Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

COMMENTARY

Use of the CPRD Aurum Database: Insights Gained from New Data Quality Assessments

Susan Jick[1,2], Catherine Vasilakis-Scaramozza[1], Rebecca Persson[1], David Neasham[3], George Kafatos[3], Katrina Wilcox Hagberg[1]

[1]Epidemiology, Boston Collaborative Drug Surveillance Program, Lexington, MA, USA; [2]Epidemiology, Boston University School of Public Health, Boston, MA, USA; [3]Center for Observational Research, Amgen Ltd, Uxbridge, UK

Correspondence: Susan Jick, Boston Collaborative Drug Surveillance Program, 11 Muzzey Street, Lexington, MA, 02421, USA, Tel +1 781 862 6660, Fax +1 781 862 1680, Email sjick@bu.edu

Abstract: Ongoing evaluation of any electronic health data source is critical to assess suitability for its use in medical research. In addition, familiarity with a data source's history and recording practices is important for making informed data source selection, study design choices, and interpretation of results. In this commentary, the authors discuss three studies that assessed different aspects of the quality and completeness of information contained in Clinical Practice Research Datalink (CPRD) Aurum compared to the well-established CPRD GOLD and to other linked data sources, with the aim to describe insights gained through these data quality assessments. Our findings support the view that CPRD Aurum and GOLD are both valuable tools for studies based on information recorded in primary care but should not be used without critical consideration of strengths and limitations. Further, use of linked data should be considered for some studies, after taking into account all relevant factors.

Keywords: clinical practice research datalink, CPRD Aurum, CPRD GOLD, validation, data quality

Introduction

Clinical Practice Research Datalink (CPRD) GP Online Database (GOLD), a longitudinal United Kingdom (UK) population-based electronic medical record database originally created for research purposes, has for decades been an important source of primary care data. Participating GPs contribute deidentified data, including medical diagnoses, symptoms, referrals, demographic information, and outpatient prescriptions. CPRD GOLD's strengths and limitations for use in medical and public health research are well described. In recent years, there has been a reduction in the number of general practices that use the Vision data platform upon which CPRD GOLD is based. In 2018, a new longitudinal population-based electronic medical record database (CPRD Aurum) was introduced based on a different data platform (EMIS) with high coverage in England.

Ongoing evaluation of any data source is critical to assess suitability for its use in medical research. Familiarity with the influences of a data source's history and key characteristics on data recording (ie, patient management software, coding systems) provides important context for researchers to make informed data source selection and study design choices, and to interpret results appropriately. This current issue of Clinical Epidemiology includes three studies that assessed different aspects of the quality and completeness of information contained in CPRD Aurum compared to the well-established CPRD GOLD, as well as to linked data sources. The aim of this communication is to describe insights gained on the use of CPRD databases based on the findings of these studies and other prior evaluations.

Key Findings

As described in Hagberg et al[1] and in Table 1, there are similarities and differences in the CPRD GOLD and CPRD Aurum databases that should be taken into consideration when planning studies (eg, GP software, coding systems, geographical coverage, population size, recency of data). In the comparisons of CPRD Aurum to CPRD GOLD for breast

Received: 10 August 2023
Accepted: 29 November 2023
Published: 16 December 2023

Table 1 Data Use Considerations for Studies Using CPRD Aurum, CPRD GOLD, and Linked Data

CPRD Aurum and CPRD GOLD data coverage	• CPRD Aurum contains more patients than GOLD, particularly among currently contributing practices[a] • CPRD Aurum covers primary English practices (data from 1989 to present)[a] • CPRD GOLD covers practices from all UK countries; however, currently contributing practices are primarily in Scotland and Wales (data from 1989 to present)[b]
CPRD Aurum and CPRD GOLD data quality	• Similar data quality in GP records for CPRD Aurum and GOLD, particularly after 2004, though there is variability in quality and completeness over time • Validation efforts should be an ongoing component of research using either database
Hospital Episode Statistics and ONS Death Registration	• HES and ONS linkage are available for practices in England • Most CPRD Aurum practices have linkage to HES and ONS • Very few currently contributing CPRD GOLD practices have linkage to HES and ONS data (due to linkage availability for English practices only) • HES APC started in 1997 and ONS death registration data started in 1998, after the start of CPRD GOLD and Aurum data • HES APC and ONS are updated approximately yearly[c] • HES OP has limited capture of diagnosis information
Cancer Registry	• Cancer Registry linkage available for practices in England • Most CPRD Aurum practices have linkage to the Cancer Registry • Very few currently contributing CPRD GOLD practices have linkage to the Cancer Registry (due to linkage availability for English practices only) • Cancer Registry data include cancer registration data from 1990, with a lag in data availability [d] • Establishment of a new national standard for reporting cancer in England in 2013 has improved data capture; completeness of data fields varies significantly by tumour type and calendar time

