open	12	4·01%	2.9	BMJ Publishing Group

(*Continued*)

Table 1 (Continued).

Rank	Journal	No. of Papers	Percentage	Latest Impact Factor	Publisher
4	Respiratory Medicine	11	3·68%	4.3	W.B. Saunders Ltd
5	Chest	10	3·34%	9.6	American College of Chest Physicians
F) Publishers of ILD manuscripts					
1	Respiratory Medicine	3	10·71%	4.3	W.B. Saunders Ltd
2	Advances in Therapy	2	7·14%	3.7	Springer Healthcare
3	American Journal of Respiratory and Critical Care Medicine	2	7·14%	24.7	American Lung Association
4	Chest	2	7·14%	9.6	American College of Chest Physicians
5	British Journal of General Practice	1	3·57%	5.9	Royal College of General Practitioners

Notes: A) Publishers of cardiovascular research. B) Publishers of respiratory research. C) Publishers of COVID-19 research. D) Publishers of asthma manuscripts. E) Publishers of COPD manuscripts. F) Publishers of ILD manuscripts.
Abbreviations: COPD: Chronic Obstructive Pulmonary Disease, ILD: Interstitial Lung Disease.

manuscripts, two of the top journals publishing research were specific COPD journals whilst two were respiratory journals (Table 1E). Manuscripts regarding ILDs were published in a range of journals, including pharmacological and respiratory journals (Table 1F).

Across the three respiratory conditions, we identified the five authors who were named on the most research articles per disease. The majority of authors are professors based at higher education institutions in either the UK or the USA (Supplementary Table 1). It is clear there are many collaborations between authors in both the asthma and COPD fields (Figure 2). However, of the few ILD publications there were four distinct clusters of researchers who collaborate with one another.

Of the cardiovascular papers published in 2021 and 2022, the mean number of citations as of the time of data extraction was fifteen and four respectively. Respiratory manuscripts published in 2021 had a mean number of thirteen citations, comparatively of the mean number of citations of manuscripts published in 2022 was six. Of the Covid-19 manuscripts published in 2021 the mean number of citations was 19, of the manuscripts published in 2022 the mean number of citations was six. Of the manuscripts regarding asthma, the mean number of citations was 39; a total 12 manuscripts (3·1%) currently had zero citations. Of the COPD manuscripts, the mean number of citations was 40 and a total of 21 papers (7·0%) did not have any citations. Of the ILD manuscripts, the mean number of citations was 64, one (3·6%) manuscript had zero citations.

Of the keywords available from asthma manuscripts, the most common keyword relating to medical conditions was COPD which was present in 24 manuscripts; the most common keyword pertaining to medication was inhaled corticosteroid (Figure 3). Asthma was the most used condition keyword in COPD manuscripts and was present in 20 manuscripts. Exacerbations was the second most commonly occurring keyword in COPD manuscripts, present in a total of 16 manuscripts keywords (Figure 4a). Similarly, to what was seen in the keyword analysis of pharmacological research in asthma, inhaled corticosteroids were the most commonly occurring medication keyword in the COPD manuscripts (Figure 4b). The most used condition keyword in ILD manuscripts was idiopathic pulmonary fibrosis which was present in eight manuscripts (Figure 5a). Twelve of the ILD manuscripts listed medications within their keywords; proton pump inhibitors were specifically stated in the keywords of 2 manuscripts (Figure 5b).

