



Nonindependence of patient data in the Clinical Practice Research Datalink: a case study in atrial fibrillation patients

Cormac J Sammon¹, Thomas P Leahy¹ & Sreeram Ramagopalan^{*,2}

¹PHMR Ltd, Berkeley Works, London, UK

²Centre for Observational Research & Data Sciences, Bristol-Myers Squibb, Uxbridge, UK

*Author for correspondence: s.ramagopalan@lse.ac.uk

Aim: The impact of different strategies to handle patients with data recorded under multiple Clinical Practice Research Datalink (CPRD) identifiers (IDs) is unknown. **Patients and methods:** Six approaches to handling patients appearing under multiple CPRD IDs were defined. The impact of the approaches was illustrated using a case study describing the clinical characteristics of a population of nonvalvular atrial fibrillation patients. **Results:** 5.6% of patients had more than one CPRD ID. Across all six approaches implemented, no material difference in the characteristics of nonvalvular atrial fibrillation patients were observed. **Conclusion:** While strategies which longitudinally append patient registration periods under different CPRD IDs maintain independence while using all available data, their implementation had little impact on the results of our case study.

First draft submitted: 11 December 2019; Accepted for publication: 6 February 2020; Published online: 14 February 2020

Keywords: anticoagulant • Hospital Episode Statistics • linkage

The Clinical Practice Research Datalink (CPRD) GOLD database, formerly the General Practice Research Database, is a large electronic health record database containing details of general practitioner records for 6.9% of the UK population [1]. The database has a long history of use in pharmacoepidemiologic research with over 2200 publications in the past 20 years [2]. In recent years, a routine linkage between the patients attending English practices contributing to CPRD GOLD and their secondary and tertiary care records contained in the Hospital Episode Statistics (HES) database has become available [1,3]. This has allowed for more accurate data on exposures, outcomes and covariates to be obtained for pharmacoepidemiologic studies using the linked databases and for studies investigating the utilization and cost of healthcare to also be carried out.

A limitation of the CPRD GOLD data is that it does not contain lifetime (cradle to grave) follow-up of patients, that is, the population in the CPRD is dynamic with patients entering the databases when they register with a contributing general practitioner (GP) practice and exiting the database when they leave that practice. Specific approaches are therefore taken in designing CPRD studies in an effort to limit the extent to which this impacts results, for example by requiring a disease-free period of follow-up of a certain length to be observed prior to a diagnosis in order to assume that diagnosis is incident [4]. An added difficulty in working with the dynamic CPRD GOLD population, is that it is not possible to track the movement of patients between contributing practices, thus the same patient may contribute two periods of follow-up to the database under different patient identifiers (IDs; [Figure 1](#)). This poses issues for analyses that require assumptions regarding independence of observations in studies using CPRD GOLD data.

In contrast, patients in HES maintain the same ID throughout their time in the database ([Figure 1](#)) [5]. The linkage of CPRD data with HES data therefore provides a potential opportunity to identify cases where the same patient has registered at multiple CPRD GOLD practices and appeared under different IDs. That is, where multiple patient IDs in the CPRD GOLD dataset link to a single patient ID in the HES dataset, this suggests that the same patient has registered at multiple CPRD GOLD practices. Despite this, we are not aware of any studies that have

