

Validity of diagnostic coding within the General Practice Research Database: a systematic review

Nada F Khan, Sian E Harrison and Peter W Rose

ABSTRACT

Background

The UK-based General Practice Research Database (GPRD) is a valuable source of longitudinal primary care records and is increasingly used for epidemiological research.

Aim

To conduct a systematic review of the literature on accuracy and completeness of diagnostic coding in the GPRD.

Design of study

Systematic review.

Method

Six electronic databases were searched using search terms relating to the GPRD, in association with terms synonymous with validity, accuracy, concordance, and recording. A positive predictive value was calculated for each diagnosis that considered a comparison with a gold standard. Studies were also considered that compared the GPRD with other databases and national statistics.

Results

A total of 49 papers are included in this review. Forty papers conducted validation of a clinical diagnosis in the GPRD. When assessed against a gold standard (validation using GP questionnaire, primary care medical records, or hospital correspondence), most of the diagnoses were accurately recorded in the patient electronic record. Acute conditions were not as well recorded, with positive predictive values lower than 50%. Twelve papers compared prevalence or consultation rates in the GPRD against other primary care databases or national statistics. Generally, there was good agreement between disease prevalence and consultation rates between the GPRD and other datasets; however, rates of diabetes and musculoskeletal conditions were underestimated in the GPRD.

Conclusion

Most of the diagnoses coded in the GPRD are well recorded. Researchers using the GPRD may want to consider how well the disease of interest is recorded before planning research, and consider how to optimise the identification of clinical events.

Keywords

database management systems; meta-analysis; sensitivity and specificity; and systematic review.

INTRODUCTION

Information collected routinely from primary care can provide a cost-effective source of data for epidemiological research.¹ One of the main benefits of using automated databases for research lies in the ability to access data from large patient populations across a wide population coverage. While these datasets represent a valuable tool for conducting research, it is important to remember that the data are collected primarily for clinical and routine use rather than specifically for research purposes. Data quality and reliability must be considered by researchers using these resources.

The General Practice Research Database (GPRD) is a widely used UK-based database of clinical primary care records, and has been extensively used in both primary care and pharmacoepidemiological research. It is the world's largest source of anonymised longitudinal data from primary care, and currently contains information on 3.6 million active patients from 450 general practices in the UK.^{2,3} Practices participating in the GPRD are remunerated for recording data on clinical diagnoses, test results, prescriptions, and referral data. Clinical data are captured using the Read/OXMIS (Oxford Medical Information System) coding framework, which is based

NF Khan, BSc, MSc, Macmillan Cancer support research Fellow; SE Harrison, BSc, ClinPsyD, research officer; PW Rose, MD, FRCGP, university lecturer, Department of Primary Health Care, University of Oxford, Oxford.

Address for correspondence

Nada F Khan, Rosemary Rue Building, Old Road Campus, University of Oxford, Oxford, OX4 1PR, UK.
E-mail: nada.khan@dphpc.ox.ac.uk

Submitted: 21 May 2009; **Editor's response:** 11 June 2009; **final acceptance:** 24 July 2009.

©British Journal of General Practice

This is the full-length article of an abridged version published in print. Cite this article as: *Br J Gen Pract* 2010; DOI: 10.3399/bjgp10X483562.

on the International Classification of Diseases, (Ninth Revision, Clinical Modification: ICD-9-CM) and is widely used in British primary care. Each practice is issued with a set of GPRD recording guidelines, describing how to record all significant morbidity events in each patient's medical history.⁴ The raw data provided from each practice undergo extensive quality control and validity checks by a research team based at the Medicines and Healthcare products Regulatory Agency before release. These data are assessed by an 'up to standard' audit, confirming data recording in several key areas. Practices meeting this standard are included in the GPRD data warehouse. Patient-level data are also assessed, with patients considered 'acceptable' for inclusion in the GPRD if recorded details are internally consistent in four areas: age, sex, registration details, and event recording.⁵

While the checks conducted by the GPRD team may provide an overall evaluation of which practices are providing good-quality data meeting certain standards, they do not specifically assess the validity and completeness of individual patient records. Quality of data in disease registers can be assessed by considering two issues. First, are the data accurate; do the codes on the register represent the diagnosis under question? Second, are the data complete; what proportion of all true cases are recorded on the database?⁶

As the GPRD is increasingly used for academic research, it is important to consider the quality of the data available for study. The aims of this paper are to conduct a systematic review of the literature to present a description of the accuracy and completeness of recording in the GPRD. Specifically, the paper will consider both the recording of clinical diagnoses and comparisons between the GPRD and other databases or national statistics.

METHOD

Literature search

Six electronic databases were searched (MEDLINE, Embase, British Nursing Index, PsycINFO, Health Management Information Consortium, Social Sciences Citation Index) from inception to May 2009, using search terms relating to the GPRD in association with terms synonymous with validity, accuracy, concordance, and recording. See Appendix 1 for an example of the MEDLINE search. Two reviewers screened the titles and abstracts of all papers in the initial search, and excluded citations that did not meet the inclusion criteria. Disagreements at this stage were resolved through discussion; however, the full-text paper was ordered when either reviewer was uncertain about inclusion. One reviewer independently assessed the retrieved full-text papers for relevance and inclusion, and hand searched the reference lists of all

How this fits in

The UK General Practice Research Database (GPRD) is increasingly used for academic primary care researchers. Data collected for automated databases such as the GPRD are subject to a number of quality checks, but the validity and accuracy of specific disease coding may vary. This paper describes a systematic review of studies that conducted validation of diagnoses in the GPRD or compared rates of disease in the GPRD against other databases. The majority of diagnoses were reliably coded; however, investigators conducting research using the GPRD should consider how information is captured in primary care and how the variation in recording practices for different diseases and prevalent conditions may affect research.

retrieved papers. The same reviewer also conducted a hand search of the GPRD bibliography provided on the GPRD website,⁷ and included any potentially relevant papers in the review process.

