

2 Description of Data Sources

Summary

The following sources of data were used:

Mortality statistics

Hospital Episode Statistics (HES)

General Practice Research Database (GPRD)

Fourth Morbidity Survey in General Practice (MSGP4)

Health Survey for England 1995 (HSE95)

Population estimates including population weighted population density

Each of these are discussed in the following sections, outlining how the data was extracted.

Mortality statistics

Hospital admissions

Primary care data – the General Practice Research Database (GPRD)

Health Survey for England, 1995

Population estimates

Mortality statistics

Extraction of data

Mortality data were obtained for all England from the Office for National Statistics for the years 1991-1995 for all deaths from respiratory diseases (ICD9 codes 460-519, to 4 digits) and for cystic fibrosis (ICD9 277.0), sarcoidosis (ICD9 135) and tuberculosis (ICD9 010-012). Variables included age in years, sex, date of death, regional health authority of residence and district health authority of residence. The steps used to extract the data are outlined in Figure 2.1, overleaf.

Overlap years

Mortality data are collected by calendar year using information recorded on death certificates. Figures were collected by year in which the death was registered until 1992 and by year in which the death occurred from 1993. Data for the overlap period for deaths occurring in 1992 but registered in 1993 were obtained and included; the district boundaries used for this overlap data were those existing in April 1993.