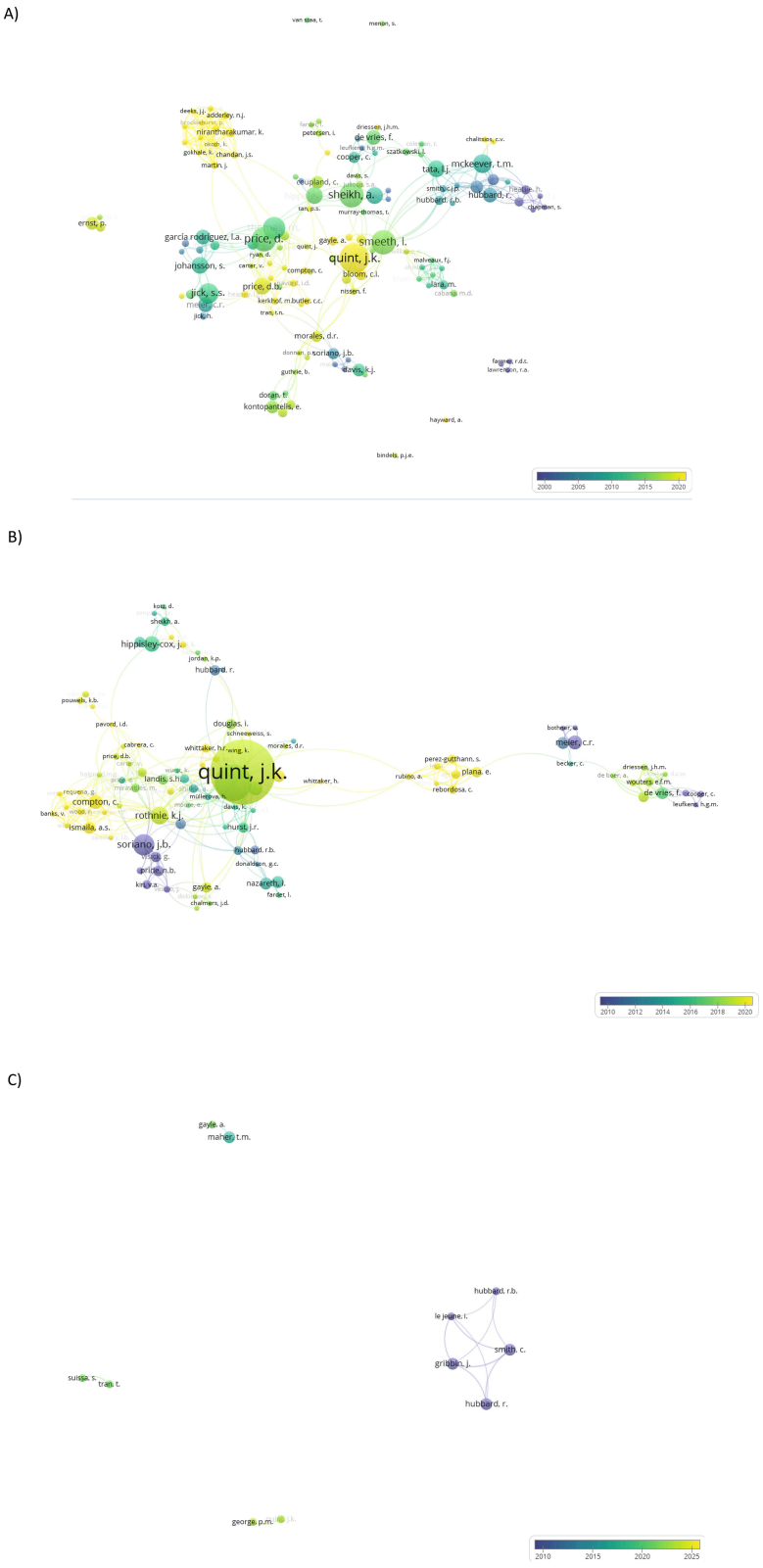

Figure 2 Network of author collaborations (**A**) Asthma, (**B**) COPD, (**C**) ILD. This figure was created using the VOSviewer application.[13]