Inclusion criteria

Studies that specifically considered the accuracy of recording of data in the GPRD were included, and those that looked at the accuracy of recording of data as part of a larger study. Papers written in English and non-English languages were considered for inclusion. Papers that did not look specifically at the GPRD were excluded, as well as those that were not original research and papers that did not conduct validation or comparison of information held on the GPRD. Database comparison studies were also excluded if the study compared prevalence rates over different time periods. Two investigators independently reviewed articles meeting the inclusion criteria and abstracted relevant data onto a standardised data extraction form. A third reviewer resolved any uncertainties in relation to the main outcomes.

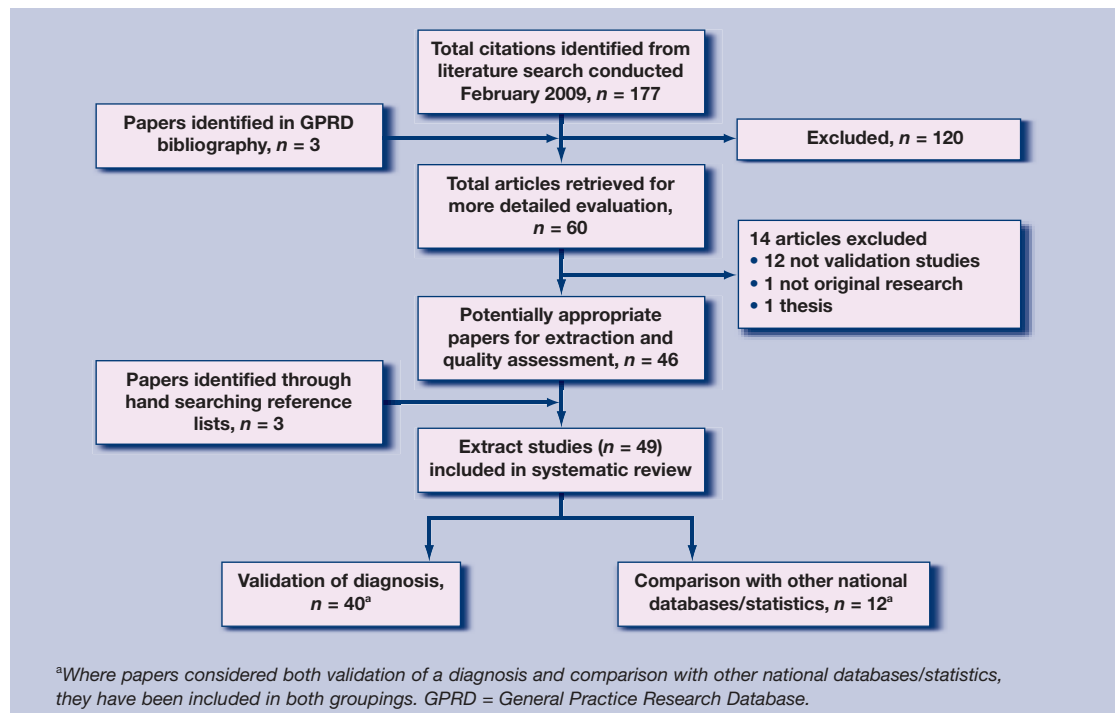
Assessment of accuracy and completeness of data

This review considers both the accuracy and completeness of data recording in the GPRD. The accuracy of diagnoses was achieved by comparing the GPRD-coded data in the electronic patient record against the gold standard defined in each paper. Assessing the completeness of data was achieved by comparing disease prevalence and prescription rates between the GPRD and other national datasets and determining the level of under-reporting or over-reporting in the GPRD.

Statistical analysis

Positive predictive value (PPV) was defined as the proportion of GPRD-coded diagnoses validated as true cases against a gold standard (GP questionnaire, primary care medical records, or hospital letters). Where the investigators used a GP questionnaire or request for hospital notes, the percentage validated

Figure 1. Flowchart of literature search results.



was calculated using the returned questionnaires as the denominator. Stata (version 10.0) was used to calculate binomial exact 95% confidence intervals (95% CIs) for each PPV, and to produce a forest plot of PPVs for correct coding of clinical diagnoses. A pooled PPV was not calculated, due to the variability in the included studies.

RESULTS

Literature search

A total of 46 papers were identified through the literature search and included in this review. Three additional papers were identified through hand searching the reference lists of the final 46 papers.⁸⁻¹⁰ Therefore, a total of 49 papers entered the review (Figure 1). Results are presented in the following main groupings: (1) validation of clinical diagnoses and (2) comparisons between GPRD and other national statistics or databases.

