Received: 14 June 2018 | Revised: 30 November 2018 | Accepted: 14 February 2019

DOI: 10.1002/pds.4765

ORIGINAL REPORT

WILEY

Limitations for health research with restricted data collection from UK primary care

Helen Strongman | Rachael Williams | Wilhelmine Meeraus | Tarita Murray-Thomas | Jennifer Campbell | Lucy Carty | Daniel Dedman | Arlene Gallagher | Jessie Oyinlola | Antonis Kousoulis | Janet Valentine

Clinical Practice Research Datalink (CPRD), MHRA, London, UK

Correspondence
J. Valentine, Clinical Practice Research Datalink (CPRD), MHRA, 151 Buckingham Palace Road, London SW1W 9SZ, UK.
Email: janet.valentine@mhra.gov.uk

Abstract

Purpose: UK primary care provides a rich data source for research. The impact of proposed data collection restrictions is unknown. This study aimed to assess the impact of restricting the scope of electronic health record (EHR) data collection on the ability to conduct research. The study estimated the consequences of restricted data collection on published Clinical Practice Research Datalink studies from high impact journals or referenced in clinical guidelines.

Methods: A structured form was used to systematically analyse the extent to which individual studies would have been possible using a database with data collection restrictions in place: (1) retrospective collection of specified diseases only; (2) retrospective collection restricted to a 6- or 12-year period; (3) prospective and retrospective collection restricted to non-sensitive data. Outcomes were categorised as unfeasible (not reproducible without major bias); compromised (feasible with design modification); or unaffected.

Results: Overall, 91% studies were compromised with all restrictions in place; 56% studies were unfeasible even with design modification. With restrictions on diseases alone, 74% studies were compromised; 51% were unfeasible. Restricting collection to 6/12 years had a major impact, with 67 and 22% of studies compromised, respectively. Restricting collection of sensitive data had a lesser but marked impact with 10% studies compromised.

Conclusion: EHR data collection restrictions can profoundly reduce the capacity for public health research that underpins evidence-based medicine and clinical guidance. National initiatives seeking to collect EHRs should consider the implications of restricting data collection on the ability to address vital public health questions.

KEYWORDS

bias, electronic health records, pharmacoepidemiology, primary care

1 | INTRODUCTION

It is well established that unrepresentative patient populations or missing data can pose significant obstacles when conducting public health research.[1-5] The scope and quality of national electronic health record (EHR) data sources that are accessible for research purposes varies significantly between countries, depending on the health care system and primary reason for collection. As a universal health care provider, free at the point of delivery, the UK National Health Service (NHS) encompasses over 90% of the UK population. The UK primary care EHR contains demographic, diagnostic, treatment, referral, and lifestyle information, thus creating a continuous record of an individual's health and medical care throughout their lifetime. Consequently, UK primary care data provide a rich source of longitudinal, comprehensive health care data for public health research.

Several databases exist that enable research access to anonymised EHR data, collected from a subset of the total UK primary care general practices (GPs).[6-8] A recent analysis evaluating the publication output from the three largest primary care EHR databases in the United Kingdom showed close to a 20% annual growth rate, rising from seven publications in 1995 to 171 throughout 2015.[9] Moreover, approximately 30% of research conducted using UK primary care databases is by international institutions, including research groups from the United States, Canada, Australia, and Europe.[9] The Clinical Practice Research Datalink (CPRD), supported by the Medicines and Healthcare products Regulatory Agency and National Institute of Health Research, is a UK government research service that has been providing anonymised UK primary care data for public health studies since the late 1980s.[6] CPRD data are used extensively by academics, regulators, and the pharmaceutical industry worldwide to investigate drug and vaccine safety,[10-17] assess uptake and effectiveness of public health policy and clinical guidance,[18-23] characterise the prevalence of diseases and associated risk factors,[24] and improve health care delivery.[25-28] Twenty-five National Institute for Health and Clinical Excellence (NICE) guidance documents covering 12 disease areas have used CPRD data, including recommendations for suspected cancer referrals that drew exclusively on studies using CPRD data.[29]

The accuracy and generalisability of research using CPRD data is underpinned by the population level coverage of the NHS, the longitudinal nature of the database, and the quality of the CPRD data made available for research: EHR data from a total of 22 million patients across the United Kingdom from 1987 onwards are available for public health studies. The composition of the dataset aligns with the overall UK population with respect to age and sex.[6] To provide more information about patient care pathways and disease management across multiple settings, CPRD data are linked to other health care data sources, including inpatient and outpatient hospital care data.[30]

Longitudinal data enable simultaneous retrospective and prospective analyses[31] and are invaluable when there is insufficient time to carry out a randomised clinical trial in response to safety concerns. In 2014, near real-time data were used to evaluate the safety of a new national pertussis immunisation programme for pregnant women, introduced to combat an outbreak of whooping cough in newborns.[11]

KEY POINTS

- UK primary care provides a rich source of data for public health research, but the impact of government-proposed restrictions on data collection has not previously been studied.
- This study analysed the extent to which high-impact studies would have been possible using a database with data collection restrictions in place.
- Overall, 91% of studies were deemed compromised if repeated with all restrictions in place, and 56% of studies were unfeasible even with design modification.
- Findings from this study can be widely used to promote better understanding of the patient and public health benefits of data sharing.