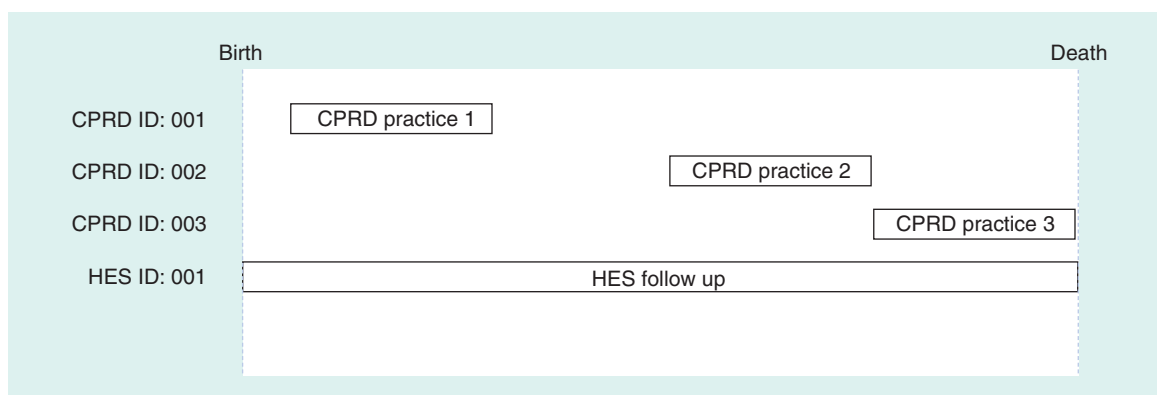


Figure 1. Schematic illustrating how the lifetime follow-up of a single patient could appear in a Clinical Practice Research Datalink-Hospital Episode Statistics linked dataset. The patient registers with three CPRD contributing General practitioner practices in their lifetime and therefore appears under three different CPRD IDs in the database. In contrast, the patient is issued an NHS number at birth which is used to track all of their inpatient and outpatient attendances over their lifetime thereby allowing for lifetime follow-up of patients in this data source. CPRD: Clinical Practice Research Datalink; HES: Hospital Episode Statistics; ID: Identifier.

utilized the linked data to explore this issue in detail, with few published studies providing any detail on how they handle cases where multiple CPRD IDs are linked to a single HES ID.

In light of this, the current study utilized a CPRD GOLD-HES linked dataset of patients with a diagnosis of atrial fibrillation (AF) to describe the frequency of one-to-many matches typically observed between HES and the CPRD GOLD, outline a number of possible approaches to handle one-to-many matches and explore the impact of these approaches on features of the AF study population. As such, the results of the study provide future investigators with information on the extent of the issue and the potential impact the different approaches to handling the issue can have on study results, thereby allowing them to make more informed choices in their study.

Methods

Data source

The study utilized data from the UK Clinical Practice Research Datalink (CPRD GOLD or CPRD for conciseness) and the HES. These databases can be linked through pseudorandomized patient identification numbers. CPRD, formally the General Practice Research Database, is a large primary care database forming a representative sample of 6.9% of the UK general population [1]. HES is a secondary care database in England and records patient data related to presentations in National Health Service (NHS) hospitals or private healthcare institutions where the NHS provides partial funding [6]. CPRD-HES linked data used in this study was acquired and analyzed in line with the Independent Scientific Advisory Committee approved protocol (protocol number: 18_271R).

Study population

This case study used a descriptive and exploratory population-based case-control design. The population of interest was patients who were newly diagnosed with nonvalvular atrial fibrillation (NVAf). Patient follow-up began at the latest of 1 January 2013, 1 year following registration at a GP practice or the date a patient's GP practice met the minimum recording standards for the CPRD. Patient follow-up ended at the earliest of death, 31 December 2017, the date a patient's GP practice last contributed data or the date a patient transferred out of their GP practice.

The study population consisted of individuals with a diagnosis of AF recorded during patient follow-up. As AF diagnosis may occur in secondary care, or lead to referral to secondary care, the index date for each patient was defined as occurring 1 year after their AF diagnosis date. This allowed adequate time for individuals diagnosed in secondary care to be discharged and come back under the care of their GP for the identification of treatment status. The index date was the same for all study populations.

Patients were excluded if they had at least one of the following criteria: aged under 45 years old at their index date, had an AF code recorded prior to their start of follow-up (to capture incident AF cases only), had codes indicative of valvular AF on or ever before their end of follow-up, had less than 1 year follow-up following their AF diagnosis or were ineligible for CPRD and HES linkage.

Table 1. Clinical Practice Research Datalink and Hospital Episode Statistics linkage strategies.

