

3 Data Quality

Summary

The following sections look in detail at the quality of data from the following sources:

Mortality data

Hospital Episode statistics (HES)

General Practice Research Database (GPRD)

Health Survey for England 1995 (HSE95)

and the coding used for urban-rural classification

Mortality data

Quality checks on source data

Overall, the quality of mortality statistics was good. All sex and age fields appeared valid and less than 1% of records had missing or invalid place of residence. This is documented in the extraction flow chart in the data sources section (Figure 2.1 in Section 2).

Cystic fibrosis deaths in older people: It is likely that deaths from cystic fibrosis in people older than 50 are miscoded. However, only 13 (2%) of 548 deaths were in people aged over 50 (six men, seven women) and nine were aged over 60. The deaths were not concentrated in any one year.

Changes in collection and coding of data by ONS from 1993 onwards

(i) Rule 3

The period covered by the study, 1991-1995 covers a change in the interpretation of rule 3 by OPCS (now ONS), which is used in the coding of death certificates.[1] Prior to 1984 and from 1993 onwards, when automated coding was introduced, conditions listed in part 1 of the death certificate were used as the underlying cause of death. However, from 1984 to 1992, a major condition in part II of the death certificate was selected as the cause of death in preference if part I (immediate cause of death) was one of 11 conditions considered to be terminal (including pneumonia, pulmonary embolism, venous thrombosis and embolism, cardiac or hepatic failure and cardiac arrest).

This interpretation of rule 3 had a major impact on pneumonia leading to a halving of the number of deaths attributed to pneumonia between 1984 and 1992[2,3] and reverted to previous levels in 1993 i.e. there was an abrupt jump in pneumonia deaths between 1992 and 1993 related to coding changes. The changes in rule 3 interpretation might be expected to result in slight decreases in deaths coded to other respiratory conditions.

(ii) Other changes in data collection and coding in 1993

Medical enquiries for further information on the cause of death were discontinued and this was thought at least partly responsible for the observed increase in deaths related to symptoms and signs [2].

Another change was the use of automated coding which interpreted phrases such as 'infectious exacerbation of chronic bronchitis' as 'infection (unspecified)'. [2] Steps were taken to prevent this from happening from the 1995 data year.

Both of these changes might be expected to result in slight decreases in deaths coded to the respiratory conditions in the study from 1993 onwards.

Hospital Episode Statistics

Episodes, admissions and readmissions

Hospital Episode Statistics (HES) data, as the name implies, relate to 'Hospital Episodes'. A 'Finished Consultant Episode' (FCE) is defined as 'the period of care under a consultant'[4] and is an indirect measure of admissions - if the patient is transferred to another consultant during the same admission, this will be recorded as another consultant episode. Overall, around 5% of patients are transferred from one Consultant to another during a hospital admission,[4] but this may vary by condition and by District Health Authority. For example, a study of hospital admission for hip fracture in Wessex Region in the financial year 1994/5[5] and found the overall number of FCEs overestimated the number of admissions by 17%, but that this varied from 1% to 56% between District Health Authorities. To do this the authors constructed a unique patient identifier (using a combination of postcode, date of birth and sex).

It is possible to approximate the number of admissions by selecting only the first episode (episode number =1) of consultant care in a spell in hospital – the method used in this study. In the study mentioned above,[5] this resulted in a 2% overestimate in the number of admissions, largely because of readmissions for the same condition. These cannot be identified in HES without using a unique patient identifier. Postcoded data were not available for this study, so it was not possible to construct a patient identifier to estimate the numbers of readmissions.

Incomplete ascertainment – adjustments in published volumes of HES

HES data may be incomplete because (i) a full HES record may not be received for every episode of treatment (for example, those requiring regular, routine visits) and (ii) a proportion of diagnostic codes are missing.[6] To allow for this when making comparisons by region or by diagnosis, two grossing factors have been developed and are used to adjust data in published volumes of HES data[6] to help with geographical comparisons:

(i) Coverage grossing factor

The number of FCEs in HES is compared with the contracting return called KP70. This return provides a simple count of the number of FCEs by speciality of admitting consultant (as assigned by the Trust concerned). For the purposes of grossing KP70 returns are assumed to be complete. The numbers of FCEs are compared with the numbers of KP70 returns for each combination of speciality (for which 19 categories are used), region and type of admission (considered to be either day case or ordinary) and the FCEs adjusted to equal the KP70 returns. Exceptions are where the ratio between HES and KP70 is less than 50% or greater than 200%, when the data are considered to be incompatible and no adjustments are made.

Initially KP70 contracting returns provided a good check on completeness of HES, but their use has changed over time with the development of contracting and in recent years they may not fully reflect clinical activity (source: discussion with Department of Health). Their use is likely to differ between different Trusts and different geographical areas.

(ii) Grossing factor for missing diagnostic codes

The second type of grossing factor is calculated as the inverse of the percentage of missing diagnostic codes (coded as 799.9 in ICD9) in each combination of speciality (19 categories), region and type of admission (two types).

Patients with respiratory diseases might be admitted under consultants whose speciality is listed as “general medicine”, “other medicine” (which includes thoracic medicine and infectious disease), “paediatrics” or “geriatric medicine” depending on the age of the patient and the speciality assigned to the consultant by the Trust. Assignment of speciality of consultant varies between providers (personal communication, Department of Health); for example, a general physician with an interest in chest medicine might be coded as general medicine or thoracic medicine

If missing codes are concentrated in certain specific areas, it may not be valid to apply these grossing factors directly to a specific diagnosis. For example, the quality report for Northern RHA for 1992/3 stated that one of the most common invalid diagnoses was the inclusion of the neonate nursing codes V29 in the primary diagnosis field. It

follows that it would not be valid to adjust the numbers of emergency admissions for asthma for children aged 5-14 by multiplying by the inverse of the percentage of

missing diagnoses for the speciality of paediatrics in Northern region in 1992/3. Also, the percentage of missing codes is calculated for all admissions rather than for emergency admissions separately which may be more useful.

The percentage of records with both missing primary diagnosis codes and missing speciality is not specified. These cannot be included in the grossing factors, no information is given on this in data quality reports and it must be hoped that the number of such records is small.

The Department of Health advised us that it is usually more appropriate for epidemiological analyses to be performed using a 100% sample of raw HES data, commenting on the completeness rather than using the data in published volumes. The latter use a 25% sample with grossing factors applied which may introduce unrecognisable distortions into the data.