Notes: [a]CPRD Aurum contains ~13.3 million patients in currently contributing practices primarily in England (>99%) as of the time of this publication. [b]CPRD GOLD contains ~3 million patients in currently contributing practices in England (4%), Scotland (56%), Wales (30%), and Northern Ireland (10%) as of the time of this publication. [c]At the time of this publication, HES APC and ONS data were available through March 2021 (HES OP April 2003–October 2020). [d]At the time of this publication, Cancer Registry data were available through December 2018. Process of protocol approval to data delivery can take 12–18 months.

Abbreviations: GP, general practitioner; HES, Hospital Episode Statistics; HES APC, HES Admitted Patient Care (inpatient) data; HES OP, HES Outpatient data; ONS, Office of National Statistics.

cancer[1] and rheumatoid arthritis (RA)[2] published in this issue, the recording of diagnoses, treatments, and prescription drugs were similar. Slight differences were found for a few of the data elements, but overall, there was consistency between the two databases, particularly for the most informative clinical details. In addition, the age-standardized incidence rates of breast cancer and RA were similar between databases. Also, compared to external linked hospital and cancer registry data, the correctness and completeness of breast cancer diagnoses recorded in CPRD Aurum were high and similar compared to CPRD GOLD, providing further reassurance in the use of CPRD Aurum for research on breast cancer.[3]

Other studies indicate that the accuracy of diagnosis information present in CPRD Aurum is high compared to linked Hospital Episode Statistics (HES) Admitted Patient Care (APC) for pulmonary embolism, acute myocardial infarction, and specified malignant cancers.[4–6] Other studies have reported high internal consistency between diagnoses and presence of relevant clinical codes (labs, prescriptions, and other clinical care) for type 2 diabetes, hyperlipidemia, anemia,[7] and RA.[8] Indication for use of benign prostatic hyperplasia drugs was present for a large proportion of patients, suggesting that CPRD Aurum may adequately capture this important detail for drugs prescribed by GPs.[9] Finally, the number of recorded prescriptions for the most common antibiotics was similar in CPRD GOLD and CPRD Aurum,[10] though drugs prescribed in specialty care may have low capture.[11] Overall, these studies suggest that, where present, the clinical information recorded in CPRD Aurum is of high validity for use in research.

On the other hand, studies to date have indicated that completeness (ie, non-missingness) of coded diagnosis information in CPRD Aurum varies depending on the condition under study.[4–8] Additional breast cancer cases (beyond those recorded in CPRD Aurum or GOLD) were captured in HES APC (inpatient) and, to a smaller extent, the Cancer Registry. Few additional cases were added through HES Outpatient (OP) alone,[3] which is consistent with reports that estimate only 5% of outpatient hospital visits captured in HES OP have a recorded diagnosis.[12] Thus, the addition of HES OP alone will rarely improve case capture. For conditions treated in specialty or hospital settings, linked data may improve case capture, an important consideration for certain conditions and study designs. However, linkages are not available for all diseases of interest. Chronic diseases primarily treated in specialist outpatient settings, such as RA, do not necessarily have relevant supporting linked data.

While valuable data sources, it is important to note that HES and the Cancer Registry data are dynamic and have changed over time due to variations in policies, funding, and other administrative impacts. Neither offers a true gold standard as neither provides complete case capture.[13] Researchers should be aware that each data resource has limitations (Table 1).

Researchers should also be aware of practical impacts of linked data use, including reduced sample size (not all practices have linked data), geographic generalizability (virtually all linked practices in CPRD are based in England), and considerable lag in data availability depending on the data source (CPRD Linkage website). There are also significant practical considerations, including additional cost of linkages, extra approval requirements, and prolonged time to data acquisition for some linkages (eg, Cancer Registry).

Conclusion

CPRD Aurum and GOLD are valuable tools for studies based on information recorded in primary care but should not be used without critical evaluation. Collectively, the three studies published in this issue of Clinical Epidemiology add to the current body of literature that describes the quality and completeness of CPRD Aurum data through comparison to other data resources. These studies, along with earlier evaluations, provide valuable information for researchers planning to use these data resources. For diagnoses expected to be treated in hospital or specialty settings, linked data may complement and expand on the GP data, as well as improve case capture. However, use of linkages should be based on the requirements of each study and with knowledge of data recording practices (not based on coding potential which may not be realized) and balanced against other logistical considerations. CPRD Aurum, GOLD, and the data linkages have changed over time and are influenced by external factors such as National Health Service financial incentives, NICE guidelines and Quality Outcome Frameworks, and policy and software changes; thus, no single data source has recorded data consistently over time and none provide a true gold standard. Therefore, it is critical for the conduct of valid research, to evaluate computerized medical information in any data source on an ongoing basis.