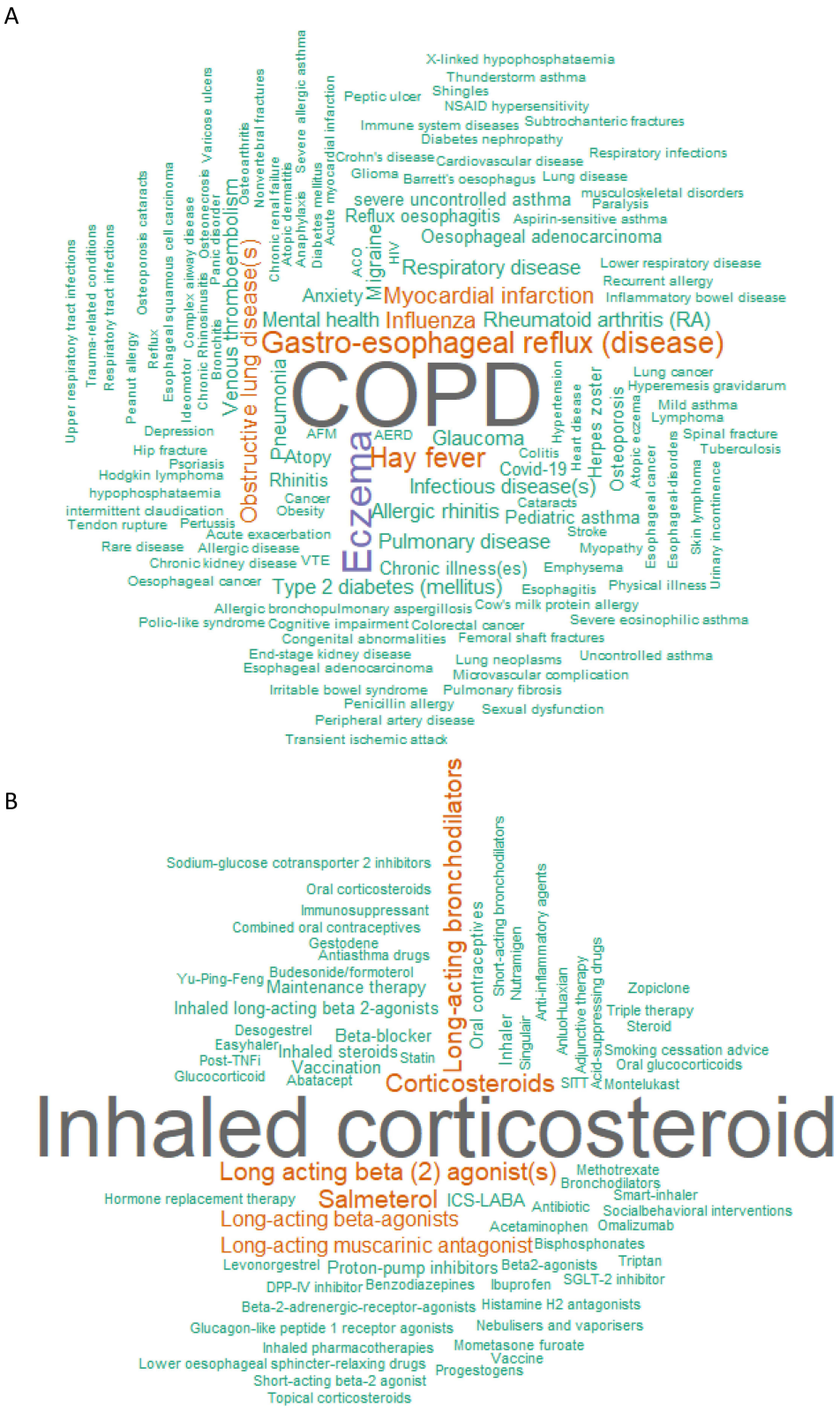

Figure 3 Keywords frequently used in asthma manuscripts. (**A**) Condition keywords, (**B**) Medication keywords.