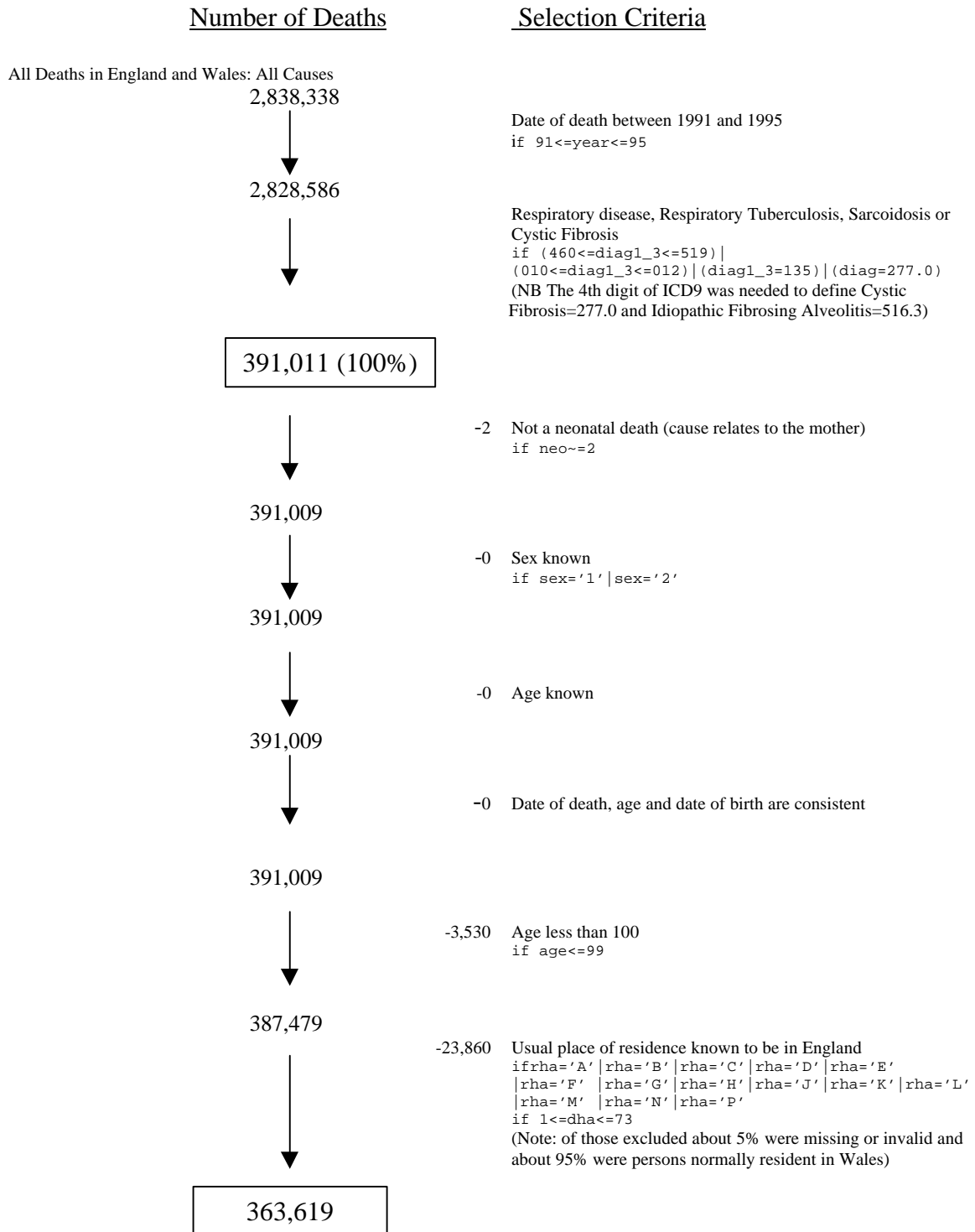
Coding

For the years of interest, deaths were coded using ICD9. This will change to ICD10 in the year 2001 and dual coding using both ICD9 and ICD10 will take place for all deaths occurring in the preceding year.

Geographical identifier

The lowest level of geographical identifier obtained was district health authority (DHA). Any changes to DHA boundaries generally take effect at the beginning of the financial year (April) All DHA boundaries used for mortality statistics relate to boundaries at the end of the calendar year in question. An urban-rural code with four ordered categories (conurbation, urban, mixed and rural) was assigned based on the DHA of residence (see Appendix A4 for details).

Figure 2.1 Extraction of data from mortality files



Hospital Admissions

Hospital Episode Statistics (HES) files for Finished Consultant Episodes (FCEs) for the 10 specified respiratory conditions were obtained from ONS for the financial years 1990/1, 1991/2, 1992/3 and 1993/4 and from Data Sciences UK Limited (under license to the Department of Health) for the years 1994/5, 1995/6 and 1996/7 via the MIDAS computer at Manchester.

Data extraction

Data were extracted as outlined in Figures 2.2 and 2.3. HES data are collected by financial year (April to March), but analysis was performed in calendar years, to allow comparability with other data sources. The years 1995/6 and 1996/7 were not included in the analysis because of the change from ICD9 coding to ICD10 in April 1995: ICD code changes have been recognised to lead to artefactual changes in rates of diseases[1] and dual coding for years preceding a change of codes (allowing the estimation of the size of artefactual changes in occurrence due to the coding change) is not routinely carried out for hospital data. This meant that only 1991-1994, four calendar years of HES data, could be analysed. There were problems obtaining HES data for the year 1990/91 for respiratory tuberculosis, sarcoidosis and cystic fibrosis, so the period January to March 1991 for these conditions was not included in the analysis.

Analysis was performed on admissions rather than on discharges and only the first consultant episode in each admission was extracted (i.e. where the variable episode number equalled “1” or where the episode number equalled “&” and the admission date was the same as the start date for the episode). The number of admissions for ICD9 respiratory disease codes 460-519 by calendar year and by financial year is shown in Table 2.1.

It is not possible to identify and exclude patients readmitted for the same condition in that year using HES data.[2]

Table 2.1 Total number of hospital admissions by financial and calendar year for all respiratory diseases (ICD9 460-519) for all admission methods for 1991 to 1994

Calendar year of admission	HES datasets available					Total
	1990/91	1991/92	1992/93	1993/94	1994/95	
1991	112,376	399,205	232	70	15	511,898
1992		135,081	390,654	212	54	526,001
1993			147,638	453,023	175	600,836
1994				147,423	414,536	561,959
Total	112,376	534,286	538,524	600,728	414,780	2,200,694

Source: HES data

Figure 2.2 Extraction of data from HES files for conditions coded in the respiratory chapter (ICD9 460-519, i.e. excluding tuberculosis, sarcoidosis and cystic fibrosis)

