BJCP British Journal of Clinical Pharmacology

DOI:10.1111/j.1365-2125.2009.03537.x

Validation and validity of diagnoses in the General Practice Research Database: a systematic review

Emily Herrett, Sara L. Thomas,[1] W. Marieke Schoonen,[2] Liam Smeeth & Andrew J. Hall[1]

Non-communicable Disease Epidemiology Unit and [1]Infectious Disease Epidemiology Unit, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, and [2]Department of International Epidemiology, Amgen Ltd, Uxbridge, UK

Correspondence
Miss Emily Herrett, Non-communicable Disease Epidemiology Unit, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.
Tel: +44 (0)20 7927 2524
Fax: +44 (0)20 7580 6897
E-mail: emily.herrett@lshtm.ac.uk

Keywords
diagnosis, GPRD, positive predictive value, Primary Care Database, systematic review, validity

Received
9 April 2009
Accepted
6 August 2009

AIMS
To investigate the range of methods used to validate diagnoses in the General Practice Research Database (GPRD), to summarize findings and to assess the quality of these validations.

METHODS
A systematic literature review was performed by searching PubMed and Embase for publications using GPRD data published between 1987 and April 2008. Additional publications were identified from conference proceedings, back issues of relevant journals, bibliographies of retrieved publications and relevant websites. Publications that reported attempts to validate disease diagnoses recorded in the GPRD were included.

RESULTS
We identified 212 publications, often validating more than one diagnosis. In total, 357 validations investigating 183 different diagnoses met our inclusion criteria. Of these, 303 (85%) utilized data from outside the GPRD to validate diagnoses. The remainder utilized only data recorded in the database. The median proportion of cases with a confirmed diagnosis was 89% (range 24–100%). Details of validation methods and results were often incomplete.

CONCLUSIONS
A number of methods have been used to assess validity. Overall, estimates of validity were high. However, the quality of reporting of the validations 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 in a highly selected group of cases. We make recommendations for methodology and reporting to strengthen further the use of the GPRD in research.

Introduction

Computerized databases of medical records are increasingly used in biomedical research. The General Practice Research Database (GPRD) is a primary care database containing anonymized patient records for about 6% of the UK population. The GPRD's strengths as a research tool include its size, representativeness of patient and practice characteristics [1], and a virtually complete medical history of patients due to the recording of referral to secondary care [2]. The GPRD has been widely used for observational studies, with over 550 studies published to date in peer-reviewed journals [3].

A typical dataset from the GPRD contains information on a patient's sex, age, year of birth and registration details. Participating general practices are required to record (i) each episode of illness, or new occurrence of a symptom, and (ii) all significant morbidity events, e.g. all significant clinical contacts, all significant diagnoses and abnormal test results, all referrals to outpatient clinics and hospital admissions [3]. In order to enter computerized information, the general practitioner (GP) types a descriptive term for the symptom or diagnosis and chooses the most appropriate entry from a drop-down list of possible choices, with corresponding Oxford Medical Information Systems (OXMIS) and Read codes. Therapeutic information includes

prescriptions using codes from the Prescription Pricing Authority, with the corresponding date, dosage and method of administration. Additional information is provided on vaccinations, weight and blood pressure measurements, laboratory test results and on some aspects of lifestyle. All information is entered by practice staff and is anonymized prior to central collection.

The validity of research based on GPRD data depends on the quality and completeness of data recorded. For example, following the publicity about a possible link between measles, mumps and rubella vaccination and autism, studies were undertaken using GPRD [4–6]. The high validity of a recorded diagnosis of autism demonstrated in these studies was an important factor in helping ensure the study findings were accepted. The GPRD carries out a series of ongoing checks to ensure that the data are 'up to standard'; this comprises assessment of both patient data (age, gender, registration details and event dates) and the completeness, continuity and plausibility of electronic data recording in key areas at the practice level (for example, ensuring a minimum specified percentage of deaths have cause of death recorded, a minimum referral rate per 100 patients, and a minimum number of prescriptions per patient per month). Prescription data in the GPRD are known to be well documented: the GP uses the computer to generate prescriptions and these are automatically recorded in the database. The therapy file is therefore virtually complete, except for prescriptions issued in secondary care and for drugs that are purchased over the counter [7]. In contrast, new diagnoses must be manually recorded on the computer and although all significant diagnoses should be included, they may be incomplete. Additionally, conditions may be misdiagnosed or miscoded in GP records, e.g. codes selected mistakenly or tentative diagnoses coded using 'definite' clinical codes. To examine these possibilities, investigators have assessed the validity of certain computerized diagnoses by conducting validation studies.

Specific validation studies have suggested high validity of diagnoses recorded in the GPRD, reporting strong measures of positive predictive value (PPV), sensitivity and specificity [8–14]. However, there has not been a systematic review of all validation studies of diagnoses to assess the totality of evidence. Here we report a systematic review of studies that assessed quality of morbidity and mortality data available in the GPRD. The aims of the review were to investigate the range of methods used to validate diagnoses in the GPRD, summarize the findings of these studies and assess the quality of reporting of validation methods and results.