Initial results from observational studies using UK EHRs were available within 6 months, a time frame that would be impractical for randomised clinical trials to generate results.[11]

The NHS England programme, care.data, initiated in 2013 and closed in 2016, sought to capture health care data from all general practices in England.[32] Data were to be collected by the national Health and Social Care Information Centre (HSCIC), now known as NHS Digital and made available by HSCIC for the secondary purposes of commissioning, service planning, and research.

A set of limitations on data collection were proposed during development of the programme:

1. Restricting the retrospective collection of data to specified diseases, risk factors, and conditions defined by Read codes (Table S2). Specified disease areas primarily covered cardiovascular disease, diabetes, respiratory disease, mental health, cancer, and neurodegenerative disease.
2. Restricting retrospective data collection to either 6 years (proposed to begin in 2016, ie, 2010 onwards for the purpose of this study) or 12 years (2004 onwards) (time-limited retrospective data)
3. Restricting the prospective and retrospective collection of sensitive data; the exclusion of legally restricted or particularly sensitive data, eg, in vitro fertilisation, abortion, gender reassignment, sexually transmitted infections, and HIV status (Table S3).

The impact of such restrictions on the output of public health research is not known.

This study aims to systematically review the potential impact of restricting the scope of EHR data on observational and pharmacovigilance research outcomes in accordance with proposed care.data programme data limitations of time, sensitive data, and specific diseases, using the CPRD database as an exemplar.

2 | METHODS

The consequences of the proposed restrictions on observational research were modelled by examining the feasibility of repeating previously conducted CPRD studies under a restricted data collection model.

2.1 | Search strategy

Published, high-impact, observational research studies conducted using anonymised, longitudinal primary care data from CPRD and from its predecessor the General Practice Research Database (GPRD) were identified for review using a systematic approach (Figure 1).

High-impact research was defined as being referenced in a UK clinical guidance document and/or published in a top five journal according to impact factor for the relevant field of study, (a) Pharmacology and Pharmacy (b) Medicine, General, and Internal, and (c) Public Environmental and Occupational Health (Table S1). Impact factors were identified through the Journal Citation Reports database on the ISI Web of Knowledge platform.[33]

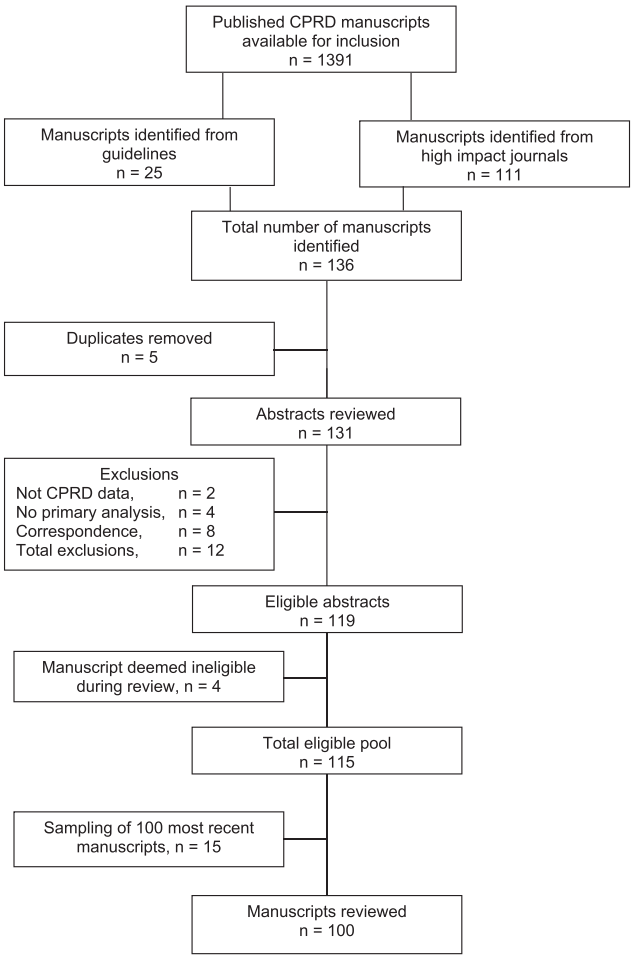

FIGURE 1 Flow diagram of manuscript identification

Studies referenced in UK clinical guidance documents were identified from a recent systematic review.[34] Studies published in a top five journal were identified by searching the CPRD bibliography, which is compiled through systematic searches of PubMed (https://www.ncbi.nlm.nih.gov/pubmed) using the term ("CPRD") OR ("clinical practice research datalink") OR ("GPRD") OR ("general practice research database").

2.2 | Eligibility criteria

Studies were included if they were referenced in a clinical guidance document or had been published in a high-impact journal on or before 20 July 2015 and used CPRD or GPRD data. Studies were excluded if they were published in a language other than English, did not include any primary research, or were not an original report/full journal article. This excluded reviews, meta-analyses, editorials, letters, commentaries, and research letters. Publications on interventional studies were also excluded. No restrictions were imposed on the period covered by the respective publication, and the methodological quality of studies was not considered as part of the eligibility criteria for inclusion in the review.

Eligibility for inclusion in the study was ascertained by independent review of published abstracts by at least two researchers. Duplicate and related publications were flagged and removed. A sample of 100 publications from the total eligible pool was selected for review and data extraction. Publications were sampled with priority given to the most recently published manuscripts. Sampling effectively added a time criterion to the review so that only studies published on or after January 2000 were included.