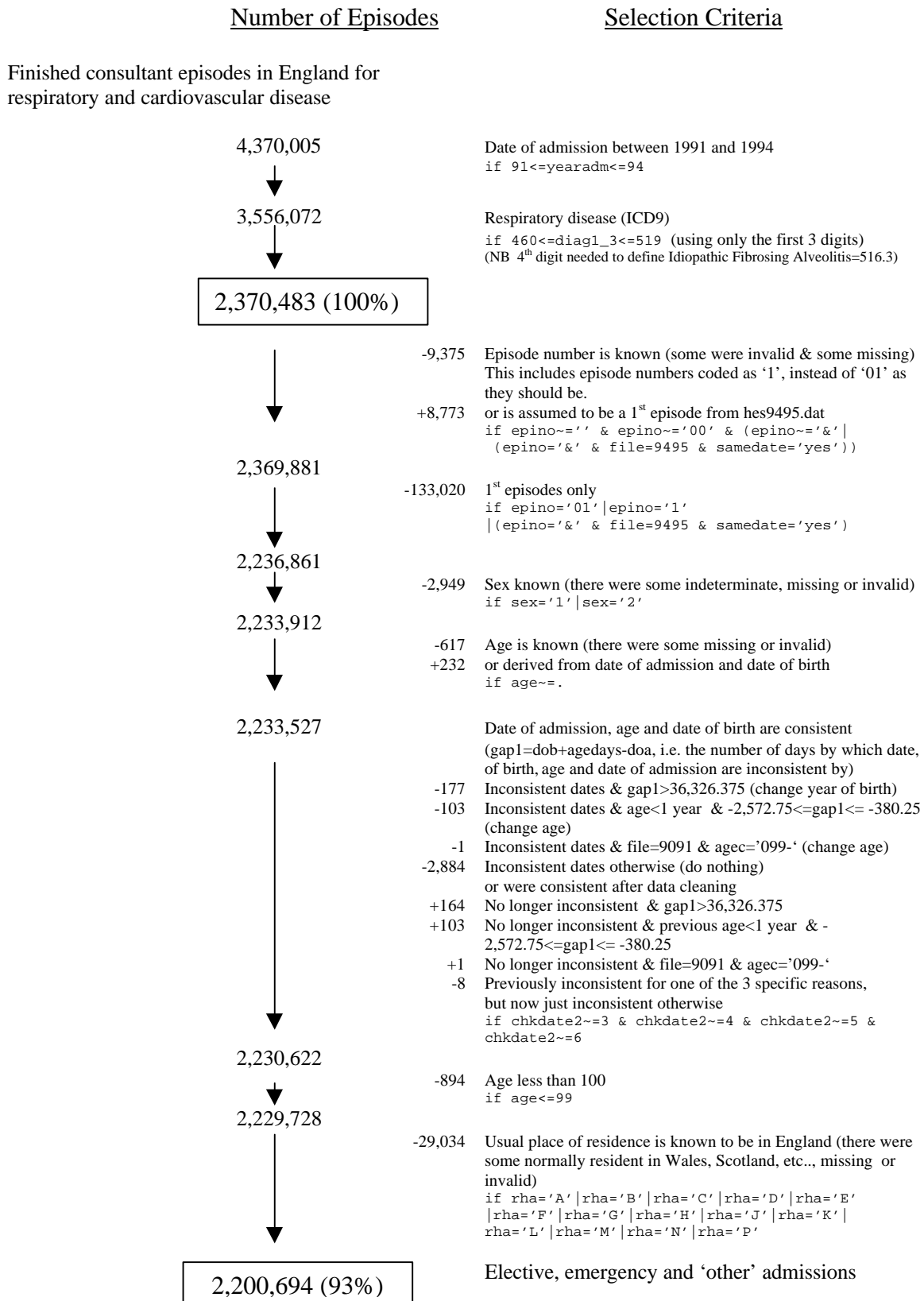
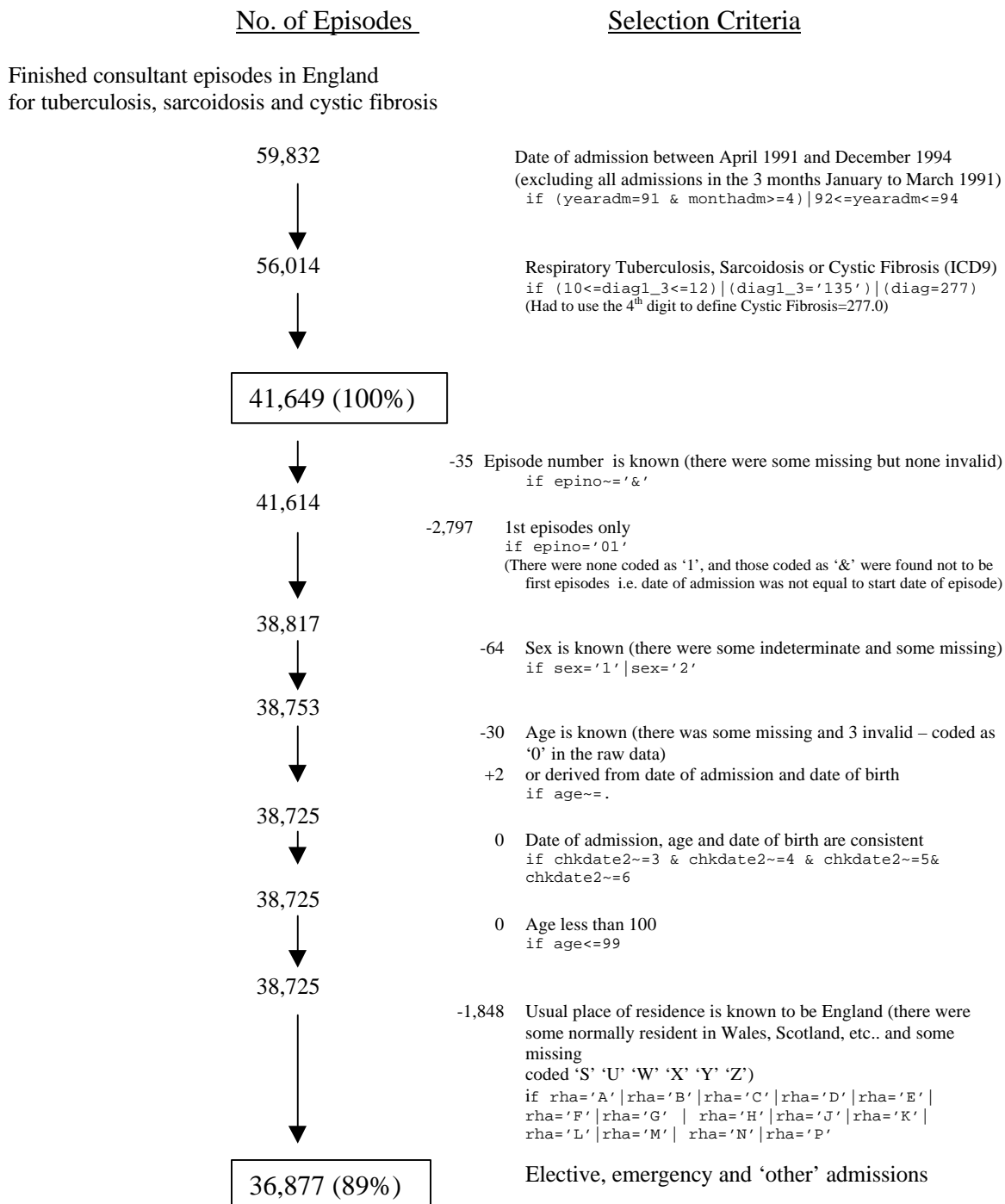