Papers validating a diagnosis or patient characteristic

A total of 40 papers conducted validation of a diagnosis or patient characteristic coded in the GPRD database. Most of these validation exercises involved sending a questionnaire to the patients' GPs ($n = 19$) and conducting independent verification of diagnoses against hospital letters or medical records in practice ($n = 16$). Some studies involved both sending a GP questionnaire and conducting verification against medical records ($n = 5$). The PPV of the accuracy of GPRD clinical codes from these validation exercises is summarised in Figure 2. The majority of papers report

PPVs over 50%, and simply required patients' GPs to confirm the diagnosis on the GPRD database. Five of the seven papers reporting a PPV under 50% considered acute outcomes including drug-induced liver injury, pancreatitis, or renal failure.¹¹⁻¹⁵ The studies considering acute conditions all used strict diagnostic criteria to validate cases, which included confirmation of the diagnosis via biochemical tests or specialist confirmation.

Studies in this review reported high PPVs over 90% for the recording of anorexia, bulimia,¹⁶ cataract,¹⁷ congenital heart defects (including ventricular septal defects, tetralogy of Fallot, and coarctation of the aorta),¹⁸ inflammatory bowel disease,¹⁹ cerebrovascular disease, diabetes, respiratory tract infection,¹⁹ Paget's disease,²⁰ hip fracture,²¹ upper gastrointestinal bleeding,²² non-affective and non-organic psychosis,²³ venous leg ulcer,²⁴ and pressure ulcer.²⁵

Recording of psoriasis,²⁶ venous thromboembolism,²⁷ schizophrenia,²³ dementia, and Alzheimer's disease²⁸ was relatively accurate, with PPVs between 80% and 90%. Other diagnoses, including cardiovascular events and thromboembolic disease,^{10,29} irritable bowel syndrome,³⁰ chronic obstructive pulmonary disease,³¹ chronic atrial fibrillation,³² and cardiac arrhythmia,³³ were not as accurately recorded, with reported PPVs lower than 80%.

Several papers validated the same diagnosis. Two papers considered the recording of autism in the GPRD.^{8,34} Both studies validated the diagnosis directly against hospital and patient medical records, and report that autism is well recorded in the GPRD.

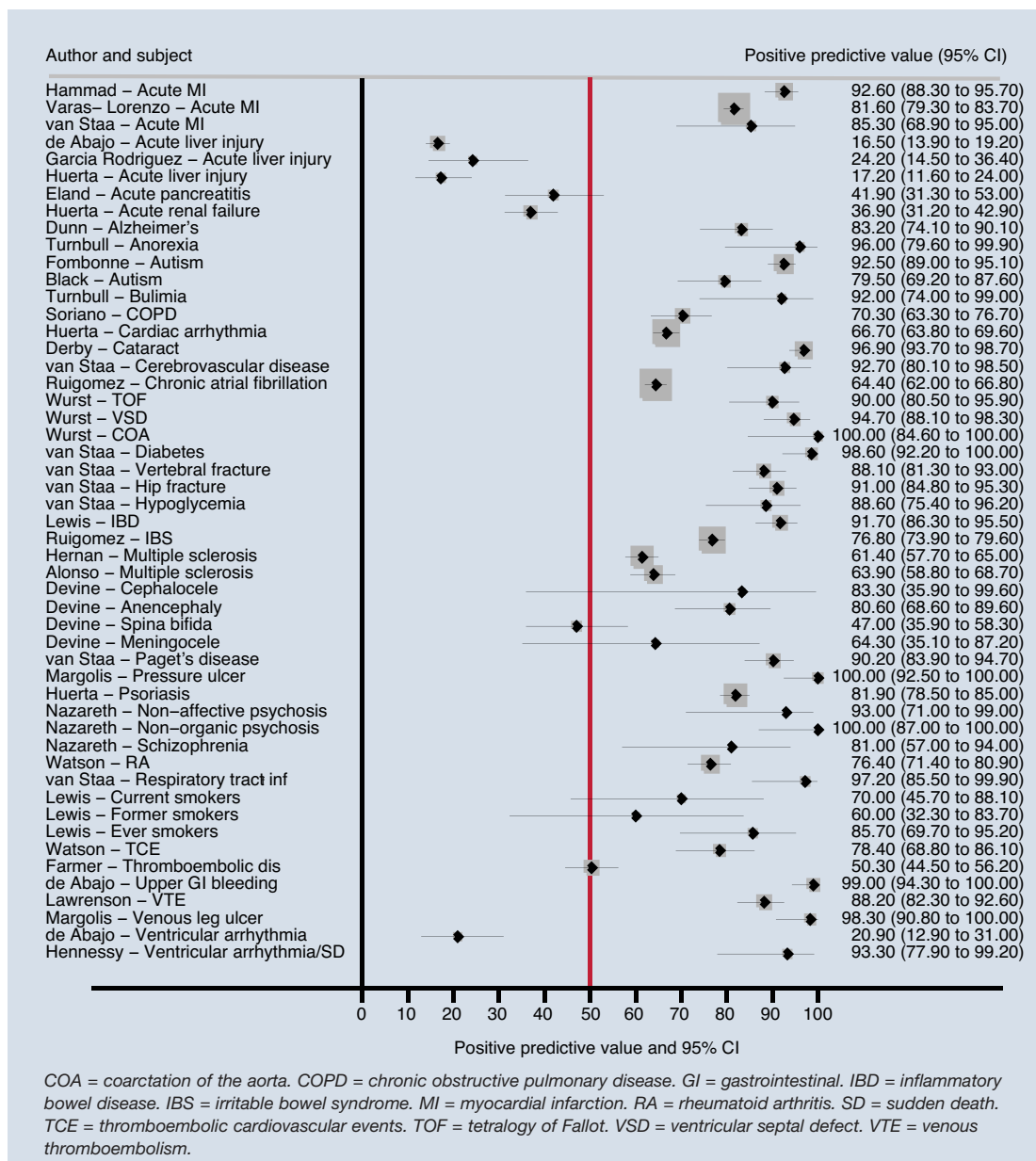


Figure 2. Forest plot reporting positive predictive values of diagnoses in the General Practice Research Database.