Coverage of HES in 1990/1 to 1994/5

Information on coverage (the number of FCEs expressed as a percentage of the KP70 contracting returns) was obtained from published volumes. This was available for all years in the study (1991/2 to 1994/5) by regions, but all England figures were not given for 1994/5. Comparisons were made for day cases (defined as “a patient, admitted electively during the course of a day for care or treatment which can be completed in a few hours, who does not require a hospital bed overnight”[7]) and for ordinary admissions (defined as “an admission where the patient is expected to remain in hospital for at least one night”[7]).

For all England ordinary admissions (Table 3.1), the number of FCEs was extremely close to the number of KP70 contracting returns for all admissions and for those in specialities likely include the majority of respiratory admissions.

For day case admissions, the HES and KP70 totals were similar for all specialities combined. However, in specialities likely to include the majority of respiratory admissions, HES was slightly higher for ‘general medicine’ admissions, lower for ‘other medicine’, higher by 20% or more for three of the years for ‘geriatric medicine’ and variable for ‘paediatrics’. There was no apparent year on year trend.

Table 3.1 HES as a percentage of KP70 for general medicine, other medicine, geriatrics and paediatrics from published volumes – for England, 1990/1 to 1993/4

(i) England - Ordinary admissions

	All specialities	General medicine	Other medicine	Paediatrics	Geriatric medicine
1990/1	101%	102%	99%	99%	102%
1991/2	98%	99%	100%	99%	98%
1992/3	99%	100%	99%	100%	101%
1993/4	100%	100%	100%	102%	101%

(ii) England - Day cases

	All specialities	General medicine	Other medicine	Paediatrics	Geriatric medicine
1990/1	100%	106%	90%	83%	144%
1991/2	94%	112%	78%	113%	76%
1992/3	97%	118%	80%	97%	120%
1993/4	99%	115%	86%	99%	147%

Coverage of HES by region in 1990/1 to 1994/5

In general, more variability between FCEs and KP70 returns was seen in the regional tables (Appendix A2) than in the England tables (Table 3.1). However, as in the England tables, the ordinary admissions showed closer agreement (between FCEs and KP70 returns) than the day cases.

The regions with closest agreement across all five years in both ordinary and day case admissions (all or all except one figure within 10% of 100% coverage measured as number of FCEs divided by number of KP70 returns expressed as a percentage) were South West Thames, South Western and North Western. Mersey had good agreement if 1990/1 day cases were excluded.

For ordinary admissions for all specialities, there was good agreement to within 5% for each of the four years 1990/1 to 1993/4 for all except three regions. Of the three exceptions, exclusion of 1990/1 in North East Thames and exclusion of 1991/2 in West Midlands gave agreement within 3% for the remaining years. The remaining Health Authority was the Special Health Authority where FCEs were 3-12% higher than contracting returns.

For ordinary admissions assigned to specialities likely to include most respiratory admissions, there was greater variability than with all specialities combined. The regions with closest agreement (all or all except one figure within 10% of 100% coverage) across all five years were South West Thames, South Western, Mersey and North Western. In two districts, poorer agreement was confined to one year of data (1994/5 in Trent, 1991/2 in West Midlands). In East Anglia and Wessex poorer agreement was confined to paediatrics, while in the other districts poorer agreement was generally in a mixture of other medicine, paediatrics and geriatric medicine but not in general medicine. Some of the variation may be due to chance because of smaller numbers, but no indication of the numbers on which percentages were based was provided.

Percentage of missing clinical codes for 1990/1 to 1994/5

Information on missing codes at Health Authority and Trust level was obtained from the Department of Health (Statistics Division 2: SD2 HES) for all specialities and for the specialities likely to include most admissions for respiratory disease.

From 1990/1 to 1994/5, there was a decline in the overall percentage of FCEs with missing primary diagnosis codes from 13% to 4% (Table 3.2).

Table 3.2 Percentage of all FCEs coded as missing (ICD9 799.9) in all England for 1990/1 to 1994/5

Year	Total number of FCEs	Number of FCEs coded as 799.9	% of FCEs coded as 799.9	Minimum and maximum RHA for % FCEs coded as 799.9 (excluding SHAs)
1990/1	8,898,744	1,141,526	13.0%	2.6%-26.2%
1991/2	9,091,323	463,098	5.1%	0.8%-13.5%
1992/3	8,584,880	478,876	5.6%	0.9%-17.2%
1993/4	10,097,884	415,825	4.1%	0.9%-8.0%
1994/5	10,313,379	439,768	4.3%	0.6%-14.7%

Source: Department of Health

However, there was marked variation both between regions and within regions. While the average of missing clinical codes was small at national level, greater variation was seen at regional level and within regions, wide variation was seen at Trust and DHA level and between clinical specialities (Tables 3.3 to 3.7).

There was generally close agreement within regions between the percentages of missing primary diagnosis codes for the specialities likely to contain most admissions for respiratory disease and the total percentage of missing codes for all specialities (Figures 3.1 & 3.2).

Table 3.3 Within and between region variation in missing diagnostic codes for all specialities and for specialities likely to include most admissions for respiratory disease in 1990/1

Region	All FCEs in 1990/1				FCEs for 'Respiratory' Specialities in 90/1			
	Number of FCEs	Number of FCEs coded as 799.9	% of FCEs coded as 799.9	Minimum DHA % total coded as 799.9	Maximum DHA % total coded as 799.9	Average % resp speciality coded as 799.9	Minimum DHA % resp speciality coded as 799.9	Maximum % resp speciality coded as 799.9
A Northern	598,402	50,770	8%	1.03%	26.39%	7.9%	0.0%	60.3%
B Yorkshire	720,682	158,468	22%	0.62%	61.65%	22.5%	0.0%	96.6%
C Trent	812,185	50,276	6%	0.18%	12.69%	6.4%	0.0%	35.7%
D East Anglia	350,901	64,320	18%	1.36%	41.89%	17.1%	0.0%	52.9%
E NW Thames	502,265	131,617	26%	9.82%	56.50%	27.7%	0.8%	82.5%
F NE Thames	886,692	148,477	17%	3.31%	43.00%	17.5%	0.8%	65.8%
G SE Thames	648,620	104,274	16%	1.26%	31.75%	19.6%	0.0%	69.2%
H SW Thames	454,737	71,095	16%	5.95%	31.10%	16.2%	0.0%	86.6%
J Wessex	521,528	30,606	6%	0.00%	25.39%	5.8%	0.0%	40.2%
K Oxford	380,581	74,276	20%	2.54%	41.47%	22.4%	0.7%	79.9%
L South West	609,837	16,068	3%	0.01%	19.18%	2.9%	0.0%	67.6%
M Midlands	931,850	115,849	12%	0.42%	26.18%	14.6%	0.0%	66.5%
N Mersey	492,581	60,746	12%	0.59%	20.93%	11.2%	0.0%	48.2%
P North West	879,214	40,920	5%	0.07%	20.05%	4.9%	0.0%	26.3%
T SHAs	108,669	23,764	22%	0.09%	60.61%	20.6%	0.3%	70.5%
All England	8,898,744	1,141,526	13%	0.00%	61.65%	13.6%		
Total A to P	8,790,075	1,117,762	13%	0.00%	61.65%	13.6%		

