Short Communication

For reprint orders, please contact: reprints@futuremedicine.com

Journal of **Comparative Effectiveness Research**

Approaches for the identification of chronic kidney disease in CPRD–HES-linked studies

Sreeram Ramagopalan*,[1], Thomas P Leahy[2], Elaine Stamp[2] & Cormac Sammon[2]

[1]Centre for Observational Research & Data Sciences, Bristol-Myers Squibb, Uxbridge, UK

[2]PHMR Ltd, Berkeley Works, London, UK

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

Aim: There are different methods to identify chronic kidney disease (CKD) in Clinical Practice Research Datalink (CPRD)-Hospital Episode Statistics (HES). **Methods:** Using CPRD-HES, nonvalvular atrial fibrillation patients were classified according to CKD category. **Results:** Using glomerular filtration rate/estimated glomerular filtration rate tests only to identify patients with CKD resulted in 3.5% stage 2, 2.7% stage 3, 0.3% stage 4 and 0.03% stage 5. Using data from diagnostic codes to identify patients with CKD resulted in 1.4% stage 3, 0.4% stage 4 and 0.3% stage 5. Using test records and codes resulted in 3.5% stage 2, 4.0% stage 3, 0.6% stage 4 and 0.4% stage 5. **Conclusion:** To identify CKD status in CPRD-HES, a combination of test records and codes should be used. Using diagnostic codes only significantly underestimates CKD prevalence.

First draft submitted: 11 December 2019; Accepted for publication: 20 February 2020; Published online: 9 March 2020

Keywords: atrial fibrillation • chronic kidney disease • CPRD • diagnostic codes • GFR • pharmacoepidemiology • test results

The Clinical Practice Research Datalink (CPRD) GOLD and Hospital Episode Statistics (HES) databases are two of the UK's most widely used electronic healthcare databases [1,2]. Data from the two databases are routinely linked and widely used to perform epidemiologic and health services research. When using the databases for research, the presence or absence of a medical condition is typically identified through the use of diagnostic codes, with diagnoses made in inpatient settings recorded in the HES dataset using International Statistical Classification of Diseases and Related Health Problems (ICD) codes, and clinically significant diagnoses seen in primary care recorded in the CPRD GOLD dataset using Read codes. Read codes are a clinical vocabulary describing a range of patient details including diagnoses, test results and demographics [3].

Given the real-world, dynamic nature of the datasets, diagnoses may be recorded prior to the start of patient follow-up, may be miscoded or may only be coded in other sources. When using the databases for research, investigators therefore often take a more exhaustive approach to case identification, utilizing data specific to the management and monitoring of the condition to supplement the diagnostic data. Diagnostic tests and treatments are two common data elements used as proxies or additional support for a diagnostic code.

Chronic kidney disease (CKD) is a prevalent and debilitating condition characterized by a progressive deterioration in kidney function. Progression of the condition leads to significant morbidity and mortality and places a considerable burden on healthcare systems. Given the progressive nature of CKD, and the variation in resource utilization and outcomes according to the stage of an individual's condition, it is typically important that epidemiological studies accurately characterize the specific stage of an individuals' condition at specific points in time.

The CPRD and HES databases have been regularly used to investigate the epidemiology of CKD and its relationship with other medical events [4–7]. The approaches used to define CKD in such studies have varied with some investigators using diagnostic codes alone, some using test data alone and other using a combination of these

Future Medicine

sources. While a number of studies have also commented on the challenges of identifying CKD in the database, particularly with diagnosis codes, few have directly investigated the issue [8–10].

In this study, we characterize the CKD stage of a cohort of patients diagnosed with atrial fibrillation (AF), investigating the relative contribution of diagnostic codes, test data and treatment data to our ability to classify their CKD stage. The combination of AF and CKD is frequently associated with adverse outcomes and an understanding of how to define CKD in an AF population is important for future studies [11]. The results of the study will support researchers using the CPRD-HES datasets to better design and execute their studies and will support readers in better understanding the potential limitations of published studies.