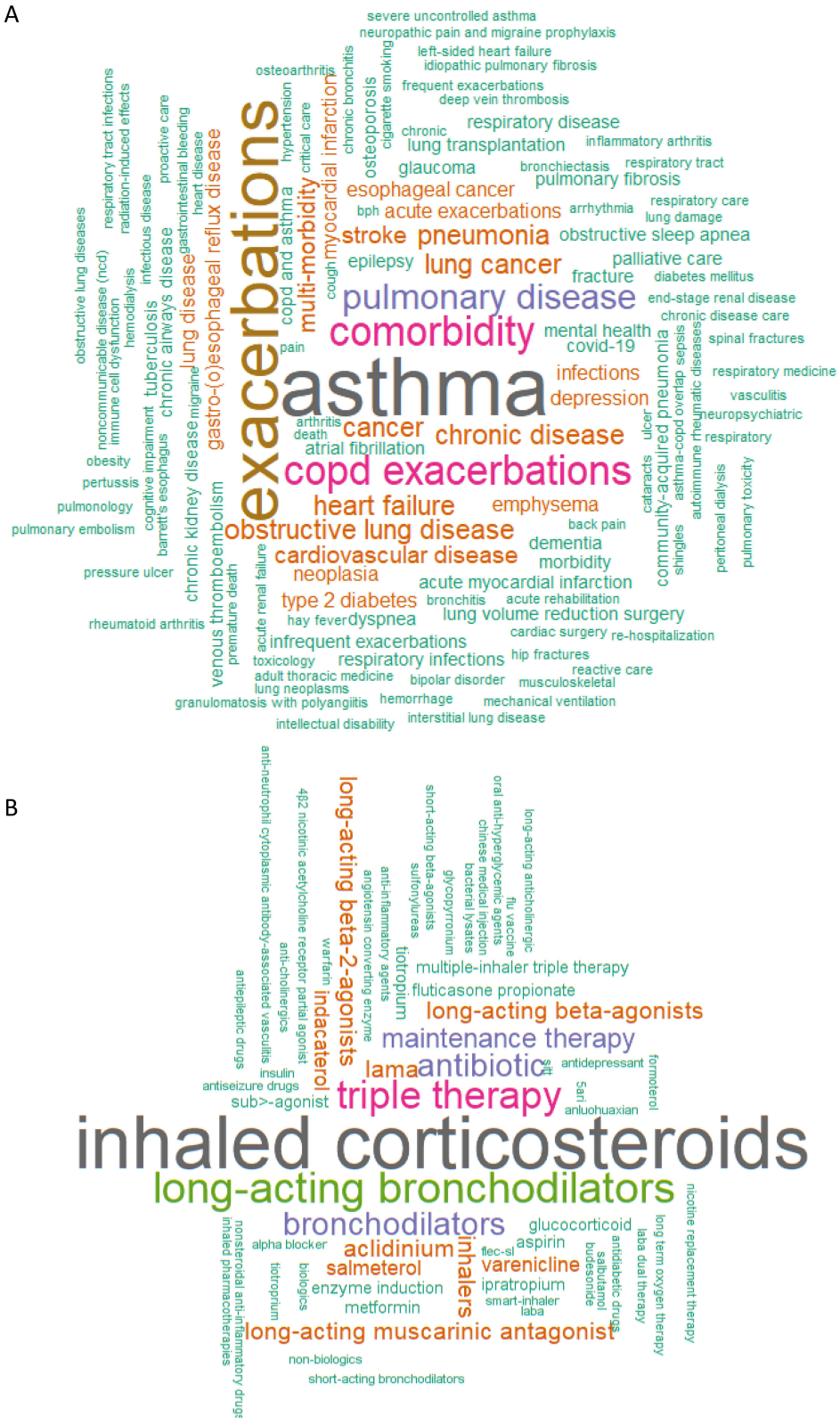

Figure 4 Keywords frequently used in COPD manuscripts. (**A**) Condition keywords, (**B**) Medication keywords.
Abbreviation: COPD, Chronic Obstructive Pulmonary Disease.