Linkage strategy		Interpretation for illustrative patient in Figure 1
1	Drop all patients with multiple CPRD IDs for a single HES ID from the study population	Drop all data for CPRD ID 001, 002 and 003 and HES ID 001 from the study population
2	Select one CPRD record at random of those that have multiple CPRD records	One of CPRD ID 001, 002 or 003 selected at random for inclusion in the study population and HES ID 001 data from the equivalent period included
3	Keep all patients CPRD linked to a HES ID in study population and treat them as unique patients	Keep CPRD ID 001, 002 and 003 in the study population as separate patients, and include HES ID 001 data from the equivalent periods for each
4	Keep only the CPRD patients with most recent current registration date	Keep CPRD ID 003 in the study population, and HES ID 001 data from the equivalent period
5	Keep only the CPRD patient with the most recent current registration date, unless the gap between the records is short enough to assume continuous capture of GP contacts (30 days) – in which case the records are appended	Keep CPRD ID 003 and 002 in the study population, and HES ID 001 data from the equivalent period
6	Keep only the CPRD patient with the most recent current registration date, unless the gap between the records is short enough to assume continuous capture of GP contacts (30 days) – in which case the records are appended. If the gap is too large to assume continuous capture of GP contacts, use older patient ID to define medical history (but not incident events).	Keep CPRD ID 003 and 002 in the study population and use CPRD ID 001 to support the identification of medical history of events. HES ID 001 data from the equivalent periods included

CPRD: Clinical Practice Research Datalink; GP: General practitioner; HES: Hospital Episode Statistics; ID: Identifier.

Several study populations were defined based on the various options available for handling situations where multiple CPRD IDs were linked to a single HES ID (Table 1).

Cases & controls

Cases were identified as NVAf patients who had an oral anticoagulant (OAC) prescription for a vitamin K antagonist, Dabigatran, Apixaban, Rivaroxaban or Edoxaban between the date of their NVAf diagnosis and index date. Controls were all remaining NVAf patients (i.e. those without an OAC prescription within this period).

Data analysis

For each study population defined based on the approaches outlined in Table 1, covariate prevalence was summarized using descriptive statistics (mean and standard deviation or frequency and percentage). Univariable logistic regression models were implemented to estimate odds ratios and 95% confidence intervals to determine the association between each covariate and OAC use for each of the linkage approaches. The total number of patients for each linkage approach is also reported.

For strategy 5 (Table 1), periods of registration of patients with the same HES ID but more than one CPRD ID are longitudinally appended if the gap between the transferred-out date of a previous registration to the current registration date of the next registration period is less than 30 days. The same is true for strategy 6, however for any registration periods with a gap of greater than 30 days between transferred out date and current registration date, instead of disregarding them, those periods are used to help define patient history but not to define the study population, in this case study, those periods are not used to define incidence of NVAf.

Results

There were 358,101 patients in the CPRD-HES linked dataset with unique HES IDs and AF recorded at some point in their HES or linked CPRD records. Out of these, 338,403 (94.5%) were linked to a single CPRD patient ID, 18,477 (5.16%) of patients were linked to two CPRD patient IDs and approximately 1221 (0.4%) were linked to three or more CPRD patient IDs. There were 11,843 gaps in follow-up between CPRD IDs linked to the same HES ID, out of which 50.5% were 30 days or less and 49.5% were greater than 30 days, with an average gap of 2.63 years and standard deviation of 4.1 years. The magnitude of the mean gap compared with the proportions less than or greater than 30 days reflects the skewness of the gap distribution.

For each of the 6 strategies used to handle situations where multiple CPRD IDs link to a single HES ID, the relevant inclusion and exclusion criteria were applied. The resultant number of patients in each of the populations are provided in Table 2, with the population where patients with multiple CPRD IDs for a single HES ID are dropped (Link 1) having the smallest population of 13,684 compared with keeping all patients and treating them as unique individuals (Link 3) having the largest population of 14,648.

Table 2. Summary of population characteristics for the six linkage strategies.