Material & methods

Data sources

The CPRD is a large UK primary care database of anonymized medical records collected from general practitioners. It collects a variety of information on demographics, diagnoses, referrals, medications, tests and immunizations and is deemed to be representative of the UK population [2,12]. The CPRD covers approximately 6.9% of the UK population accruing information from 674 practices with approximately 50% of practices in England eligible for HES linkage [2].

HES data from admitted patient care and out-patients were also used. HES admitted patient care data are collected on all admissions to NHS hospitals in England and has an approximate 98% coverage of all hospital admissions [1]. HES out-patients data consists of any outpatient appointments in NHS hospitals in England. In the financial year 2017–2018, there were more than 119 million outpatient appointments recorded in HES [13].

Read and ICD-10 codes used in this study are provided in the Supplementary Data.

Study population

Patient follow-up began at the latest of 1 Januray 2013, an age of 45 years, 1 year following the registration at a CPRD practice or a practice Up-To-Standard date. Patient follow-up ended at the earlier of death, 31 December 2017 or the date a patient transferred out of the practice. The population was restricted to those individuals with a diagnosis of AF recorded in the CPRD or HES during patient follow-up and no AF codes recorded prior to the start of patient follow-up, in other words incident AF cases.

Individuals with codes indicative of valvular AF were identified and excluded to create a cohort of nonvalvular AF (NVAF) patients. Individuals ineligible for linkage with HES were also excluded. Individuals with a sex other than male or female were excluded, including missing information, as this variable is required to calculate the primary outcome measure.

Baseline CKD stage

The CKD stage of a patient was determined using test records of their glomerular filtration rate (GFR), dialysis codes and specific CKD diagnostic codes. Additional to specific GFR test records, the estimated GFR (eGFR) was calculated using the CKD Epidemiology Collaboration equations using available data on age, sex, serum creatinine and race [14].

Each patient's CKD stage at their NVAF diagnosis date was determined using the following criteria. For specific CKD diagnostic codes, the most recent code in the year prior to the index date was used. For GFR/eGFR, the most recent two consecutive tests of the same category in the year prior to the index date greater than 3 months apart in line with Kidney Disease Improving Outcomes and National Institute for Health Care and Excellence [15,16]. For dialysis codes, baseline was defined as two dialysis codes greater than 14 days apart but less than 6 months apart. A dialysis code was deemed to be reflective of GFR category 5 [17]. The minimum separation of dialysis codes was designed to prevent the misclassification of acute kidney injury as CKD. Patients whose baseline CKD category could not be determined were excluded from the study.

Variance in the measurement of serum creatine is not an issue in this population as in the UK, laboratory-specific standardization was phased in from 2006 for the measurement of serum creatine and started to calibrate creatine assays to a reference assay using isotope-dilution mass spectrometry [18]. By the start of the follow-up period in this study (1 January 2013) serum creatine assays are expected to be fully standardized [19].

Table 1. The renal function categorized by glomerular filtration rate category of the nonvalvular atrial fibrillation population at their diagnosis date.

Severity	GFR/eGFR (tests)	Diagnosis codes	Dialysis codes	GFR/eGFR (tests) + diagnosis codes + dialysis codes
Stage 1	153 (0.6%)	<5	–	162 (0.6%)
Stage 2	936 (3.5%)	<20	–	938 (3.5%)
Stage 3	724 (2.7%)	384 (1.4%)	–	1,065 (4.0%)
Stage 4	66 (0.3%)	106 (0.4%)	–	159 (0.6%)
Stage 5	8 (0.03%)	86 (0.3%)	33 (0.1%)	97 (0.4%)
Not available/present†	24,740 (92.9%)	26,031 (97.8%)	26,594 (99.9%)	24,206 (90.9%)

At index date (n = 26,627).
†Patients whose GFR category could not be determined at baseline.
GFR/eGFR: Glomerular filtration rate/estimated glomerular filtration rate.

Statistical analysis

Summary table showing the number and percentage of patients in each GFR category at baseline (index date) were calculated. This table is divided into four, showing the GFR categories for GFR/eGFR test records only, specific CKD diagnostic codes only, dialysis codes only and a combination of the previous three. Only those individuals who contributed a full year of follow-up prior to each time point were included in the calculations for that time point. Table cells comprising 0–4 patients were recorded as '<5' in line with CPRD small cell policy.