Methods

Search strategy

We searched PubMed and Embase for publications using the GPRD data published between 1987 and April 2008. Bibliographies on the websites of the GPRD (http://www.gprd.com/bibliography/) and the Boston Collaborative Drug Surveillance Program (http://www.bcdsp.net/publications.html) were scrutinized to identify additional articles. Selected International Society of Pharmacoepidemiology conference proceedings, issues of Health Statistics Quarterly, and back issues of *Pharmacoepidemiology and Drug Safety* that were not incorporated into PubMed were hand-searched. Reference lists of identified articles were examined. Our first search linked a comprehensive list of free text terms and exploded thesaurus terms to identify GPRD publications in which a diagnostic validation was reported. This preliminary search showed that terms indicating case validation were not mentioned in the title, abstract or keywords in many published papers. Therefore, we broadened our search strategy to identify all publications reporting the use of GPRD data.

Study selection

We examined the full text of all publications identified via the search strategy that possibly used GPRD data and were published in English. A publication was considered for initial inclusion when a set of medical codes for a syndrome, disease diagnosis or death, defined by the investigators as a 'case', was verified using one of the methods summarized in Table 1. Such methods use data either entirely contained within the database (internal validations) or from outside the database (external validations).

Inclusion criteria

Publications using methods 1, 2, 4 and 5 (see Table 1) were included only if a quantitative estimate of validity (e.g. the proportion of cases with a confirmed diagnosis) was described or could be calculated. Publications using method 3 were included when results of the sensitivity analyses were reported. We did not include validations verifying only the date of diagnosis, idiopathic diagnoses (e.g. reviewing records of venous thromboembolism cases to identify those with idiopathic disease rather than a clear cause [15]), severity of diagnosis, or studies not validating diagnoses (e.g. prescriptions, procedures, smoking) or were set up to distinguish incident from prevalent diagnoses. Eligibility was assessed by three reviewers (W.M.S., S.L.T., E.H.); all disagreements were resolved after discussion.

Data extraction

Data extraction was conducted by two reviewers (W.M.S., E.H.) using a standardized data extraction sheet, and a third reviewer (A.J.H.) assessed a random sample of 10% of the articles to verify the extraction process. Data extracted included the disease validated, the validation method(s) used and, where appropriate, the proportion of cases with a confirmed diagnosis. Details regarding the quality of the validation exercise including GP response rates to requests for information, the number and proportion of total

Table 1

Methods used in validations of diagnoses in the General Practice Research Database (GPRD)

		Method	Description	Example
Internal	1	Diagnostic algorithm	A diagnosis was validated by the presence of codes indicating specific symptoms/signs, prescriptions for disease-specific drugs and/or confirmatory test results	Andersohn [23] validated acute myocardial infarction by choosing only cases with codes for troponin tests, treatment with fibrinolytic drugs, coronary intervention or a hospital stay of >3 days
	2	Manual review of anonymized free text on computerized records	The complete computer records (including the anonymized free text) for individuals with a diagnosis were assessed for confirmatory evidence of disease status	Yang [15] validated colorectal cancer cases by reviewing the computerized records to look for clinical events to confirm the diagnosis
	3	Sensitivity analysis	In an analytical study, a comparison of the measure of effect using a broad set of disease/therapeutic codes with that of a restricted set more likely to represent true cases	Gupta [26] varied the definition of multiple sclerosis to examine how its association with inflammatory bowel disease changed
External	4	Questionnaire to GP	A questionnaire investigating various aspects of the computerized diagnosis was sent to the GP	Garcia Rodriguez [27] validated prostate cancer by comparing the computerized diagnosis with answers to a questionnaire filled by the GP regarding the diagnosis
	5	Record request to GP	Request to the GP to provide anonymized copies of paper medical records, hospital discharge summaries or death certificates. Obtained copies were examined to validate the diagnosis, using further diagnostic criteria	Hall [28] requested medical records of lung cancer patients in order to verify the cancer diagnosis made in the computerized records
	6	Comparison of rates	Measures of disease incidence, prevalence or patterns (e.g. time trends) from GPRD data were compared with a non-GPRD, UK-based data source	Hollowell [7] compared the incidence of chicken pox, allergic rhinitis, asthma and diabetes with external rates

eligible cases that underwent validation, blinding of reviewers and case selection methods were also recorded. The specific OXMIS, Read or International Classification of Diseases (ICD) codes used to identify each condition were not extracted, as describing the validity of a single disease or group of diseases was not the purpose of this review.

Data analysis

Each validation study was categorized as internal or external and then by the validation method used. If a publication validated more than one diagnosis, each diagnosis was analysed separately. Also, if a publication used more than one method to validate a diagnosis, each method was considered a separate validation. The proportion of cases with a confirmed diagnosis was calculated overall, by disease group (categorized by ICD 10 chapter where possible) and for each validation method. The quality of reporting was analysed by validation method; the median or mean value for each data quality variable was calculated.

Results

We identified a total of 1515 nonduplicate abstracts from the PubMed, EMBASE and website searches, of which 806 were definitely not GPRD studies after reviewing title and abstract (Figure 1). Reviewing papers and hand-searching relevant journals and conference proceedings identified another 222 publications. After reviewing the full text, we included 212 of these 931 publications. The main reasons for exclusion were: no validation of the diagnosis under investigation (n = 478), data source used was not GPRD (n = 62), the publication reported a previously published validation (n = 47), or no diagnosis was investigated (n = 132), e.g. study of prescriptions or procedures.