However, the electronic record was not detailed enough to provide sufficient differentiation between subtypes of pervasive developmental disorders. Acute myocardial infarction was also well recorded in the GPRD, with three studies reporting PPVs above 80%.^{9,35,36} There was good agreement in two papers that validated the recording of incident multiple sclerosis against hospital and medical records; however, both papers reported relatively low PPVs of around 60%.^{37,38}

Two papers considering the validity of coding for ventricular arrhythmia reported markedly different PPVs.^{39,40} Using a GP questionnaire as the standard for validation resulted in a PPV of 93% (95% CI = 78 to 99%) for cases of sudden death or ventricular arrhythmia, whereas more stringent criteria requiring objective evidence of ventricular arrhythmia from

specialist clinics and absence of recent angina or myocardial infarction resulted in a PPV of only 20.9% (95% CI = 13 to 31%). However, the investigators using a GP questionnaire to validate ventricular arrhythmia also report that only 23% (95% CI = 10 to 42%) of these diagnoses originated from outpatient events, which was their main outcome of interest. Two papers consider the validity of coding for rheumatoid arthritis; however, one study reported four validation categories (valid, invalid, possible, and unclassifiable) from which a PPV could not be derived.^{10,41}

Two studies assessed the completeness of GP recording of diagnoses made by hospital consultants. In these studies, diagnoses were transcribed from the hospital discharge letter onto patients’ electronic medical files in a high proportion (~90%) of cases.^{42,43}

Devine *et al* conducted a validation study using an algorithm to identify children with neural tube defects. They reported an overall PPV of 71% (95% CI = 63 to 78%); however, the PPV varied considerably according to the specific neural tube defect diagnosis.⁴⁴ While anencephaly and cephalocele were generally well recorded, spina bifida was not. The Read Code algorithm used by the authors located spina bifida in the mother and not the child in 37% of cases.

One paper considered the recording of smoking status in the GPRD.⁴⁵ Although current smoking is generally well recorded, former smoking is not as well recorded. Appendectomy is also under-recorded in the GPRD; in a study of ulcerative colitis, the self-reported rate of appendectomy was 13% in the random sample of patients; however, the GPRD-coded rate of appendectomy in the same study was only 3.5%.⁴⁶

Three papers reported results relating to accuracy of the date of diagnoses. There were discrepancies in date recording in 45/95 (47%) of dementia cases.²⁸ The differences were generally small, with an interquartile range (IQR) of -7 to 0 weeks. In a study of the validity of inflammatory bowel disease recording, the median difference in the first date reported by the GP and the first inflammatory bowel disease diagnosis in the electronic record was -8 days (IQR = -81 to 0 days). However, for 33 of the 53 patients included in the study, the first recorded diagnosis of inflammatory bowel disease in the electronic record was within 30 days of the date reported by the GP.¹⁹ Recording of the date of acute myocardial infarction was also generally reliable; only 31/201 (15%) of confirmed cases had a GP-reported date that was inconsistent with the electronic record. The differences in dates were generally small; 28/31 (90%) of the GP-reported dates were within 15 days of the date in the electronic record.³⁵

The GPRD compared with other databases or statistics

In total, 12 papers compared GPRD database prevalence or consultation rates with other primary care databases or national statistics registers. All compared the GPRD with UK data, except for one comparison with a US-based database.⁴⁷

Three papers compared GPRD consultation or prevalence rates against the Morbidity Statistics from General Practice 1991-92 (MSGP4), a UK-wide survey of consultation patterns in primary care.⁴⁸⁻⁵⁰ The MSGP4 itself has been evaluated;⁴⁸ 96% of all consultations in the GP surgery were recorded, suggesting it contains good-quality data on consultation patterns in the UK. There was good agreement between the GPRD and MSGP4 for 11 common respiratory conditions. However, consultation rates and prevalence of diabetes and musculoskeletal

conditions were underestimated in the GPRD compared with the MSGP4.^{49,50}

Three studies compared the GPRD with the Doctors' Independent Network (DIN), a UK-based primary care database that has been collecting routine data from over 300 practices distinct from the GPRD since 1989.⁵¹⁻⁵³ Generally, there was good agreement between the two databases for common childhood conditions, hay fever, ischaemic heart disease, and prescribing for skin emollients.

