A comparison of the recording of 30 common childhood conditions in the Doctors' Independent Network and General Practice Research Databases

S De Wilde, I M Carey and S A Bremner, *Department of Community Health Sciences, St George's Hospital Medical School, London,* **N Richards,** *CompuFile Ltd, Send, Woking, Surrey* **S R Hilton, D P Strachan and D G Cook,** *Department of Community Health Sciences, St George's Hospital Medical School, London*

In this article we compare the recording of 30 common childhood conditions in two general practice databases of anonymised computerised medical records based on fundamentally different systems – the Doctor's Independent Network (DIN) database (Torex system) and the General Practice Research Database (GPRD) (In Practice Systems). Analysing the records of all children born 1990–1993 and followed for 5 years we found comparable results for most conditions, but differences between the hierarchical structures of the diagnostic coding systems (Read in DIN, OXMIS in GPRD) led to some differences between the databases. Practice variation was marked, but comparable between databases. Variation was greatest in conditions that are poorly defined clinically.

Introduction

Common childhood illnesses form a major part of general practice (GP) workload.[1] The illnesses themselves may lead to loss of educational time and there may also be consequential loss of parental employment time when ill children require care at home.[2] Some common childhood illnesses may also have longer-term morbidity associated with them. It has long been recognised that it is important to understand the incidence and prevalence of these conditions in order to facilitate the planning of health care provision.

The most recent detailed information on the prevalence of such illnesses in the UK is to be found in the 4th National Morbidity Survey from General Practice (MSGP4).[3] The Office of Population Censuses and Surveys (now part of the Office for National Statistics), jointly with the Royal College of General Practitioners and the Department of Health, conducted the fourth of a series of studies of Morbidity Statistics from General Practice in 1991–1992. Data collection was from 60 general practices in England and Wales with half a million patients, recording all face-to-face contacts over a period of a year. MSGP4 is now more than 10 years old, and no comparable survey has been undertaken more recently. It is unlikely that there will be an MSGP5, but large primary care databases have the potential to repeat and extend the previous National Morbidity Surveys.

Large-scale databases routinely collecting data from general practice computer systems have existed for more than 10 years. The largest of these is the General Practice Research Database (GPRD),[4] which has increasingly been used to provide data on disease prevalence and prescribing as well as for other types of research.[5] Other databases exist, which might complement GPRD or have some advantages over it.[6]

Routinely collected prevalence data differ from data collected in surveys such as MSGP4 in that they are collected as part of the process of patient care using GP computer systems that may differ from each other in design philosophy and coding systems.[7] Individual GPs will vary in the degree and manner in which they use their computers,[8] and differences between computer systems will influence this. If such data are to be used effectively, it is important to consider the degree of variation in data recording between GPs at least at the practice level, and also to consider the manner in which data recording is influenced by the use of different computer systems.

In this article, we use birth cohorts in GPRD and DIN (The Doctors' Independent Network database) to study the cumulative incidence, to age five, of 30 common childhood conditions, contrasting recording levels in each database and paying particular attention to the issue of practice variation.

Methods

Background to DIN and GPRD

DIN is an on-going, anonymised computerised database from practices that use Torex (formerly Meditel) software from 1989 onwards. The DIN data on which we base this article are from 142 practices selected as high quality data providers.[9] The GPRD based on In Practice Systems (formerly VAMP) software is well established in epidemiological research,[10] and has been collecting data over the same period. Our analyses are limited to data collected up to 1998 from 464 practices in GPRD.

Coding systems and record structure

There are two fundamental differences between the databases. The first is that DIN has used Read Codes (GP 4-Byte set) for recording diagnoses, while for much of this period practices in GPRD used the OXMIS (Oxford Medical Information System) coding system (with a few exceptions where some practices used Read Codes [Unified 5-Byte Version 2 Set] for diagnoses from 1997). Read has a hierarchical structure, starting with high level codes with less precise meanings, and increasing in specificity as the hierarchy branches. OXMIS is not hierarchical, and tends to contain more GP oriented diagnoses.

DIN and GPRD also have differently structured medical records, again arising from the two different systems on which they are based. Torex (formerly Meditel) System 5, underlying DIN, was based on the concept of the Problem Orientated Medical Record (POMR),[11] which was designed to present the medical record as a chain of intertwined but discrete problems, with prescriptions being linked to diagnoses under problem headings. In contrast VAMP Medical software, underlying GPRD, presented the notes as a series of discrete episodes. Previous work comparing the databases found similar levels of prescribing but fewer diagnostic codes in DIN. However, once linkage was taken into account the level of recording of diagnoses was similar in the two databases.[9] By basing the present analyses on cumulative incidence rates within birth cohorts any differences between the underlying systems in the way diagnoses are recorded are minimized.

Defining birth cohorts in each database

A previous article described how we set up birth cohorts in DIN and GPRD.[12] These comprised children who were registered within three months of their date of birth and followed continuously for 5 years. One advantage of using only these children is that all consultations should have been recorded in real-time with no reliance on retrospectively entered data.

A total of 40,183 children in DIN (born 1989–1997) and 76,310 (born 1989–1993) in GPRD fulfilled these criteria. Due to suspect quality of data recording in the 1989 birth cohort in GPRD it was decided to exclude these births (n=13,772) as well as the small number of corresponding births from DIN (n=511) for completeness. To compare the cohorts over the same period we also excluded children born after 1993 on DIN (n=20,034). This left 19,638 children in DIN cohort (from 123 practices) and 62,538 in the GPRD cohort (462 practices). In DIN, these children represent 69 per cent of all births in these practices over this time.

Definition of the 30 conditions

Data from MSGP4 were used to identify ICD-9 codes for the most frequent diagnoses in children aged 0–4. Initially we selected ICD groups with a consultation rate of greater than 500 per 10,000 person years at risk, but added a few other conditions to this if we felt that they were of clinical importance in general practice (such as 'helminthiasis' and 'pneumonia and influenza'). As ICD is not a general practice coding system, we had to identify clinical interpretations for each grouping in the context of 0- to 4-year-olds e.g. the ICD subgroup 'ill defined intestinal infections' was interpreted as 'diarrhoea and vomiting'. Some subgroups were broken down further, where we felt it was clinically meaningful to do so e.g. 'Other viral exanthemata' was clinically interpreted by us as 'viral rash' and further subdivided as 'measles', 'rubella', 'hand, foot and mouth' and 'other viral rash'.

