OPEN ACCESS

GAVIN PUBLISHERS

Journal of Surgery

Moussa OM, et al. J Surg: JSUR-1124.

Research Article

DOI: 10.29011/2575-9760. 001124

Diagnostic Accuracy of a Large Primary Clinical Database in the UK: A Narrative Review

Osama M. Moussa*, Chanpreet S. Arhi, Ara Darzi, Sanjay Purkayastha, Paul Ziprin

Division of Surgery, Department of Surgery & Cancer, Imperial College London, UK

***Corresponding author:** Osama M. Moussa, Division of Surgery, Department of Surgery & Cancer, Imperial College London, Academic Surgical Unit, 10th Floor QEQM, St Mary's Hospital, Praed street, London, W2 1NY, UK. Tel: +442033126666; Fax: +442033126309; Email: omoussa@ic.ac.uk

Citation: Moussa OM, Arhi CS, Darzi A, Purkayastha S, Ziprin P (2018) Diagnostic Accuracy of a Large Primary Clinical Database in the UK: A Narrative Review. J Surg: JSUR-1124. DOI: 10.29011/2575-9760. 001124

Received Date: 05 April, 2018; **Accepted Date:** 16 March, 2018; **Published Date:** 23 April, 2018

Abstract

Background: The Clinical Practice Research Datalink (CPRD) database is extensively utilised in observational studies, clinical epidemiology and outcomes research. Quality and completeness of the obtained data varies.

Aim: To conduct a narrative review of the evidence on accuracy and completeness of diagnostic coding in the (CPRD). The aim was to review studies that compared the CPRD with endorsed internal or external validation means.

Design and Setting: This manuscript was set up on validating a large clinical primary care database in the UK.

Methods: A systematic review was implemented through PubMed, EMBASE and Medline for relevant publications between 1997 and 2017. A total of 1720 non-duplicate abstracts were sourced. This was reproducible among authors.

Results: Of the 1720 abstracts, 927 were eliminated following review. A further 310 studies were identified. The factors that led to a study being excluded were: having no validation of the diagnosis being investigated (n = 652) if the data source used was not CPRD (n = 98). There were 21 publications where validation was the major focus of the research. Majority of the validations (85%) were external, with use of a questionnaire to the GP being the most frequently used (56%) and rate comparison in 33% of the 310 validations. Internal validation methods were used in 52 studies.

Conclusion: Several methods have been used to assess validity. The quality of reporting validation results was often inadequate to permit a clear interpretation. Not all methods provided a quantitative estimate of validity and most methods considered only the positive predictive value of a set of diagnostic codes. How this fits in The Clinical Practice Research Datalink has been increasingly used in epidemiologic research and have become most used source of information in pharmacoepidemiology. A key feature in the selection of a computerized database for research is completeness and validity of the data. As Khan, et al. highlight, [1] researchers should investigate their information source and how well it covers the diagnosis under study.

Introduction

The UK Clinical Practice Research Datalink (CPRD), which was known as the General Practice Research Database (GPRD) until March 2012, is a computerized database containing longitudinal medical records from primary care that has been anonymized. As of March 2011, there were over 12 million patient records that were translated to over 64 million years of prospectively collected data. With the transition from GPRD to CPRD, the volume of patient records has been estimated to increase to 52 million [2]. The information available through CPRD include patient demographic data, symptoms, signs, referrals, immunization history, behavioral factors, diagnostic tests, medical diagnosis, prescription history, as well as health outcomes [3]. The CPRD is constantly assembling anonymized data from millions of individuals, currently approaching almost 10% of the UK population, with consistent research standard data [4]. Patients that are registered with a participating primary care practice are included, unless the patient has requested not to be part of the data sharing [5]. The CPRD database is extensively utilized in observational studies such as

research on clinical epidemiology, disease patterns, drug utilization, and outcomes research, producing over 800 publications [4]. The major advantage of CPRD as a research tool is its large volume of records, attributes of patient visits as well as practice features [6], along with a past medical history (however, it suffers from missing data of patients owing to the fact of voluntary input [6] For this reason, it is useful to use the CPRD as an apparatus for epidemiological research.