The six remaining papers compare the GPRD with a variety of other primary and secondary care databases. Three papers report similar rates of disease among the GPRD and other databases. The UK-based MediPlus primary care database, which covers about 150 practices across the UK, provided similar crude incidence rates to the GPRD for venous thromboembolic disease.⁵⁴ Derby *et al* compared rates of suicide in the GPRD to a US-based database held by the Group Health Cooperative, and report that the overall rate of suicide among users of antidepressants was similar to the rate in the Group Health Cooperative.⁴⁷ A comparison of the GPRD with the Hospital Episodes Statistics demonstrates comparable overall and age-specific incidence of Guillain-Barré syndrome.⁵⁵

There were some differences between disease coding in the GPRD and other datasets. A comparison of the GPRD and the Living in Britain National Household Survey from 1996 suggests that current smoking rates in the GPRD are 79% of the expected rate. The rates for ex-smokers were substantially underestimated; the GPRD rate for ex-smoking was 29% of what was expected according to the National Household Survey.⁴⁵ Frischer *et al* describe under-reporting of drug misuse recording in the West Midlands Regional Drug Misuse Database compared to the GPRD.⁵⁶ The prevalence of congenital heart defects was higher in the GPRD than in the National Congenital Anomaly System.⁵⁷ The same authors also validated heart defect diagnoses using a GP questionnaire, and reported an overall PPV of 0.935, suggesting that the GPRD is a good source of information on congenital heart defects.¹⁸

DISCUSSION

Summary of main findings

A systematic review of literature was carried out to validate the accuracy and completeness of the UK GPRD. The studies included in this review considered the accuracy of diagnostic codes in the GPRD and the completeness of data compared with other databases and national statistics.

Most of the diagnoses coded in the GPRD electronic record were well recorded when compared against GP questionnaire responses, medical records held at the

GP practice, or hospital letters. However, it seems that acute diagnoses were not as well recorded. The studies in this review used a variety of 'gold standard' references to ascertain the accuracy of diagnoses, which may explain some of the differences in accuracy of diagnosis recording, especially in the validation of acute conditions.

When a questionnaire is sent to a GP to support the accuracy of a diagnosis, the GP has several options for verifying the diagnosis, including checking through the computerised medical record or free-text information, looking for supporting evidence for a diagnosis from test results or hospital discharge letters, or relying on memory alone. Investigators conducting independent validation of a diagnosis can also request copies of patient medical records, hospital discharge letters, or correspondence or biochemical test results. This extra information can be used to find key words relating to a diagnosis, or conduct expert validation of a diagnosis.

Several studies assessed the accuracy of recording against an objective standard as defined by an external body; for instance, acute liver injury defined as an increase of more than two times the upper limit of normal in alanine aminotransferase by international consensus statement, or evidence of specific behavioural or cognitive symptoms as described in the Diagnostic and Statistical Manual of Mental Disorders.^{15,23,34} The number of patient records validated may depend on the level of evidence required to assess the accuracy of recording.

Smoking status is an important risk factor and confounder in many epidemiological studies, and although current smoking may be recorded well enough for the purposes of epidemiological research, data on former smoking may need to be independently validated.⁴⁵

Only three papers looked specifically at the differences between the date of onset of disease in the GPRD electronic medical record and the GP-reported date.^{19,28,35} Although there were some inconsistencies in date recording, the differences were small. Investigators who require precise dates of onset of disease may need to be aware that there could be a slight difference in the date recorded in the electronic record and, if necessary, conduct further validation.

Generally, there is good agreement in disease prevalence rates between the GPRD and other national databases and statistics; however, there were some differences identified in this review. There is no 'gold standard' measure against which data from one database can be compared, or to suggest which database contains the most accurate measure. Discrepancies between two data sources do not necessarily mean that one database is right and one is wrong. There may be geographical differences or

disease coding system variability that will lead to systematic differences in disease prevalence or data recording.⁴⁶ It is important to consider these differences when conducting research using these datasets.

There are two reasons why the GPRD may be systematically different from other datasets. First, not all consultations for chronic diseases need to be recorded in the GPRD; the GPRD recording guidelines state that the GP should make at least one entry in the medical history for each episode of illness or new occurrence of a symptom.⁴ The requirement to record only the first instance of disease may partially explain why consultation rates and prevalence of diabetes and chronic musculoskeletal conditions were underestimated in the GPRD compared with the MSGP4.^{49,50} Second, it is important to consider that practices supplying early years of data to the GPRD provided OXMIS-coded data. Other databases may use different coding systems; for instance DIN practices have always used Read Codes for recording diagnoses and prescriptions under a problem-orientated medical record, which presents each medical record as a set of intertwined but separate problems.^{51,52} Investigators attributed many of the differences between DIN and the GPRD to the Read and OXMIS coding systems used in the respective databases.⁵³

Strengths and limitations of the study

This is the first study to search for studies systematically and to combine studies that consider the accuracy and completeness of the GPRD. By using broad search terms it was possible to find a wide range of literature covering a range of diagnoses. This review provides vital information to aid researchers and clinicians who are planning to conduct research using the GPRD. However, very few of the papers in this review gave results that were directly comparable. A wide range of diagnoses were considered and many of the investigators used different criteria to assess the validity of diagnoses, making it difficult to compare directly PPVs across studies even when diagnoses were the same. There was often a lack of an objective standard for comparison of data recording, and the papers in this review often used a variety of methods to judge the accuracy of clinical diagnoses in the patient electronic records. Finally, many of the studies included in this review only validated a small number of patient records, due to the expense of conducting validation of diagnoses via GP questionnaire or independent evaluation of hospital letters or medical records.