The clinical interpretations of the ICD subgroups were then mapped across to the Read 4 byte codes used in DIN and the OXMIS and Read 5 byte codes used in GPRD (Table 2, data not shown for Read 5). Only 'trauma' was excluded due to difficulties in the mapping between coding systems. We also made the post-hoc decision to combine 'URTI' and 'Common Cold' as it was apparent that underlying differences in the structure of coding systems made it difficult to distinguish between the conditions. In OXMIS, a single code for 'URTI' exists and was frequently used. However, in Read the precise codes for 'URTI' were not as apparent, and it appeared that many GPs in DIN were not making a consistent clinical distinction between them, with many using a high-level code instead ('H1..' – Acute respiratory infections). Because this high level code includes both 'Acute bronchitis/Chest infection' and 'URTI' it was excluded from both definitions. Inevitably this would result in lower rates in DIN than in GPRD and we therefore added a further condition ('Any RTI') that included all codes for acute respiratory infections (including 'H1..') to facilitate comparison between databases.

Within the birth cohorts in each database, the identified codes for the 30 conditions were electronically searched for. Presence of an appropriate code enabled a condition to be identified, and contributed to the numerator for one- and five-year cumulative incidence rates for each condition.

Assessing practice variation

For each condition the relative odds in DIN compared to GPRD was derived from a logistic regression with practice as the unit of analysis, allowing for extra binomial variation (the 'random effect' of practice). The model was fitted using *Proc NLMIXED* in SAS version 8.1 for Solaris (SAS Institute, Cary, NC).

Variation between practices in the cumulative incidence of each condition are summarised using box and whisker diagrams with the boxes indicating the median, lower and upper quartiles and the whiskers extend to the practice immediately proceeding 1.5 times the interquartile range from the median. Practices lying outside this range are individually plotted. To minimise the impact of chance variation, practices contributing fewer than 25 births in our data are excluded from these plots (n=9 DIN, n=34 GPRD).

To assess the extent to which practice variation existed beyond that which might be expected by chance, simulations were carried out; these assumed there was no clustering effect of practice. For each database an overall probability p was calculated for each disease. In a single simulation, every child in that database was given the same risk of disease *p*, and allocated a status of 'Diseased' or 'Not Diseased'. They were then randomly divided into groups representing the known practice sizes (n_i). The number of diseased in each practice was counted (r_i), such that the probability of disease in practice *i* was $p_i = r_i / n_i$. The simulation was repeated 500 times to produce a distribution of "Expected" practice means, which could be compared to the 'Observed' set. We present the observed and expected distributions for two of the thirty conditions (eczema and sore throat), separately for DIN and GPRD.

Results

Table 1 summarises the birth cohorts from the two databases. Cumulative incidences (one- and five-year) for each of the 30 conditions by database are summarised in Table 3. Generally there is good agreement between databases, with 25 conditions having an OR (odds ratio) between databases of 0.7–1.3 and 19 conditions having the OR between 0.8–1.2. There were however notable exceptions. Higher in DIN were influenza (OR=2.30) and ringworm (OR=2.25); higher in GPRD were upper respiratory tract infections (OR=0.19), acute bronchitis (OR=0.60) and viral illness (OR=0.14). The discrepancy between databases within respiratory infections improved when the combined category 'Any RTI' was used, but it was still lower overall in DIN (OR=0.70).

Table 4 summarises the five-year cumulative incidence rates by sex. Where there are differences between boys and girls, they are consistent across databases. Boys appear to have higher incidence of respiratory tract infections (such as bronchiolitis and laryngitis), atopic conditions (asthma and hayfever) and behavioural problems. Girls have higher rates of candidiasis, urinary tract infections, nappy rash and infestations (headlice and threadworms).

While there is a high level of variation in cumulative incidence across practices in both databases, the level of variation does not appear to be different in the two databases (Figure 1). The conditions that show the most variation tend to be those diagnoses that are more varied in their presentation or aetiology such as URTI, acute bronchitis, influenza, diarrhoea and sore throat. In many situations achieving a differential diagnosis may be difficult or unimportant clinically. More specific diagnoses such as hay fever, urinary tract infection or ringworm show much less variation between practices. Our estimates of the additional variation due to the effect of practice showed a similar pattern in the two databases (data not shown) and Figure 2 gives two examples of this. The top graphs show the distribution of observed and expected practice percentages in DIN and GPRD for eczema and dermatitis, where the variation is not too dissimilar from that which would be expected if there were no clustering by practice. The bottom graphs show this for sore throat, where there is clearly much more variation than would be expected.

Discussion

In this article we have utilised birth cohorts in two large-scale GP databases to investigate the cumulative incidence of 30 common childhood conditions. Whilst our estimates for incidence have shown there to be reasonable agreement between the databases, they have also highlighted some differences. These may be due to the different computing and coding systems used by the GPs providing data to the databases as well as to geographical and social class differences. We have also shown that some of the estimates show considerable variation between the individual practices providing data.

Explaining differences between databases

There was good agreement between the two databases in the mean/median cumulative incidence rates for most conditions. However there were major differences for 'URTI', acute bronchitis and viral illness which were higher in GPRD, and for influenza and ringworm which were higher in DIN. All except ringworm appear to be attributable to the two different coding systems.

Differences within respiratory tract infections would have been even greater if we had tried to distinguish between 'URTI' and 'Common Cold' as it was apparent that GPs using their IPS or Torex systems used the diagnoses almost interchangeably. Therefore choice of diagnosis was largely dependent on the ease of coding, which in turn depended on the coding system and the clinical system being used. Slightly higher rates were also seen for lower respiratory infections in GPRD, while influenza was higher in DIN. These differences between the databases are readily explained by many GPs in DIN failing to distinguish between upper and lower respiratory tract infections. Instead many were using the high-level code instead ('H1..' – Acute respiratory infections). Because this high level code includes both 'Acute bronchitis/Chest infection' and 'URTI' it was excluded from both definitions. The combined category of any respiratory tract infection shows better comparability between the databases both for one- and five-year cumulative incidence.