Factor	Linkage options					
	Link 1	Link 2	Link 3	Link 4	Link 5	Link 6
n	13,684	14,340	14,648	14,183	14,252	14,234
Person years per person, mean (SD)	15.09 (6.10)	14.80 (6.19)	14.78 (6.18)	14.92 (6.14)	14.87 (6.17)	14.98 (6.14)
Age, mean (SD)	75.46 (11.21)	75.32 (11.29)	75.41 (11.30)	75.42 (11.25)	75.33 (11.27)	75.40 (11.24)
Gender:						
– Male	7382 (53.9%)	7753 (54.1%)	7902 (53.9%)	7660 (54.0%)	7703 (54.0%)	7686 (54.0%)
– Female	6302 (46.1%)	6587 (45.9%)	6746 (46.1%)	6523 (46.0%)	6549 (46.0%)	6548 (46.0%)
Weight (kg), mean (SD)	81.36 (20.87)	81.42 (20.91)	81.37 (20.90)	81.40 (20.87)	81.42 (20.85)	81.40 (20.89)
BMI, mean (SD)	28.63 (6.50)	28.66 (6.50)	28.65 (6.50)	28.64 (6.49)	28.65 (6.48)	28.64 (6.49)
Stroke/TIA/TE†:						
– No	10,860 (79.4%)	11,364 (79.2%)	11,599 (79.2%)	11,235 (79.2%)	11,317 (79.4%)	11,329 (79.6%)
– Yes	2824 (20.6%)	2976 (20.8%)	3049 (20.8%)	2948 (20.8%)	2935 (20.6%)	2905 (20.4%)
Chronic heart failure†:						
– No	10,987 (80.3%)	11,518 (80.3%)	11,763 (80.3%)	11,394 (80.3%)	11,449 (80.3%)	11,429 (80.3%)
– Yes	2697 (19.7%)	2822 (19.7%)	2885 (19.7%)	2789 (19.7%)	2803 (19.7%)	2805 (19.7%)
Vascular disease†:						
– No	8989 (65.7%)	9420 (65.7%)	9612 (65.6%)	9316 (65.7%)	9356 (65.6%)	9355 (65.7%)
– Yes	4695 (34.3%)	4920 (34.3%)	5036 (34.4%)	4867 (34.3%)	4896 (34.4%)	4879 (34.3%)
Hypertension†:						
– No	858 (6.3%)	901 (6.3%)	912 (6.2%)	887 (6.3%)	889 (6.2%)	882 (6.2%)
– Yes	12,826 (93.7%)	13,439 (93.7%)	13,736 (93.8%)	13,296 (93.7%)	13,363 (93.8%)	13,352 (93.8%)
Diabetes†:						
– No	10,791 (78.9%)	11,302 (78.8%)	11,538 (78.8%)	11,179 (78.8%)	11,234 (78.8%)	11,219 (78.8%)
– Yes	2893 (21.1%)	3038 (21.2%)	3110 (21.2%)	3004 (21.2%)	3018 (21.2%)	3015 (21.2%)
CHADS2 score, mean (SD)	2.33 (1.28)	2.33 (1.28)	2.33 (1.28)	2.33 (1.28)	2.33 (1.28)	2.33 (1.27)
CHADS2 score:						
– 0–1	3891 (29.9%)	4102 (30.1%)	4154 (29.9%)	4026 (29.9%)	4062 (30.0%)	3993 (29.5%)
– 2–3	7159 (55.1%)	7485 (55.0%)	7667 (55.2%)	7420 (55.1%)	7465 (55.2%)	7525 (55.7%)
– 4–5	1782 (13.7%)	1858 (13.6%)	1902 (13.7%)	1848 (13.7%)	1841 (13.6%)	1834 (13.6%)
– 6+	162 (1.2%)	171 (1.3%)	176 (1.3%)	168 (1.2%)	167 (1.2%)	166 (1.2%)
CHADS2 VASC score, mean (SD)	3.96 (1.72)	3.95 (1.73)	3.96 (1.73)	3.96 (1.72)	3.96 (1.72)	3.97 (1.72)
CHA2DS2 VASC score:						
– 0–1	1047 (7.7%)	1105 (7.7%)	1117 (7.6%)	1083 (7.6%)	1093 (7.7%)	1075 (7.6%)
– 2–3	4509 (33.0%)	4772 (33.3%)	4853 (33.1%)	4700 (33.1%)	4726 (33.2%)	4674 (32.8%)
– 4–5	5829 (42.6%)	6059 (42.3%)	6210 (42.4%)	6012 (42.4%)	6043 (42.4%)	6078 (42.7%)
– 6–7	2046 (15.0%)	2136 (14.9%)	2185 (14.9%)	2120 (14.9%)	2123 (14.9%)	2133 (15.0%)

† Ever before index.