Funding

This study was funded by Amgen, Inc.

Disclosure

Kafatos and Neasham are employees of Amgen Ltd, which uses CPRD data, and own shares of Amgen Inc. Jick, Vasilakis-Scaramozza, Persson, and Hagberg are employees of Boston Collaborative Drug Surveillance program which receives industry funding to conduct research using CPRD data. The authors report no other conflicts of interest in this work.

References

1. Hagberg KW, Vasilakis-Scaramozza C, Persson R, Kafatos G, Neasham D, Jick S. Presence of Breast Cancer Information Recorded in United Kingdom Primary Care Databases: Comparison of CPRD Aurum and CPRD GOLD (Companion Paper 1). *Clin Epidemiol*. 2023;15:1183–1192. doi:10.2147/CLEP.S434795
2. Vasilakis-Scaramozza C, Hagberg KW, Persson R, et al. Comparison of Rheumatoid Arthritis Information Recorded in UK CPRD Aurum and CPRD GOLD Databases (Companion Paper 3). *Clin Epidemiol*. 2023;15:1207–1218. doi:10.2147/CLEP.S434831.
3. Hagberg KW, Vasilakis-Scaramozza C, Persson R, Neasham D, Kafatos G, Jick S. Correctness and Completeness of Breast Cancer Diagnoses Recorded in UK CPRD Aurum and CPRD GOLD Databases: Comparison to Hospital Episode Statistics and Cancer Registry (Companion Paper 2). *Clin Epidemiol*. 2023;15:1193–1206. doi:10.2147/CLEP.S434829
4. Jick S, Hagberg KW, Persson R, et al. Quality and completeness of diagnoses recorded in the new CPRD Aurum Database: evaluation of pulmonary embolism. *Pharmacoepidemiol Drug Saf*. 2020;29(9):1134–1140. doi:10.1002/pds.4996
5. Persson R, Sponholtz T, Vasilakis-Scaramozza C, et al. Quality and completeness of myocardial infarction recording in Clinical Practice Research Datalink Aurum. *Clin Epidemiol*. 2021;13:745–75. 10.2147/CLEP.S319245.

6. Hagberg KW, Vasilakis-Scaramozza C, Persson R, et al. Quality and completeness of malignant cancer recording in United Kingdom Clinical Practice Research Datalink Aurum compared to Hospital Episode Statistics. *Annals of Cancer Epi*. 2022;6:6. doi:10.21037/ace-22-4
7. Persson R, Vasilakis-Scaramozza C, Hagberg KW, et al. CPRD Aurum database: assessment of data quality and completeness of three important comorbidities. *Pharmacoepidemiol Drug Saf*. 2020;29:1456–1464. doi:10.1002/pds.5135
8. Vasilakis-Scaramozza C, Hagberg KW, Persson R, et al. Quality of rheumatoid arthritis recording in United Kingdom Clinical Practice Research Datalink Aurum. *Pharmacoepidemiol Drug Saf*. 2023;32:73–77. doi:10.1002/pds.5551
9. Persson R, Hagberg KW, Vasilakis-Scaramozza C, et al. Presence of codes for indication for use in Clinical Practice Research Datalink Aurum: an assessment of benign prostatic hyperplasia treatments. *Clin Epi*. 2022;Volume 14:641–652. doi:10.2147/CLEP.S360843
10. Gulliford MC, Sun X, Anjuman T, Yelland E, Murray-Thomas T. Comparison of antibiotic prescribing records in two UK primary care electronic health record systems: cohort study using CPRD GOLD and CPRD Aurum databases. *BMJ Open*. 2020;10:e038767.
11. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epi*. 2019;1740. doi:10.1093/ije/dyz034
12. Medicines & Healthcare Products Regulatory Agency. Hospital Episode Statistics (HES) Outpatient Care and CPRD primary care data documentation. Available from: https://cprd.com/sites/default/files/2022-02/Documentation_HES_OP_set21.pdf. Accessed November 30, 2023.
13. Strongman H, Williams R, Bhaskaran K. What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? A concordance and validation study using linked English electronic health records data. *BMJ Open*. 2020;10:e037719. doi:10.1136/bmjopen-2020-037719