2.3 | Assessing the impact of restrictions

The impact of restricting data collection on study feasibility and internal validity was evaluated using the questionnaire shown in Box 2. For each publication, two researchers independently assessed full texts of eligible manuscripts using the questionnaire to determine whether the study could be repeated under the restriction scenarios outlined. Researchers assumed a study start date of 2016, in line with the proposed care.data launch date. The impact of restricting retrospective data collection to a limited time period was considered from two perspectives: (a) the impact of restricted data collection over a fixed length of time, ie, either 6 or 12 years of historical data and (b) over a restricted calendar period, ie, not having events recorded before either 01/01/2010 or 01/01/2004 (Table 3).

The questionnaire was initially piloted by five researchers to assess internal consistency and completeness of capture. Discordance between reviewers at the development stage and data abstraction stage was resolved with discussion or by a third reviewer when a consensus could not be reached.

Outcomes were categorised as either unfeasible, ie, not able to be reproduced without the introduction of major bias (eg, selection bias, detection bias, and misclassification); compromised, ie, feasible but requiring study design modification; or unaffected, ie, no impact of the restrictions. No assessment of publication bias was performed.

Box 1. Questionnaire

Study information

1. Publication reference no.
2. Lead author
3. Year of publication
4. Title of publication
5. Study type—Select one of following options only: Adverse drug reaction/drug safety, drug utilisation, disease epidemiology, drug effectiveness, pharmacoeconomics, methodological, health/public health services research, other
6. Multi-database study? (Select one of the following options: Y—other non-linked UK database, Y—other foreign database, Y—other non-linked UK database and other foreign database, N) If Y, list databases
7. Linked data sources (Y, N) If Y, tick all that apply: ONS mortality, HES APC, HES OP, CR, SES-practice, SES-patient, MINAP, other/bespoke linkage
8. Study population or denominator population (ie, all adults over age 18 years)
   a. Select one category (Supplementary Table AAA) that best describes the study population/denominator
   b. Describe study population/denominator including important inclusion and exclusion criteria, time periods, and any other important details
9. Primary study exposure (ie, smoking), if appropriate
   a. Select one category (Supplementary Table AAA) of exposure
   b. Describe exposure
10. Primary study outcome (ie, lung cancer), if appropriate
    a. Select one category (Supplementary Table AAA) of outcome
    b. Describe outcome
11. Important covariates/confounders (defined as covariates/confounders mentioned in introduction or discussion of paper), if applicable. List covariates/confounders
12. Key findings (Copy results and conclusions directly from publication abstract)

Feasibility and bias assessment

Restricting sensitive data

With retrospective and prospective restricted to nonsensitive data, would it be possible to fully define the following, where fully defined means "as defined in the CPRD study"?

1. Study population/denominator: (Select one of: Y, N—because definition includes legally restricted codes, N—because definition includes abortion and gender codes, N—because definition includes other sensitive codes, n/a)
   If N, provide additional details, for example whether codes used as inclusion or exclusion criteria
2. Primary exposure: (Select one of: Y, N—because definition includes legally restricted codes, N—because definition includes abortion and gender codes, N—because definition includes other sensitive codes, n/a)
3. Primary outcome: (Select one of: Y, N—because definition includes legally restricted codes, N—because definition includes abortion and gender codes, N—because definition includes other sensitive codes, n/a)
4. Important covariates/confounders—select N if the answer for any of the covariates/confounders is N: (Select one of: Y, N—because definition includes legally restricted codes, N—because definition includes abortion and gender codes, N—because definition includes other sensitive codes, n/a)
   If N, list variables

If N to any of Q1-Q4, could the study objectives have been met, without major bias, if the definition of the following were changed?

5. Study population/denominator: (Select one of: Y—change the definition in primary care, Y—change definition by adding/using linked HES/ONS data (assuming no restrictions of sensitive date), N, n/a)
6. Primary exposure: (Select one of: Y—change the definition in primary care, Y—change definition by adding/using linked HES/ONS data (assuming no restrictions of sensitive date), N, n/a)
7. Primary outcome: (Select one of: Y—change the definition in primary care, Y—change definition by adding/using linked HES/ONS data (assuming no restrictions of sensitive date), N, n/a)
8. Important covariates/confounders—select N if the answer for any of the covariates/confounders is N: (Select one of: Y—change the definition in primary care, Y—change definition by adding/using linked HES/ONS date (assuming no restrictions of sensitive date), N, n/a)

Excluding historical data prior to 2010

In answering these questions, assume that there would be full coded historic data. The aspect of incomplete/limited historic data capture for specific conditions, risk factors, etc is assessed in Q7 and Q8. All the questions below should be answered as if the CPRD study were to be conducted on January 1, 2016, using only 6 years of historic, retrospective data.

9. Restricting historic information to events recorded on or after to 01/01/2010, what would be the impact of replicating the study using only 5 years of follow-up (mark all that apply)?

a. No impact
b. Bias could be introduced (eg, bias due to misclassification)
c. The likelihood of ascertaining the association between exposure and outcome would be limited because of lag between first exposure and outcome ascertainment (eg, drug exposure and cancer outcome)
d. The ability to explore temporal trends would be limited
e. Study would be unfeasible

10. Restricting historic information to events recorded on or after to 01/01/2010, what would be the impact of not having data prior to 2010 (mark all that apply)?
    a. None
    b. Study not feasible because exposure/outcome not available after 2010
    c. Study not feasible because event of interest (eg, guideline released) occurred prior to 2010
    d. Study not feasible for other reason(s)

Excluding historical data prior to 2004

In answering these questions, assume that there would be full coded historic data. The aspect of incomplete/limited historic data capture for specific conditions, risk factors, etc is assessed in Q7 and Q8. All the questions below should be answered as if the CPRD study were to be conducted on January 1, 2016, using only 12 years of historic, retrospective data.