‡ 1 year before index.

BMI: Body mass index; CCI: Charlson comorbidity index; SD: Standard deviation; TE: Thromboembolism; TIA: Transient ischemic attack.

Table 2. Summary of population characteristics for the six linkage strategies (cont.).

Factor	Linkage options					
	Link 1	Link 2	Link 3	Link 4	Link 5	Link 6
– 8+	253 (1.8%)	268 (1.9%)	283 (1.9%)	268 (1.9%)	267 (1.9%)	274 (1.9%)
CCI, mean (SD):	5.75 (3.04)	5.74 (3.05)	5.75 (3.05)	5.75 (3.04)	5.75 (3.05)	5.77 (3.05)
CCI:						
– 1–3	3102 (22.9%)	3277 (23.1%)	3325 (22.9%)	3223 (22.9%)	3245 (23.0%)	3190 (22.6%)
– 4–6	5758 (42.4%)	6014 (42.3%)	6136 (42.3%)	5948 (42.3%)	5974 (42.3%)	5989 (42.4%)
– 7–9	3167 (23.3%)	3309 (23.3%)	3395 (23.4%)	3290 (23.4%)	3294 (23.3%)	3308 (23.4%)
– 10+	1543 (11.4%)	1616 (11.4%)	1666 (11.5%)	1604 (11.4%)	1618 (11.5%)	1630 (11.5%)
Bleed [‡] :						
– No	8620 (63.0%)	9034 (63.0%)	9210 (62.9%)	8946 (63.1%)	8979 (63.0%)	8952 (62.9%)
– Yes	5064 (37.0%)	5306 (37.0%)	5438 (37.1%)	5237 (36.9%)	5273 (37.0%)	5282 (37.1%)
Major bleed [‡] :						
– No	12,379 (90.5%)	12,976 (90.5%)	13,248 (90.4%)	12,839 (90.5%)	12,897 (90.5%)	12,880 (90.5%)
– Yes	1305 (9.5%)	1364 (9.5%)	1400 (9.6%)	1344 (9.5%)	1355 (9.5%)	1354 (9.5%)
Liver disease [†] :						
– No	13,289 (97.1%)	13,920 (97.1%)	14,216 (97.1%)	13,771 (97.1%)	13,837 (97.1%)	13,821 (97.1%)
– Yes	395 (2.9%)	420 (2.9%)	432 (2.9%)	412 (2.9%)	415 (2.9%)	413 (2.9%)
Renal disease [†] :						
– No	10,002 (73.1%)	10,516 (73.3%)	10,717 (73.2%)	10,376 (73.2%)	10,435 (73.2%)	10,392 (73.0%)
– Yes	3682 (26.9%)	3824 (26.7%)	3931 (26.8%)	3807 (26.8%)	3817 (26.8%)	3842 (27.0%)
HAS-BLED risk:						
– Low	5126 (37.5%)	5401 (37.7%)	5484 (37.4%)	5318 (37.5%)	5348 (37.5%)	5306 (37.3%)
– High	8558 (62.5%)	8939 (62.3%)	9164 (62.6%)	8865 (62.5%)	8904 (62.5%)	8928 (62.7%)
Cancer [†] :						
– No	9837 (71.9%)	10,328 (72.0%)	10,544 (72.0%)	10,196 (71.9%)	10,258 (72.0%)	10,224 (71.8%)
– Yes	3847 (28.1%)	4012 (28.0%)	4104 (28.0%)	3987 (28.1%)	3994 (28.0%)	4010 (28.2%)
Cancer [‡] :						
– No	12,327 (90.1%)	12,917 (90.1%)	13,190 (90.0%)	12,770 (90.0%)	12,846 (90.1%)	12,827 (90.1%)
– Yes	1357 (9.9%)	1423 (9.9%)	1458 (10.0%)	1413 (10.0%)	1406 (9.9%)	1407 (9.9%)
Dementia [†] :						
– No	12,743 (93.1%)	13,351 (93.1%)	13,608 (92.9%)	13,192 (93.0%)	13,264 (93.0%)	13,232 (93.0%)
– Yes	941 (6.9%)	989 (6.9%)	1040 (7.1%)	991 (7.0%)	988 (6.9%)	1002 (7.0%)

† Ever before index.