Forty of the 212 publications validated a single diagnosis using a combination of methods. For example, Ruigomez [16] performed three validations of atrial fibrillation: first a manual review of computerized records, then by a questionnaire to the GP and finally a comparison of disease incidence to an external source. Twenty-nine papers validated more than one diagnosis, e.g. Hollowell [7] validated allergic rhinitis, diabetes, chicken pox and asthma by comparing GPRD rates of each with external sources.

The 212 publications reported 357 validations, verifying 183 different diagnoses (Figure 1). There were 14 publications where validation was the main focus of the study. Table 2 shows the frequency of use of the different validation methods. Eighty-three percent of validations were external; use of a questionnaire to the GP and/or a request for copied medical records were the most commonly used (160/357, 45%), while a comparison of rates was carried out in 40% of the 357 validations. Of the 31 validations that used internal methods, a minimum of 24 used a manual review; in the remaining seven it was not possible to determine whether a manual review of computerized records or a diagnostic algorithm had been conducted.

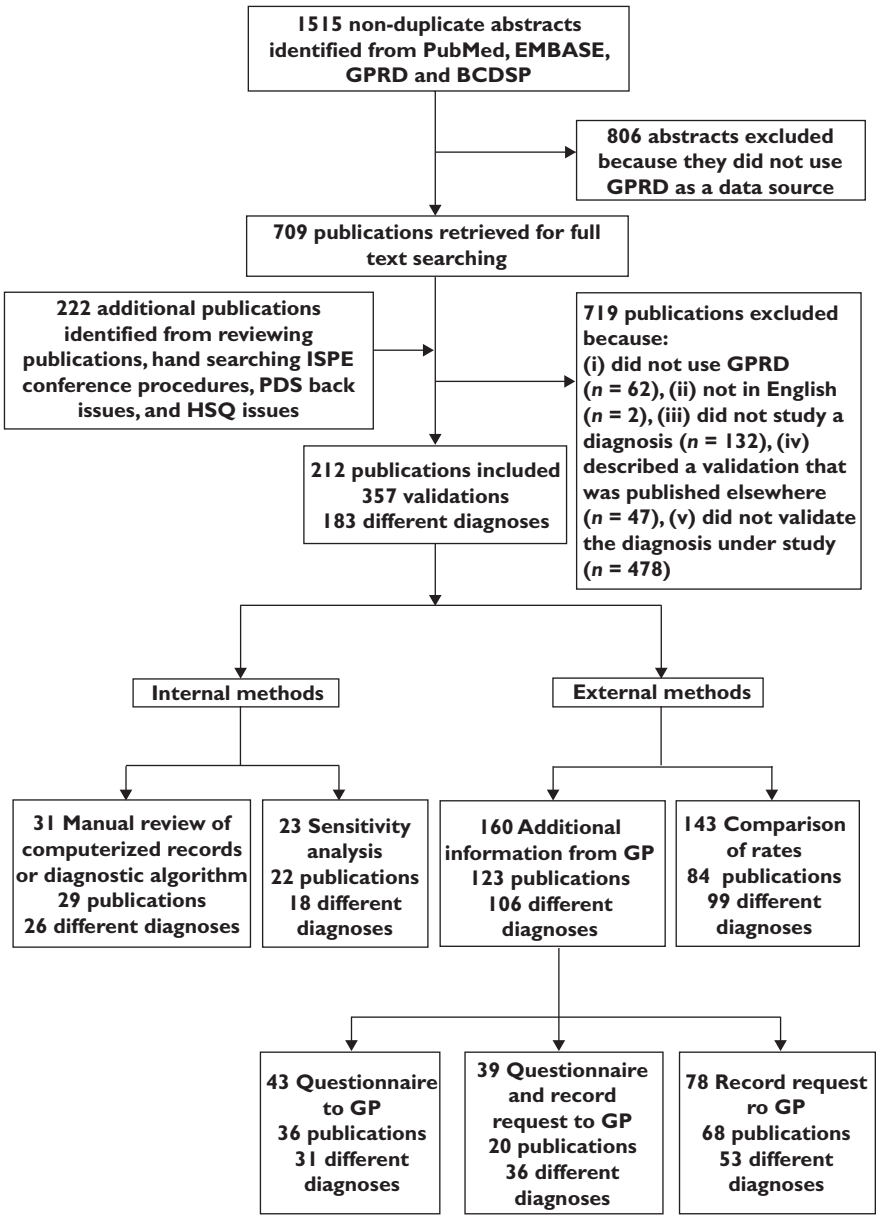

Figure 1

Stream diagram of article search, retrieval and review process. GPRD, General Practice Research Database; BCDSP, Boston Collaborative Drug Surveillance Program; ISPE, International Society of Pharmacoepidemiology; PDS, Pharmacoepidemiology Drug Safety; HSQ, Health Statistics Quarterly

Estimates of validity

Overall, a high proportion of cases were confirmed for all diseases with a median of 89% and range 24–100% (Table 2), i.e. 89 of 100 cases with a computerized diagnosis were confirmed based on additional internal or external information. Within each disease category and method the proportion of cases confirmed varied widely (Table 3) but the median proportion was >80% for most categories.