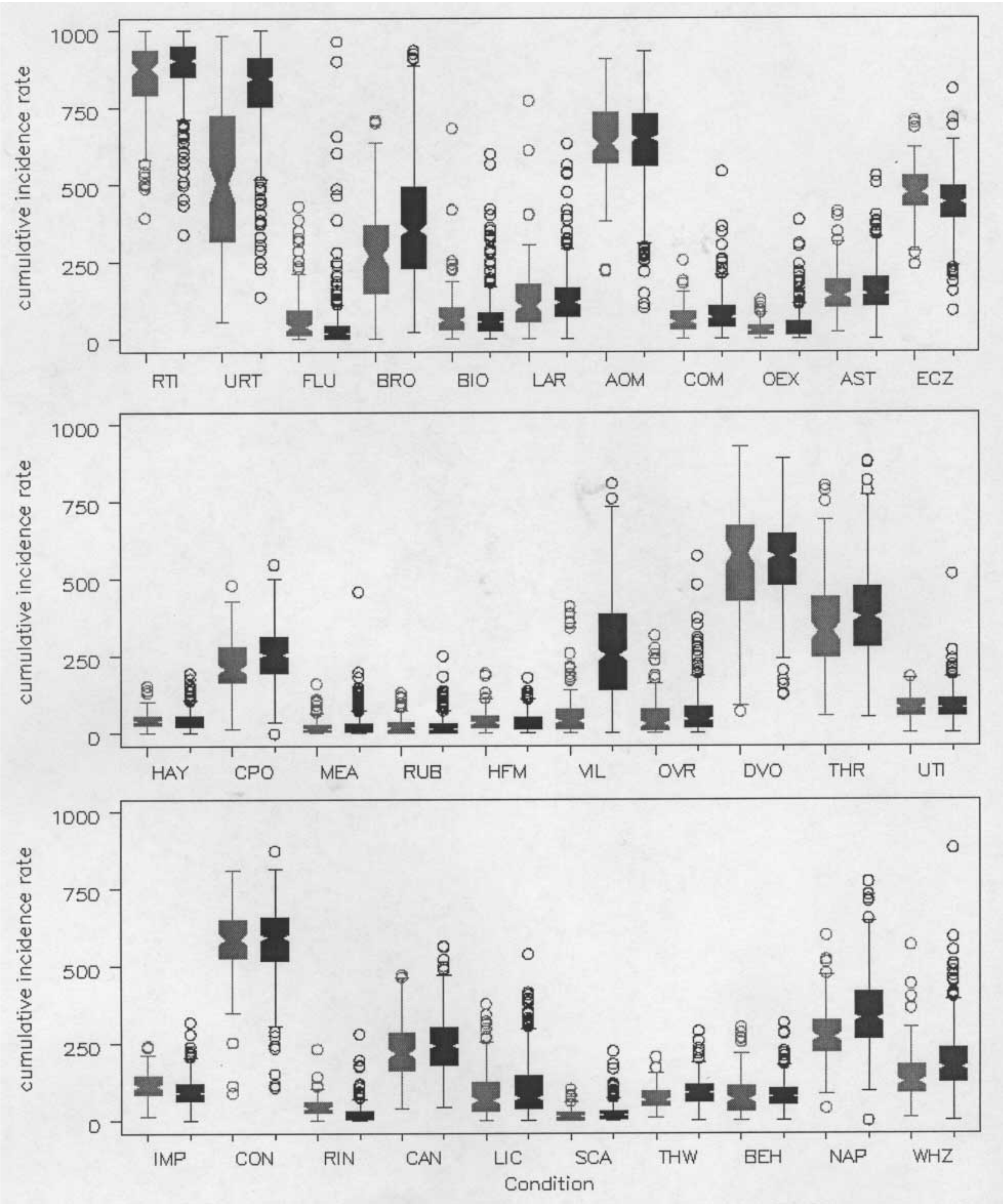
The higher rates of 'viral illness' in GPRD may be explained similarly. OXMIS has a precise code for this common non-specific presentation in general practice, while 4-byte Read does not. The higher level Read code 'A4..' (viral diseases) is not included in our definition as it also covers other specific viral illnesses (e.g. chickenpox). However, it seems likely that GPs in DIN are using this high level code to represent some of these occurrences.

This highlights the importance of considering the effects that different coding and clinical systems can have on coding the diagnoses of the same clinical entities. Comparison between different databases, or between coding systems in the same database (as in the case of GPRD, which uses both OXMIS and Read codes) involves more than a simple cross-mapping table to convert codes between systems. The meanings intended by users of apparently similar codes in different systems may be quite different. We also observed that the deeply hierarchical structure of the Read codes increases the possibility of ambiguous coding, as the coding system contains many necessarily less precise high-level codes. Clearly any research comparing databases or coding systems requires careful validation of the codes chosen.

Other smaller differences between the databases may be due to differing geographic coverage and social class structure of the practice populations. Thus we have previously shown that rates of treated ischaemic heart disease are similar to but slightly lower in DIN than GPRD are entirely explained by the southern dominance in DIN (60% of practices in the south) compared to the northern dominance in GPRD (60% of practices in the north).[13] There are also likely to be socio-economic differences, with DIN being slightly overrepresented in more prosperous areas.[13] For example, such differences may explain the slightly lower rates of infestations in DIN.

Finally we need to consider whether the differences between databases may be explained by differences in data quality control. Certainly GPRD has more explicit criteria for when diagnostic codes should be recorded.[14, 15] However, this seems an unlikely explanation for differences as several codes were more common within DIN, and we have previously demonstrated good comparability in the overall volume of diagnostic codes within the DIN and GPRD birth cohorts.[9]

Figure 1 **Box and whisker plots of practice mean cumulative incidence rates (per 1,000) at 5 years for 30 common childhood conditions in DIN (grey boxes) and GPRD (black boxes)**



Footnote:
Boxes indicate the median, lower and upper quartiles (Grey=DIN, Black=GPRD). Whiskers extend to the practice immediately prceeding 1.5 times the interquartile range from the median. Practices lying outside this range are individually plotted.
Abbreviations for conditions are as follows: TRI - Any RTI, URT - URTI (incl. common code), FLU - Influenza, BRO - Acute bronchitis/Chest infection, BIO - Bronchiolitis, LAR - Laryngitis and croup, AOM - Acute otitis media, COM - Chronic OM and glue ear, OEX - Otitis externa, AST - Asthma, ECZ - Eczema and dermatitis, HAY - Hayfever/allergic rhinitis, CPO - Chickenpox, MEA - Measles, RUB - Rubella, HFM - Hand, Foot and Mouth, VIL - viral illness, OVR - Other Viral rash, DVO - Diarrhoea and vomiting, THR - Sore throat, URTI - Urinary tract infections, IMP - Impetigo, CON - Conjunctivitis, RIN - Ringworm, CAN - Candidiasis, LIC - Headlice, SCA - Scabies, THW - threadworms, BEH - Behavioural problems, NAP - Nappy Rash, WHZ - Wheeze.