‡ 1 year before index.

BMI: Body mass index; CCI: Charlson comorbidity index; SD: Standard deviation; TE: Thromboembolism; TIA: Transient ischemic attack.

Across all the linkage options, there is little variability in the study population characteristics (Table 2). Population demographics were very similar with the only difference being the number of patients in each of the study populations. Percentages of those with or without prior stroke risk factors and bleeding risk factors are similar across different linkage strategies. Risk scores (CHADS₂, CHA₂DS₂-VASc, HAS-BLED) are also similar. The Charlson Comorbidity index (CCI) also shows little variation between study populations.

More variability is shown in the adjusted odds ratios describing the association between patient characteristics and initiation of OAC treatment (Table 3), however not to the extent that would alter the clinical interpretation of results.

Discussion

In this study, we detail and explore six different methods to handle nonindependence due to the repeated registration of a single patient under different patient IDs in the CPRD GOLD database. The methods were explored using a case study investigating predictors of OAC treatment in a cohort of NVAf patients. Overall, this study found that the prevalence of the issue is low, and that the method used to handle patients appearing under multiple CPRD IDs did not have a significant impact on the results.

The issue explored in this paper has received little focus in the published literature to date, despite the fact that it is an issue encountered in almost all HES-linked CPRD GOLD studies. The lack of recognition of the issue may be due to the expectation that it has little impact on study results given its low prevalence in typical study populations. While our results support this view, we believe it is important that the nuances of real-world data sources are explored, and their impact illustrated empirically rather than making such assumptions *a priori*. As such, our results provide investigators with some reassurance that the approach used to handle situations where multiple CPRD IDs are linked to a single HES ID is likely to have a small impact on results.

While we found that the approach used to handle one-to-many linkages did not overly influence the results of our case study, other research questions might be more sensitive to the inclusion/exclusion of patients with multiple CPRD IDs, or to the misclassification of covariate status. In such cases the approach used may have a greater impact. For example, in a study investigating a rare outcome, exclusion or misclassification of a small number of patients can have a large impact on the results. Careful consideration should therefore be given to how this group of patients is handled in defining a study population. In this context, linkage strategy 5 and 6 make the best use of all available data while excluding nonindependent observations. Options 1, 2 and 4 are more straightforward to implement than options 5 and 6 and would also maintain independence, these may therefore represent more pragmatic options but will result in the omission of some potentially relevant data. Previous work suggests that the use of all available historic data may be the optimum approach when defining time-invariant dichotomous covariates [7–9], therefore options 5 and 6 will likely represent best practice in most study settings. Option 3 would result in the inclusion of nonindependent observations and will therefore violate the assumptions of a number of analytic approaches, including the logistic regression analysis presented in this paper, and should not be used.

Some of the one-to-many linkages encountered in this study may result from linkage errors, where one of the CPRD IDs is erroneously linked to the wrong HES ID. CPRD data are linked to HES data by NHS Digital using an 8-stage deterministic methodology, however, in order to be included in standard research datasets patients must be linked on one of the first 5 sets of linkage criteria, all of which require linkage on NHS number and gender, date of birth and/or postcode [3]. As a result, we expect the occurrence of linkage error to be low and the majority of one-to-many linkages between HES and the CPRD to result from the registration of the same patient under multiple CPRD IDs. To our knowledge there is no published literature examining linkage errors between the CPRD and HES.

Finally, the lack of recognition of the issue in the literature may also derive from a lack of awareness of the issue among investigators inexperienced with the CPRD GOLD database. Our results therefore also serve the purpose of raising awareness of the issue among such investigators and detailing a set of potential approaches that they can use to address the issue. At a minimum, our results should lead to an improvement in the reporting of methodology used to address the issue in published studies, in line with reporting guidelines in this field [10].

Conclusion

While the strategies which longitudinally append patient registration periods under different CPRD IDs maintain independence while using all available data, their implementation had little impact on the results of our case study. More pragmatic options which maintain the independence of observations while applying basic inclusion/exclusion

Table 3. Adjusted odds ratios for each of the population characteristics for each study population of those treated or untreated with an oral anticoagulant.