Respiratory specialities defined as 300 General Medicine, 340 Thoracic Medicine, 350 Infectious Diseases, 410 Rheumatology, 420 Paediatrics, 430 geriatric medicine
Total number of admissions in 'respiratory specialities' = 2,888,968 Percentages based on total of <10 FCEs were ignored DHA = DHA of treatment

Table 3.4 Within and between region variation in missing diagnostic codes for all specialities and for specialities likely to include most admissions for respiratory disease in 1991/2

Region	All FCEs in 1991/2			FCEs for 'Respiratory' Specialities in 91/2				
	Number of FCEs	Number of FCEs coded as 799.9	% of FCEs coded as 799.9	Minimum DHA % coded as 799.9	Maximum DHA % coded as 799.9	Average regional resp specialities coded as 799.9	Minimum DHA % resp specialities coded as 799.9	Maximum % resp speciality coded as 799.9
A Northern	645,721	40,834	6.3%	0.3%	16.4%	6.3%	0.1%	14.9%
B Yorkshire	736,974	98,558	13.4%	0.7%	32.8%	17.6%	0.4%	29.8%
C Trent	850,104	8,580	1.0%	0.0%	2.2%	0.8%	0.0%	1.8%
D East Anglian	385,602	9,083	2.4%	0.2%	7.0%	2.5%	0.3%	7.3%
E NW Thames	560,028	75,443	13.5%	1.1%	40.3%	11.7%	0.9%	40.0%
F NE Thames	746,252	55,622	7.5%	1.6%	19.8%	6.4%	1.7%	16.6%
G SE Thames	677,098	52,861	7.8%	0.1%	20.5%	9.7%	0.0%	24.0%
H SW Thames	499,643	29,844	6.0%	0.3%	22.2%	6.6%	0.0%	26.4%
J Wessex	545,991	19,135	3.5%	0.1%	11.6%	3.9%	0.1%	11.5%
K Oxford	394,172	15,372	3.9%	0.1%	9.3%	5.4%	0.1%	11.6%
L South Western	630,389	5,177	0.8%	0.0%	1.8%	0.8%	0.0%	3.3%
M Midlands	861,728	3,089	0.4%	0.0%	2.1%	0.5%	0.0%	4.1%
N Mersey	516,825	33,454	6.5%	1.1%	12.9%	6.8%	0.5%	13.9%
P North Western	923,091	7,605	0.8%	0.0%	3.3%	0.9%	0.0%	4.8%
T SHAs	117,705	8,442	7.2%	0.0%	22.5%	10.6%	0.1%	23.8%
All England	9,091,323	463,098	5.1%	0.0%	40.3%	5.5%	0.0%	29.8%
A to P	8,973,618	454,656	5.1%	0.0%	40.3%	5.5%	0.0%	29.8%

Respiratory specialities defined as 300 General Medicine, 340 Thoracic Medicine, 350 Infectious Diseases, 410 Rheumatology, 420 Paediatrics, 430 geriatric medicine
Total number of admissions in 'respiratory specialities' = 2,945,091 Percentages based on total of <10 FCEs were ignored DHA = DHA of treatment

Table 3.5 Within and between region variation in missing diagnostic codes for all specialities and for specialities likely to include most admissions for respiratory disease in 1992/3

Region	All FCEs in 1992/3			FCEs for 'Respiratory' Specialities in 92/3				
	Number of FCEs	Number of FCEs coded as 799.9	% of FCEs coded as 799.9	Minimum DHA % coded as 799.9	Maximum DHA % coded as 799.9	Average regional resp specialities coded as 799.9	Minimum DHA % resp specialities coded as 799.9	Maximum % resp speciality coded as 799.9
A Northern	678,046	39,289	5.79%	0.2%	11.4%	7.1%	0.0%	36.9%
B Yorkshire	787,041	134,974	17.15%	0.2%	87.0%	19.5%	0.0%	99.3%
C Trent	887,978	15,082	1.70%	0.1%	4.3%	1.3%	0.0%	23.5%
D East Anglian	400,975	12,025	3.00%	0.6%	7.2%	3.4%	0.0%	14.2%
E NW Thames	580,770	77,999	13.43%	1.6%	45.6%	11.7%	0.5%	86.4%
F NE Thames	752,753	50,833	6.75%	1.9%	19.2%	6.4%	0.0%	28.4%
G SE Thames	710,931	61,401	8.64%	0.0%	24.2%	9.6%	0.0%	61.4%
H SW Thames	519,450	20,387	3.92%	0.4%	17.0%	5.4%	0.0%	57.9%
J Wessex	574,558	10,236	1.78%	0.0%	3.6%	1.6%	0.0%	6.4%
K Oxford	428,514	12,829	2.99%	0.1%	5.6%	5.8%	0.0%	46.5%
L South Western	670,967	4,674	0.70%	0.0%	2.0%	0.8%	0.0%	7.8%
M West Midlands	1,029,628	52,025	5.05%	0.2%	5.1%	4.3%	0.1%	51.2%
N Mersey	535,252	20,837	3.89%	0.6%	11.4%	5.2%	0.0%	31.1%
P North Western	932,729	8,502	0.91%	0.0%	2.0%	0.7%	0.0%	10.6%
T SHAs	124,829	9,807	7.86%	0.1%	42.8%	10.1%	0.0%	22.6%
DHA Not known	87	1	1.15%			0 (n=5)		
All England	8,584,880	478,876	5.58%	0.0%	87.0%	5.8%	0.0%	99.3%
A to P	8,459,964	469,068	5.54%	0.0%	87.0%	5.8%	0.0%	99.3%

Respiratory specialities defined as 300 General Medicine, 340 Thoracic Medicine, 420 Paediatrics, 430 geriatric medicine