Table 1 Comparison of the DIN and GPRD birth cohorts

	DIN (123 practices)		GPRD (462 practices)	
	Number of children	Percentage	Number of children	Percentage
Total with 5-year follow-up	19,638		62,538	
Sex				
Boys	10,013	51.0	31,908	51.1
Girls	9,625	49.0	30,630	48.9
Birth Year				
1990	3003	15.3	18,946	30.2
1991	4421	22.5	17,532	28.0
1992	5412	27.6	14,646	23.4
1993	6802	34.6	11,414	18.3
Contribution by Practice				
0–24 children	9	7.3	34	7.4
25–100 children	36	29.3	178	38.5
100–200 children	43	35.0	144	31.1
200–500 children	33	26.8	104	22.5
Over 500 children	2	1.6	2	0.4

Figure 2 Observed and expected distributions of practice mean cumulative incidence at five years for eczema and sore throat

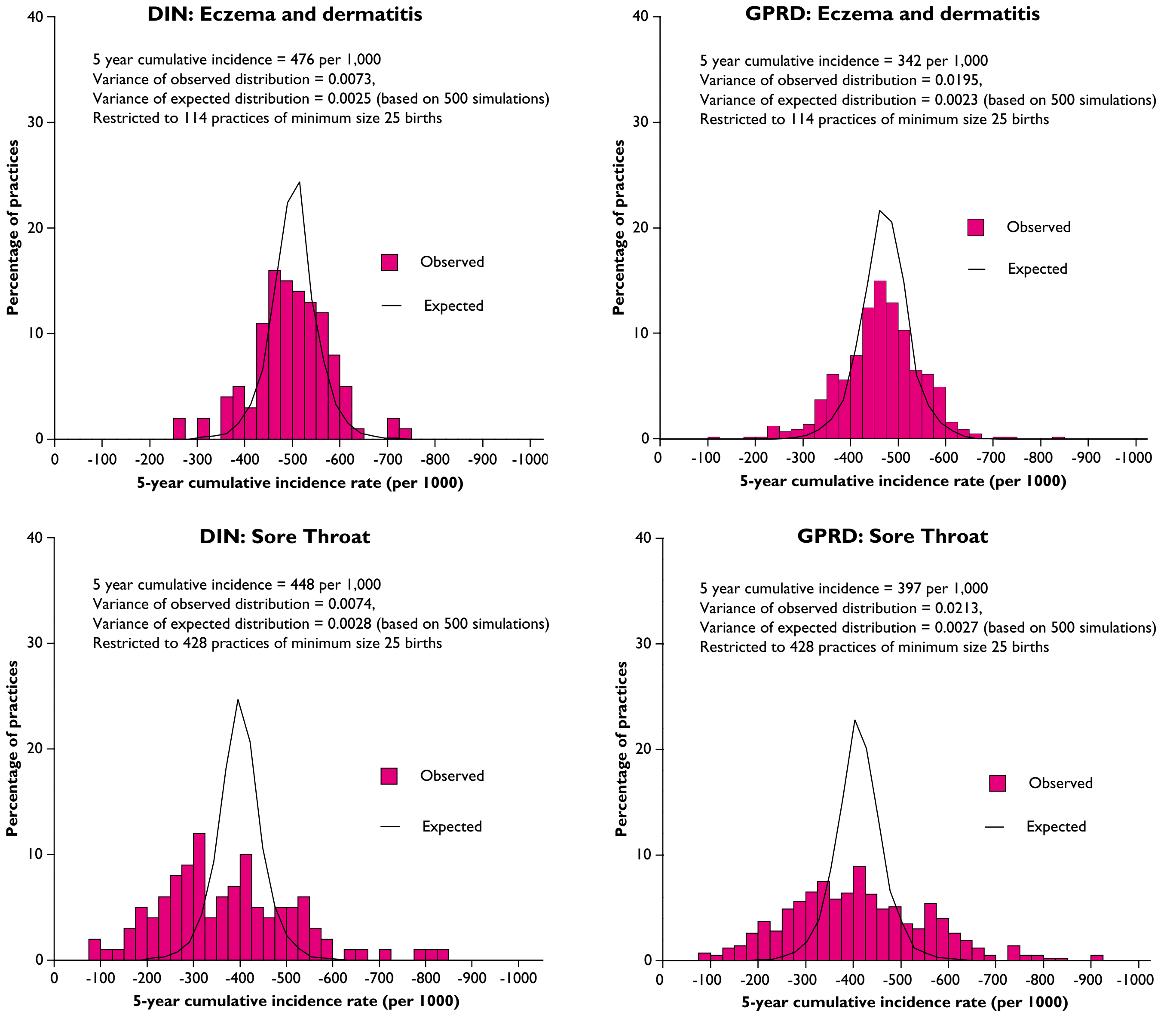


Table 2 Definitions of 30 conditions using Read and OXMIS codes (respiratory tract infections and ear infections)

