

4 Analysis

Summary

Four types of descriptive analysis were performed by disease group: a simple description, a geographical analysis, a seasonal analysis and a geographical correlation across data sources (Table 4.1). An assessment of consistency across data sources was made by comparing the results of the descriptive, geographical and seasonal analyses and by the strength of the geographical correlation coefficients across data sources.

Table 4.1 Types of analyses conducted by data source

	Data source			
	Mortality (1991-5)	HES (1991-4)	GPRD (1991-5)	HSE95 (1995 only)
1. Descriptive analysis: crude rates for ages 0-99 – Age and sex distribution – Unadjusted year on year time trends – Cohort analysis	✓	✓	✓	asthma, COAD, hayfever
2. Geographical analysis: standardised event ratios for ages 0-99 and 0-84 – By urban-rural grouping and RHA – Adjusted year on year trends	✓	✓	✓	asthma, COAD, hayfever
3. Seasonal analysis: crude rates By year, week and age for age-groups 0-4, 5-14, 15-44, 45-64, 65-99	✓	✓	✓	asthma, COAD, hayfever
4. Geographical correlation across data sources: standardised event ratios for ages 0-84 – 1991 data by region and urban rural group – 1994 data by region	✓	✓	✓	asthma, COAD, hayfever

✓ = conducted for all disease groups

All analyses were conducted using the statistical package Stata version 5.0.[1]

Population rates were calculated using:

- number of events as the numerator and ONS population estimates as the denominator for mortality and HES data
- number of patients as the numerator and person years at risk as the denominator for GPRD data
- number of patients as the numerator and population base (total number of respondents) as the denominator for the Health Survey for England 1995.

Small numbers of events limited some geographical correlations across data sources.

The following sections refer to

Descriptive analyses

Age-sex, year on year and cohort

Seasonality

Geographical analyses

Geographical comparisons across and within data sources

Adjustments for confounders such as smoking and social class

Format of standard output graphs

Descriptive analyses

Simple descriptive age-sex and seasonality analyses were undertaken for each data source for all ten disease groups and the consistency between data sources for each disease was noted. More than one outcome was used for asthma, COPD and hayfever in the GPRD (diagnosis, therapy and therapy plus diagnosis) and for asthma in the HSE95 (Tables 4.2 & 4.3).

Age-sex, year on year trend and cohort analysis

Age and sex patterns were examined using both single years of age and by broader age-groups for all years combined and by year. Cohort analyses were performed using both single years of birth and five year groups of year of birth for males and females.

Table 4.2 Descriptive analyses performed by condition for each data source

Condition	Mortality	HES	GPRD	HSE95
Asthma symptoms	x	x	x	✓
Asthma diagnosis	✓	✓	✓	✓
Inhaler therapy plus asthma diagnosis (excluding concurrent COPD diagnosis)	x	x	✓	✓
COPD symptoms	x	x	x	✓
COPD diagnosis	✓	✓	✓	x
Inhaler therapy plus COPD diagnosis	x	x	✓	✓
Inhaler prescribed	x	x	✓	✓
Hay fever and allergic rhinitis diagnosis	✓	✓	✓	x
Hay fever symptoms	x	x	x	✓
Hayfever therapy plus hayfever diagnosis	x	x	✓	x
Pneumonia	✓	✓	✓	x
Acute bronchitis and bronchiolitis	✓	✓	✓	x
Tuberculosis	✓	✓	✓	x
Cystic fibrosis	✓	✓	✓	x
Sarcoidosis	✓	✓	✓	x
Fibrosing alveolitis	✓	✓	✓	x
Pneumothorax	✓	✓	✓	x

GPRD analyses included analyses by diagnosis only, by treatment only and by a combination of diagnosis and treatment (Table 4.2). A comparison of GPRD data with MSGP4 data ([2] or see Section 3) suggested that the most appropriate measures of the disease burden in primary care were:

- (i) for COPD: inhaler prescription plus diagnosis of COPD
- (ii) for asthma: either diagnosis only or inhaler prescription plus diagnosis
- (iii) for acute bronchitis or bronchiolitis: 'acute bronchitis or bronchiolitis' OR 'chest infection or bronchitis not otherwise specified'
- (iv) for hayfever: therapy for hayfever plus diagnosis of hayfever
- (v) for all other conditions: diagnosis only

Seasonal analyses

Seasonal analyses were performed for the ten respiratory conditions by week for all years combined and by each year separately (Table 4.3). Weeks were calculated as seven day intervals from 2nd January in each year. The first of January in each year and 31st December in 1992, a leap year, were excluded from the analysis. The analysis was conducted for all ages and using broad age-bands.