11. Restricting historic information to events recorded on or after 01/01/2004, what would be the impact of replicating the study using only 11 years of follow-up (mark all that apply)?
    a. None
    b. Bias could be introduced (eg, due to misclassification)
    c. The likelihood of ascertaining the association between exposure and outcome would be limited because of lag between first exposure and outcome ascertainment (eg, drug exposure and cancer outcome)
    d. The ability to explore temporal trends would be limited
    e. Study would be unfeasible
12. Restricting historic information to events recorded on or after to 01/01/2004, what would be the impact of not having data prior to 2004 (mark all that apply)?
    a. None
    b. Study not feasible because exposure/outcome not available after 2004
    c. Study not feasible because event of interest (eg, guideline released) occurred prior to 2004
    d. Study not feasible for other reason(s)

Limiting historical data to specific diseases, risk factors, conditions

In answering the following questions, imagine you want to conduct a study on January 1, 2016. At this point, only historical data would be available and that historical data will be restricted to the codes defined in Table S2. The focus in answering these questions should be around whether particular code groups are available to conduct the study.

13. With restricted collection of historic information to codes for specific diseases, risk factors and conditions, would it be possible to fully define the following where fully defined means "as defined in the CPRD study"?
    a. Study population/denominator: (Select one of: Y, N, n/a)
    b. Primary exposure: (Select one of: Y, N, n/a)
    c. Primary outcome: (Select one of: Y, N, n/a)
    d. Important covariates/confounders—select N if the answer for any of the covariates/confounders is N: (Select one of: Y, N, n/a)
14. If N to any of Q13, could the study objectives have been met, without major bias, if the definition of the following were changed
    a. Study population/denominator: (Select one of: Y, N, n/a)
    b. Primary exposure: (Select one of: Y, N, n/a)
    c. Primary outcome: (Select one of: Y, N, n/a)
    d. Important covariates/confounders—select N if the answer for any of the covariates/confounders is N: (Select one of: Y, N, n/a)

3 | RESULTS

Figure 1 shows the process of identification of manuscripts eligible for inclusion. From a total pool of 1391 studies, 111 publications were identified by impact factor criteria, and 25 were identified by a recently published systematic review of studies using CPRD or GPRD data that were subsequently included in UK clinical guidelines.[29] Following removal of duplicates, 131 unique manuscript abstracts were reviewed for eligibility, of which 119 manuscripts met the full inclusion/exclusion criteria and were eligible for full review. Full review further identified four ineligible manuscripts. Due to resourcing restraints, a sample of 100 publications from the total eligible pool of

115, with priority given to manuscripts with the most recent publication date, were systematically reviewed and included in the analysis (Table S4).

3.1 | Study characteristics

An aggregate summary of the major characteristics of the included manuscripts is shown in Table 1. Of the 100 studies analysed, 75 (75%) used CPRD primary care data alone, and 17 (17%) used CPRD primary care data linked to other datasets, primarily Hospital Episode Statistics Admitted Patient Care (HES APC) data. More than half of the studies (52%) were published in the *British Medical Journal* (BMJ). The majority focused on drug safety analysis (43%), disease epidemiology (31%), or drug effectiveness (14%). Over a third of studies (34%) researched circulatory system diseases. Neoplasms, mental disorders, digestive system diseases, and endocrine- or immunity-related diseases were also covered in the top five disease areas.

3.2 | Retrospective restriction according to disease area

The hypothetical outcomes of repeating the included studies with retrospective data collection restricted to specified diseases, risk factors, conditions, and treatments are shown in Table 2.

The majority of studies (74%) were deemed unfeasible or compromised if conducted under the restricted data collection, often with every area of study design affected. Of the affected studies, 69% (51/74) were unfeasible even with study design modification. Among studies found to be unfeasible, even with study modification, the majority (61%, [31/51]) were due to limitations on the primary outcome; ie, data collected under the restriction did not include or allow for a complete investigation of the primary outcome necessary to answer the research question. Similarly, a large proportion of studies that were compromised but could be modified to meet the study objective required modification to the primary outcome (87%, [20/23]).

3.3 | Retrospective restriction according to time period

A marked proportion of studies were deemed either unfeasible or compromised with restriction of retrospective data collection to a 6- (67%) or 12-year period (22%).

Studies were largely compromised because of the impact of restricted data collection over a limited period of either 6 (65%) or 12 years (18%). Misclassification leading to the introduction of bias was a major limitation; of the studies affected by limited follow-up time, 79% (6 years, 51/65) and 83% (12 years, 15/18) would have suffered from misclassification of exposure and/or outcome (Table 3).