Factor	Ref.	Linkage options (adjusted odds ratios)					
		Link 1	Link 2	Link 3	Link 4	Link 5	Link 6
Age, mean (SD)		1.03 (1.02, 1.04)	1.03 (1.02, 1.04)	1.03 (1.02, 1.04)	1.03 (1.02, 1.04)	1.03 (1.02, 1.04)	1.03 (1.02, 1.04)
Gender	Male	1.07 (0.93, 1.23)	1.06 (0.92, 1.22)	1.05 (0.92, 1.21)	1.05 (0.92, 1.21)	1.05 (0.91, 1.21)	1.06 (0.92, 1.22)
Weight (kg), mean (SD)		1.02 (1.01, 1.03)	1.02 (1.01, 1.03)	1.02 (1.01, 1.03)	1.02 (1.01, 1.03)	1.02 (1.01, 1.03)	1.02 (1.01, 1.03)
BMI, mean (SD)		0.99 (0.97, 1.01)	0.99 (0.97, 1.01)	0.99 (0.97, 1.01)	0.99 (0.97, 1.01)	0.99 (0.97, 1.01)	0.99 (0.97, 1.01)
Stroke/TIA/TE† (yes/no)	No	1.75 (1.52, 2)	1.75 (1.53, 1.99)	1.77 (1.55, 2.02)	1.76 (1.54, 2.01)	1.73 (1.52, 1.98)	1.72 (1.5, 1.97)
Systemic thromboembolism† (yes/no)	No	1.89 (1.14, 3.13)	1.85 (1.12, 3.03)	1.71 (1.06, 2.75)	1.81 (1.1, 2.98)	1.89 (1.15, 3.13)	1.89 (1.14, 3.12)
Chronic heart failure† (yes/no)	No	1.57 (1.38, 1.79)	1.55 (1.36, 1.76)	1.58 (1.39, 1.79)	1.58 (1.39, 1.8)	1.56 (1.37, 1.78)	1.57 (1.37, 1.79)
Vascular disease† (yes/no)	No	0.7 (0.63, 0.78)	0.7 (0.63, 0.78)	0.69 (0.62, 0.76)	0.7 (0.63, 0.78)	0.7 (0.63, 0.78)	0.7 (0.62, 0.78)
Hypertension† (yes/no)	No	2.72 (2.12, 3.5)	2.68 (2.1, 3.43)	2.66 (2.09, 3.4)	2.7 (2.11, 3.46)	2.7 (2.11, 3.46)	2.71 (2.12, 3.48)
Diabetes† (yes/no)	No	0.89 (0.79, 1)	0.9 (0.8, 1.01)	0.9 (0.81, 1.01)	0.89 (0.79, 0.99)	0.9 (0.81, 1.01)	0.9 (0.8, 1.01)
Bleed‡ (yes/no)	No	1.27 (1.13, 1.43)	1.3 (1.16, 1.46)	1.31 (1.17, 1.47)	1.28 (1.14, 1.43)	1.29 (1.15, 1.45)	1.28 (1.14, 1.44)
Major bleed‡ (yes/no)	No	0.55 (0.46, 0.66)	0.54 (0.45, 0.65)	0.55 (0.46, 0.66)	0.55 (0.46, 0.66)	0.55 (0.46, 0.66)	0.55 (0.46, 0.66)
Liver disease† (yes/no)	No	0.38 (0.28, 0.51)	0.4 (0.3, 0.54)	0.41 (0.31, 0.55)	0.41 (0.3, 0.55)	0.4 (0.3, 0.54)	0.38 (0.28, 0.51)
Renal disease† (yes/no)	No	0.93 (0.83, 1.05)	0.94 (0.84, 1.06)	0.93 (0.83, 1.05)	0.94 (0.84, 1.06)	0.94 (0.84, 1.06)	0.94 (0.84, 1.06)
Cancer† (yes/no)	No	0.92 (0.8, 1.06)	0.92 (0.8, 1.05)	0.91 (0.8, 1.04)	0.92 (0.8, 1.05)	0.92 (0.81, 1.05)	0.92 (0.81, 1.06)
Cancer‡ (yes/no)	No	0.63 (0.52, 0.77)	0.63 (0.52, 0.76)	0.63 (0.52, 0.76)	0.62 (0.51, 0.75)	0.62 (0.51, 0.76)	0.64 (0.52, 0.77)
Dementia‡ (yes/no)	No	0.37 (0.3, 0.47)	0.36 (0.29, 0.45)	0.36 (0.29, 0.44)	0.36 (0.29, 0.44)	0.37 (0.29, 0.45)	0.37 (0.29, 0.46)

† Ever before index.