A dataset obtained from the CPRD typically contains data on a patient's gender, age, year of birth and details of registration. General practices that are participating in the database share the details of every episode of illness along with any new symptom; as well as every pertinent morbidity event, such as most clinical contact, most significant diagnoses and test results, every outpatient clinic attendance and hospital they have been referred to and admitted in [7]. For the General Practitioner (GP) the most suitable diagnosis is within a drop-down list of possible options, which corresponds to the Oxford Medical Information Systems (OXMIS) and Read codes. The therapeutic data obtained from CPRD includes prescriptions with the utilization of codes from the Prescription Pricing Authority, complete with the date, dosage and method of administration of that medication. Other data in the database include vaccinations, body weight and blood pressure values, and results of laboratory analysis as well as information on lifestyle.

The authors set out to assess quality and completeness of the obtained data to help appreciate the validity of research results derived from the CPRD. For example, it was the data obtained from the CPRD that provided insight into the probable association between measles, mumps and rubella vaccination and autism [8,9]. The number and high validity of a recorded diagnosis of autism shown in such studies was a deciding factor that facilitated to enforce that results of the study were accredited. The CPRD engages in several ongoing validation that the information is compatible with a minimal standard of completeness and quality; this is made up of patient data (e.g. age, sex, details of registration and dates the events occurred), extent of completeness, continuity and plausibility of electronic data recording in key areas at the practice level (for instance, making certain that a minimum specified percentage of deaths comes with the recorded cause of death, a minimum referral rate per 100 patients, and a minimum number of prescriptions per patient per month) [4]. Furthermore, prescription information in the CPRD is well documented as the GP uses the system to produce electronic prescriptions that are automatically recorded in the database. This marks the therapy file as comprehensive [10], with the exception of prescriptions that were issued in secondary care as well as drugs that were bought over the counter [11]. On the other hand, new diagnoses may be manually recorded on the system and even though it is required that every significant diagnosis must be included, sometimes they may not be complete. Also, certain conditions may be misdiagnosed or miscoded in GP records, provisional diagnoses coded as if they are certain. To explore the veracity of this claim investigators have evaluated the validity of certain computerized diagnoses through validation studies.

Studies that have investigated the validity of diagnosis on the CPRD have postulated that there is a high validity of diagnoses that are recorded in the CPRD, as well as reporting to have found strong measures of Positive Predictive Value (PPV), sensitivity and specificity [12,13]. However, there isn't a systematic review of all validation studies of diagnoses that aims to evaluate if the evidence presented is accurate. This review will detail a systematic review of studies that explore the quality of diagnosis data available in the CPRD. It is the aim of this study to conduct a systematic review of the literature with the goal of determining how accurate and complete the data regarding diagnosis is recorded in the CPRD. Furthermore, we will evaluate the methodology used to validate diagnoses in the CPRD, summarize the findings of these studies and evaluate the quality of reporting of validation methods and results.

Methods

Search Strategy

PubMed and Embase were searched for publications using the CPRD data published between 1997 and April 2017. Bibliographies that were found on the website of the CPRD (http://www.cprd.com/bibliography/) were also examined to identify additional articles. The authors manually searched PubMed to manually pick journal articles. Furthermore, the reference lists of identified articles were scrutinized to see if they were relevant to the present study. The results of the first search were associated with a comprehensive list of free text terms and expanded the thesaurus terms to identify CPRD publications where a diagnostic validation was reported. The findings from the initial search revealed that terms showing case validation were not mentioned in the title, abstract or keywords in the relevant published papers.

Study Selection

The full manuscript of the relevant publications that were identified systematically through the search strategy were reviewed, and we identified studies that used CPRD data and were published in English. A study was considered for inclusion if it utilized a set of medical codes for a syndrome, diagnosis, which the researchers defined as a condition, was verified using one of the methods summarized in (Table 1).

	Method	Description
Internal	Diagnostic algorithm	Validation of diagnosis using codes showing specific symptoms/signs, prescriptions for disease-specific drugs and/or confirmatory test results
	Manual review of anonymized free text on computerised records	Complete computer records (inclusive of anonymized free texts) for diagnosed cases with a diagnosis were assessed for confirmation of disease status
	Sensitivity analysis	Involves disease incidence measurement, patterns or prevalence from CPRD data comparing non-CPRD, UK-based data source
External	Questionnaire to GP	Questionnaire on based on computerised diagnosis was given to GPs in clinics to fill.
	Record request to GP	GPs asked to provide anonymized medical records, hospital discharge summaries or alternate death certificates which necessitated. further diagnostic criteria
	Comparison of rates	Disease incidence, prevalence or patterns measurement was obtained from GPRD data base and compared to non-CPRD UK-soured databases

Table 1: Methods employed to validate diagnoses in Clinical Practice Research Database.