Rate comparisons and sensitivity analyses did not confirm cases individually, but provided additional evidence of a high validity of diagnoses in the GPRD. Disease incidence and prevalence estimates based on GPRD data were comparable to other UK population-based sources, with a few exceptions. For example, the incidence rate of rheumatoid arthritis (RA) based on GPRD data was 50% higher than previous estimates, which was attributed to over-ascertainment of RA by GPs in the GPRD compared with rheumatologists [17]. Conversely, the prevalence of musculoskeletal diseases in the GPRD was underestimated by the GPRD when compared with other general practice databases [18]. Most sensitivity analyses showed no differences between measures of effect calculated with broad

Table 2

Methods of validation and median percentage of cases confirmed using each method

Validation method	*n* validations carried out	Percentage of cases confirmed by validation Median (Range)
Internal		
Manual review of computerized records or diagnostic algorithm	31	86.2 (33–100)
Sensitivity analysis	23	n/a
External		
Comparison of rates	143	n/a
Additional information from GP	160	88.6 (24–100)
Questionnaire to GP only	43	91.7 (26–93)
Questionnaire and record request to GP	39	90.0 (41–100)
Record request to GP only	78	82.7 (24–100)
Total	357	88.6 (24–100)

It was a requirement for inclusion in our review for each study (with the exception of comparison of rates and sensitivity analyses) to provide a numerical estimate of the proportion of cases confirmed.

sets of codes and those with more restrictive sets of codes, suggesting that the majority of cases included in the original definition were verified by the stricter criteria.

Quality of reporting

The medical codes used to define cases were seldom reported in papers, although some researchers indicated that they would make code lists available. Occasionally, medical codes from the GPRD coding system (OXMIS and Read codes) were mapped onto ICD codes, but no details were given about which OXMIS/Read codes corresponded to specific ICD codes.

Table 4 shows the results of the data quality assessment. The number of cases undergoing validation varied according to the method used (Table 4, column B). Information requests typically validated fewer cases, but numbers were highly variable [range 1–3010 (median 100 cases)]. Comparisons of incidence or prevalence were often based on large numbers of cases and with a common disease reached as many as 200 000 [19], with a median of 1984 cases for this method.

The proportion of identified cases undergoing validation also varied (Table 4, column C). Some investigators reported validating 100% of identified cases. However, these were often a very small and highly selected group of cases chosen using strict inclusion criteria. Many validations included only cases that (i) already had some supporting evidence in their records or (ii) were restricted, e.g. by age, presence of comorbid conditions or by having had specific therapies. Publications with broader inclusion criteria validated far fewer of the total cases and in several publications the proportion validated was <0.5%. Few of these studies reported how they sampled cases for validation from all eligible cases.

Eighty-four percent (134) of 160 validations requesting additional information from GPs reported GP response rates; 55–100% (median 90%) of requests were met by GPs (Table 4, column D). In general, details were not provided on how many patient records were unavailable (e.g. due to patient transfer or death). Similarly, most validations did not report blinding of reviewers during review of patient records.

Discussion

This review identified 212 publications in which 183 different diagnoses were validated. Given the breadth of our search strategy, we feel that we are likely to have captured the majority of validations of GPRD diagnostic data published in the specified time period.

The majority of validations were external, and most frequently were requests to GPs to provide additional information. Relatively few publications documented use of internal validations. Overall, quantitative estimates of validity were high (median 89% of cases confirmed) and qualitative evidence from external rate comparisons and sensitivity analyses supported the validity of diagnoses. However, we are reluctant to draw conclusions regarding the overall validity of diagnoses in the database, for three reasons. First, despite their strengths, the methods presented here have limitations: questionnaires to GPs, record requests, algorithms and manual reviews predominantly examine PPV, whereas sensitivity analyses and comparisons of rates cannot provide quantitative estimates of validity. Even where quantitative validations are carried out, it may only be possible to categorize some coded cases as 'possible cases' based on the extra information given in the case notes. Second, the quality of reporting of many validations was insufficient to assess the possibility of bias and generalizability of validity estimates across the GPRD. Finally, it is possible that validation

Table 3

Proportion of cases confirmed, by disease group, and number of validations in each disease category, by validation method

Disease group	Internal		External Request additional information from GP					Median (range) proportion of cases confirmed†	
	Manual review or diagnostic algorithm	Sensitivity analysis	Questionnaire to GP only	Questionnaire and record request to GP	Record request to GP only	Comparison of rates	Total		
Infectious and parasitic	0	1	1	0	0	15	17	93.00	(n/a)
Neoplasms	7	0	5	2	5	7	26	95.25	(74–100)
Blood and blood-forming organs	0	0	0	1	3	1	5	57.61	(31–89)
Endocrine, nutritional and metabolic	4	0	0	5	1	7	17	87.70	(53–100)
Mental and behavioural disorders	1	0	1	5‡	8	10	20	83.00	(52–100)
Nervous system	1	4	1	1	6	12	25	81.00	(39–100)
Eye and adnexa	0	0	0	2	3	3	8	89.47	(75–97)
Ear and mastoid process	0	0	0	0	0	2	2	n/a	
Circulatory system	10	3	7‡	10	20	12	55	85.30	(48–100)
Respiratory system	2	0	3	2	2	25	34	88.00	(26–100)
Digestive system	1	6‡	13	1	19	12	46	87.35	(24–100)
Skin and subcutaneous tissue	1	1	5	1	0	5	13	94.55	(82–100)
Musculoskeletal system and connective tissue	2	6‡	2	1	4	10	19	80.00	(33–97)
Genitourinary system	0	1	0	4	1	1	7	91.00	(28–100)
Pregnancy, childbirth and the puerperium	0	0	0	0	0	0	0	n/a	
Perinatal period	0	0	0	0	0	0	0	n/a	
Congenital	0	0	3‡	0	0	2	2	93.50	(71–100)
Injury and poisoning	1	1	2	2	3	3	12	89.52	(73–100)
External causes of morbidity and mortality	0	0	0	1	0	8	9	100.00	(n/a)
Other*	1	0	0	1	3‡	8	10	90.00	(45–100)
Total	**31**	**11**	**33**	**34**	**75**	**143**	**327**	**88.6**	**(24–100)**