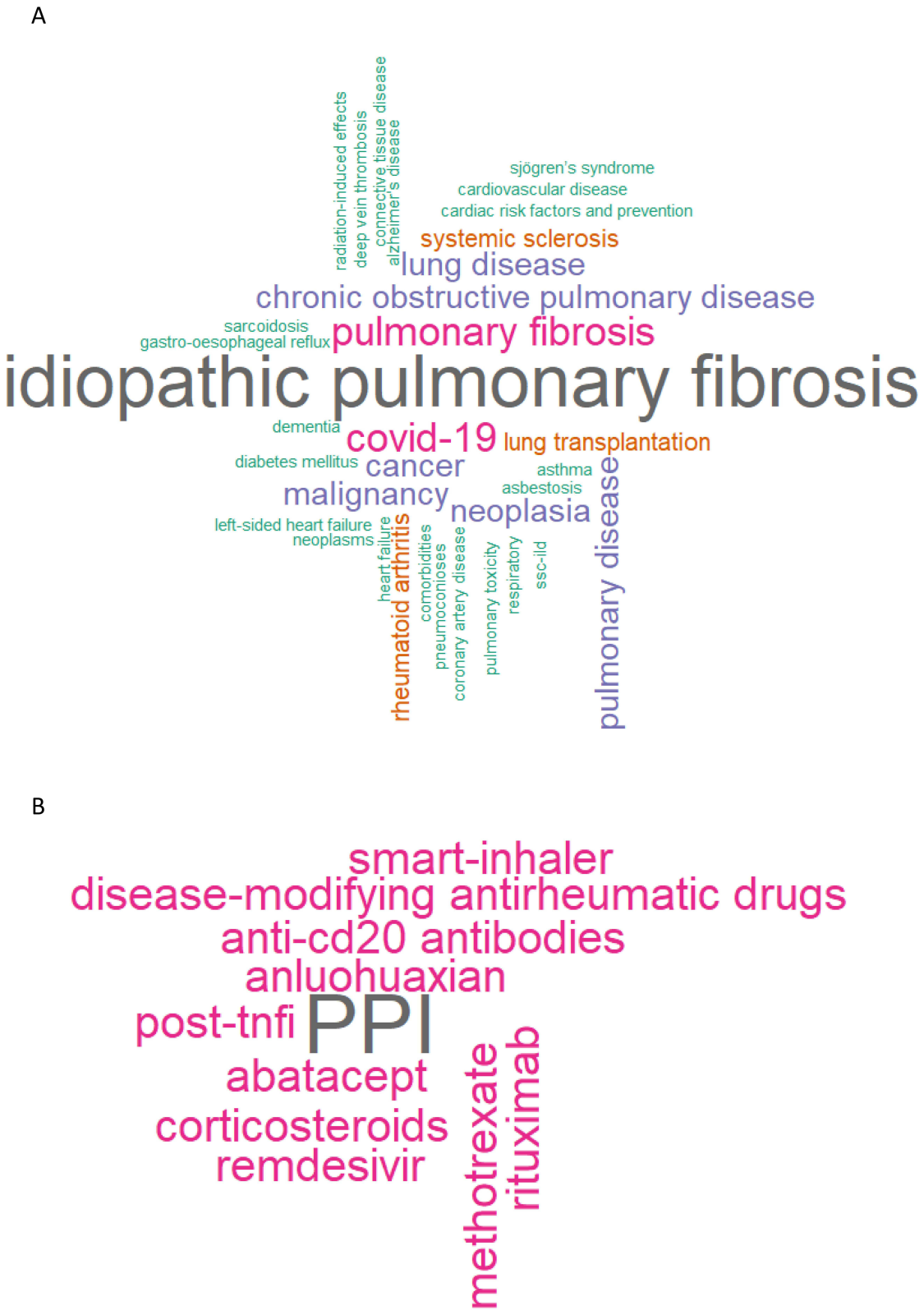

Figure 5 Keywords frequently used in ILD manuscripts. (**A**) Condition keywords, (**B**) Medication keywords.
Abbreviation: ILD, Interstitial Lung Disease.

Discussion

We conducted a scientometric study of research using UK routinely collected EHRs. Quantitative studies of bibliometric data are important in understanding the broader aspects of research, unlike systematic reviews which focus on the results of research. By examining the bibliometric data, a greater understanding of the scientific output can be gained: measuring research impact, understanding the citation process and which journals commonly publish research within a specific field. This type of analysis can be used to map the knowledge structure and evolutions of published scientific research. Scientometric studies are broad and often capture most of the research within the specified field which enables the depiction of core, overarching concepts.

In investigating the use of EHRs to research three common chronic respiratory conditions, we have shown that these data have been extensively used to research both asthma and COPD, with a recent increase in the use of these data to investigate ILDs. These data have been used to investigate comorbidities, off-target effects of medication, as well as assessing disease incidence and prevalence. We have shown that throughout the COVID-19 pandemic, EHRs were used extensively to conduct research into this novel viral communicable disease. However, the amount of published research using these data for cardiovascular and respiratory research has gradually increased in the past decade. In 2022, 168 papers were published which used EHRs to research respiratory disease, this was superseded by COVID-19 research for which there were 178 publications, a total of 265 manuscripts were published which used these data for cardiovascular disease research. We found similar citation averages between both respiratory disease research and cardiovascular research. We have shown that many of the journals which regularly publish research which use EHRs have a low JIF, irrespective of whether the research was within respiratory or cardiovascular disease.

Predominantly publishing in low JIFs could demonstrate that research using EHRs to conduct epidemiological studies is underappreciated and not always of interest to the broader scientific community, however it is ultimately the choice of authors alone to decide where to submit works. This is incredibly surprising considering that EHRs are currently the best databases to study the incidence and prevalence of diseases given the large quantity of people whose data are included and the granular detail, which are available to be studied, it is often hard to locate up to date figures for disease incidence and prevalence even though these are relatively easy to locate from EHRs. This untapped wealth of data, which can be used to conduct broad research which is insightful with regards to current patient populations has great potential to further understand disease epidemiology.