‡ 1 year before index.

BMI: Body mass index; SD: Standard deviation; TE: Thromboembolism; TIA: Transient ischemic attack.

criteria may perform as well in most study settings. Regardless of the approach used, more transparent reporting of the methodology employed in published studies is warranted.

Summary points

- Patients in the Clinical Practice Research Datalink (CPRD) may appear in the database multiple times under different patient identifiers, potentially introducing bias.
- In contrast, patients typically appear under one identifier in Hospital Episode Statistics (HES) therefore linkage with HES offers an opportunity to explore the issue.
- Among 358,101 patients in our atrial fibrillation case study, 18,477 (5.16%) appeared under two CPRD patient identifiers and 1221 (0.4%) were linked to three or more CPRD patient identifiers.
- We defined six approaches to handle the patients with multiple CPRD identifiers when encountered in studies, none of which had a significant impact on results of our case study.
- We recommend that future studies using CPRD GOLD-HES linked data make clear the strategy used to handle these cases and consider its impact on the interpretation of study results.

Author contributions

CJ Sammon provided substantial contributions to the conception, design of the work and interpretation of data for the work, drafting the work or revising it critically for important intellectual content; final approval of the version to be published. TP Leahy provided substantial contributions to the analysis and interpretation of data for the work, drafting the work; final approval of the version to be published. S Ramagopalan provided substantial contributions to the conception or design of the work and interpretation of data for the work; final approval of the version to be published; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgments

The authors thank the two anonymous reviewers and the editor for their revisions that improved the overall quality of the manuscript.

Financial & competing interests disclosure

CJ Sammon and TP Leahy are employed by PHMR, LLC, who received consulting fees from Bristol Myers Squibb. S Ramagopalan reports personal fees from Bristol-Myers Squibb outside the submitted work. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Open Access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

References

Papers of special note have been highlighted as: ●● of considerable interest

1. Herrett E, Gallagher AM, Bhaskaran K *et al*. Data resource profile: clinical practice research datalink (CPRD). *Int. J. Epidemiol.* 44(3), 827–836 (2015).
- **Data resource profile.**
2. CPRD: bibliography (2019). <https://cprd.com/bibliography?s=year&co=desc>
3. 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. *Eur. J. Epidemiol.* 34(1), 91–99 (2019).
- **Data resource profile related to the linkage of Clinical Practice Research Datalink with other health-related patient data.**
4. Lewis JD, Bilker WB, Weinstein RB, Strom BL. The relationship between time since registration and measured incidence rates in the General Practice Research Database. *Pharmacoepidemiol. Drug Saf.* 14(7), 443–451 (2005).
5. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: hospital episode statistics admitted patient care (HES APC). *Int. J. Epidemiol.* 46(4), 1093–1093i (2017).

•• Data resource profile.

6. NHS UK. Hospital Episode Statistics (HES) (2019). <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics>
7. Brunelli SM, Gagne JJ, Huybrechts KF *et al.* Estimation using all available covariate information versus a fixed look-back window for dichotomous covariates. *Pharmacoepidemiol. Drug Saf.* 22(5), 542–550 (2013).
8. Nakasian SS, Rassen JA, Franklin JM. Effects of expanding the look-back period to all available data in the assessment of covariates. *Pharmacoepidemiol. Drug Saf.* 26(8), 890–899 (2017).
9. Conover MM, Stürmer T, Poole C *et al.* Classifying medical histories in US Medicare beneficiaries using fixed vs all-available look-back approaches. *Pharmacoepidemiol. Drug Saf.* 27(7), 771–780 (2018).
10. Benchimol EI, Smeeth L, Guttman A *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* 12(10), e1001885 (2015).