The outlined methods utilized data either completely from the database (internal validations) or outside the database (external validations).

Method

Internal

i. Diagnostic algorithm

- **Description:** The presence of codes for specific signs/ symptoms, prescriptions, and/or confirmatory test results were used to validate a diagnosis
- **Example:** Eastwood et al. [14] (2016) validated diabetes by using medication, hyperglycaemia, diabetes medication, blood tests, diabetes complications, Cardiovascular Disease (CVD) risk factors.

ii. Manual review of anonymized free text on computerised records.

- **Description:** The entire computer records (including the anonymized free text) for persons with a diagnosis were evaluated to confirm evidence of disease status.
- **Example:** Wang, et al. [14,15] (2012) was able to validate ovarian cancer by reviewing the computerised records to search for clinical events to confirm the diagnosis

Sensitivity analysis

- **Description:** An analytical study was used to identify the measurement of effectiveness using a broad set of disease/ therapeutic codes and their counterpart validation method.
- **Example:** Charlton, et al. [15,16] (2017) analyzed the risk of Neurodevelopmental Disorders (NDDs) following prenatal Antiepileptic Drug (AED) exposure in children born to Women with Epilepsy (WWE).

External

Questionnaire to GP

- **Description:** Use of a questionnaire to investigate the several aspects of the computerized diagnosis was sent to GPs.
- **Example:** Rodriguez [17] (1998) used a questionnaire sent to GPs to validate prostate cancer by comparing answers with computerized diagnosis.

i. Record request to GP

- **Description:** GPs were requested to provide anonymized hard copies of medical records, hospital discharge summaries or death certificates. The results obtained were used to examine and validate the diagnosis, by utilizing more diagnostic criteria.
- **Example:** Hall, et al. [18] (2005) sought for medical records of lung cancer patients to verify the cancer diagnosis made in the computerized records.

ii. Comparison of rates

- **Description:** Measures of disease incidence, prevalence or patterns (e.g. Time trends) from CPRD data were compared

with a non-CPRD, UK-based data source Bhatnagar, et al. [18,19] (2015) compared the mortality, morbidity and treatment of cardiovascular diseases in England with those of Ireland and Scotland.

- **Inclusion criteria:** Using the methods outlined in table 1 above, studies that were included in this review must have carried out a quantitative estimate of validity, which can be described or calculated. Studies that used sensitivity analyses were included in the breakdown if reported. Validity studies that only verified the date of diagnosis or unknown diagnoses, those that were aimed at differentiating between incident from prevalent diagnoses were excluded from the analysis.

Data Extraction

Data extraction was conducted by the author using a standardized data extraction sheet. Afterwards, about 10% of the extracted studies were evaluated to ensure that the extraction process was done appropriately. Examples of extracted information included which disease was validated, the method of validation and, where necessary, the number of cases with a confirmed diagnosis. Other information that were obtained included the quality of the validation process such as the rate of GP response to requests for information, the total number of eligible cases that were validated, how the reviewers were blinded, and method used to select the cases. However, the specific OXMIS, Read or International Classification of Diseases (ICD) codes that were used to identify each condition were not extracted, as it was not the aim of this review to describe the validity of a single disease or group of diseases.

Data Analysis

As described in the method, all validation studies were divided as internal or external. The studies were also divided by the validation method used. For studies that validated more than one diagnosis, each of the diagnosis was analyzed differently). Furthermore, if a study utilized more than a single method of validating a diagnosis, each method was considered separately. The number of cases that had a confirmed diagnosis was calculated and sorted by disease group as well as validation method. The quality of each study was assessed by a validation method, and the median or mean for each data quality variable was calculated.

Results

A total of 1720 non-duplicate abstracts were sourced from the PubMed, EMBASE and website searches, of which 927 were not CPRD studies following review of the title and abstract. Furthermore, reviewing articles and thorough search of related journals and conference proceedings produced a further 310 studies. The factors that led to a study being excluded were: having no validation of the diagnosis being investigated (n = 652), if the data source used was not CPRD (n = 98), if the source included a repeat diagnosis validation (n = 85), or if a diagnosis was not investigated (n = 181), e.g. study that did not include prescriptions or procedures. Fifty-eight of the 310 publications carried out a validation if a single diagnosis utilizing a combination of methods. For example, Ruigomez [20] (2005) carried out three validations of atrial fibrillation: initially, a manual review of computerised records, followed by a questionnaire to the GP and finally comparing incidence of the disease incidence to an external source. Thirty-five papers validated more than one diagnosis, e.g. Hippisley-Cox, et al. [21] (2014) validated cardiovascular disease, is chaemic stroke, type 2 diabetes, osteoporotic fracture and hip fracture, moderate and severe kidney failure, venous thromboembolism as well as intracranial bleed and upper gastrointestinal haemorrhage. There were 21 publications where validation was the major focus of the research. Majority of the validations (85%) were external, with use of a questionnaire to the GP being the most frequently used (56%) and studies that compared the rates of conditions being 33% of the 310 validations. With regards to internal methods, 52 studies utilized this method, with several of them (30) using a manual review.