Given that EHRs are routinely collected in clinical practice, they can provide great insight into both patient care and patient experience. The amount of data available continues to grow; CPRD Aurum, for example, now contains the EHRs of 15,608,621 people who are currently contributing data (23.28% of the UK population) (September 2023 build), this has grown significantly since the introduction of CPRD Aurum in 2017 which contained the data of 7.1 million current patients.[5,14] Though these data are plentiful, there are barriers which must be overcome when working with them, ranging from study specific limitations to the costs of conducting a study. Accessing EHRs can be expensive and appropriately there is significant governance surrounding the data. The minimum cost of using QResearch data (primary care only) for a singular study is, for example, £30,000.[15] A basic user license for CPRD data allows access to two databases (GOLD and Aurum) and covers the use of two users is £45,000, for multiple users to submit applications which include accessing linkages to secondary care data the cost is £363,000 for a full license per year.[16]

Routinely collected EHRs can be used to assess the association between prescribed medications and potential negative outcomes of licensed drugs. As clinical trials commonly include individuals who do not have comorbidities and are therefore not always representative of the true patient population with the disease of interest, potential drug interactions with these conditions are not assessed. EHRs are a great source of data to study the effectiveness of drugs, these data can be used to understand whether an intervention is having beneficial effects on a patient; clinical trials often only measure efficacy, a measure of the desired drug effect. Effectiveness of treatment can also be measured in trials, but does not necessarily provide the same information as EHRs.[17,18] Whilst randomized control trials remain the gold standard for pharmaco-epidemiological studies, findings are only representative to the population studied, whilst used routinely collected data provides a representative sample to analyse the effects of a drug within a disease population.[19,20] Clinical trials are moving towards utilising routinely collected data in their studies to assess clinical endpoints as well as

assessing eligibility to enroll into trials, this will lead to an increase in the amount of publications which use EHR data as this field continues to incorporate the data source.[21,22]

There has been an increase in the use of EHRs in producing prediction models.[23,24] These models could greatly benefit the patient population should an intervention be possible to decrease the risk of the predicted outcome. However, it is important that clinicians who would be using these models are well versed in their design, enabling them to feel confident in explaining how the prediction was produced should a patient ask. Therefore, it is important that before the development of prediction models has commenced, clinicians are involved in the development of the study question to ensure its usefulness which in turn would increase the likelihood of implementing the information outputted by the prediction model in practice. Both validation work and impact assessments are crucial in assessing the potential utility of prediction models as well as the validity and accuracy.

As EHRs are not collected with the intention for use in research, there are many caveats that need to be understood when using and interpreting these data. For example, when conducting studies of medications, it is not possible to say that the person received their prescribed drugs let alone adhered to the medication. Dispensing data can help to understand who are collecting medications, though in many instances this information is not routinely collected. However, if a person is regularly prescribed a medication, it increases the likelihood that they are using the medication as they would have had to reattend/ contact the practitioner to be re-prescribed the medication and are therefore more likely to be engaged in their healthcare. Clinicians (or clinical coders in secondary care) record data using both clinical codes and free-text, however free-text is not commonly available to researchers meaning information can be lost, the cost of using free-text. There are many clinical codes which can be used to record the presence of a condition or symptom; it is pivotal to define covariates using standardised methodologies, which are reproducible and produce a cohort of people which is representative of the people who are commonly affected by the exposure of interest. To better understand the results of studies using EHR data, it is necessary to include information regarding how the cohort and covariates were defined by making codelists freely available and accessible. Currently, there is no gold standard for the recording of codelists with many repositories in existence, however the transparency regarding methodology is vital in improving the validity and trustworthiness of studies.[25,26] Recently, there have been publications detailing the derivation of codelists for both medical codes and medication codes.[25,27,28] It should be a requirement implemented during the publishing process that papers using electronic healthcare records report the codelists used to define exposures, outcomes and covariates, as described previously in the RECORD reporting guidelines.[29,30]

To increase the use of EHRs in research, a greater appreciation (within the research community and public) of the benefits of using these routinely collected data is required. Involving members of the public has an important place within research, ensuring that research is beneficial to the population who are affected by the condition that is to be researched; there are increasing requirements for involving and consulting with the public in grant applications and more recently applications for data. It is important that the general population are made aware that this research takes place and how their data are used. This transparency would increase the amount of support that a grant application would receive from the involved patient participation group, minimising the amount of time taken to educate the participants and instead building the most beneficial grant possible and a stronger business case, in turn increasing the likelihood of it being funded.