*Includes multiple disease groups, ill-defined conditions, miscellaneous, stillbirth and mortality. †Based on GP record request, algorithm and manual review. ‡Investigated the validity of a composite outcome combining more than one diagnosis.

Table 4

Quality of the validation studies*

Validation method	*n* validations carried out	(A) Number of cases identified with study inclusion criteria				(B) Number of cases chosen to undergo validation				(C) Proportion of all cases identified with study inclusion criteria that were chosen to undergo validation				(D) Response rate to GP questionnaire/record request			
		Median	(Range)	*n* (%) validations where reported		Median	(Range)	*n* (%) validations where reported		Mean	(Range)	*n* (%) validations where reported		Median	(Range)	*n* (%) validations where reported	
Internal																	
Manual review of computerized records or diagnostic algorithm	31	1268	(50–78 172)	27	(87)	314	(57–61 097)	30	(97)	89.6	(0.12–100)	27	(87)	n/a		n/a	
Sensitivity analysis	23	4732	(21–36 702)	19	(83)	n/a				n/a				n/a		n/a	
External																	
Comparison of rates	143	1984	(5–200 000)	70	(49)	n/a				n/a				n/a		n/a	
Additional information from GP	160	226	(1–51 688)	139	(87)	100	(1–3 010)	153	(96)	68.2	(0.15–100)	138	(86)	90.4	(55–100)	134	(83)
Questionnaire to GP only	43	1562	(10–51 688)	34	(79)	159	(10–2 040)	40	(93)	48.6	(0.39–100)	34	(79)	92.5	(55–100)	31	(72)
Questionnaire and record request to GP	39	51.5	(10–24 131)	36	(92)	40	(10–795)	38	(97)	78.9	(0.41–100)	36	(92)	90.4	(68–100)	35	(88)
Record request to GP only	78	199	(1–22 195)	69	(88)	88	(1–3 010)	75	(96)	72.3	(0.15–100)	68	(87)	90.0	(56–100)	68	(88)
Total	**357**	**468**	**(1–200 000)**	**255**	**(71)**	**104**	**(1–61 097)**	**183**	**(96)***	**71.0**	**(0.12–100)**	**165**	**(86)**	**n/a**		**n/a**	

*For example, in an external validation of glaucoma [29], 24 131 glaucoma patients were identified based on the study inclusion criteria (column A). Of these patients, 100 were chosen to undergo validation by questionnaire and record request (column B). This means that the proportion of cases chosen to undergo validation was just 0.41% of the total (column C). Of these 100 cases, the GP responded in 95, giving a response rate of 95% (column D).

		Gold standard diagnosis		
		Disease	No disease	Total
GPRD database	Disease	A (true cases correctly identified in GPRD)	B (non-cases wrongly coded as cases in GPRD)	A+B
	No disease	C (true cases not identified in GPRD)	D (true non-cases correctly identified in GPRD)	C+D
	Total	A+C	B+D	A+B+C+D

Figure 2

Measures of validity of categorical data. Sensitivity: A/(A+C); specificity: D/(B+D); positive predictive value: A/(A+B); negative predictive value: D/(C+D)

studies that found low validity of diagnoses were not published and that this publication bias could have affected our results.

The most robust method of validation may be to request additional information from the GP, since this method uses information external to the database to verify disease status of individual cases. Most such validations were restricted to establishing the proportion of cases with specific diagnostic codes that were confirmed by medical record review or responses to questionnaires, thus providing an estimate of the PPV of that set of codes (Figure 2). Although a useful measure, PPV varies with disease prevalence, so use of historical validations may not be justified if disease incidence has changed over time.

Information for cases alone does not allow calculation of sensitivity (the proportion of true cases correctly identified in the GPRD data), specificity (the proportion of individuals without the disease identified as such in the database), or negative predictive value. Even if PPV is high, other measures of validity could be low. These other measures require additional sampling of individuals without the diagnostic codes of interest (Figure 2). In most validations, the sensitivity, specificity and negative predictive value are not assessed, and this may be partly explained by the fact that for rare diseases, sampling from the vast number of individuals without the code of interest is particularly daunting. A handful of publications have successfully investigated sensitivity and specificity of diagnoses, demonstrating high validity for certain GPRD diagnoses [20, 21]. For example, Nazareth [8] estimated sensitivity and PPV of schizophrenia and psychosis diagnoses.