Estimates of Validity

Overall, a high number of cases were confirmed for all diseases with a median of 86%, with a range 24-100%. This means that 86 of 100 cases that had a computerized diagnosis were confirmed with further internal or external information. However, in every disease co-morbidity the frequency of cases confirmed varied, even though the median proportion was greater than 83% for majority of the categories. The findings could not individually confirm the cases through rate comparisons and sensitivity analyses, but offered further evidence of a high validity of diagnoses in the CPRD. Albeit very few cases, the rate of disease incidence and prevalence based on CPRD data were in line with other UK population-based datalinks. For example, Watson et al (2003) [22] reported that based on data from the CPRD, the incidence rate of rheumatoid arthritis (RA) was 50% higher than previous studies, and this was because GPs in the CPRD were certain of an RA compared with rheumatologists. On the other hand, Jordan, et al. [23] reported that the prevalence of musculoskeletal diseases in the CPRD was lower and probably underestimated in comparison to other general practice databases. Majority of the sensitivity analyses did not show variation in the measures of effect calculated with a wide range of codes and those with limited set of codes, showing that many of the cases that were part of the original definition were verified using firmer standards.

Discussion

Summary

With the extensive strategy that was utilized for this study, this study intended to capture as much validation of the CPRD

diagnostic data that was published within period of interest. The most valid technique of validation is likely to ask for further information from the GP, because this method utilizes data that external validation means to clarify the status of the disease of individual cases. Many of these validations were limited to evaluating the frequency of cases with diagnostic codes that were acknowledged reviewing the medical record or reviewing the responses GPs provided to the questionnaires, thus providing an estimate of the Positive Predictive Value (PPV) of that set of codes. Even though the PPV is a measure, it differs depending on disease prevalence, thus if the disease incidence has not altered over time, utilizing historical validations may not be wholly correct.

Strengths and Limitations

There may be a difficulty with the generalization of the findings of validation studies, since there are certain CPRD practices that do not give consent to research studies. So, even though a high number of practices comply with researchers, the observed PPV will only be obtained from cases within a subgroup of practices only. By doing so, practices that do not take part in validation studies may end up providing data for solitary cases. For example, Thomas, et al. [24] found that certain practices refused to provide copies of very large case files, plausibly leading selection bias among researchers.

Comparison with Existing Literature

A comparison of rates of validation provides a quick indication of the validity of the CPRD, individual case review. Such comparisons do not validate separate cases or offer a statistically significant estimate of validity. In studies comparing prevalence rates, the CPRD may show decreased lower prevalence since it is not necessary for GPs to code prevalent diseases after every consultation [24]. Even though the findings are essential for descriptive purposes, comparing the rates of disease conditions lacks the ability to identify data or cases that have been misclassified between varying diagnoses [4]. Thus relying on this technique to ascertain the validity of a diagnosis in the CPRD should be done carefully and it will not be useful in analytic studies that require individual validity.

Implications for Research

In the same manner, while sensitivity analysis indicates the quality of diagnosis, it is not a significant validation of the data. Nested case-control studies make up majority of the research done with CPRD data. Thus, future researches using case-control studies need to engage similar inclusion and exclusion criteria. On the other hand, validation studies that are based only on cases may deliver more insightful criteria for cases than for controls.

Conclusion

The CPRD is a very useful and effective tool for researching morbidity as recorded in primary care, even though the quality of studies using the information is dependent on the validity of data input. It is therefore imperative for researches to carry out certain forms of validation before using the data. Currently, robust validations seeking further clarification from GPs are limited in size owing to the cost involved, thus compromising the generalizability of the findings owing to decline of many practices to participate in researches. The database is also being updated to expand the CPRD as a genuine tool for controlled randomized trials and as a sampling frame in other to get genetic data. By linking the CPRD with other healthcare databases, morbidity registers and death certificates will enable researchers to synchronize diagnoses made in the hospital with no alternative to seeking further medical records. On the other hand, the utilization of such associations will bring up questions regarding how to solve the problem of unrelated or missing diagnoses in the two databases. It is hoped that this study will provide further discussion about how best to evaluate the quality of the database to further improve the validity and the effectiveness of the CPRD in future research studies.