It was only possible to conduct this analysis in the non-survey sources – mortality statistics, HES and GPRD data. The consistency of the seasonal patterns across these three data sources for each disease was noted.

Table 4.3 Seasonal analyses performed by condition for each data source

Condition	Mortality	HES	GPRD
Asthma diagnosis	✓	✓	✓
Inhaler therapy plus asthma diagnosis (excluding concurrent COPD diagnosis)	NA	NA	✓
COPD diagnosis	✓	✓	✓
Inhaler therapy plus COPD diagnosis	NA	NA	✓
Inhaler therapy	NA	NA	✓
Hay fever and allergic rhinitis	NA	✓	✓
Pneumonia	✓	✓	✓
Acute bronchitis and bronchiolitis	✓	✓	✓
Tuberculosis	✓	✓	✓
Cystic fibrosis	✓	✓	✓
Sarcoidosis	✓	✓	✓
Idiopathic fibrosing alveolitis	✓	✓	✓
Pneumothorax	✓	✓	✓

Geographical analyses

The analyses concentrated on the use of data at a health district level in line with the original invitation to tender. Age-sex standardised event ratios for mortality, incidence of hospital admission and prevalence of disease in the GPRD and HSE95 were calculated by regional health authority and by degree of urbanisation. The patterns for each disease in the different data sources were compared as part of the assessment of the consistency of geographical patterns of respiratory disease. Analyses were conducted for 1991-1995 combined using 1995 boundaries and also for the single year 1993.

Indirect standardisation

Both the geographical analyses (described below) and the geographical correlations across data sources (described later in this section) compared standardised event ratios. These are very commonly used to make allowance for variables such as age and sex that could otherwise confound the comparison of interest. The standardised event ratio (SER) was calculated as:

$$\text{SER} = \frac{\text{observed number of events}}{\text{expected number of events}} \times 100$$

where the expected number of events was derived from the age and sex specific rates in all regions and the particular age and sex distribution of the region concerned. If the observed number of events in a region is exactly equal to the expected number of events, the SER will be 100. If the observed number is half that expected, the SER will be 50 and if the observed number is exactly double the expected, the SER will be 200. This method is known as indirect standardisation. When applied to mortality, this is referred to as standardised mortality ratios or SMRs.

In the geographical analysis (described below), the regional comparison was based on comparing SERs standardised for age, sex and the urban rural classification of the district and also year if based on more than one year of data. The urban rural comparison was based on comparing SERs standardised for age sex and region and also year if based on more than one year of data. The yearly trend was based on comparing SERs standardised for age, sex, region and urban rural classification.

In the geographical correlations across data sources (described later in this section), all SERs were standardised for age and sex.

Age groups used in geographical analyses

Age groups used were 0-84 and 0-99 (Table 4.4). District level population estimates were only available by five year age bands to age 84 and then for the group aged 85+ treated as a single band. This meant that regional analyses for ages 0-99 had slightly higher denominators and therefore slightly lower rates than they should have been. This introduced a lack of precision, but in practice made minimal difference to the results.

Results presented generally relate to 1991-1995 data and ages 0-99 unless otherwise specified. Results for both age groups can be found in the standard output graphs in Appendices A6 to A15.

Table 4.4 Age groups used for geographical analyses

Data source	1991-5	1993 only	1995 only
Mortality	Ages 0-99 Ages 0-84	Ages 0-99 Ages 0-84	-
HES	Ages 0-99 Ages 0-84	Ages 0-99 Ages 0-84	-
GPRD	Ages 0-99	Ages 0-99	-
HSE95	-	-	Ages 0-99

Confidence intervals for the geographical analyses

Confidence intervals were calculated for the standardised event ratios (SERs) for combined years of data and for a single year of data, chosen to be the middle year of 1993. However, calculating confidence intervals requires observations to be independent. While this condition will hold for mortality, it was only valid for single years within the GPRD data we extracted (the same patient may have consulted and been counted again in subsequent years). Within HES, it was not possible to distinguish patients readmitted for the same condition, although the number readmitted in the same year is likely to be small (see section on Data quality). However, confidence intervals were generally wider in the single year because of fewer observations.