Total number of admissions under 'respiratory specialities' = 2,999,058

Percentages based on total of <10 FCEs were ignored

DHA = DHA of treatment

Table 3.6 Within and between region variation in missing diagnostic codes for all specialities and for specialities likely to include most admissions for respiratory disease in 1993/4

Region	All FCEs in 1993/4			FCEs for 'Respiratory' Specialities in 93/4				
	Number of FCEs	Number of FCEs coded as 799.9	% of FCEs coded as 799.9	Minimum DHA % coded as 799.9	Maximum DHA % coded as 799.9	Average regional resp specialities coded as 799.9	Minimum DHA % resp specialities coded as 799.9	Maximum % resp speciality coded as 799.9
A Northern	711,691	35,203	4.95%	1.3%	14.4%	6.1%	0.0%	58.7%
B Yorkshire	810,165	65,014	8.02%	1.8%	12.4%	8.2%	0.2%	28.0%
C Trent	948,694	26,008	2.74%	0.1%	7.0%	2.8%	0.0%	9.6%
D East Anglian	421,978	7,927	1.88%	0.2%	3.5%	2.4%	0.1%	16.3%
E NW Thames	633,404	35,026	5.53%	1.3%	12.5%	5.6%	0.2%	41.3%
F NE Thames	753,266	46,122	6.12%	2.1%	14.3%	6.1%	1.0%	23.7%
G SE Thames	729,024	41,832	5.74%	0.0%	21.0%	6.9%	0.0%	60.4%
H SW Thames	543,748	18,143	3.34%	0.2%	13.1%	3.6%	0.0%	38.0%
J Wessex	623,306	23,579	3.78%	0.0%	15.5%	4.2%	0.0%	25.8%
K Oxford	456,860	15,001	3.28%	0.5%	6.7%	7.2%	0.0%	48.6%
L South Western	711,514	11,798	1.66%	0.0%	4.3%	2.1%	0.0%	9.2%
M West Midlands	1,072,471	57,212	5.33%	0.1%	22.8%	5.0%	0.1%	38.5%
N Mersey	571,723	18,507	3.24%	0.3%	11.3%	4.0%	0.0%	31.7%
P North Western	976,249	8,357	0.86%	0.0%	1.3%	0.7%	0.0%	12.4%
T SHAs	133,791	6,096	4.56%	0.0%	37.8%	4.2%	0.0%	20.0%
All England	10,097,884	415,825	4.12%	0.0%	37.8%	4.6%	0.0%	60.4%
A to P	9,964,093	409,729	4.11%	0.0%	22.8%	4.6%	0.0%	60.4%

Respiratory specialities defined as 300 General Medicine, 340 Thoracic Medicine, 420 Paediatrics, 430 geriatric medicine

Total number of admissions under 'respiratory specialities' = 3,195,312

Percentages based on total of <10 FCEs were ignored

DHA = DHA of treatment

Table 3.7 Within and between region variation in missing diagnostic codes for all specialities and for specialities likely to include most admissions for respiratory disease in 1994/5

Region	All FCEs in 1994/5			FCEs for 'Respiratory' Specialities in 94/5				
	Number of FCEs	Number of FCEs coded as 799.9	% of FCEs coded as 799.9	Minimum DHA % coded as 799.9	Maximum DHA % coded as 799.9	Average regional combined resp specialities coded as 799.9	Minimum DHA % combined resp specialities coded as 799.9	Maximum % resp speciality coded as 799.9
A Northern	697,014	37,030	5.31%	1.4%	11.7%	4.3%	0.1%	14.8%
B Yorkshire	835,091	62,207	7.45%	0.3%	18.1%	7.7%	0.0%	26.2%
C Trent	823,999	17,398	2.11%	0.1%	4.9%	2.3%	0.0%	9.8%
D East Anglian	526,665	77,579	14.73%	0.1%	79.1%	17.4%	0.0%	99.0%
E NW Thames	640,752	16,337	2.55%	1.2%	6.0%	2.6%	0.2%	12.4%
F NE Thames	788,523	57,672	7.31%	1.6%	16.8%	5.8%	0.0%	38.2%
G SE Thames	756,113	42,064	5.56%	0.8%	27.0%	5.6%	0.0%	59.4%
H SW Thames	574,417	15,777	2.75%	1.2%	5.4%	4.0%	0.0%	33.8%
J Wessex	667,285	31,592	4.73%	0.7%	11.2%	5.2%	0.0%	23.7%
K Oxford	438,266	11,290	2.58%	0.1%	6.4%	6.3%	0.0%	45.5%
L South Western	739,326	4,590	0.62%	0.0%	1.3%	0.6%	0.0%	3.1%
M West Midlands	1,145,333	29,892	2.61%	0.2%	9.2%	2.7%	0.0%	14.7%
N Mersey	598,486	11,904	1.99%	1.1%	3.0%	2.2%	0.1%	50.7%
P North Western	1,082,109	24,436	2.26%	0.1%	10.1%	2.1%	0.0%	11.8%
All England (A to P)	10,313,379	439,768	4.26%	0.0%	79.1%	4.6%	0.0%	99.0%

Respiratory specialities defined as 300 General Medicine, 340 Thoracic Medicine, 420 Paediatrics, 430 geriatric medicine

Total number of admissions under 'respiratory specialities' = 3,176,277

Percentages based on total of <10 FCEs were ignored

DHA = DHA of treatment

SHAs ceased to exist from April 1994

Figure 3.1 Comparison of missing codes as percentage of total for all specialities and for 'respiratory' specialities in HES data for 1990/1 to 1993/4 by region (from Tables 3.3 to 3.6)