Grouped Heading	ICD subgroups in MSGP4	Clinical interpretation	Rubrics for 4 byte Read codes (%)*		Rubrics for OXMIS codes (%)*	
Respiratory Tract Infections	Acute upper respiratory tract infection (including acute nasopharyngitis)	Upper respiratory tract infection (including common cold)	H11.	Acute nasopharyngitis (71%)	465	URTI (Upper respiratory tract infection) (85%)
			H1Z.	Acute resp. infection NOS (20%)	460DC	Coryza (8%)
			H17.	Acute upper respiratory tract infection (4%)	460CT	Catarrh acute (3%)
			H2..	Other upper respiratory tract disease (2%)	460D	Common cold (2%)
			H2Z1	Upper respiratory tract infection (2%)	465A	URI (Upper respiratory infection) (1%)
	Influenza	Influenza	H35Z	Influenza NOS (91%)	470F	Flu (62%)
			H3..	Pneumonia and influenza (5%)	460C	Influenza-like illness (27%)
			H35.	Influenza (3%)	470	Influenza (11%)
	Acute bronchitis and bronchiolitis	Acute bronchitis/ Chest infection	H161	Acute bronchitis (66%)	519EE	Infection chest (87%)
			H6ZA	Chest infection NOS (28%)	490	Bronchitis (12%)
			H6Z1	Bronchitis NOS (2%)		
			H16.	Acute bronchitis/bronchiolitis (2%)		
			H163	Acute low respitract infection (1%)		
		Bronchiolitis	H162	Acute bronchiolitis (100%)	466A	Bronchiolitis (96%)
					466B	Bronchiolitis acute (4%)
	Acute laryngitis and tracheitis	Laryngitis and croup	H155	Croup (68%)	464E	Croup (83%)
			H152	Acute tracheitis (15%)	464D	Tracheitis (9%)
			H153	Acute laryngotracheitis (12%)	464B	Laryngitis (5%)
			H151	Acute laryngitis (3%)	464LT	Layngotracheitis (2%)
Ear Infections	Suppurative and unspecified otitis media	Acute otitis media	F621	Acute nonsuppurative otitis media (51%)	3819	Otitis media (71%)
			F631	Acute suppurative otitis media (18%)	3810MR	Otitis media acute right (6%)
			F62.	Nonsuppurative otitis media (13%)	3810ML	Otitis media acute left (6%)
			F679	Otalgia – earache (8%)	384AB	Earache (6%)
			F625	Otitis media NOS (2%)	3819E	Otitis media bilateral (5%)
			F63.	Suppurative otitis media (2%)	3819N	Otitis (4%)
			2D95	O/E – tympanic membrane red (2%)	3819A	Infection ear (2%)
			2D96	O/E – tympanic membrane pink (1%)		
			1C3.	Earache symptoms (1%)		
	Non-suppurative otitis media	Chronic otitis media and glue ear	F623	Chronic mucoid otitis media (40%)	3811B	Glue ear (31%)
			7432	Myringotomy and grommet inserted (19%)	3811TR	Secretory otitis media (18%)
			743.	Tympanic membrane operation (7%)	K1933	Insertion grommet ear (16%)
			F624	Eustachian tube dysfunction (7%)	384B	Eustachian catarrh (12%)
			F622	Chronic serous otitis media (7%)	K1931M	Myringotomy (7%)
			313A	Tympanogram abnormal (4%)	3811AT	Otitis media catarrhal (7%)
			F633	Chronic purulent otitis media (4%)	384BD	Eustachian tube dysfunction (2%)
			2D9A	O/E – otoscopy:fluid-middle ear (4%)	3811D	Otitis media chronic suppurative (2%)
			F626	Eustachian catarrh (3%)	3811EF	Effusive ear (1%)
			2D9B	O/E – tympanic membr retracted (3%)		
	Disorders of the external ear	Otitis externa	F612	Infective otitis externa (81%)	380	Otitis externa (99%)
			F616	Otitis externa NOS (10%)		
			F613	Non-infective otitis externa (9%)		
Mainly Atopic Conditions	Asthma	Asthma	H43.	Asthma (47%)	493	Asthma (88%)
			H43Z	Asthma NOS (25%)	493AA	Asthma acute (3%)
			H431	Extrinsic asthma – atopy (11%)	493KB	Asthma exacerbation (2%)
			H432	Intrinsic asthma (5%)	493AB	Asthma attack (2%)
			663U	Asthma management plan given (2%)	493BR	Bronchial asthma (2%)
			173A	Exercise induced asthma (2%)		
			14B4	H/O: asthma (2%)		
			6630	Asthma not disturbing sleep (1%)		
			H434	Asthma attack NOS (1%)		
	Atopic dermatitis and related conditions	Eczema and dermatitis	L2..	Dermatitis/eczemas (22%)	6929CE	Eczema (72%)
			L233	Infantile eczema (22%)	709L	Skin dry (13%)
			L2Z.	Dermatitis/eczemas NOS (17%)	691B	Infantile eczema (3%)
			L23.	Atopic eczema/dermatitis (10%)	6929C	Dermanitis (3%)
			2F13	O/E – dry skin (8%)	6869MI	Eczema infected (3%)
			L232	Flexural eczema (5%)	691C	Cradle cap (2%)
			L22.	Seborrhoeic dermatitis/eczema (4%)	691EC	Eczema atopic (1%)
			L221	Seborrhoea capitis (3%)	690A	Seborrhoeic eczema (1%)
			L23Z	Atopic eczema/dermatitis NOS (2%)		
			L22Z	Seborrhoeic dermat/eczema NOS (2%)		
	Allergic rhinitis	Hayfever/allergic rhinitis	H28.	Hay fever (82%)	507HF	Hay fever (61%)
			H29.	Allergic rhinitis NOS (16%)	5070S	Hayfever symptoms (27%)
			14B1	H/O: hay fever (3%)	507AN	Rhinitis allergic (12%)

Table 2 continued **Definitions of 30 conditions using Read and OXMIS codes (respiratory tract infections and ear infections**