Figure 2.3 Extraction of data from HES files for respiratory tuberculosis (ICD9 codes 010-012), sarcoidosis (ICD9 135) and cystic fibrosis (ICD9 277.0)

These files contained records for non-respiratory TB and for ICD9 codes 277.1 - 277.9 which had to be excluded. Episodes completed before April 1991 were unavailable. There were a few admissions in the first 3 months of 1991 that were completed on or after 1st April 1991, but these were excluded. Admissions for the other conditions included these three months



Variables

Variables used included age in years, sex, hospital admission dates. Regional health authority of residence and district health authority (DHA) of residence (rather than the health authority of treatment – where the treatment was given) were used for consistency with other data sources (these are derived from postcode data). Persons usually living outside England were excluded. An urban-rural code with four ordered categories (conurbation, urban, mixed and rural) was assigned based on the DHA of residence (see Appendix A4 for details).

Hospital stays which straddled HES years

Full clinical details are not entered into HES records until the end of the ‘episode of care’ (generally when the patient is discharged, dies or is transferred to the care of another consultant).[2] A small number of episodes start in one financial year and end in another and the details of the admission appear in the statistics for the year in which they finished i.e. information on episodes not finished by the end of March is incomplete (for example, no diagnosis).[2] These episodes will be coded to the District existing at the end of the episode – if boundary changes have occurred this may not be the same DHA as admission. These only constituted approximately 300 of 400,000 total admissions for respiratory codes (i.e. excluding tuberculosis, cystic fibrosis and sarcoidosis) and only 14 of the emergency admissions in 1991. Since there were so few of these, attempts to retrieve them from later years of HES data (1995/6 onwards) were not attempted.