As described in Table 4, the proportion of identified cases that underwent validation was highly variable. Where this proportion is low, the precision of the validity estimate is reduced; most studies did not report confidence intervals around the PPV. Where only a proportion of total cases have been validated, it would be useful to compare those cases found to be valid with all other cases in terms of age, sex and other descriptive variables to look for systematic differences between them. One reason for small sample sizes in many validations is the high financial cost of record retrieval from GPs (currently averaging £70 per single set of notes).

Some GPRD practices do not participate in research studies, raising the question of generalizability of validation findings. For example, in a study by Van Staa [13], 719 practices contributed to the database during the study period but only 295 (41%) were known to provide additional information. Thus, even if compliance in providing records is high, the observed PPV may be applicable only to cases from a subgroup of practices. Practices who do participate in validation studies may only send information for certain cases, e.g. refusing to copy very large case files [22]; this may result in selection bias. Many publications did not report response rates clearly (with a complete lack of reporting in 16% of validations), making it impossible to assess whether selection bias could have affected their validation results. Where practices did respond to requests there were three possible outcomes: (i) notes were unavailable due to patient transfer or death, (ii) notes were returned with incomplete and/or inconclusive details of disease diagnosis, (iii) notes were returned with sufficient detail to verify the diagnosis. Since nonresponse, inadequate notes and exclusion because of patient death/transfer could bias assessment of validity in different ways, it would be useful to report them separately.

Given the high cost of record retrieval and GP questionnaires, manual review of the computerized records is cost effective but is also time consuming and takes away much of the advantage of having automated data. Less than half of the validations using this method specified the criteria used to determine 'true' cases. Without prespecified case criteria, there is scope for bias arising from judgements by individual physicians, which may vary over time and between physicians. Furthermore, recording of symptoms, results of diagnostic procedures and feedback from

secondary care may not be complete in computerized records, thereby limiting the usefulness of this approach.

Many investigators develop internal diagnostic algorithms to identify cases, but few use these to validate specific diagnostic codes (e.g. a medical code for acute myocardial infarction was validated by the presence of supportive evidence, e.g. codes for chest pain, fibrinolytic therapy, coronary intervention, troponin test results or hospitalization [23]). This method is quick and incurs no extra cost, so could be used more widely to validate diseases for which specific treatments are given universally. However, use of such algorithms may exclude less severe cases that do not require treatment, and the inclusion of test results in these algorithms is problematic since not all test results are recorded in the GPRD.

Comparison of rates gives a quick indication of the validity of the GPRD without the effort of individual case review. These comparisons do not validate individual cases or provide a measurable estimate of validity. Where prevalence rates are being compared, the GPRD may have a lower prevalence because GPs are not required to code prevalent conditions in each consultation [18]. Although results are reassuring for descriptive purposes, comparable rates of disease cannot identify potential balanced misclassifications between different diagnoses (i.e. the situation sometimes seen in death certification where the loss of deaths from cause A because of misclassification is balanced by the inclusion of people dying of cause B but misclassified to cause A). Reliance on this method to establish the validity of a diagnosis in the GPRD should be approached with caution and is not appropriate in analytic studies where individual validity is required. Similarly, sensitivity analysis is not a true validation of the data but does give an indication of the quality of diagnoses.

Most studies carried out using GPRD data are nested case–control studies. When conducting such a study, it is important to apply the same inclusion and exclusion criteria to cases and controls. However, validation studies which focus solely on cases may produce more detailed criteria for cases than for controls. For example, Garcia Rodriguez [24] investigated the relation between exposure to nonsteroidal anti-inflammatory drugs and acute liver injury. The investigators retrieved medical records of acute liver injury cases to verify their computerized diagnosis and excluded 16 of 166 potential cases (10%) from further analyses due to alcoholism. No further details on alcohol consumption by controls were retrieved, which may have led to bias. Validating a sample of noncases should ensure that control patients are subject to the same criteria as cases, although this would increase the financial cost of the research.

An alternative approach is the method that we recently applied to validate GPRD diagnoses of RA [22]. We used external medical records to validate RA diagnoses, but did not simply assess the overall PPV of an RA code. Instead, we identified characteristics in the computerized records of RA-coded patients that were associated with a valid diagnosis (e.g. specific prescriptions), and carried out multivariable analyses of these characteristics (using a valid RA diagnosis as the outcome) to develop a data-derived diagnostic algorithm of characteristics that could be used to identify valid cases in the database [25]. This method could be adapted to develop algorithms for a wide range of GPRD diagnoses.

Although considerable effort is often made to validate cases, the lack of detailed description of validation methods hinders interpretation of results. In some publications, lack of reporting was due to space constraints, which could be overcome by providing the relevant data as a web supplement. It is also helpful to make accessible a table of the medical OXMIS and Read codes used for diagnosis (or the mapping of these codes to specific ICD codes), so that others studying the disease can replicate case identification criteria. Figure 3 summarizes other information that could be made available to aid interpretation of validations.

Conclusion

The GPRD is an enormously powerful tool for the study of morbidity as recorded in primary care, but the quality of research using the data depends on the validity of the computerized information. It is therefore important for studies to perform some form of validation. At present, robust validations requesting additional information from GPs are limited in size due to costs, and their generalizability is compromised by nonparticipation of many practices. Careful use of internal diagnostic algorithms overcomes these concerns, and is a cost-effective method of identifying valid individual cases.