Comparison with existing literature

Several UK-based studies consider the quality of morbidity coding in general practice, and a

systematic review of these studies shows that morbidity coding in general practice is variable. However, the investigators suggest that conditions with clear diagnostic features are better recorded than conditions with more subjective criteria.⁵⁸ In their paper, Jordan *et al* include eight GPRD studies which are also assessed in the current review. The sensitive search strategy used in this study, which specifically considered the GPRD, made it possible to find and consolidate information from a larger number of papers validating a diagnosis in the GPRD. Thiru *et al* investigated the quality of data in primary care; however, their review focused more on how well GPs record the outcome of a consultation on electronic patient records.⁵⁹ Their review also found that studies report consistently high PPVs, indicating that data on the patient record were valid. As noted in the present review, the authors point out that variability in the assessment of data quality made it difficult to compare results directly between studies.

Implications for future research

Investigators conducting research using the GPRD need to consider carefully how information is recorded in primary care, and how GPs may use different Read/OXMIS codes to represent the same diagnosis. Some diagnoses may be recorded differently from others. This review suggests that researchers can be confident about case validity when using the GPRD for research into most chronic conditions. However, research into acute conditions may need additional validation. It may not be feasible to conduct validation studies of diagnoses in the GPRD for every project, as this can be expensive; current prices start at £60 per patient for a questionnaire or request for additional information from a practice.

One approach to ensure better identification of cases is to construct Read/OXMIS code diagnostic algorithms comprising several codes to identify events and diagnoses in the GPRD. Often, these diagnoses can then be internally validated using evidence within the GPRD to support the diagnosis; for instance, a Read/OXMIS code for an acute myocardial infarction may be followed by a referral to cardiology, details of a discharge letter from hospital, and relevant medication.

A study validating the recording of neural tube defects found that in some cases, the diagnosis represented a condition in the mother and not in the child. Birth defect researchers using the GPRD may wish to search the mother's medical history to determine whether the code relates to a diagnosis in the mother or the child. This supplemental information can be obtained from within the GPRD to improve the reliability of diagnostic codes.

Prescription data are well recorded in the GPRD because prescriptions for patients are generated

directly from the computer, and details on drug type and dosage are digitally recorded in this automated process. Therefore, prescribing data can be used to verify clinical diagnoses, or to capture additional cases. For instance, use of inhalers was used as a proxy for asthma diagnoses.⁴⁸ However, investigators should be cautious about using drug prescribing as a proxy for disease, and ensure that the prescribed drug is specific to the diagnosis of interest.

One of the future strengths of the GPRD lies in planned linkages with other national databases, including the Hospital Episodes Statistics, and Office for National Statistics databases, and the National Cancer Intelligence Network. These linkages will allow investigators to access more detailed clinical information relating to inpatient and outpatient hospital attendances and diagnoses, death registration, cause of death, and cancer diagnoses and treatment. This additional information will be a source of accurate and complete information on many of the clinical outcomes occurring outside of primary care.

Funding body

Nada Khan has undertaken this work as part of a research training fellowship funded in the UK by Macmillan Cancer Support. This work was also supported by Cancer Research UK (CR-UK) grant number C23140/A8854.

Competing interests

The authors have stated that there are none.

Discuss this article

Contribute and read comments about this article on the Discussion Forum: <http://www.rcgp.org.uk/bjgp-discuss>

REFERENCES

- Jordan K, Porcheret M, Kadam UT, Croft P. The use of general practice consultation databases in rheumatology research. *Rheumatology (Oxford)* 2006; **45**(2): 126–128.
- Walley T, Mantgani A. The UK General Practice Research Database. *Lancet* 1997; **350**(9084): 1097–1099.
- Garcia Rodriguez LA, Perez GS. Use of the UK General Practice Research Database for pharmacoepidemiology. *Br J Clin Pharmacol* 1998; **45**(5): 419–425.
- Medicines and Healthcare products Regulatory Agency. *GPRD recording guidelines for vision users*. London: Crown Copyright, 2004.
- Medicines and Healthcare products Regulatory Agency. *Data quality assessment in GPRD*. London: Crown Copyright, 2007.
- Goldberg J, Gelfand HM, Levy PS. Registry evaluation methods: a review and case study. *Epidemiol Rev* 1980; **2**: 210–220.
- GPRD Group. *Bibliography*. London: GPRD. <http://www.gprd.com/bibliography/> (accessed 20 Jan 2010).
- Black C, Kaye JA, Jick H. Relation of childhood gastrointestinal disorders to autism: nested case-control study using data from the UK General Practice Research Database. *BMJ* 2002; **325**(7361): 419–421.
- Van Staa TP, Abenheim L. The quality of information recorded on a UK database of primary care records: a study of hospitalizations due to hypoglycemia and other conditions. *Pharmacoepidemiol Drug Saf* 1994; **3**: 15–21.
- Watson DJ, Rhodes T, Cai B, Guess HA. Lower risk of thromboembolic cardiovascular events with naproxen among patients with rheumatoid arthritis. *Arch Intern Med* 2002; **162**(10): 1105–1110.
- de Abajo FJ, Montero D, Madurga M, *et al*. Acute and clinically relevant drug-induced liver injury: a population based case-control study. *Br J Clin Pharmacol* 2004; **58**(1): 71–80.
- Eland IA, Alvarez CH, Stricker BH, *et al*. The risk of acute pancreatitis associated with acid-suppressing drugs. *Br J Clin Pharmacol* 2000; **49**(5): 473–478.