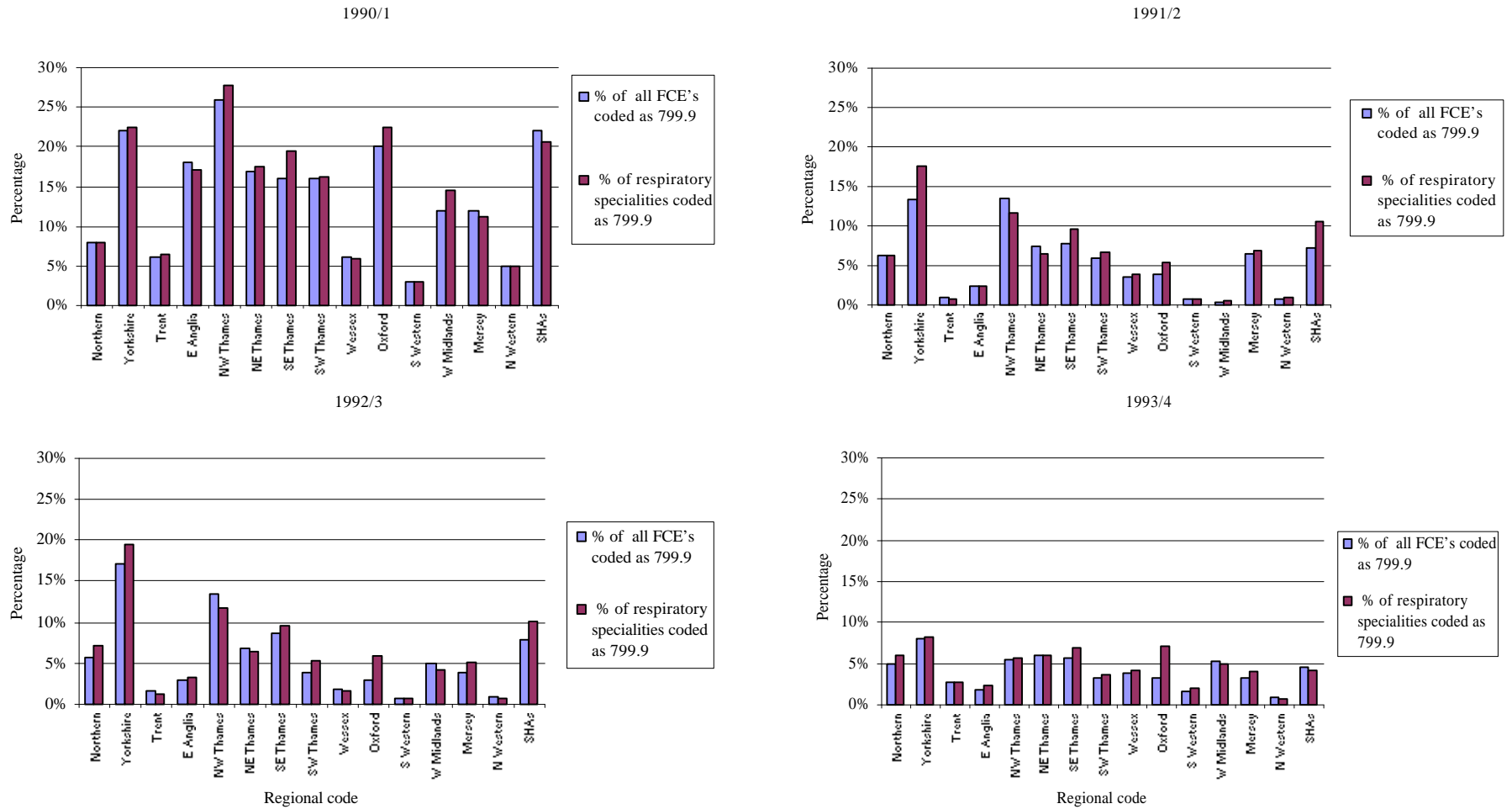
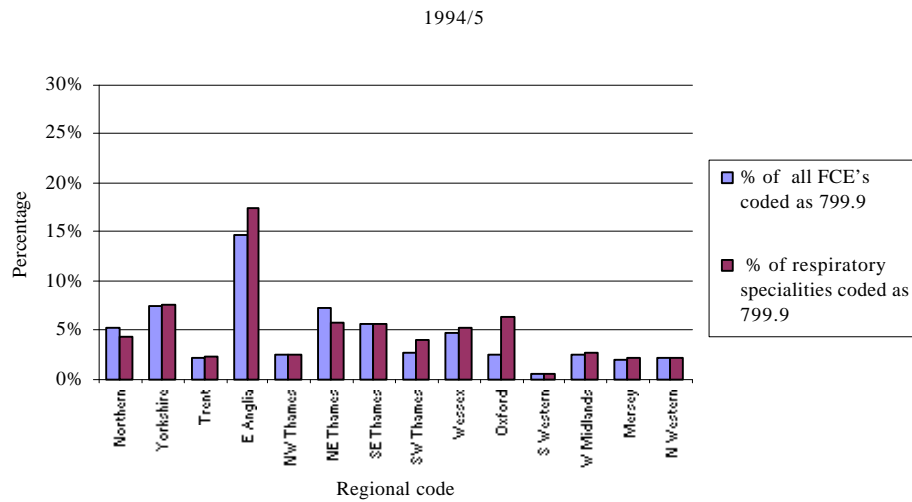


Figure 3.2 Comparison of missing codes as percentage of total for all specialities and for 'respiratory' specialities in HES data for 1994/5 by region (from Table 3.7)



Accuracy of coding of clinical diagnosis

It was not possible to make an assessment of the quality of coded primary diagnoses for this project. Anecdotally, from 1995/6 onwards (i.e. beyond the scope of this study) the training provided to clinical coding clerks in Trusts (responsible for the coding of admission, transfer and discharge details supplied by clinical staff) for the change from ICD9 to ICD10 is reported to have resulted in an improvement in the quality of coding and the reduction in the number of missing codes for data.

One study examining this issue[8] considered the quality of coding at two hospitals in North West Thames region for patients discharged between April 1991 and January 1993. Discharges for asthma, diabetes, appendicitis, fractured femur and a general sample of diagnosis were coded by both internal and external coders for each hospital and the results were compared. Agreement was good (86% and 91% in the two hospitals) for asthma and specific diagnoses, but poor for the general sample. A second study[9] looked at the diagnosis of acute stroke in routine hospital activity data from Oxford city hospitals with an episode start date between August 1994 and January 1995 compared with information from a stroke register and found 88% of acute strokes had been correctly coded.

However, clinical coding clerks can only use the information provided and no information is routinely provided on the accuracy of diagnoses made. In respiratory disease, a common area of confusion is the diagnosis of asthma, as illustrated by studies assessing the accuracy of death certificates.[10-12]

It is unlikely that diagnostic fashions have changed dramatically over the course of the study, but hospital admissions have been rising between 1958 and the late 1980s which has been partly attributed to a general increase in willingness to diagnose asthma combined with changes in instructions to coders.[12-14] It is possible that these factors also vary geographically.