Method of admission

At the time of admission to hospital, the ‘method of admission’ is recorded. This is coded into one of 13 codes, which can be grouped into emergency, elective and other (such as transfer from another healthcare provider).[2] The number and percentage of different types of admission by condition were calculated. Further analysis was conducted on elective and on emergency admissions to find the percentage which were day cases, defined as an episode whose end date was the same as the start date.

All further analyses concentrated on emergency admissions.

Primary care data – the General Practice Research Database (GPRD)

There are several sources of computerised routine information on general practice morbidity and prescribing in England, but the largest is the General Practice Research Database (GPRD, formerly the VAMP database). This contains data from 1987 up to the present and in 1994 covered 5.6% of the population of England & Wales.[3]

Defined outcomes: Diagnostic and prescription groups

The diagnostic coding system used in the GPRD is OXMIS which can be cross-referenced to Read codes, but do not map directly to ICD9. We defined 12 diagnostic groups using 206 Oxmis codes (Appendix A1):

1. Asthma
2. Chronic bronchitis, emphysema and obstructive airways disease (excluding asthma)
3. Pneumonia
4. Acute bronchitis and bronchiolitis
5. Hay fever and allergic rhinitis
6. Tuberculosis
7. Cystic fibrosis
8. Sarcoidosis
9. Fibrosing alveolitis
10. Pneumothorax
11. Symptoms including wheeze/cough
12. Chest infection and bronchitis not otherwise specified

The GPRD also holds data on prescriptions issued. We defined three relevant prescription groups:

1. Inhalers: BNF chapters 3.1 to 3.3 inclusive
2. Hayfever therapy: BNF chapters 3.4.1 plus 12.2 (nosedrops)
3. Tuberculosis therapy: BNF chapter 5.1.9

Data extraction

Data were extracted from the GPRD database held at ONS by calendar year. Using diagnosis and prescription data, we derived the following outcomes for the first recorded mention within that year for the respiratory diagnoses of interest and for relevant prescriptions for obstructive airways diseases (including asthma), tuberculosis and hayfever for practices within England:

- (i) The number of patients with a diagnosis at any time during the year
- (ii) The number of patients with a relevant prescription at any time during the year
- (iii) The number of patients with a relevant prescription at any time during the year plus a corresponding diagnosis (e.g. hayfever therapy for hayfever). The diagnosis was recorded at the same time as the prescription, OR prior to the prescription (including previous years which may also have included diagnoses made prior to registration with the practice), OR subsequent to the

- prescription and at any time up to the end of the *year* in which the prescription was issued.
- (iv) (For seasonality – week of year – analyses) The number of patients in each week of the year with a first ever recorded diagnosis. First ever was defined as no mention in previous years including information predating the patient's registration with that practice.
 - (v) (For seasonality – week of year – analyses) The number of patients during the year and week with a relevant prescription and a corresponding diagnosis. The timing of the diagnosis could be prior to the prescription, OR at the same time as the prescription OR subsequent to the prescription at any time up to the end of the *week* in which the prescription was issued.
 - (vi) (For seasonality – week of year – analyses) The number of patients during the year and week with a non-repeat relevant prescription and a corresponding diagnosis. The timing of the diagnosis, as above, could be prior to the prescription, OR at the same time as the prescription, OR subsequent to the prescription at any time up to the end of that *week*.

We also extracted the person years at risk within the time period (year or week of year). We did not link patient data between years.

Calculation of period prevalence rates for diagnostic and therapy variables

Counts (as defined in (i) to (vi) above) were divided by the person years at risk within that time period to give a measure of period prevalence (unlike the hospital admissions and mortality which were measures of incidence). The period prevalence rate for each diagnosis can therefore be defined as a retrospective assessment of the numbers of persons (recurrent and incident cases) consulting for a specific diagnosis within the calendar year in question divided by the person years at risk during that period. Weighted average GPRD rates were calculated ($1/3 \times 1991$ rates + $2/3 \times 1992$ rates) for comparison with the fourth Morbidity Survey in General Practice (MSGP4) conducted in the last four months of 1991 and the first eight months of 1992.[4]

Other variables

Other variables extracted were age in years (calculated as year of observation minus year of birth) and region in which the GP practice was situated in 1991 - the GPRD does not permit access to the patient's place (and therefore district and region) of residence. An urban-rural code based on the district health authority (DHA) in which the GP practice was situated in 1991 (see Appendix A4 for details) was assigned to the practice. Researchers themselves are not allowed access to the DHA code in which the practice is situated because the confidentiality of participating practices is guaranteed by the GPRD and this information could potentially be used to identify the practice. Because of these confidentiality restrictions, data are only made available for geographical groupings containing at least three practices.