TABLE 1 Characteristics of included studies

Reason for inclusion	%
Published in a high-impact factor journal	79
Referenced in a UK clinical guidance document	17
Both	4
Year of publication	%
2000	8
2001	5
2002	4
2003	5
2004	7
2005	6
2006	4
2007	8
2008	3
2009	6
2010	10
2011	3
2012	6
2013	11
2014	8
2015	2
Journal title	%
British Medical Journal (BMJ)	52
Other	21
Journal of the American Medical Association (JAMA)	10
Lancet	8
International Journal of Epidemiology (IJE)	5
New England Journal of Medicine (NEJM)	3
Annals of Internal Medicine	1
Study type	%
Adverse drug reaction/drug safety	43
Disease epidemiology	31
Drug effectiveness	14
Health/public health services research	8
Other	2
Drug utilisation	1
Methodological	1
Data source	%
CPRD primary care data only	75
CPRD primary care linked data	17
Other nonlinked UK database	6
Other non-UK database	2
Type of CPRD primary care linked data	%
Any	17
HES admitted patient care	10

(Continues)

TABLE 1 (Continued)

Reason for inclusion	%
ONS mortality	7
Deprivation	3
HES outpatient	2
Myocardial Ischaemia National Audit Project (MINAP)	2
Cancer registry	1
Other	1
Five most frequently studied disease areas	%
Circulatory system diseases	34
Mental disorders	16
Neoplasms (ie, cancer related)	13
Digestive system diseases	12
Endocrine, nutritional, metabolic, and immunity disorders	11

Abbreviations: CPRD, Clinical Practice Research Datalink; HES, Hospital Episode Statistics; ONS, Office for National Statistics.

Restriction by calendar period adversely affected studies to a lesser, but still notable, extent compared with restricted follow-up time. Overall, 10% (2010 onwards) and 8% (2004 onwards) of studies were affected by restricted calendar time, primarily because of exposures, outcomes, or events of interest having occurred prior to that point (Table 3).

3.4 | Restricted collection of sensitive data

Restricting the retrospective and prospective collection of sensitive data limited the ability to conduct 10% of all studies included (Table 4), primarily because of limitations on the study population (80%, [8/10]). Of the affected studies, 50% were restricted because they required legally restricted Read codes and 50% because they required abortion or gender Read codes. In several cases, the study population could not be defined because patients with potential immunosuppression due to HIV and/or hepatitis could not be excluded.

Two studies remained unfeasible despite modifications (Table 4). One focused on vaccine exposures during pregnancy and crucially relied on information on pregnancy loss (termination/abortion/miscarriage). The second focused on trends in sexually transmitted infections, many of which are considered legally restricted terms.

Restricting collection of sensitive data had no impact (or was not applicable) on any of the studies when considering the primary exposure and important covariates used in the original research studies.

3.5 | Impact of full restriction

Individually, restricting retrospective data collection to specified disease areas, or to a period of 6 years, had the greatest impact on the feasibility and overall internal validity of the studies reviewed.

When all restrictions were in place, over half (56%) of high-impact observational research studies were deemed unfeasible without the introduction of major biases, and 35% of studies were deemed compromised. A total of 91% of studies were affected by the restrictions proposed.

4 | DISCUSSION

4.1 | Summary

Large population health care databases that are used in public health research exist worldwide.[35-37] However, data quality and the extent to which the database reflects the overall population can vary significantly.[38-40] Medical care in many countries may be funded altogether or in part by medical insurers, meaning patient data are generally not held centrally, are not continuous or representative, and cannot be readily linked to other data sources. In contrast, the UK NHS primary care EHR captures health data on over 90% of the UK population, creating a single continuous medical record for each patient over their lifetime. It follows that UK primary care records, with their ability to be linked to other data sets, are extensively used by regulators and academic and pharmaceutical researchers worldwide.

The aim of the English care.data programme to collect data from all general practices in England for the purposes of secondary uses including research was laudable. Access to a larger patient dataset of comparable research quality to those currently available from existing UK research databases would further extend the utility of UK primary care data, such as for research into rare diseases or events. The limitations on the availability of primary care data for research proposed by

TABLE 2 Impact on published studies due to retrospective data collection restricted to specified diseases only

	Aspect of Study Affected by Restriction				
	Study Population	Exposure	Primary Outcome	Important Covariates	Study as a Whole[a]
Total unfeasible or compromised studies, % (n = 100)	28	25	51	30	**74**
Studies with potential for modified design, % (n = 100)	12	7	20	15	**23**
Studies that could not be done, even with design changes, % (n = 100)	16	18	31	15	**51**

[a]Studies may have more than one reason for being unfeasible or compromised.

Key overall results highlighted in bold.

TABLE 3 Impact on published studies due to restrictions on the time period of data collection[a]

	Historical Data Available From 01/01/2010 (6 y of Data)	Historical Data Available From 01/01/2004 (12 y of Data)
Total unfeasible or compromised studies according to year, % (n = 100)	**67**	**22**
Compromised	55	14
Unfeasible	12	8
Studies compromised because of restricted follow-up, % (n = 100)	65	18
Bias introduced (eg, misclassification)	51	15
Unable to measure exposure-outcome association	25	1
Cannot examine temporal trends	12	3
Study considered unfeasible	6	2
Studies compromised because of restricted calendar time, % (n = 100)	10	8
Exposure/outcome not available	1	0
Event of interest not available	9	8
Unfeasible for other reason	1	0

[a]Some studies are restricted by both follow-up and calendar time and thus included in both categories.

Key overall results highlighted in bold.