Strengths and Limitations

This study has described the use of EHRs in the broad disciplines of cardiovascular and respiratory disease as well as COVID-19 as well as within three specific respiratory diseases. We have summarized bibliographic data to understand the current landscape of the use of UK electronic healthcare records in research both broadly and with respect to individual disease domains. It is possible that some records may not have been included due to the nonexistence of keywords we have been able to gather a large quantity of studies which represent their respective fields. We sought to investigate the use of UK EHRs for research, therefore the findings are only representative of these databases, however it would be interesting to further investigate this question with respect to other databases from other countries such as the USA and Korea. Also, this work specifically looked at the use of electronic healthcare records, though we do acknowledge other datasources are available which contain healthcare data, including disease registries.

Conclusions

We have shown that EHRs have been and are increasingly used extensively to conduct respiratory research, and this was accelerated during the COVID-19 pandemic to conduct epidemiological research to better understand the virus and its effects. As many countries have now transitioned from paper-based records to EHRs, and the now considerable experience of using these digitised health infrastructures to enable operational, epidemiological, and experimental research, there is a pressing need to improve methodological quality and transparent reporting to ensure confidence in the robustness of findings. This responsibility falls to both the authors and journals who publish the research. Existing guidelines for the reporting of observational research should be adhered to.

Ethics Statement

Ethical approval was not required for this study.

Acknowledgments

Olivia Blamires, Megan Grainger, Max Matta, Rachel Monica Gyemfuah Twumasi, Tanvi Joshi, Alex Laity, Elena Nakariakova, and Thilaksana Thavaranjan are co-second authors for this study. This research was supported by the NIHR Imperial Biomedical Research Centre (BRC).

Disclosure

GMM, JKQ & AS have published using EHRs in relation to respiratory, cardiovascular and COVID-19 research, including some of the papers identified in this analysis. The authors report no other conflicts of interest in this work.

References

1. Kulaylat AS, Schaefer EW, Messaris E, Hollenbeak CS. Truven health analytics marketscan databases for clinical research in colon and rectal surgery. *Clin Colon Rectal Surg*. 2019;32(1):54–60. doi:10.1055/s-0038-1673354
2. Price LE, Shea K, Gephart S. The veterans affairs's corporate data warehouse: uses and implications for nursing research and practice. *Nurs Adm Q*. 2015;39(4):311. doi:10.1097/NAQ.0000000000000118
3. Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. *J Intern Med*. 2013;274(6):547–560. doi:10.1111/joim.12119
4. Alvarez-Madrazo S, McTaggart S, Nangle C, Nicholson E, Bennie M. Data resource profile: the Scottish national Prescribing Information System (PIS). *Int J Epidemiol*. 2016;45(3):714–715f. doi:10.1093/ije/dyw060
5. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) aurum. *Int J Epidemiol*. 2019;48 (6):1740–1740g. doi:10.1093/ije/dyz034
6. Quint JK, Millett ERC, Joshi M, et al. Changes in the incidence, prevalence and mortality of bronchiectasis in the UK from 2004-2013: a population based cohort study. *Eur Respir J*. 2016;47(1):186–193. doi:10.1183/13993003.01033-2015
7. Stone PW, Osen M, Ellis A, Coaker R, Quint JK. Prevalence of Chronic Obstructive Pulmonary Disease in England from 2000 to 2019. *Int J Chron Obstruct Pulmon Dis*. 2023;18:1565–1574. doi:10.2147/COPD.S411739
8. Gayle AV, Axson EL, Bloom CI, Navaratnam V, Quint JK. Changing causes of death for patients with chronic respiratory disease in England, 2005-2015. *Thorax*. 2019;74(5):483–491. doi:10.1136/thoraxjnl-2018-212514
9. Quint JK, O'Leary C, Venerus A, Holmgren U, Varghese P, Cabrera C. Development and validation of a method to estimate COPD severity in multiple datasets: a retrospective study. *Pulm Ther*. 2021;7(1):119–132. doi:10.1007/s41030-020-00139-0
10. Using CPRD primary care data CPRD [Internet]; 2024. Available from: https://www.cprd.com/using-cprd-primary-care-data. Accessed Mar 8, 2024.
11. Williams S, Sheikh A, Campbell H, et al. Respiratory research funding is inadequate, inequitable, and a missed opportunity. *Lancet Respir Med*. 2020;8(8):e67–8. doi:10.1016/S2213-2600(20)30329-5
12. UK Clinical Research Collaboration. UK Health Research Analysis 2018. 2020.
13. van Eck NJ, Waltman L. Software survey: vOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84 2:523–538 doi: 10.1007/s11192-009-0146-3.
14. CPRD, Medicines & Healthcare products Regulatory Agency. CPRD Aurum Accessed September 2023 dataset; 2023.doi : 10.48329/6j2c-nh78.
15. QResearch. Information for Researchers [Internet]. Available from: https://www.qresearch.org/information/information-for-researchers/. Accessed August 8, 2024.
16. Clinical Practice Research Datalink. Pricing [Internet]. 2023. Available from: https://cprd.com/pricing. Accessed August 8, 2024.
17. Anglemyer A, Horvath HT, Bero L Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials.Cochrane Database Syst Rev [Internet]. 2014;4. Available from: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.MR000034.pub2/full. Accessed Oct 26, 2023.
18. Burches E. Efficacy, Effectiveness and Efficiency in the Health Care: the Need for an Agreement to Clarify its Meaning; 2023. Available from: https://clinmedjournals.org/articles/iaphcm/international-archives-of-public-health-and-community-medicine-iaphcm-4-035.php#ref9. Accessed October 26, 2023.