Health Survey for England, 1995

The Health Survey for England is an annual survey covering a representative sample of the population of England which includes information on symptoms and potential confounders such as smoking and socio-economic class. The years 1995 and 1996 both concentrated on respiratory disease and included data on IgE levels and on lung function. The main findings have been published, the 1996 volume containing pooled data on respiratory disease from both the 1995 and the 1996 survey. To date, an anonymised version of the 1995 individual level dataset has been released.

The Health Survey for England 1995 was obtained from the Data Archive held at Essex University. A smaller dataset for the purposes of this analysis (Table 2.2) was created from the main dataset of individual variables and contains a combination of supplied variables and new, derived variables. Quality checks on the Health Survey for England, which are described in Section 3, used additional variables which are not listed in Table 2.2.

Table 2.2 Variables extracted from the Health Survey for England

Variable	Description
1. Identifier	
pserial	Unique identifier
2. Geographical identifiers	
Region and district	
rha8	Eight NHS regions
rha	Number of the 14 regions
dha	Number given to the district health authority
rha_dha (derived variable)	1993/4 boundary DHAs, eg A05, D11
Urban-rural of household	
urbrur	Whether household is in an urban or rural setting
3. Demographic details	
Sex	
sex	Male/female
Age	
age	Age last birthday
age5	Age in 5 year bands (2-4, 5-9, 10-14, 15-19 etc)
age65	Age in 5 bands (2-4, 5-14, 15-44, 45-64, 65-99, 100)
4. Outcome measures	
Self-reported diagnosis	
num1 (derived variable)	Self-reported asthma as a long-standing illness (where HSE variable illsm(1-6)=23)
num2 (derived variable)	Self-reported bronchitis/emphysema as a long-standing illness (where HSE variable illsm(1-6)=22)
num5 (derived variable)	Self-reported hayfever as a long-standing illness (where HSE variable illsm(1-6)= 24)
Asthma Symptoms	
twewz	Wheezing or whistling in last 12 months (multi-stage question, NA interpreted as no)
Chronic bronchitis symptoms	
coflem	Cough/phlegm 3+ months in winter
Hayfever symptoms	
eyeit	Itchy watery eyes <u>and</u> sneezing/blocked nose without a cold in the last 12 months (multi-stage question, NA interpreted as no)
Inhaler therapy	
asthdrug (derived variable)	Patients with either (i) wheeze/whistling in the chest who have had an attack within the last 5 years OR (ii) with doctor diagnosed asthma who have used any inhaler within the last 12 months
num8 (derived variable: the subset 'twewz NOT asthdrug')	Patients who have had symptoms of wheeze/whistling in the chest within the last 12 months but have not used an inhaler
5. Potential confounders	
Smoking	
cigsmk2	ages 15+, cigarette smoking classification (current, ex regular, never regular)
Social class	
schstx3	social class head of household (I&II, IIINM, IIIM, IV&V, other)

Outcome measures

Outcome measures were available for three diseases: asthma, COPD and hayfever.

COPD

One specific outcome measure was used for COPD in the analysis: whether a person had cough or phlegm for 3 or more months in the winter.

Hayfever

One specific outcome measure was used for hayfever: whether a person had had sneezing or a runny or blocked nose when they did not have a cold in the previous 12 months accompanied by itchy, watery eyes.

The outcome measures of self-reported bronchitis and self-reported hayfever as a long-standing illness were available, but could not be used in the analyses, because they identified very small numbers of cases.

Asthma

Three outcome measures were selected for the analyses for asthma

- (i) symptoms of wheezing and whistling in the chest in the last 12 months – as a measure of current asthma
- (ii) inhaler use in the previous 12 months (by those with doctor diagnosed asthma or who had had an attack of asthma/wheeze/whistling in the chest within the last five years) – as a measure of current treated asthma
- (iii) asthma self-reported as a long-standing illness – as a measure of current troublesome asthma

However, the first two asthma measures obviously had some overlap with COPD.