TABLE 4 Impact on published studies due to restrictions on collection of sensitive data

	Aspect of Study Affected by Restriction		
	Study Population	Primary Outcome	Total
Total unfeasible or compromised studies, % (n = 100)	8	2	**10**
Requiring legally restricted Read codes	4	1	5
Requiring abortion or gender Read codes	4	1	5
Studies with potential for modified design, % (n = 100)	7	1	**8**
By changing definition of study population	5	1	6
By using linked data from other sources	2	0	2
Studies that could not be done, even with design changes, % (n = 100)	1	1	**2**

Key overall results highlighted in bold.

the care.data programme therefore serve as a working model to understand whether the benefits of increasing the size of the UK population source are outweighed by curtailing the content and longitudinal nature of the information within the EHR. By applying this model, this study found that more than 90% of high-impact studies conducted using CPRD data would be compromised and greater than half would be unfeasible to conduct with all data collection restrictions in place.

Retrospective collection restricted to prespecified disease areas had a major impact on the number of research studies that could be undertaken. Overall, 74% of all studies were affected by the restricted database of which greater than half (51%) were not feasible even with modifications. Loss of feasibility was primarily due to the inability to identify primary outcomes of interest to the research question. Research on musculoskeletal diseases, infectious diseases, and vaccine efficacy would be severely limited, if not impossible under the proposed restrictions. In a recent analysis, Vezyridis and Timmons identified the top keywords and disease areas associated with the output of research publications conducted using UK primary care EHRs since 1995.[6] Smoking, diabetes, cardiovascular disease, mental health, and cancer were among the most frequent health areas researched and appear among the proposed areas included in the present analysis. In contrast, other top keywords and health areas including pregnancy, fracture, gastrointestinal diseases, and vaccination would not be captured with the proposed restrictions.

Studies deemed unfeasible to conduct under the restriction included an investigation published in *The Lancet* confirming the safety of the combined measles, mumps, and rubella vaccination.[17] The investigation followed the Wakefield study (1998), which raised concerns over a link between the vaccine and the development of autism in children and was instrumental in restoring confidence in the safety of the vaccine.[41] Without retrospective collection of vaccination data, future surveillance of vaccine safety, uptake, and effectiveness would be significantly compromised.

Longitudinal research studies rely on long-term follow-up to ensure the accurate differentiation between newly diagnosed (incident) and established (prevalent) conditions. Sufficient follow-up time maintains statistical power when clinical outcome measures do not occur for several years following risk factor exposure or intervention. It is expected that constraining the collection of data to a specified period would substantially affect exposure/outcome classification and the feasibility of research on conditions with long lead time.

In line with this, more than two-thirds of studies analysed were deemed to be limited by reducing the retrospective data available to a 6-year period. Even with the extension of this period to 12 years, almost a quarter of studies were still affected. Misclassification and the consequent introduction of bias was a major limitation imposed by the restricted database, with up to 79% of affected studies falling into this category. This was particularly evident with exposures and risk factors, which may be infrequently recorded or recorded only at the point when a patient registers with their primary care provider, such as smoking status, weight, and body mass index. Additionally, populations that are used as controls in these studies often do not have repeated or recent measures in their records. Such a restriction would have significant implications for future development of accurate risk algorithms and risk predictions tools. For example, an

algorithm to predict the risk of blindness and amputation in individuals with diabetes was developed using data from QResearch (and validated using CPRD data). This study used 16 years of patient data and a range of clinical information to derive a reliable algorithm.[42] Under the proposed longitudinal restrictions, the accuracy of the algorithms and, ultimately, the usefulness of the predictor in clinical practice would be significantly reduced.

Up to one in 10 studies were limited in feasibility because of restriction of data collection to a specific calendar time point, ie, 2010 onwards or 2004 onwards. Among the studies affected included an investigation into the risk of stroke in individuals prescribed antipsychotic medication, prior to changes to safety recommendations in 2004, ie, prior to the hypothetical data collection restriction time frames.[43] Only making data available from a specified time point in the recent past limits the ability to evaluate changes in clinical practice and drug safety guidance in a timely manner.

Protecting patient anonymity is fundamental when conducting research using sensitive information, such as studies on pregnancy, abortion, gender, sexually transmitted infections, and HIV status. The findings here show that one in 10 research studies would potentially be affected by restricting the collection of such sensitive or legally restricted data. Only two studies were deemed to be unfeasible despite design changes, but these included an important public health study establishing an increased risk of first venous thromboembolism among pregnant women admitted to hospital, unrelated to delivery admissions.[44] Research studies investigating HIV or sexually transmitted diseases would be significantly impaired by restrictions on sensitive data. In addition, studies investigating cancer or immunity may be subject to selection bias and effect modification if immunocompromised individuals, such as those with HIV, could not reliably be excluded from the analysis. The inability to carry out research on specific patient groups could have unintended consequences on patient care if critical information cannot be taken into account.

4.2 | Strengths and limitations

This study has several key strengths including the application of a robust review methodology. The review of study eligibility and subsequent assessment was conducted by two independent researchers, as was the assessment of the impact of restrictions. A large sample of studies was included across a broad range of disease areas. However, studies published prior to 2000 were excluded. It should also be noted that this study assessed the initial impact of restricting the collection of retrospective data, which would reduce as the database matured.

4.3 | Comparison with existing literature

The accuracy and degree of completeness of health care data sources is an essential consideration when conducting epidemiological research and pharmacovigilance studies.[38-40] A reduction in data completeness can reduce the applicability and usefulness of the data source for pharmacovigilance and epidemiological studies.[38-40] In particular, data may become less suitable to support pharmacovigilance studies that are dependent on a complete and continuous health record over time.[39] It may therefore be preferable to prioritise data quality over total patient numbers when considering the value of health care database use for research purposes. This study suggests that the value of complete and representative UK EHR databases above larger, less representative databases available in other countries may be significantly diminished when the scope of collection is restricted.