In future, it is likely that results from external clinical investigations and letters from specialists will be better captured in electronic records. These advances, along with the introduction of the Quality of Outcomes Framework, will greatly strengthen validations and are likely to improve the quality of the data (with fewer data entry errors and improved completeness). Work is also underway to extend the use of GPRD as a basis for randomized trials and as a sampling frame to obtain genetic data.

Linkage of GPRD with other healthcare databases (e.g. linkage to Hospital Episode Statistics), disease registers and death certificates will allow researchers to corroborate diagnoses made in hospital without the need to request medical records, and linkage to cancer registry data and the national audit of myocardial infarction (MINAP) will provide further opportunities. However, use of such linkages raises questions of how to resolve discordant or missing diagnoses in the two data sources. We hope that this study generates further debate about how best to assess the quality of the database and that this will further

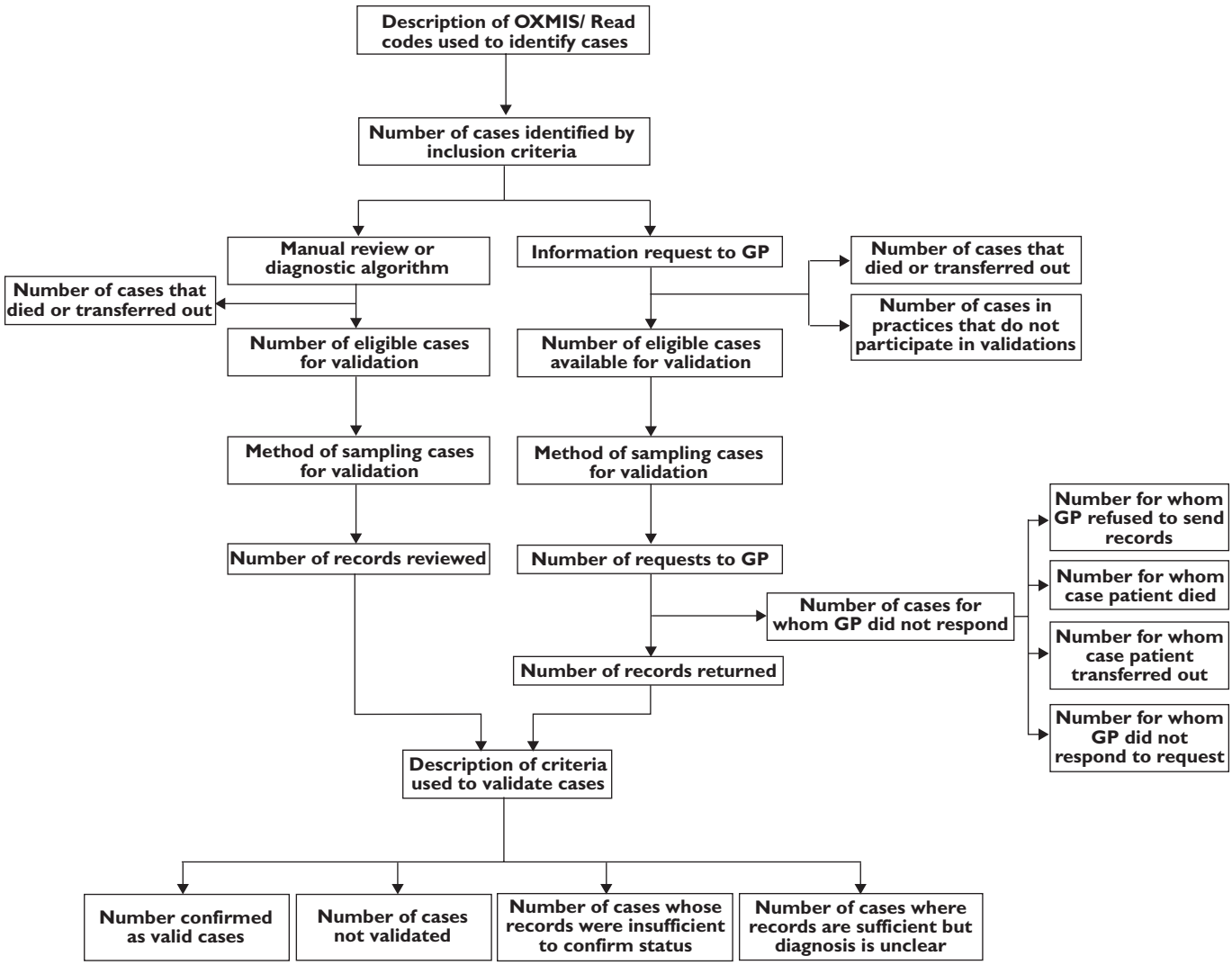

Figure 3

Stream diagram showing the information from General Practice Research Database (GPRD) validation studies that could be made available to researchers

enhance the reputation and the strength of the GPRD for use in research.

Competing interests

None declared.

L.S. was supported by a Wellcome fellowship.

REFERENCES

1 Lawrenson R, Williams T, Farmer R. Clinical information for research; the use of general practice databases. J Public Health Med 1999; 21: 299–304.