Grouped Heading	ICD subgroups in MSGP4	Clinical interpretation	Rubrics for 4 byte Read codes (%)*		Rubrics for OXMIS codes (%)*	
Viral Infections	Chickenpox	Chickenpox	A43. 141A	Chicken pox (98%) H/O: chickenpox (2%)	052 052A	Chickenpox (95%) Varicella (5%)
	Other viral exanthemata	Measles	A46. R212 A46Z	Measles (76%) Morbilliform rash (20%) Other measles (4%)	055 055C	Measles (96%) Modified measles (3%)
		Rubella	A47. R211	Rubella (53%) Rubelliform rash (47%)	056 056G	Rubella (87%) Measles German (13%)
		Hand, foot and mouth	A4D2	Hand, foot and mouth disease (100%)	0749C 794	Hand foot and mouth disease (97%) Foot and mouth disease (3%)
	Other disease due to viruses or chlamydia	Viral illness	A4Z. A4..	Viral diseases NOS (52%) Viral diseases (48%)	0799VI 0799E 0799ES	Viral illness (77%) Viral infection (19%) Viral symptoms (3%)
		Other viral rash	A48. R210	Other viral exanthemata (65%) Exanthem (35%)	0579AB 0579 0578NF 0570A 0570	Viral rash (51%) Virus exanthem (21%) Roseola infantum (15%) Fifth disease (9%) Erythema infectiosum (4%)
Other Infections	Ill defined intestinal infections	Diarrhoea and vomiting	A121 19F2 1992 19F. 199. A12. 19FZ R701 A1Z. A12Z	Gastroenteritis – viral and NOS (30%) Diarrhoea (20%) Vomiting (11%) Diarrhoea symptoms (9%) Vomiting (9%) Viral and ill-defined GIT infectious (9%) Diarrhoea symptom NOS (3%) Vomiting (3%) Intestinal infectious dis. NOS (3%) Infectious diarrhoea NOS (1%)	0091A 0092C 0091B 0091AB 0092	Diarrhoea (56%) Gastroenteritis (22%) Vomiting & diarrhoea (16%) Stools loose (2%) Enteritis (1%)
	Acute pharyngitis/ tonsillitis	Sore throat	H14. H13. 1C92 1C9. H141	Acute tonsillitis (56%) Acute pharyngitis (35%) Has a sore throat (3%) Sore throat symptom (3%) Acute recurrent tonsillitis (1%)	463A 462AR 462AB 463 462C	Tonsillitis (63%) Throat soreness (23%) Pharyngitis (8%) Tonsillitis acute (3%) Pharyngitis (1%)
	Other disease of urethra and urinary tract	Urinary tract infections	J2B1 J261 J26.	Urinary tract infection NOS (91%) Acute cystitis (6%) Cystitis (2%)	599A 595	UTI(Urinary tract infection) (94%) Cystitis (5%)
	Impetigo	Impetigo	L15.	Impetigo (100%)	684A 684AD 6869MP	Impetigo (88%) Impetigenous dermatitis (11%) Eczema impetiginous (1%)
	Disorders of conjunctiva	Conjunctivitis	F5B1 F5B7	Acute conjunctivitis (97%) Allergic conjunctivitis (2%)	360A 361P 360G 0399A 360E	Conjunctivitis (77%) Sticky eye (17%) Conjunctivitis acute (4%) Conjunctivitis bacterial (1%) Conjunctivitis allergic (1%)
	Dermatophytoses	Ringworm	A716 A715 A71. A714 A712 A711 A71Z A713	Tinea of body – ringworm (39%) Tinea pedis – athlete's foot (22%) Dermatophytosis-tinea/ringworm (12%) Tinea of groin/perianal (8%) Tinea of scalp (8%) Pityriasis versicolor (4%) Other dermatophytosis (3%) Tinea of nail – onychomycosis (3%)	110D 110A 110C 110AC 110F 110B 1110T	Ringworm (47%) Tinea (27%) Tinea pedis (10%) Tinea corporis (6%) Tinea capitis (4%) Tinea cruris (4%) Tinea versicolor (1%)
	Candidiasis	Candidiasis	A721 O255 A723 A722 A72. A72Z A724	Oral Thrush (40%) Neonatal monilia (26%) Perianal candida (13%) Candidal vulvovaginitis (9%) Candidiasis (7%) Other candidiasis (3%) Penile candida (2%)	112B 112A 112BC 112G 6928NC 112BM	Thrush stomatitis (45%) Thrush (42%) Oral candidiasis (5%) Urogenital thrush (2%) Nappy rash candidal (1%) Oral moniliasis (1%)

Table 2 continued **Definitions of 30 conditions using Read and OXMIS codes (respiratory tract infections and ear infections**

Grouped Heading	ICD subgroups in MSGP4	Clinical interpretation	Rubrics for 4 byte Read codes (%)*		Rubrics for OXMIS codes (%)*	
Infestations	Pediculosis and phthirus infestation	Headlice	A831 A83.	Pediculus capitis – head lice (98%) Pediculosis and other lice (2%)	132B 132	Pediculosis capitis (97%) Lice (3%)
	Other intestinal helminthiases	Threadworms	A812 19E9 A81	Threadworms – enterobiases (84%) Worms in faeces (10%) Helminthiases (6%)	1271T 1289	Threadworms (71%) Worms (28%)
Other	Disturbance of conduct not elsewhere classified	Behavioural problems	1B11 E4A. 1B1B E4F. 1B15 E4FZ 19E2 E4H. E2Z1	Crying, excessive (53%) Sleep disorders (13%) Cannot sleep – insomnia (8%) Disturbance of conduct NOS (8%) Irritable (6%) Disturbance of conduct NOS (5%) Soiling – encopresis (3%) Overactive child syndrome (3%) Infantile autism (1%)	3064A 3069CP 307TA 308BH 3064D 308RT 308CR 7856L 308B 308	Sleep inability to (34%) Behaviour problem (11%) Screaming attacks (8%) Breath holding attacks (7%) Sleep disorder (7%) Irritable infant (5%) Crying infant (5%) Soiling(faecal) (3%) Temper tantrums(childhood) (3%) Disorder behaviour childhood (3%)
	None	Nappy rash	L231	Nappy rash : dermatitis (100%)	6928NR 6928NP	Rash napkin(diaper) (91%) Rash diaper (8%)
	None	Wheezing	1737 2326 R609 173B	Wheezing (44%) O/E – expiratory wheeze (27%) Wheezing (16%) Nocturnal cough / wheeze (12%)	7832A 490WH 7832AB	Wheezing (70%) Wheezy bronchitis (28%) Wheezy bronchial (2%)

* Percentages represent the % of all codes used in our definition to age 5 years for each condition in each database (excluding Read 5 practices in GPRD). Only those contributing more than 1% are shown in the table.