Results

Among 26,627 patients diagnosed with AF it was possible to classify 2421 (9.1%) of them as having CKD of a certain stage at the time of diagnosis using at least one of the identification approaches.

Using GFR/eGFR test data to classify CKD stage at diagnosis resulted in the classification of 1887 individuals at a CKD stage (Table 1). The majority of these individuals were classified as having stage 1 or stage 2 CKD at diagnosis (4.1%), with 2.7% classified as stage 3, 0.3% classified at stage 4 and 0.03% classified at stage 5. Using diagnostic codes allowed for the classification of only 596 patients at a CKD stage at diagnosis. Less than 25 (<0.1%) of patients classified using diagnostic codes were assigned to stage 1 and stage 2, with 1.4% classified at stage 3, 0.4% at stage 4 and 0.3% at stage 5. Diagnostic codes allowed for the classification of an additional 33 patients at stage 5 at diagnosis.

Patient characteristics stratified by CKD stage and by CKD classification method is contained in Supplementary Table 1.

Discussion

In this study, we compared the use of different sources of data (diagnostic, test and treatment) to classify NVAF patients in the CPRD-HES according to their CKD stage. The study found that using CKD diagnostic codes alone to determine CKD stage in the database is not a feasible approach as stage-specific diagnostic codes are recorded in less than 32% of individuals who have tests indicative of CKD. Using test records alone allowed for the classification of CKD stage in a larger number of individuals and captured most of the cases identified through diagnostics, however it underestimated the number of patients with the most severe level (stage 5).

There are a number of explanations for the poor performance of diagnostic codes in identifying the stage of CKD in an individual at a given moment in time. In order to ensure that we had a contemporary measure of an individual's CKD stage, we required a diagnostic code to be recorded within 1 year prior to their NVAF diagnosis. However, if an individual entered that CKD stage more than 1 year prior to their NVAF diagnosis they are unlikely to have their diagnosis re-recorded within the year prior. Even among those who were first diagnosed with CKD in the year prior to their NVAF diagnosis, a large number of individuals are likely to have had their diagnosis recorded under a more general CKD code which does not specify the disease stage. This points to an important coding issue in the CPRD where, despite the availability of very specific diagnostic codes, recording under less specific codes appears to be common practice. The lack of stage 1 and stage 2 CKD diagnostic codes was also particularly notable and is likely indicative of the perceived low clinical significance of these diagnoses. While the reasons underlying the under-recording appear clear, any study using the CPRD-HES is likely to be subject to these issues and diagnostic

codes are therefore not a viable standalone option for the classification of populations according to their CKD stage.

The use of GFR and eGFR data alone to identify CKD stage appeared to perform relatively well compared with the other approaches with the exception of the more acute stages 4 and 5. This is aligned with previous suggestions that when investigating CKD using electronic health record databases, test records of GFR or eGFR are used as the gold standard for determining renal function and hence GFR category [9,20–23]. Despite this, our study finds that the use of GFR/eGFR test data only did result in the misclassification of a number of severe CKD cases, particularly those with stage 4 and 5 CKD. Notably, the absolute number of additional stage 5 CKD cases identified using the combined approach was (n = 89), this represents more than a 11-fold increase in the number of stage 5 cases over the number identified using test records only and therefore could have a significant impact on study results. This finding likely reflects the fact that the CPRD does not capture diagnostic tests carried out in inpatient and outpatient settings, the settings in which more severe CKD is typically confirmed and managed. The fact that some of these cases are captured using diagnostic codes only is unsurprising, given that upon receiving a discharge summary or specialist letter confirming a stage 4 or 5 CKD diagnosis, a GP practice is likely to deem it clinically important and code it in the database. Similarly, individuals on dialysis in inpatient and outpatient settings are more likely to have their GFR estimated in these settings than by their GP. While GFR recording appears to identify a reasonable distribution of CKD patients, its use in the absence of diagnostic data may lead to the identification of a less severe sample of CKD patients.