Quality of coding of sex

Checks on the quality of coding for sex were run on the original HES dataset (respiratory disease and ischaemic heart disease FCEs) prior to selection for respiratory disease or first episode (Table 10). Males consistently formed a higher percentage of FCEs (57% vs 43%). The percentage of FCEs by year with unknown sex was less than 0.5%, but most of these FCEs were concentrated in two districts:

- The majority of the ‘3’ codes (indeterminate sex or sex change) were in Oxford region (2,381: 67% out of the total of 3,541) and concentrated in Northamptonshire District Health Authority which came into being in April 1994 (2,112: 37% of the FCEs for 1994/5 for this DHA).
- The majority of the ‘&’ codes (sex missing – first used in 1994/5) were in Trent region (2,131: 97% out of the total of 2,204) and concentrated in Sheffield District Health Authority (1,797: 3% of the FCEs for 1990/1 to 1994/5 for this DHA).

Theoretically, the problems with coding for sex in these two health authorities have the potential to affect rates of some of the rarer respiratory conditions. For example, if the missing codes are concentrated in ethnic minorities, whose gender may be more difficult to determine from their name for English coders, the rates of tuberculosis could be artefactually lowered.

Table 3.8 Coding of sex in HES dataset pre-selection (all completed FCEs for respiratory disease in England, all cardiovascular disease in Thames regions and all ischaemic heart disease in England excluding Thames regions) by year

Coding for sex	1990/1	1991/2	1992/3	1993/4	1994/5
1 (Male)	414,029 (57.4%)	474,295 (57.3%)	487,421 (57.7%)	533,677 (57.0%)	579,816 (55.8%)
2 (Female)	306,556 (42.5%)	353,294 (42.7%)	357,300 (42.3%)	402,625 (43.0%)	455,222 (43.8%)
Blank, 0, 3 or &	139 (0.0%)	520 (0.1%)	299 (0.0%)	232 (0.0%)	4,580 (0.4%)
Total	720,724 (100%)	828,109 (100%)	845,020 (100%)	936,534 (100%)	1,039,618 (100%)
Unknown codes					
Blank	0	0	4	0	0
0	21	0	0	0	0
3	118	520	295	232	2,376
&	0	0	0	0	2,204

GPRD data

Geographical coverage

This project focussed on the years 1991-1995, which had the largest number of participating GPs in the GPRD. In 1994, the GPRD covered 5.6% of the population of England & Wales.[15] More than 5% of the population were covered in all RHAs except Yorkshire, but there was lower coverage of 3% of inner London.[16] GPs in GPRD are more representative than another major source of general practice data, the Morbidity Surveys in General Practice (MSGP) both in terms of geographical spread and in percentage of single-handed practice: 17% of GPRD practices were single-handed compared with 12% in the Fourth Morbidity Survey in General Practice (MSGP4) and 31% nationally (unpublished data).

Quality of recording

The GPRD was designed to record all prescriptions issued, the indication for all new prescriptions and all "significant" events such as consultations resulting in a referral and "events which the Partner will require to be reminded of at a later date"[17] (for example, diagnoses such as cystic fibrosis and tuberculosis and information from hospital letters and coroners' reports). There is no requirement to enter diagnoses for minor consultations or to record follow-up consultations for chronic conditions unless the consultation leads to new therapy or to a referral. Paperless GPRD practices are likely to record more consultations than those who enter paper records into the database.

Not all practices using the VAMP Medical practice software are included on the database - for inclusion practices must satisfy standard validation checks to ensure good capture of data.[18] Recording of consultations resulting in a prescription is good: a study looking at first time use of non-steroidal anti-inflammatory drugs suggested that the indication for first-time prescribing was recorded in 96% of cases.[18] The extent to which consultations not resulting in a prescription are recorded is less certain. A study in 1990 looking at psychoses in 11 practices using VAMP found that 95% of prescriptions but only 73% of consultations in the written notes were entered on the computer.[19]

There is no requirement to record a definitive diagnosis: for example, the indication for an inhaler prescription may be entered as wheezing or cough. This means that information recorded may closely reflect clinical practice. However, it is possible that any prescription given will influence the diagnosis recorded in GPRD: for example, if a person with COAD consults with a 'chest infection' and requires antibiotics they may be recorded as 'chest infection', but if they require a prescription of inhalers as well the most appropriate indication would be 'chronic bronchitis'.

Validity of the GPRD

The GPRD was compared with the Fourth Morbidity Survey in General Practice (MSGP4) conducted in 1991/2, a widely referred to source of information on consultations in primary care, for eleven respiratory conditions[20]: asthma; hayfever or allergic rhinitis (referred to as "hayfever"); chronic bronchitis, emphysema, or obstructive airways disease excluding asthma ("COPD"); tuberculosis; pneumonia; acute bronchitis or bronchiolitis; chest infection or bronchitis not otherwise specified ("chest infection"); cystic fibrosis; sarcoidosis; fibrosing alveolitis; and pneumothorax.

MSGP4 patient consultation rates came from published data[21] supplemented by special analyses to separate pneumonia from influenza and to combine codes relating to chronic obstructive airways disease.

Despite different recording and coding requirements, GPRD rates of persons with a diagnosis or prescription plus diagnosis closely approximated rates of persons consulting for that illness in the MSGP4 for most respiratory diseases except for acute bronchitis and bronchiolitis. Here, the combination category within the GPRD of {'acute bronchitis or bronchiolitis' or 'chest infection or bronchitis not otherwise specified'} showed a similar pattern and order of magnitude to MSGP4 consultation rates for 'acute bronchitis or bronchiolitis' and was used as the identifier for this condition in GPRD comparisons with the data from other sources.

We concluded[20] that GPRD appeared valid for epidemiological studies in primary care by comparison with the MSGP4 and that it offered advantages in terms of large size, a longer time period covered and ability to link prescriptions with diagnoses. However, the GPRD is a complex database requiring careful interpretation, for example, not all consultations are recorded and the coding system used contains terms which do not directly map to ICD codes.

Misclassification

On first analysis, 23 women but no men, aged 45-64 were recorded as consulting with cystic fibrosis, which was inconsistent with the known epidemiology of the disease. Further manual review of records showed that a total of 57 women (of all ages) with fibrocystic disease of the breast had been miscoded as having cystic fibrosis. Corrected figures were used for further analyses. The misclassification identified in cystic fibrosis was probably related to the OXMIS coding scheme structure as both "cystic fibrosis" and "fibrosis cystic" are grouped together under a single code.