References

1. Khan NF, Harrison SE, Rose PW (2010) Validity of diagnostic coding within the General Practice Research Database: A systematic review. Br J Gen 60: 199-206.
2. Kousoulis AA, Rafi I, De Lusignan S (2015) The CPRD and the RCGP: Building on research success by enhancing benefits for patients and practices. Br J Gen Pract 65: 54-55.
3. Williams T, van Staa T, Puri S, Eaton S (2012) Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. Ther Adv drug Saf 3: 89-99.
4. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ (2010) Validation and validity of diagnoses in the General Practice Research Database: a systematic review 2010.
5. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, et al. (2015) Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol 44: 827-836.
6. Lawrenson R, Williams T, Farmer R (1999) Clinical information for research; the use of general practice databases. J Public Health Med 21: 299-304.
7. Y Lis and Mann RD (1995) The VAMP research multi-purpose database in the U.K. J Clin Epidemiol 48: 431-443.
8. Williams T and Puri S (2010) The General Practice Research Database Background to GPRD. Presentation 2010.
9. Black C, Kaye JA, Jick H (2002) Relation of childhood gastrointestinal disorders to autism: nested case-control study using data from the UK General Practice Research Database. BMJ 325: 419-421.

10. Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM, et al. (2009) Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. PLoS One 4: e7168.

11. Fombonne E, Heavey L, Smeeth L, Rodrigues LC, Cook C, et al. (2004) Validation of the diagnosis of autism in general practitioner records. BMC Public Health 4: 5.

12. Hollowell J (1997) The General Practice Research Database: quality of morbidity data. Popul Trends 87: 36-40.

13. Nazareth I, King M, Haines A, Rangel L, Myers S (1993) Accuracy of diagnosis of psychosis on general practice computer system. BMJ 307: 32-34.

14. Eastwood S V, Mathur R, Atkinson M, Brophy S, Sudlow C, et al. (2016) Algorithms for the capture and adjudication of prevalent and incident diabetes in UK Biobank. PLoS One 11.

15. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, et al. (2012) Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS One 7.

16. Charlton RA, Mc Grogan A, Snowball J, Yates LM, Wood A, et al. (2017) Sensitivity of the UK Clinical Practice Research Datalink to Detect Neurodevelopmental Effects of Medicine Exposure in Utero: Comparative Analysis of an Antiepileptic Drug-Exposed Cohort. Drug Saf. Springer International Publishing 40: 387-397.

17. García Rodríguez LA and Pérez Gutthann S (1998) Use of the UK General Practice Research Database for pharmacoepidemiology. Br J Clin Pharmacol 45: 419-425.

18. Hall GC, Roberts CM, Boulis M, Mo J, MacRae KD (2005) Diabetes and the Risk of Lung Cancer. Diabetes Care 28: 590-594.

19. Bhatnagar P, Wickramasinghe K, Williams J, Rayner M, Townsend N (2015) The epidemiology of cardiovascular disease in the UK 2014. Heart 101: 1182-1189.

20. Ruigómez A, Johansson S, Wallander M-A, García Rodríguez LA (2005) Predictors and prognosis of paroxysmal atrial fibrillation in general practice in the UK. BMC Cardiovasc Disord 20.

21. Hippisley-Cox J, Coupland C, Brindle P (2014) The performance of seven Q Prediction risk scores in an independent external sample of patients from general practice: a validation study. BMJ Open 4: e005809.

22. Watson DJ, Rhodes T, Guess HA (2003) All-cause mortality and vascular events among patients with rheumatoid arthritis, osteoarthritis, or no arthritis in the UK General Practice Research Database. J Rheumatol 30: 119-202.

23. Jordan K, Clarke AM, Symmons DP, Fleming D, Porcheret M, et al. (2007) Measuring disease prevalence: A comparison of musculoskeletal disease using four general practice consultation databases. Br J Gen Pract 57: 7-14.

24. Thomas SL, Edwards CJ, Smeeth L, Cooper C, Hall AJ (2008) How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the general practice research database? Arthritis Care Res 59: 1314-1321.