As pointed out above, GFR/eGFR testing of patients with more severe CKD will often be performed in secondary and tertiary healthcare and will not be captured in the databases. While this does not impact the validity of our recommendations for research, we take the opportunity to highlight it as a limitation of the databases which should be carefully considered in any study investigating CKD. Also, while the results suggest that there may be under-recording of CKD diagnoses in healthcare records, we recommend that the results are only used to inform best practice in research. There are a number of reasons for this recommendation, including the above note regarding the lack of outpatient data and the fact that a large amount of data in primary care are recorded in free text fields which are not available within the CPRD. As a result, a patient's CKD stage may be adequately recorded for clinical practice, but in data fields not available to us for research. Further sensitivity in the results can arise from the definition of CKD. Although literature of the prevalence of CKD in a NVAF population is limited, the reported prevalence of CKD in a NVAF population varies [24–26]. Some of this variation can be due to the method of defining CKD. For example, in this study, if the CKD guideline definition of tests occurring greater than 3 months apart was relaxed, 12,616 patients CKD status could be calculated based on GFR/eGFR tests compared with 1,887 when applying the guidelines CKD definition. Finally, we have based our recommendations on an internal comparison of approaches to identify CKD stage, we have not sought to compare the prevalence we observe with that observed in other cohorts of NVAF patients.

In terms of implications for future studies, we recommend that all available information on diagnoses, tests and treatment should be utilized to obtain a reflective distribution of CKD category. This will help to avoid biases generated by the lack of test records in the most/least severe GFR categories. Beyond the design of studies, we recommend that those reading CPRD-HES studies bear these findings in mind when interpreting the results of such studies.

Conclusion

This study finds that within CPRD and HES, the CKD stage of a population should be determined by using both GFR/eGFR test records and specific CKD diagnostic codes when available; using diagnoses alone would substantially underestimate CKD.

Author contributions

S Ramagopalan – 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. TP Leahy – substantial contributions to the analysis and interpretation of data for the work, drafting the work; final approval of the version to be published. E Stamp – manuscript revision and final approval of the version to be published. C Sammon – 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.

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

This work was supported by funding from Bristol-Myers Squibb-Pfizer Alliance. S Ramagopalan reports personal fees from Bristol-Myers Squibb outside the submitted work. TP Leahy, E Stamp and C Sammon are employed by PHMR, LLC, who received consulting fees from Bristol Myers Squibb. 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.

Summary points

- There are several approaches to define chronic kidney disease (CKD) status, using only one approach can potentially induce bias.
- Among 26,627 patients diagnosed with atrial fibrillation it was possible to classify 2,421 (9.1%) of them as having CKD using at least one of the identification approaches.
- Using diagnostic codes allowed for the classification of 596 CKD patients whereas using glomerular filtration rate/estimated glomerular filtration rate allowed for the classification of 1,887 CKD patients.
- Combining glomerular filtration rate/estimated glomerular filtration rate test records and specific CKD diagnostic codes is found to be the best approach to identify CKD stage.

References

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

1. 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.**

2. 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.**

3. Bentley T, Price C, Brown P. Structural and lexical features of successive versions of the Read Codes. Presented at: *Proceedings of the Annual Conference of the Primary Health Care Specialist Group*. Worcester, UK (1996).

4. Mc Donald HI, Thomas SL, Millett ER, Nitsch D. CKD and the risk of acute, community-acquired infections among older people with diabetes mellitus: a retrospective cohort study using electronic health records. *Am. J. Kidney Dis.* 66(1), 60–68 (2015).

5. Iwagami M, Mansfield KE, Hayes JF *et al.* Severe mental illness and chronic kidney disease: a cross-sectional study in the United Kingdom. *Clin. Epidemiol.* 10, 421–429 (2018).

6. Hamada S, Gulliford MC. Multiple risk factor control, mortality and cardiovascular events in type 2 diabetes and chronic kidney disease: a population-based cohort study. *BMJ Open* 8(5), e019950 (2018).