4.4 | Implications for research and practice

Safeguarding the confidentiality and privacy of individuals when sharing health data is paramount.[45] It is crucial to gain public support for responsible data sharing to generate an accurate source of information on which decision makers depend. There have been several initiatives in the United Kingdom aimed at understanding public attitudes to data sharing; it is hoped that the findings from this study can be widely used to promote better understanding of the patient and public health benefits of data sharing.[46-48] Despite scepticism towards government-led health interventions from some sectors, research has indicated that the majority of people do accept the use of health data for public benefit.[48] Furthermore, data show that more than one in 30 UK citizens voluntarily take part in health research studies.[49]

Costs of data collection and privacy concerns are key drivers of policy decisions. This study demonstrates that imposing restrictions on information collected from EHRs may at face value allay privacy concerns; however, this intervention may lead to inadvertent compromises in patient and public safety. Moreover, limiting retrospective data collection might save on data collection costs but prevent secondary use of EHR data for epidemiological or pharmacoepidemiological studies supporting essential drug safety and public health research.

The care.data programme did not progress to the stage of scaled primary care data collection and was closed in 2016. Lessons learnt from this programme can be used to inform the development of future national initiatives in the United Kingdom and beyond seeking to mandate collection of patient EHR. Experiences from the care.data programme highlight the importance of stakeholder engagement, garnering professional and public support and obtaining user input into design, prior to implementing national EHR collection for secondary uses including research.

The findings from this study have applicability for policy and health care decision makers globally who are seeking to develop and implement systems to collect EHRs. It may be politically and financially expedient to limit the scope of data collection. However, it is prudent to also understand the potential impact of these decisions on the ability to address national public health and drug safety concerns in the future.

ETHICS STATEMENT

The authors state that no ethical approval was needed.

CONFLICT OF INTEREST

All authors were employed by CPRD at the time of their contribution to the study.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTION

H.S., R.W., W.M., T.M.T., J.C., D.D., A.G., J.O., and A.K. designed the study, systematically reviewed publications, and carried out analysis. L.C., H.S., R.W., W.M., T.M.T., J.C., D.D., A.G., J.O., A.K., and J.V. drafted and revised the manuscript.

REFERENCES

1. Staff M, Roberts C, March L. The completeness of electronic medical record data for patients with type 2 diabetes in primary care and its implications for computer modelling of predicted clinical outcomes. *Prim Care Diabetes*. 2016;10(5):352-359.
2. Leonard CE, Bresinger CM, Nam YH, et al. The quality of Medicaid and Medicare data obtained from CMS and its contractors: implications for pharmacoepidemiology. *BMC Health Serv Res*. 2017;17(1):304.
3. Rothman K, Greenland S, Lash T. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
4. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. *Annu Rev Public Health*. 2016;37(1):61-81.
5. Howe CJ, Cain LE, Hogan JW. Are all biases missing data problems? *Curr Epidemiol Rep*. 2015;2(3):162-171.
6. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44 (3):827-836.
7. Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Inform Prim Care*. 2004;12:171-177.
8. Hippisley-Cox J, Stables D, Pringle M. QResearch: a new general practice database for research. *Inform Prim Care*. 2004;12:49-50.
9. Vezyridis P, Timmons S. Evolution of primary care databases in UK: a scientometric analysis of research output. *BMJ Open*. 2016;6(10): e012785.
10. Vamos EP, Pape UJ, Curcin V, et al. Effectiveness of the influenza vaccine in preventing admission to hospital and death in people with type 2 diabetes. *CMAJ*. 2016;188(14):E342-E351.
11. Donegan K, King B, Bryan P. Safety of pertussis vaccination in pregnant women in UK: observational study. *BMJ*. 2014;349:g4219.
12. Coupland C, Morriss R, Arthur A, Moore M, Hill T, Hippisley-Cox J. Safety of antidepressants in adults aged under 65: protocol for a cohort study using a large primary care database. *BMC Psychiatry*. 2013;13(1):135.
13. Man SL, Petersen I, Thompson M, Nazareth I. Antiepileptic drugs during pregnancy in primary care: a UK population based study. *PloS One*. 2012;7(12):e52339.
14. de Vries F, Setakis E, van Staa TP. Concomitant use of ibuprofen and paracetamol and the risk of major clinical safety outcomes. *Br J Clin Pharmacol*. 2010;70(3):429-438.
15. Meropol SB, Chan KA, Chen Z, et al. Adverse events associated with prolonged antibiotic use. *Pharmacoepidemiol Drug Saf*. 2008;17 (5):523-532.
16. Hardy JR, Leaderer BP, Holford TR, Hall GC, Bracken MB. Safety of medications prescribed before and during early pregnancy in a cohort of 81,975 mothers from the UK General Practice Research Database. *Pharmacoepidemiol Drug Saf*. 2006;15(8):555-564.
17. Smeeth L, Cook C, Fombonne E, et al. MMR vaccination and pervasive developmental disorders: a case-control study. *Lancet*. 2004;364 (9438):963-969.
18. Baker A, Chen LC, Elliott RA, Godman B. The impact of the 'Better Care Better Value' prescribing policy on the utilisation of angiotensin-converting enzyme inhibitors and angiotensin receptor blockers for treating hypertension in the UK primary care setting: longitudinal quasi-experimental design. *BMC Health Serv Res*. 2015;15:367.
19. Forster AS, Dodhia H, Booth H, et al. Estimating the yield of NHS Health Checks in England: a population-based cohort study. *J Public Health (Oxf)*. 2015;37(2):234-240.
20. Forster AS, Burgess C, Dodhia H, et al. Do health checks improve risk factor detection in primary care? Matched cohort study using electronic health records. *J Public Health (Oxf)*. 2016;38(3):552-559.
21. Szatkowski L, Murray R, Hubbard R, Agrawal S, Huang Y, Britton J. Prevalence of smoking among patients treated in NHS hospitals in England in 2010/2011: a national audit. *Thorax*. 2015;70(5):498-500.
22. Neal RD, Din NU, Hamilton W, et al. Comparison of cancer diagnostic intervals before and after implementation of NICE guidelines: analysis of data from the UK General Practice Research Database. *Br J Cancer*. 2014;110(3):584-592.
23. Smith CJ, Gribbin J, Challen KB, Hubbard RB. The impact of the 2004 NICE guideline and 2003 General Medical Services contract on COPD in primary care in the UK. *QJM*. 2008;101(2):145-153.
24. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*. 2008;336(7659):1475-1482.
25. Hawkins NM, Scholes S, Bajekal M, et al. Community care in England: reducing socioeconomic inequalities in heart failure. *Circulation*. 2012;126(9):1050-1057.
26. Gaitatzis A, Purcell B, Carroll K, Sander JWAS, Majeed A. Differences in the use of health services among people with and without epilepsy in the United Kingdom: socio-economic and disease-specific determinants. *Epilepsy Res*. 2002;50(3):233-241.
27. Hippisley-Cox J, Pringle M. Inequalities in access to coronary angiography and revascularisation: the association of deprivation and location of primary care services. *Br J Gen Pract*. 2000;50(455):449-454.
28. Barker I, Steventon A, Deeny SR. Association between continuity of care in general practice and hospital admissions for ambulatory care sensitive conditions: cross sectional study of routinely collected, person level data. *BMJ*. 2017;356:j84.
29. National Institute for Health and Care Excellence, Suspected cancer: recognition and referral. 2015.
30. Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *European Journal of Epidemiology*. 2018. https://doi.org/10.1007/s10654-018-0442-4;34(1):91-99.
31. Moore E, Chatzidiakou L, Jones RL, et al. Linking e-health records, patient-reported symptoms and environmental exposure data to characterise and model COPD exacerbations: protocol for the COPE study. *BMJ Open*. 2016;6(7):e011330.