Table 3 **One- and five-year cumulative incidence rates (per 1,000) for 30 common childhood conditions in DIN and GPRD**

Condition	Cumulative incidence at 1 Year (rate per 100)		Cumulative incidence at 5 Years (rate per 100)			
	GPRD	DIN	GPRD	DIN	OR*	95% CI
Respiratory Tract Infections						
Any RTI†	636	601	886	838	0.70	0.58–0.85
URTI (incl. common cold)	546	331	819	508	0.19	0.16–0.24
Influenza	7	11	40	74	2.30	1.68–3.15
Acute bronchitis/Chest inf.	174	115	391	282	0.60	0.49–0.72
Bronchiolitis	58	65	72	83	1.13	0.94–1.36
Laryngitis and croup	33	34	136	132	0.96	0.82–1.12
Ear Infections						
Acute otitis media	216	206	653	650	1.02	0.90–1.15
Chronic OM and glue ear	8	6	83	66	0.81	0.69–0.95
Otitis externa	10	8	40	31	0.81	0.67–0.97
Mainly Atopic Conditions						
Asthma	31	31	167	152	0.96	0.85–1.07
Eczema and dermatitis	214	234	448	476	1.15	1.07–1.23
Hayfever/allergic rhinitis	2	3	42	45	1.10	0.96–1.25
Viral Infections						
Chickenpox	33	30	252	214	0.80	0.72–0.89
Measles	9	8	22	22	0.91	0.71–1.17
Rubella	10	12	21	26	1.16	0.91–1.48
Hand, Foot and Mouth	2	4	39	43	1.08	0.92–1.27
Viral illness	95	21**	299	68*	0.14	0.11–0.17
Other Viral rash	21	20	66	54	0.79	0.62–1.01
Other Infections						
Diarrhoea and vomiting	286	269	564	540	0.93	0.82–1.05
Sore throat	52	53	397	366	0.84	0.73–0.96
Urinary tract infections	11	11	88	89	0.99	0.89–1.10
Impetigo	12	14	98	114	1.23	1.11–1.35
Conjunctivitis	329	307	586	583	1.00	0.92–1.10
Ringworm	3	7	22	44	2.25	1.90–2.66
Candidiasis	191	166	250	233	0.91	0.82–1.00
Infestations						
Headlice	4	2	110	89	0.79	0.64–0.98
Scabies	2	2	25	18	0.71	0.56–0.89
Threadworms	7	5	93	75	0.80	0.72–0.88
Other						
Behavioural problems	31	45	83	88	0.96	0.84–1.09
Nappy Rash	251	192	349	282	0.72	0.65–0.80
Wheeze	91	57	195	143	0.72	0.62–0.82

* This is the Odds ratio of DIN vs. GPRD adjusted for practice.
** Estimates excludes one practice in DIN that categorised 851 of its 874 (97%) children as having a viral illness.
† This includes any URTI, influenza, acute bronchitis, bronchiolitis or laryngitis plus 'H1..' code in DIN ('Acute respiratory infection').

Table 4 **Five-year cumulative incidence rates (per 1,000) for 30 common childhood conditions in DIN and GPRD by sex**

Condition	DIN Cumulative incidence at 5 years (rate per 1000)			GPRD Cumulative incidence at 5 years (rate per 1000)		
	Boys	Girls	P-value*	Boys	Girls	P-value*
Respiratory Tract Infections						
Any RTI†	844	831	0.03	892	880	<0.001
URTI (incl. common cold)	510	506	0.71	820	818	0.91
Influenza	75	72	0.34	40	39	0.03
Acute bronchitis/Chest infection	292	272	0.001	414	368	<0.001
Bronchiolitis	94	72	<0.001	84	59	<0.001
Laryngitis and croup	151	113	<0.001	157	114	<0.001
Ear Infections						
Acute otitis media	661	637	0.002	662	644	<0.001
Chronic OM and glue ear	76	56	<0.001	92	74	<0.001
Otitis externa	32	30	0.31	42	38	<0.001
Mainly Atopic Conditions						
Asthma	174	130	<0.001	195	138	<0.001
Eczema and dermatitis	487	464	0.001	456	440	<0.001
Hayfever/allergic rhinitis	52	38	<0.001	50	35	<0.001
Viral Infections						
Chickenpox	216	211	0.51	248	255	0.03
Measles	22	22	0.65	22	23	0.52
Rubella	26	26	0.97	21	22	<0.001
Hand, Foot and Mouth	47	39	0.008	43	34	<0.001
Viral illness	69**	66**	0.54	301	297	0.51
Other Viral rash	55	53	0.43	65	66	0.31
Other Infections						
Diarrhoea and vomiting	552	527	<0.001	583	544	<0.001
Sore throat	379	352	<0.001	411	383	<0.001
Urinary tract infections	53	126	<0.001	53	125	<0.001
Impetigo	118	110	0.07	103	93	<0.001
Conjunctivitis	602	563	<0.001	599	572	<0.001
Ringworm	43	45	0.70	23	20	0.02
Candidiasis	203	264	<0.001	228	273	<0.001
Infestations						
Headlice	72	107	<0.001	92	130	<0.001
Scabies	16	19	0.11	24	26	0.05
Threadworms	66	84	<0.001	83	102	<0.001
Other						
Behavioural problems	97	79	<0.001	95	70	<0.001
Nappy Rash	235	330	<0.001	303	397	<0.001
Wheeze	164	121	<0.001	229	159	<0.001

* P-values for sex differences are adjusted for practice and year of birth.
† This includes any URTI, influenza, acute bronchitis, bronchiolitis or laryngitis plus 'H1..' code in DIN ('Acute respiratory infection').
** Estimates excludes one practice in DIN that categorised 851 of its 874 (97%) children as having a viral illness.

Practice variation within databases

There were marked variations between practices for all conditions beyond what would be expected by chance. Again it was noticeable that the variation was greater where the diagnosis was less precise, as in the case of 'URTI/Common Cold', again reflecting the misleading degree of diagnostic precision implied by the categories derived from MSGP4. Conversely, a clear diagnosis such as 'Laryngitis/Croup' showed less variation. Where the diagnosis is less precise, other codes may be used to describe it.