2 Lis Y, Mann RD. The VAMP research multi-purpose database in the U.K. J Clin Epidemiol 1995; 48: 431–43.

3 General Practice Research Database [Internet]. Available at http://www.gprd.com (last accessed 28 July 2009).

4 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: 419–21.

5 Smeeth L, Cook C, Fombonne E, Heavey L, Rodrigues LC, Smith PG, Hall AJ. MMR vaccination and pervasive developmental disorders: a case–control study. Lancet 2004; 364: 963–9.

6 Fombonne E, Heavey L, Smeeth L, Rodrigues LC, Cook C, Smith PG, Meng L, Hall AJ. Validation of the diagnosis of autism in general practitioner records. BMC Public Health 2004; 4: 5.

7 Hollowell J. The General Practice Research Database: quality of morbidity data. Popul Trends 1997; 87: 36–40.

8 Nazareth I, King M, Haines A, Rangel L, Myers S. Accuracy of diagnosis of psychosis on general practice computer system. BMJ 1993; 307: 32–4.

9 Soriano JB, Maier WC, Visick G, Pride NB. Validation of general practitioner-diagnosed COPD in the UK General Practice Research Database. Eur J Epidemiol 2001; 17: 1075–80.

10 Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. BMJ 1991; 302: 766–8.

11 Jick H, Terris BZ, Derby LE, Jick SS. Further validation of information recorded on a general practitioner based computerized data resource in the United Kingdom. Pharmacoepidemiol Drug Saf 1992; 1: 347–9.

12 Jick SS, Kaye JA, Vasilakis-Scaramozza C, Garcia Rodriguez LA, Ruigomez A, Meier CR, Schlienger RG, Black C, Jick H. Validity of the general practice research database. Pharmacotherapy 2003; 23: 686–9.

13 van Staa T, Abenheim L. The quality of information recorded on a UK database of primary care records: a study of hospitalisations due to hypoglycemia and other conditions. Pharmacoepidemiol Drug Saf 1994; 3: 15–21.

14 Wurst KE, Ephros SA, Loehr J, Clark DW, Guess HA. The utility of the general practice research database to examine selected congenital heart defects: a validation study. Pharmacoepidemiol Drug Saf 2007; 16: 867–77.

15 Yang CC, Jick SS, Jick H. Statins and the risk of idiopathic venous thromboembolism. Br J Clin Pharmacol 2002; 53: 101–5.

16 Ruigomez A, Johansson S, Wallander MA, Garcia Rodriguez LA. Predictors and prognosis of paroxysmal atrial fibrillation in general practice in the UK. BMC Cardiovasc Disord 2005; 5: 20.

17 Watson DJ, Rhodes T, Guess HA. All-cause mortality and vascular events among patients with rheumatoid arthritis, osteoarthritis, or no arthritis in the UK General Practice Research Database. J Rheumatol 2003; 30: 1196–202.

18 Jordan K, Clarke AM, Symmons DPM, Fleming D, Porcheret M, Kadam UT, Croft P. Measuring disease prevalence: a comparison of musculoskeletal disease using four general practice consultation databases. Br J Gen Pract 2007; 57: 7–14.

19 van Staa TP, Dennison EM, Leufkens HG, Cooper C. Epidemiology of fractures in England and Wales. Bone 2001; 29: 517–22.

20 Soriano JB, Maier WC, Egger P, Visick G, Thakrar B, Sykes J, Pride NB. Recent trends in physician diagnosed COPD in women and men in the UK. Thorax 2000; 55: 789–94.

21 Margolis DJ, Bilker W, Knauss J, Baumgarten M, Strom BL. The incidence and prevalence of pressure ulcers among elderly patients in general medical practice. Ann Epidemiol 2002; 12: 321–5.

22 Thomas SL, Edwards CJ, Smeeth L, Cooper C, Hall AJ. How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the general practice research database? Arthritis Rheum 2008; 59: 1314–21.

23 Andersohn F, Suissa S, Garbe E. Use of first- and second-generation cyclooxygenase-2 selective nonsteroidal antiinflammatory drugs and risk of acute myocardial infarction. Circulation 2006; 113: 1950–7.

24 Garcia Rodriguez LA, Williams R, Derby LE, Dean AD, Jick H. Acute liver injury associated with nonsteroidal anti-inflammatory drugs and the role of risk factors. Arch Intern Med 1994; 154: 311–6.

25 Quigley MA, Chandramohan D, Rodrigues LC. Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. Int J Epidemiol 1999; 28: 1081–7.

26 Gupta G, Gelfand JM, Lewis JD. Increased risk for demyelinating diseases in patients with inflammatory bowel disease. Gastroenterology 2005; 129: 819–26.

27 Gonzalez-Perez A, Garcia Rodriguez LA. Prostate cancer risk among men with diabetes mellitus (Spain). Cancer Causes Control 2005; 16: 1055–58.

28 Hall GC, Roberts CM, Boulis M, Mo J, MacRae KD. Diabetes and the risk of lung cancer. Diabetes Care 2005; 28: 590–4.

29 Huerta C, Rodriguez LA. Incidence of ocular melanoma in the general population and in glaucoma patients. J Epidemiol Community Health 2001; 55: 338–9.