The diagnosis category "chest infection" in GPRD presented a problem of interpretation as it was non-specific and involved a large number of patients. Misclassification of other diagnoses as 'chest infection' could potentially have caused large biases in those conditions affecting smaller numbers of patients. Some overlap was identified: 49% of those with a diagnosis of 'chronic bronchitis, emphysema or OAD (excluding asthma)' also had a diagnosis of chest infection within the same year - part of this may reflect the natural history of the diseases with a number of chest infections occurring prior to diagnosis - but this overlap group represented only 4.5% of patients with a diagnosis of chest infection.

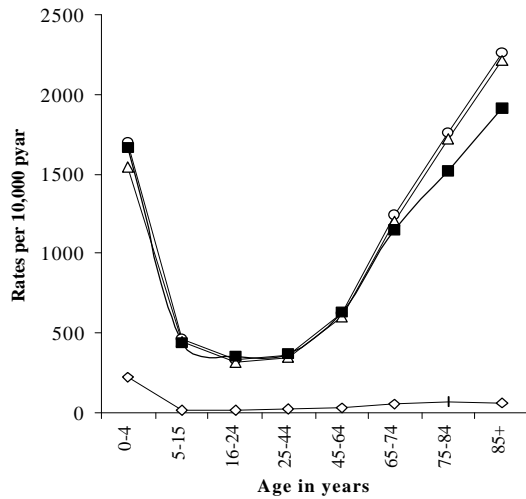
A large degree of misclassification was unlikely since GPRD rates seen for "chest infection" very closely resembled the MSGP4 rates for consultations for 'acute bronchitis or bronchiolitis' particularly in those under 65 (Figure 3.3). In the over 65 age-groups, GPRD rates were slightly higher than MSGP4 rates; some of the discrepancy may be accounted for by the increased rates of diagnoses of 'bronchitis not otherwise specified' (ICD9 490) in MSGP4 in these age-groups. However, there were some similarities with MSGP4 consultation rates for the common cold (ICD9 code 460) in ages 0-15 years.[21]

Using prescription data

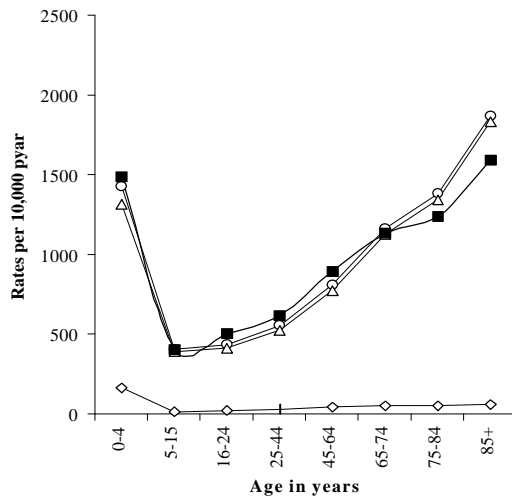
Patient prescription rates alone could not be used as a proxy for disease prevalence rates as they lacked specificity. Inhalers were prescribed for asthma, COPD and symptoms. However, the combination of inhaler plus either an asthma or COPD diagnosis was judged to be a better measure of the burden of these conditions in primary care as it would also pick up patients receiving repeat prescriptions. The 'a priori' definition of hayfever therapy that we used was not a good measure of hayfever prevalence, probably because the category used was too broad, including nose-drops which are prescribed for other reasons than hayfever. TB therapy was not a good measure of TB prevalence probably because of the use of rifampicin for meningitis prophylaxis and as a general antibiotic.

Figure 3.3 A comparison of 'acute bronchitis or bronchiolitis' in the MSGP4 and GPRD

Males



Females



- ◇ GPRD Acute bronchitis or bronchiolitis diagnosis
- GPRD Acute bronchitis, bronchiolitis or 'chest infection' diagnosis
- △ GPRD 'Chest infection' diagnosis
- MSGP4 Consultation for acute bronchitis or bronchiolitis

Health Survey for England 1995 (HSE95)

The descriptive analyses looked for missing values and a number of quality checks were performed as detailed below. Missing values had generally been coded with negative numbers, which explained the reasons that these data had not been collected.

Geographical identifiers

There were no missing values.

Demographic details

Age last birthday ranged from age 2-100 years (mean 39.3, median 38). This variable was complete. A check variable for age was created, calculated by subtracting the year of birth from 1995. Seven people had refused to give their year of birth (dobyea=-9) and four had stated they did not know it (dobyea=-8). The ages of 19,775 of the 19,777 remaining individuals were ± 1 year of the calculated age. One person had an age +2 years from the calculated age, and one had an age 21 years higher than the calculated age.

Smoking

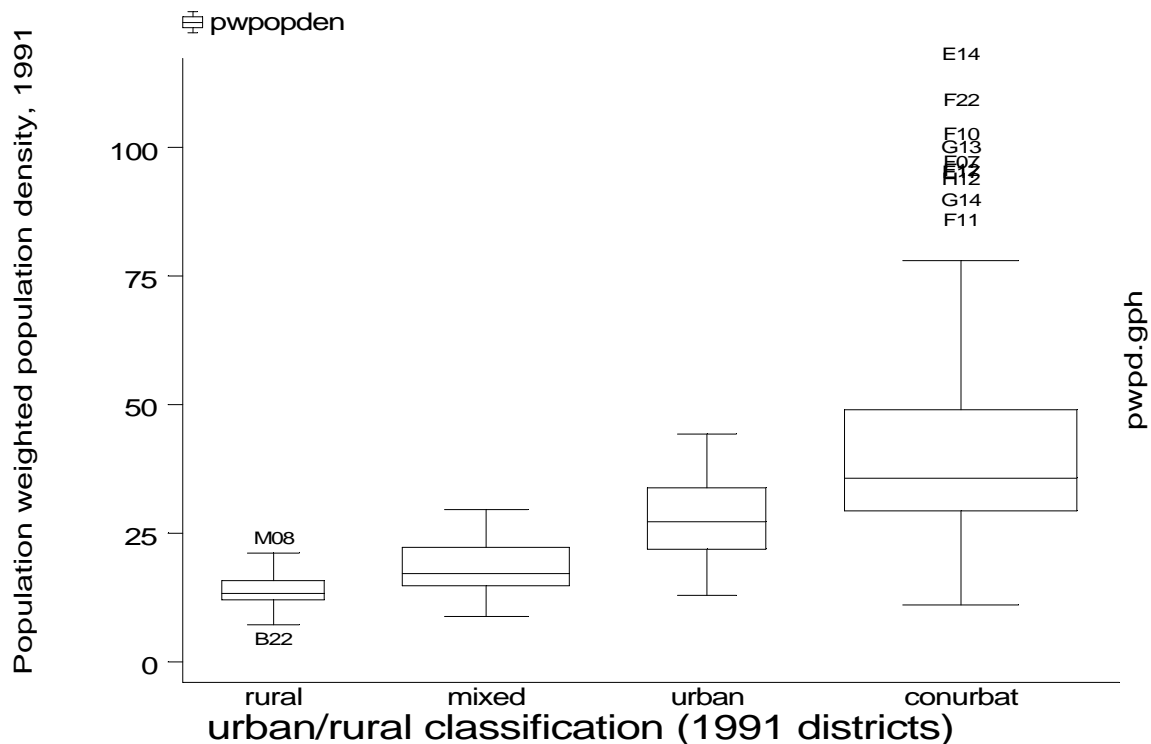
Thirteen variables relating to active smoking were initially selected from the main dataset, but only one of these, the variable *cigsmk2*, was used in the analysis and this variable was complete. A second variable identifying never smokers was derived to inform comments on data on smoking from the GPRD, which is presented in the Discussion section (Section 16). The HSE95 dataset included the variable *smokever* "Have you ever smoked" but this was only complete for ages 20 and over. Information on smoking status of those under 20 was present, but was contained in three different variables (*smokever*, *current* which contained smoking status of 222 16-19 year olds and *ksmokcig* which contained details of smoking status of children aged 8-15). These were combined with data for those aged 20 and over to give a derived binary variable *eversmok* divided into never and ever (including 'social') smokers.