13. Garcia Rodriguez LA, Duque A, Castellsague J, *et al.* A cohort study on the risk of acute liver injury among users of ketoconazole and other antifungal drugs. *Br J Clin Pharmacol* 1999; **48**(6): 847–852.
14. Huerta C, Zhao SZ, Garcia Rodriguez LA, *et al.* Risk of acute liver injury in patients with diabetes. *Pharmacotherapy* 2002; **22**(9): 1091–1096.
15. Huerta C, Castellsague J, Varas-Lorenzo C, *et al.* Nonsteroidal anti-inflammatory drugs and risk of ARF in the general population. *Am J Kidney Dis* 2005; **45**(3): 531–539.
16. Turnbull S, Ward A, Treasure J, *et al.* The demand for eating disorder care. An epidemiological study using the general practice research database. *Br J Psychiatry* 1996; **169**(6): 705–712.
17. Derby L, Maier WC, Derby L, Maier WC. Risk of cataract among users of intranasal corticosteroids. *J Allergy Clin Immunol* 2000; **105**(5): 912–916.
18. Wurst KE, Ephross SA, Loehr J, *et al.* The utility of the general practice research database to examine selected congenital heart defects: a validation study. *Pharmacoepidemiol Drug Saf* 2007; **16**(8): 867–877.
19. Lewis JD, Brensinger C, Bilker WB, *et al.* Validity and completeness of the General Practice Research Database for studies of inflammatory bowel disease. *Pharmacoepidemiol Drug Saf* 2002; **11**(3): 211–218.
20. Van Staa TP, Selby P, Leufkens HG, *et al.* Incidence and natural history of Paget's disease of bone in England and Wales. *J Bone Miner Res* 2002; **17**(3): 465–471.
21. Van Staa TPA. The use of a large pharmacoepidemiological database to study exposure to oral corticosteroids and risk of fractures: validation of study population and results. *Pharmacoepidemiol Drug Saf* 2000; **9**(5): 2000.
22. De Abajo FJ, Rodriguez LA, Montero D, *et al.* Association between selective serotonin reuptake inhibitors and upper gastrointestinal bleeding: population based case-control study. *BMJ* 1999; **319**(7217): 1106–109.
23. Nazareth I, King M, Haines A, *et al.* Accuracy of diagnosis of psychosis on general practice computer system. *BMJ* 1993; **307**(6895): 32–34.
24. Margolis DJ, Bilker W, Santanna J, *et al.* Venous leg ulcer: incidence and prevalence in the elderly. *J Am Acad Dermatol* 2002; **46**(3): 381–386.
25. Margolis DJ, Bilker W, Knauss J, *et al.* The incidence and prevalence of pressure ulcers among elderly patients in general medical practice. *Ann Epidemiol* 2002; **12**(5): 321–325.
26. Huerta C, Rivero E, Rodriguez LA, *et al.* Incidence and risk factors for psoriasis in the general population. *Arch Dermatol* 2007; **143**(12): 1559–1565.
27. Lawrenson R, Todd JC, Leydon GM, *et al.* Validation of the diagnosis of venous thromboembolism in general practice database studies. *Br J Clin Pharmacol* 2000; **49**(6): 591–596.
28. Dunn N, Mullee M, Perry VH, *et al.* Association between dementia and infectious disease: evidence from a case-control study. *Alzheimer Dis Assoc Disord* 2005; **19**(2): 91–94.
29. Farmer RD, Lawrenson RA, Todd JC, *et al.* Oral contraceptives and venous thromboembolic disease. Analyses of the UK General Practice Research Database and the UK Mediplus database. *Hum Reprod Update* 1999; **5**(6): 688–706.
30. Ruigomez A, Garcia Rodriguez LA, Johansson S, *et al.* Is hormone replacement therapy associated with an increased risk of irritable bowel syndrome? *Maturitas* 2003; **44**(2): 133–140.
31. Soriano JB, Maier WC, Visick G, *et al.* Validation of general practitioner-diagnosed COPD in the UK General Practice Research Database. *Eur J Epidemiol* 2001; **17**(12): 1075–1080.
32. Ruigomez A, Johansson S, Wallander MA, *et al.* Incidence of chronic atrial fibrillation in general practice and its treatment pattern. *J Clin Epidemiol* 2002; **55**(4): 358–363.
33. Huerta C, Lanes SF, Garcia Rodriguez LA, *et al.* Respiratory medications and the risk of cardiac arrhythmias. *Epidemiology* 2005; **16**(3): 360–366.
34. Fombonne E, Heavey L, Smeeth L, *et al.* Validation of the diagnosis of autism in general practitioner records. *BMC Public Health* 2004; **4**: 5.
35. Hammad TA, McAdams MA, Feight A, *et al.* Determining the predictive value of Read/OXMIS codes to identify incident acute myocardial infarction in the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2008; **17**(12): 1197–1201.
36. Varas-Lorenzo C, Garcia-Rodriguez LA, Perez-Gutthann S, *et al.* Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study. *Circulation* 2000; **101**(22): 2572–2578.
37. Alonso A, Jick SS, Olek MJ, *et al.* Recent use of oral contraceptives and the risk of multiple sclerosis. *Arch Neurol* 2005; **62**(9): 1362–1365.
38. Hernan MAJ. Recombinant hepatitis B vaccine and the risk of multiple sclerosis: A prospective study. *Neurology* 2004; **63**(5): 14.
39. de Abajo FJG. Risk of ventricular arrhythmias associated with non-sedating antihistamine drugs. *Br J Clin Pharmacol* 1999; **47**(3): 307–313.
40. Hennessy S, Leonard CE, Palumbo CM, *et al.* Diagnostic codes for sudden cardiac death and ventricular arrhythmia functioned poorly to identify outpatient events in EPIC's General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2008; **17**(12): 1131–1136.
41. Thomas SL, Edwards CJ, Smeeth L, *et al.* How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the general practice research database? *Arthritis Rheum* 2008; **59**(9): 1314–1321.
42. Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *BMJ* 1991; **302**(6779): 766–768.
43. Jick H, Terris BZ, Derby L, Jick S. Further validation of information recorded on a general practitioner based computerized data resource in the United Kingdom. *Pharmacoepidemiol Drug Saf* 1992; **1**: 347–349.
44. Devine S, West SL, Andrews E, *et al.* Validation of neural tube defects in the full featured — general practice research database. *Pharmacoepidemiol Drug Saf* 2008; **17**(5): 434–444.
45. Lewis JD, Brensinger C. Agreement between GPRD smoking data: a survey of general practitioners and a population-based survey. *Pharmacoepidemiol Drug Saf* 2004; **13**(7): 437–441.
46. Derby LE, Jick H, Derby LE, Jick H. Appendectomy protects against ulcerative colitis. *Epidemiology* 1998; **9**(2): 205–207.
47. Derby LE, Jick H, Dean AD, *et al.* Antidepressant drugs and suicide. *J Clin Psychopharmacol* 1992; **12**(4): 235–240.
48. Hansell A, Hollowell J, Nichols T, *et al.* Use of the General Practice Research Database (GPRD) for respiratory epidemiology: a comparison with the 4th Morbidity Survey in General Practice (MSGP4). *Thorax* 1999; **54**(5): 413–419.
49. Hollowell J. The General Practice Research Database: quality of morbidity data. *Popul Trends* 1997; **(87)**: 36–40.
50. Jordan K, Clarke AM, Symmons DP, *et al.* Measuring disease prevalence: a comparison of musculoskeletal disease using four general practice consultation databases. *Br J Gen Pract* 2007; **57**(534): 7–14.
51. Carey IM, Cook DG, De WS, *et al.* Implications of the problem orientated medical record (POMR) for research using electronic GP databases: a comparison of the Doctors Independent Network Database (DIN) and the General Practice Research Database (GPRD). *BMC Fam Pract* 2003; **4**: 14.
52. Carey IM, Cook DG, De WS, Bremner SA, *et al.* Developing a large electronic primary care database (Doctors' Independent Network) for research. *Int J Med Inform* 2004; **73**(5): 443–453.
53. de Wilde S, Carey IM, Bremner SA, *et al.* A comparison of the recording of 30 common childhood conditions in the Doctor's Independent Network and General Practice Research Databases. *Health Stat Q* 2004; **(22)**: 21–31.
54. Farmer RD, Lawrenson RA, Todd JC, *et al.* Oral contraceptives and venous thromboembolic disease. Analyses of the UK General Practice Research Database and the UK Mediplus database. *Hum Reprod Update* 1999; **5**(6): 688–706.
55. Stowe J, Andrews N, Wise L, Miller E. Investigation of the temporal association of Guillain-Barre syndrome with influenza vaccine and influenzalike illness using the United Kingdom General Practice Research Database. *Am J Epidemiol* 2009; **169**(3): 382–388.
56. Frischer M, Norwood J, Heatlie H, *et al.* A comparison of trends in problematic drug misuse from two reporting systems. *J Public Health Med* 2000; **22**(3): 362–367.
57. Wurst KE, Ephross SA, Loehr J, *et al.* Evaluation of the General Practice Research Database congenital heart defects prevalence: comparison to United Kingdom national systems. *Birth Defects Res A Clin Mol Teratol* 2007; **79**(4): 309–316.
58. Jordan K, Porcheret M, Croft P. Quality of morbidity coding in general practice computerized medical records: a systematic review. *Fam Pract* 2004; **21**(4): 396–412.
59. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ* 2003; **326**(7398): 1070.