32. NHS England. NHS England sets out next steps of public awareness about care.data. Avilable from: https://www.england.nhs.uk/2013/10/care-data
33. Oyinlola JO, Campbell J, Kousoulis AA. The use of real world evidence to influence practice: a systematic review of CPRD studies in English guidances and guidelines. *Value Health*. 2015;18(7):A565.
34. IMS Disease Analyzer. IMS Disease Analyzer. 2017; Available from: http://www.imshealth.com/en/solution-areas/real-world-evidence/real-world-data-rwd.
35. Military Health System Data Repository. Military Health System Data Repository. Available from: https://health.mil/Military-Health-Topics/Technology/Clinical-Support/Military-Health-System-Data-Repository.
36. Centers for Medicare & Medicaid Services. On its 50th anniversary, more than 55 million Americans covered by Medicare. 2015; Available from: https://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-releases/2015-Press-releases-items/2015-07-28.html.
37. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *Int J Epidemiol*. 1996;25(2):435-442.
38. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323-337.
39. Stergachis AS. Record linkage studies for postmarketing drug surveillance: data quality and validity considerations. *Drug Intell Clin Pharm*. 1988;22(2):157-161.
40. Wakefield AJ, Murch SH, Anthony A, et al. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet*. 1998;351(9103):637-641.
41. Hippisley-Cox J, Coupland C. Development and validation of risk prediction equations to estimate future risk of blindness and lower limb amputation in patients with diabetes: cohort study. *BMJ*. 2015;351:h5441.
42. Nuffield Council on Bioethics, The collection, linking and use of data in biomedical research and health care: ethical issues. 2015.
43. Darzi A. We must convince the public that researchers need access to medical records. *BMJ*. 2015;351:h3853.
44. Understanding Patient Data. Understanding Patient Data. 2017; Available from: https://understandingpatientdata.org.uk.
45. Wellcome Trust. *The One-Way Mirror: public attitudes to commercial access to health data*. 2016.
46. Pell J, Valentine J, Inskip H. One in 30 people in the UK take part in cohort studies. *Lancet*. 2014;383(9922):1015-1016.
47. ISI Web of Knowledge. Journal Citation Reports. Available from: https://www.webofknowledge.com/JCR
48. Douglas IJ, Smeeth L. Exposure to antipsychotics and risk of stroke: self controlled case series study. *BMJ*. 2008;337:a1227.
49. Abdul Sultan A, West J, Tata LJ, Fleming KM, Nelson-Piercy C, Grainge MJ. Risk of first venous thromboembolism in pregnant women in hospital: population based cohort study from England. *BMJ*. 2013;347(nov07 15):f6099.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Strongman H, Williams R, Meeraus W, et al. Limitations for health research with restricted data collection from UK primary care. *Pharmacoepidemiol Drug Saf*. 2019;1–11. https://doi.org/10.1002/pds.4765