In 1993, 1994 and 1996 questionnaire reporting on smoking status was validated by measuring serum cotinine, a metabolite of nicotine with a half-life of 16-20 hours. Serum levels above 20ng/ml were taken to indicate regular smoking. This validation was not performed for the 1995 survey. In 1994 and 1996, if those with serum cotinine levels above 20ng/ml who say they do not smoke were reclassified as misreporters, then the prevalence of smoking in all ages rose by 2%, but in informants aged 16-24 the prevalence of smoking rose by 5%. It is likely that the same level of misreporting occurred in 1995.

Coding used for urban rural classification

The urban-rural classifications based on district health authority were compared with population weighted population densities (pwpds) as an alternative indicator of the degree of urbanisation for districts in 1991 (Figure 3.5). There was a clear increase in median pwpd for increasing level of urbanisation in the urban-rural classification and very high pwpds were seen for conurbation districts in the London area. However, the range in pwpds increased with increasing urbanisation of the urban/rural indicator and there was some overlap between the higher pwpd districts for each urban/rural classification and the lower pwpd districts for the next more urban classification. Overlap was a particularly problem for districts classed as conurbations in Northern areas of England.

Figure 3.4 Box plot* comparing urban/rural classification with pwpd for district health authorities in 1991



* Stata box plots are described in the manual in the following way:

The line in the middle of the box represents the median or 50th percentile of the data. The box extends from the 25th percentile to the 75th percentile, the so called interquartile range (IQ). The statistical package used defines the lines emerging from the box in the following way:

(i) The line emerging upwards from the box extends to the next data point which is less than or equal to [75th percentile value + (1.5 x IQ)].

(ii) The line emerging downwards from the box extends to the next data point which is greater than or equal to [25th percentile value - (1.5 x IQ)].

Any more extreme values are referred to as outside values and are individually plotted.

References

1. Devis T, Rooney C. Death certification and the epidemiologist. *Health Statistics Quarterly* 1999;01(Spring):25-33.
2. Rooney C, Devis T. Mortality trends by cause of death in England and Wales 1980-94: the impact of introducing automated cause coding and related changes in 1993. *Population Trends* 1996;86(Winter):29-35.
3. The Lung & Asthma Information Agency. Pneumonia mortality in the elderly. *The Lung & Asthma Information Agency Factsheet* 1992;92/2
4. Department of Health Statistics Section SD2 HES. *HES The Book*. London, UK: Department of Health; 1998.
5. McColl A, Roderick P, Cooper C. Hip fracture incidence and mortality in an English Region: a study using routine National Health Service data. *Journal of Public Health Medicine* 1998;20(2):196-205.
6. Government Statistical Service. *Hospital Episode Statistics, England: Financial year 1994-95*. London, UK: Department of Health; 1996.
7. Department of Health. *Hospital Episode Statistics, England: Financial year 1994-95*. London, UK: Department of Health; 1996.
8. Dixon J, Sanderson C, Elliott P, Walls P, Jones J, Petticrew M. Assessment of the reproducibility of clinical coding in routinely collected hospital activity data: a study in two hospitals. *Journal of Public Health Medicine* 1998;20(1):63-9.
9. Mant J, Mant F, Winner S. How good is routine information? Validation of coding for acute stroke in Oxford hospitals. *Health Trends* 1998;29:96-9.
10. Guite HF, Burney PGJ. Accuracy of recording of deaths from asthma in the UK: the false negative rate. *Thorax* 1996;51:924-8.
11. A subcommittee of the BTA research committee. Accuracy of death certificates in bronchial asthma. *Thorax* 1984;39:505-9.
12. Sears MR, Rea HR, de Boger G, Beaglehole R, Gillies AJD, et al. Accuracy of certification of deaths due to asthma. A national study. *American Journal of Epidemiology* 1986;124:1004-11.
13. LAIA (Lung & Asthma Information Agency). Trends in hospital admissions for asthma. *LAIA Factsheet* 1996;96/2
14. Strachan DP. Time trends in asthma and allergy: ten questions, fewer answers. *Clinical & Experimental Allergy* 1995;25(9):791-4.
15. *Key health statistics in General Practice*. London, UK: The Stationery Office; 1996.
16. Hollowell J. *General Practice Research Database (GPRD) scope and quality of data*. London, UK: OPCS; 1994.
17. Mann RD, Hall G, Chukwujindu J. Research implications of computerised primary care. *Post Marketing Surveillance* 1992;(5):259-68.
18. Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom [see comments]. *BMJ* 1991;302(6779):766-8.
19. Nazareth I, King M, Haines A, Rangel L, Myers S. Accuracy of diagnosis of psychosis on general practice computer system. *BMJ* 1993;307(6895):32-4.
20. Hansell A, Hollowell J, Nichols T, McNiece R, Strachan DP. 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):413-9.
21. McCormick A; Fleming DM; Charlton J. *Morbidity Statistics from General Practice. Fourth national study 1991-1992*. London, UK: HMSO; 1995.