Appendix 1. Sample search strategy used in MEDLINE.

MEDLINE search: 1950 to week 3, May 2009

- | | | |
|-----|---|---|
| 1. | "general practice research database".ti,ab | |
| 2. | GPRD*.ti,ab | |
| 3. | "Value added medical products".ti,ab | |
| 4. | FF-GPRD.ti,ab | |
| 5. | ffGPRD*.ti,ab | |
| 6. | 1 or 2 or 3 or 4 or 5 | GPRD or VAMP |
| 7. | valid*.ti,ab | Accuracy and completeness |
| 8. | consisten*.ti,ab | |
| 9. | sensitiv*.ti,ab | |
| 10. | specificity*.ti,ab | |
| 11. | reliab*.ti,ab | |
| 12. | "positive predictive value*".ti,ab | |
| 13. | PPV*.ti,ab | |
| 14. | concordance*.ti,ab | |
| 15. | miss*.ti,ab | |
| 16. | accura*.ti,ab | |
| 17. | variation*.ti,ab | |
| 18. | variab*.ti,ab | |
| 19. | replica*.ti,ab | |
| 20. | note*.ti,ab | |
| 21. | verify*.ti,ab | |
| 22. | recording*.ti,ab | |
| 23. | 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 | Accuracy and completeness of data |
| 24. | 6 and 23 | Accuracy and completeness of data in the GPRD |

GPRD = General Practice Research Database. VAMP = Value Added Medical Products.