7. Watanabe H, Watanabe T, Sasaki S, Nagai K, Roden DM, Aizawa Y. Close bidirectional relationship between chronic kidney disease and atrial fibrillation: the Niigata preventive medicine study. *Am. Heart J.* 158(4), 629–636 (2009).

8. Jain P, Calvert M, Cockwell P, McManus RJ. The need for improved identification and accurate classification of stages 3–5 chronic kidney disease in primary care: retrospective cohort study. *PLoS ONE* 9(8), e100831 (2014).

•• **Compares chronic kidney disease (CKD) prevelence using glomerular filtration rate laboratory test results.**

9. Winkelmayer WC, Schneeweiss S, Mogun H, Patrick AR, Avorn J, Solomon DH. Identification of individuals with CKD from Medicare claims data: a validation study. *Am. J. Kidney Dis.* 46(2), 225–232 (2005).

10. Ronksley PE, Tonelli M, Quan H *et al.* Validating a case definition for chronic kidney disease using administrative data. *Nephrol. Dial. Transplant.* 27(5), 1826–1831 (2011).

11. Kiuchi MG. Atrial fibrillation and chronic kidney disease: a bad combination. *Kidney Res. Clin. Pract.* 37(2), 103 (2018).

12. Walley T, Mantgani A. The UK general practice research database. *Lancet* 350(9084), 1097–1099 (1997).

13. NHS Digital. Hospital Outpatient Activity, 2017–18. (2019). https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/2017-18

14. Levey AS, Stevens LA, Schmid CH *et al.* A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* 150(9), 604–612 (2009).

15. Eknoyan G, Lameire N, Eckardt K *et al.* KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int.* 3(1), 5–14 (2013).

16. National Institute for Health Care Excellence. Chronic kidney disease in adults: assessment and management (2014). www.nice.org.uk/guidance/cg182

•• **CKD identification and classification guidelines.**

17. Levin A, Stevens PE, Bilous RW *et al.* Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int. Suppl.* 3(1), 1–150 (2013).

•• **CKD identification and classification guidelines.**

18. McDonald HI, Shaw C, Thomas SL, Mansfield KE, Tomlinson LA, Nitsch D. Methodological challenges when carrying out research on CKD and AKI using routine electronic health records. *Kidney Int.* 90(5), 943–949 (2016).

19. Poh N, Mcgovern AP, De Lusignan S. Improving the measurement of longitudinal change in renal function: automated detection of changes in laboratory creatinine assay. *J. Innov. Health Inform.* 22(2), 293–301 (2015).

20. Denburg MR, Haynes K, Shults J, Lewis JD, Leonard MB. Validation of The Health Improvement Network (THIN) database for epidemiologic studies of chronic kidney disease. *Pharmacoepidemiol. Drug Saf.* 20(11), 1138–1149 (2011).

21. Mathur R, Dreyer G, Yaqoob MM, Hull SA. Ethnic differences in the progression of chronic kidney disease and risk of death in a UK diabetic population: an observational cohort study. *BMJ Open* 8(3), e020145 (2018).

22. Fleet JL, Dixon SN, Shariff SZ *et al.* Detecting chronic kidney disease in population-based administrative databases using an algorithm of hospital encounter and physician claim codes. *BMC Nephrol.* 14(1), 81 (2013).

23. So L, Evans D, Quan H. ICD-10 coding algorithms for defining comorbidities of acute myocardial infarction. *BMC Health Serv. Res.* 6(1), 161 (2006).

24. Go AS, Fang MC, Udaltsova N *et al.* Impact of proteinuria and glomerular filtration rate on risk of thromboembolism in atrial fibrillation: the ATRIA study. *Circulation* 119(10), 1363 (2009).

25. Kooiman J, Van De Peppel W, Van Der Meer F, Huisman M. Incidence of chronic kidney disease in patients with atrial fibrillation and its relevance for prescribing new oral antithrombotic drugs. *J. Thromb. Haemost.* 9(8), 1652–1653 (2011).

26. Nelson SE, Shroff GR, Li S, Herzog CA. Impact of chronic kidney disease on risk of incident atrial fibrillation and subsequent survival in medicare patients. *J. Am. Heart Assoc.* 1(4), e002097 (2012).