While some of the differences may be due to variation in disease presentation, much is likely to be due to variation in recording. While the move to paperless practices is likely to reduce this variation, it seems unlikely that it will disappear. While trends over time can be studied within practices, discussions over the absolute levels of particular conditions need to consider the possibility that the practices being studied are atypical in their recording. Clearly studies based on small numbers of selected practices need to be treated with caution, given the marked variations seen. However, even large primary care databases are not immune to selection bias (possibly by selecting high quality practices). Thus while the overall comparability of DIN and GPRD is reassuring, it is necessary to consider whether their results are typical of all practices. Few data are available on this issue, but a recent report suggests that research practices differ little in patient outcome from other practices.[16]

Further developing the databases

Large-scale databases using routinely collected general practice data offer an effective way of conducting national morbidity surveys of problems presenting in general practice with several major advantages over the national morbidity surveys in general practice. The very large size of the databases and the representative geographical spread of the practices contributing data mean that findings may be generalised to the whole population with confidence. The fact that the data are routinely collected guards against selection bias and opens up the possibility of looking at secular trends in the conditions of interest. Costs of conducting surveys in this way are very much less than with custom designed methods. The older surveys do possess the advantage of being able to collect data, such as social class, that are not routinely collected in general practice databases. In DIN this problem has been remedied by linking externally collected data, such as those derived from the National Census, at an individual postcode level, to patient records. The linkage is carried out on the practice computer system to maintain patient anonymity.[17] The major disadvantage of the primary care databases, also applied to the decennial national morbidity surveys; they do not inform about problems not presenting in general practice.

Key findings

- Large-scale databases using routinely collected general practice data offer an effective way of conducting national morbidity surveys of problems presenting in general practice with several advantages over previous methodologies.
- The extent to which findings may depend on the coding systems used and the role of practice variation have not previously been studied.
- To investigate these issues we compared the cumulative incidence, to age 5, of 30 common childhood conditions in two electronic primary care databases using different underlying computer software systems.
- Although the results from GPRD and DIN were generally very similar, some differences exist that may reflect the differences between the databases and the underlying computing and coding systems used.
- As it cannot reasonably be argued that one database is right and the other database is wrong the comparison highlights areas of uncertainty.
- Our findings emphasise that any statistical analyses using GP databases need to allow for the strong clustering effect of practice.

The study also highlights the influence that coding systems can have both on the design of surveys and on the collection of routine data. Differences between GPRD and DIN can be in part accounted for by differences in the underlying coding systems (OXMIS and Read 4 byte respectively). The hierarchical structure of Read 4 clearly poses problems when it results in the use of non-specific codes and in many ways OXMIS with its rich variety of specific codes specifically developed for general practice is a better model. It would be especially valuable to carry out a comparison of the impact of transition from OXMIS codes to Read 5 Byte codes within GPRD in order to establish the effect on prevalence rates. The development of new terminologies, such as SNOMED CT, that move away from the inflexible hierarchies of older schemes, may help to decrease recording variation between clinicians.

Conclusions

We have shown that two large-scale databases routinely collecting general practice data, but based on fundamentally different systems, produce comparable results. Where discrepancies arise, they generally reflect differences in the underlying computing and coding systems used, and are helpful in revealing possible uncertainties in the data. If large primary care databases are to replace national morbidity surveys, ongoing validation of changes in coding systems and in the structure of the databases will be crucial if time trends are to be monitored, while designing systems to make it easier to standardise coding between practices is clearly a major challenge.

Funding

This work was funded by The Wellcome Trust (Grant No. 065177/Z/01/Z)

Correspondence to: Professor Derek Cook
Department of Community Health Sciences
St George's Hospital Medical School
Cranmer Terrace
London SW17 0RE

Email: d.cook@sghms.ac.uk

References

1. Saxena S, Majeed A and Jones M (1999) Socioeconomic differences in childhood consultation rates in general practice in England and Wales: prospective cohort study. *BMJ* **318**(7184), 642–646.
2. Bruijnzeels M A, Foets M, van der Wouden J C, van den Heuvel W J and Prins A (1998) Everyday symptoms in childhood: occurrence and general practitioner consultation rates. *Br J Gen Pract.* **48**(426), 880–884.
3. McCormick A, Fleming D and Charlton J (1995) *Morbidity statistics from general practice. Forth national study 1991-92*, HMSO: London.
4. Lawson D H, Sherman V and Hollowell J (1998) The General Practice Research Database. Scientific and Ethical Advisory Group. *QJM* **91**(6), 445–452.
5. Hansell A, Hollowell J, Nichols T, McNiece R and Strachan D (1999) Use of the General Practice Research Database (GPRD) for respiratory epidemiology: a comparison with the 4th Morbidity Survey in General Practice (MSGP4). *Thorax* **54**(5), 413–419.
6. Majeed A (2004) Sources, uses, strengths and limitations of data collected in primary care in England. *Health Statistics Quarterly* **21**, 5–14.
7. Lawrenson R, Williams T and Farmer R (1999) Clinical information for research; the use of general practice databases. *J Public Health Med* **21(3)**, 299–304.
8. Scobie S, Basnett I and McCartney P (1995) Can general practice data be used for needs assessment and health care planning in an inner-London district? *J Public Health Med* **17(4)**, 475–483.
9. Carey I M, Cook D G, DeWilde S, Bremner S A, Richards N, Caine S *et al* (2004) Developing the Doctors' Independent Network database for research. *International Journal of Medical Informatics* **73**, issue 5, 443–453.
10. Garcia Rodriguez L A and Perez G S (1998) Use of the UK General Practice Research Database for pharmacoepidemiology. *Br J Clin Pharmacol* **45(5)**, 419–425.
11. Weed L W (1968) Medical Records that guide and teach. *N Engl J Med* **278**, 593–600.
12. Bremner S A, Carey I M, DeWilde S, Richards N, Maier WC, Hilton S R et al (2003) Early-life exposure to antibacterials and the subsequent development of hayfever in childhood in the UK: case-control studies using the General Practice Research Database and the Doctors' Independent Network. *Clin Exp Allergy* **33(11)**, 1518–1525.
13. DeWilde S, Carey I M, Bremner S A, Richards N, Hilton S R and Cook D G (2003) Evolution of statin prescribing 1994-2001: a case of agism but not of sexism? *Heart* 89(4), 417–421.
14. Hollowell J (1997) The General Practice Research Database: quality of morbidity data. *Population Trends* **87**, 36–40.
15. Medicines & Healthcare products Regulatory Agency (MHRA). The General Practice Research Database (GPRD). http://www.gprd.com. 2003.
16. Hammersley V, Hippisley-Cox J, Wilson A and Pringle M (2002) A comparison of research general practices and their patients with other practices–a cross-sectional survey in Trent. *Br J Gen Pract* **52(479)**, 463–468.
17. Carey I M, Cook D G, DeWilde S, Bremner S A, Richards N, Caine S et al (2003) 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 Family Practice* **4(14)**.