The following measures for asthma from the HSE95 were also considered for inclusion:

1. symptoms of wheeze or whistling in the chest ever
2. symptoms of wheeze or whistling in the chest ever, without a cold
3. symptoms disrupting lifestyle such as disturbed sleep
4. ever been told by a doctor that you have asthma

The age breakdown of the three measures chosen for analyses (in bold) and the other four measures considered is shown in Table 2.3.

Table 2.3 Numbers and percentage with asthma-related outcomes in the HSE95

Question	Age 2-4 n=894	Age 5-14 n=2,602	Age 15-44 n=8,273	Age 45-64 n=4,683	Age 65-99 n=3,335
EVER					
Ever wheezed?	281 (31.4%)	743 (28.6%)	2574 (31.1%)	1602 (34.2%)	1125 (33.7%)
Ever wheeze/whistling in the chest without a cold	153 (17.1%)	504 (19.4%)	1756 (21.2%)	1021 (21.8%)	702 (21.1%)
Doctor diagnosed asthma ever	186 (20.8%)	533 (20.5%)	1097 (13.3%)	457 (9.8%)	302 (9.1%)
LAST 12 MONTHS					
Wheeze/whistling in last 12 months	198 (22.2%)	456 (17.5%)	1604 (19.4%)	982 (21.0%)	770 (23.1%)
Sleep disturbed by wheeze/whistling in the last 12 months	143 (16.0%)	247 (9.5%)	650 (7.9%)	442 (9.4%)	293 (8.8%)
Use of inhalers in last 12 months	111 (12.4%)	379 (14.6%)	790 (9.6%)	385 (8.2%)	355 (10.6%)
'NOW'					
Self-reported asthma as a long-standing illness	99 (11.1%)	290 (11.2%)	513 (6.2%)	220 (4.7%)	156 (4.7%)

Geographical identifiers

The dataset received contained a variable for the eight regions existing in 1995, which had been obtained by recoding the original information collected in the survey relating to the 14 regions existing in 1993/4. The numerical codes for district health authorities (DHAs) had been retained, but the recoding of regional health authorities meant that the combination of region and district codes was ambiguous in 20 of the 123 combinations representing 40 DHAs and 5,029 (~25%) of observations. We therefore had to go back to the HSE95 data depositors through the Data Archive to obtain the original rha (14 region) codes.

An urban-rural code with four ordered categories (conurbation, urban, mixed and rural) was assigned based on the DHA of residence (see Appendix A4 for details). An individual level urban rural code was also available in the HSE95. The household in which the respondent lives was classified as urban or rural based on observation made by the HSE interviewer. This classification was itself a recoded version of a variable called AreaType (with divisions into urban/city centre, small country town centre, suburban residential, rural residential, rural agricultural with isolated dwellings or small hamlets) which unfortunately was not stored on the dataset.

Population estimates

Population estimates were obtained from ONS for mid 1991. Breakdowns were by single years of age by sex for England and by five year age bands (for ages 0-85; people aged 85+ formed a single group) by sex for District Health Authorities. They were used to calculate population rates of disease.

Population weighted population density (pwpd) was derived from the 1991 census Small Area & Local Base Statistics (SAS/LBS) provided by the Census Dissemination Unit and accessed through the MIDAS service at Manchester university. Pwpd is thought to be a better measure of population density than simple population density (population divided by surface area) as it measures the population density at which the average person in that area lives.[5] The formula for pwpd has the form where 'i' is an electoral ward and the summation is over all electoral wards:

$$pwpd = \frac{\sum \left[\frac{population_i}{surface\ area_i} \times population_i \right]}{\sum population_i}$$

Finally, the variable pwpd was categorised into four ordered categories from lowest to highest with roughly equal populations in each.

Pwpd was used to investigate the validity of the urban-rural coding system used (as increasing degrees of urbanisation should correspond with increasing pwpd).

References

1. Marks G, Burney PGJ. Charlton J, Murphy M, editors. The Health of Adult Britain 1841-1994. London, UK: The Stationery Office; 1997; 20, Diseases of the respiratory system.
2. Department of Health Statistics Section SD2 HES. HES The Book. London, UK: Department of Health; 1998.
3. Key health statistics in General Practice. London, UK: The Stationery Office; 1996.
4. McCormick A; Fleming DM; Charlton J. Morbidity Statistics from General Practice. Fourth national study 1991-1992. London, UK: HMSO; 1995.
5. Dorling, D. and Atkins, D. Population density, change and concentration in Great Britain 1971, 1981 and 1991. Studies on Medical and Population Subjects, No.58. London, UK: HMSO; 1995.