19. Revicki DA, Frank L. Pharmacoeconomic evaluation in the real world. Effectiveness versus efficacy studies. *PharmacoEconomics*. 1999;15 (5):423–434. doi:10.2165/00019053-199915050-00001
20. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials*. 2009;10(1):37. doi:10.1186/1745-6215-10-37
21. O'Brien EC, Raman SR, Ellis A, et al. The use of electronic health records for recruitment in clinical trials: a mixed methods analysis of the harmony outcomes electronic health record ancillary study. *Trials*. 2021;22(1):465. doi:10.1186/s13063-021-05397-0
22. Macnair A, Nankivell M, Murray ML, et al. Healthcare systems data in the context of clinical trials - A comparison of cardiovascular data from a clinical trial dataset with routinely collected data. *Contemp Clin Trials*. 2023;128:107162. doi:10.1016/j.cct.2023.107162
23. Tomašev N, Harris N, Baur S, et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat Protoc*. 2021;16(6):2765–2787. doi:10.1038/s41596-021-00513-5
24. Toma M, Wei OC. Predictive modeling in medicine. *Encyclopedia*. 2023;3(2):590–601. doi:10.3390/encyclopedia3020042
25. Watson J, Nicholson BD, Hamilton W, Price S. Identifying clinical features in primary care electronic health record studies: methods for codelist development. *BMJ Open*. 2017;7(11):e019637. doi:10.1136/bmjopen-2017-019637
26. Citing and Crediting Codelists: A discussion for the research community | Bennett Institute for Applied Data Science [Internet]. 2021. Available from: https://www.bennett.ox.ac.uk/blog/2021/02/citing-and-crediting-codelists-a-discussion-for-the-research-community/. Accessed October 25, 2023.
27. Graul EL, Stone PW, Massen GM, et al. Determining prescriptions in electronic healthcare record data: methods for development of standardized, reproducible drug codelists. *JAMIA Open*. 2023;6(3):ooad078. doi:10.1093/jamiaopen/ooad078
28. Elkheder M, Gonzalez-Izquierdo A, Qummer Ul Arfeen M, et al. Translating and evaluating historic phenotyping algorithms using SNOMED CT. *J Am Med Inform Assoc JAMIA*. 2023;30(2):222–232. doi:10.1093/jamia/ocac158
29. Nicholls SG, Quach P, Guttmann A, et al.The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) statement: methods for arriving at consensus and developing reporting guidelines.*PLoS One*.2015;105:e0125620 doi:10.1371/journal.pone.0125620.
30. Massen GM, Stone PW, Kwok HHY, et al. Review of codelists used to define hypertension in electronic health records and development of a codelist for research. *Open Heart*. 2024;11(1):e002640. doi:10.1136/openhrt-2024-002640