Geographical identifiers contained in each dataset

The lowest level of geographical aggregation common to all datasets was regional health authority subdivided into urban rural categories, using a coding frame based on district health authorities which can be seen in Appendix A4. This was used as the main geographical comparator. Further details for each data source are given below:

- The mortality datasets included postcode and District Health Authority of residence. The latter identifier was used in this study.
- The HES datasets include postcode of residence, Provider, Health Authority of residence and Health Authority of treatment, but postcoded data are very rarely released because they potentially allow identification of individual patients. Postcoded data were not available for this study[3] and the geographical identifier used was the DHA of residence as this was available for all datasets except the GPRD.
- For GPRD data, ONS holds, but cannot release, the postcodes of the participating practices, but not that of the patients. For this project, the lowest geographical level for which data were available was regional health authority in which the GP practice was situated, subdivided by one of four urban-rural codes which were assigned at ONS. Data were not released at the level of individual DHAs nor for regional and urban rural sub-divisions containing less than three GP practices because of confidentiality restrictions related to the identification of participating practices.
- In the Health Survey for England 1995 (HSE95) codes for district health authority of residence and corresponding 14 regional health authorities (RHAs) existing to March 1995 were collected. However, the regional health authority (RHA) code had been recoded to the subsequently formed eight regional offices and the original 14 RHA codes had to be obtained separately from the depositors to be able to identify the District Health Authorities. Postcodes are collected but very rarely released to investigators.
- ONS Population estimates based on census estimates are available by ward, enumeration district and district and regional health authority.

Problems with using district health authority of residence as a geographical identifier

Our study used DHA as the basic geographical unit, using this to assign region and urban rural codes as mentioned above. The main problem with using DHA was the number of changes made to Regional and District Health Authority boundaries. To cope with this problem, all changes in Health Authority boundaries were identified by consulting the Health Services Year book for each year, with help from the Department of Health Statistics Division 2 (SD2). As a double-check mid-year population estimates from ONS by alphanumeric Health Authority district code were compared with the codes for each year to confirm changes. A look-up table was then constructed to convert all DHA codes given to those present in 1995 so that five years of data could be analysed together.

While it was easy to track changes occurring because of mergers, it was more difficult to deal with districts which had been split and then merged. To maintain consistency, districts involved in a split during the period 1991-1995 were analysed together as a 'conglomerate'. Six artificial 'conglomerate' districts were created for the purposes of analysis covering an estimated 8.3 million people (Table 4.5). Four of these involved two districts, one involved three and five districts were involved in west London (covering an estimated 2.7 million people: the size of a region).

Table 4.5 Region of conglomerate districts

Region code	Region name	Number of conglomerate districts at end 1995
A	Northern	0
B	Yorkshire	1
C	Trent	0
D	Anglia	1
E	North-west Thames	0
F	North-east Thames	1
G	South-east Thames	1
H	South-west Thames	0
J	Wessex	1
K	Oxford	0
L	South Western	0
M	West Midlands	0
N	Mersey	1
P	North Western	0
Total		6

Between 1991 and 1995, 144 districts merged and 10 districts were split (Appendix 2). At the end of 1995 the total population living in districts split or created from mergers with split districts was estimated at just over eight million people (8,303,402), 19% of the total population of England (taken as 48,903,440 – the sum of the total population in each region). During the changes, one DHA and parts of a further two DHAs changed region and two DHAs merged and also changed region. The most complicated changes occurred in the Thames regions, where six Health Authorities were divided and 48 mergers occurred, with major changes to the configuration of Health Authorities in most of West London (North and South West Thames regions).

Details of district and regional boundary changes can be found in Appendix A3.

Geographical analysis by urban-rural classification

The urban-rural coding used was developed by Professor Strachan and used in the 1995 COMEAP report on Asthma and Outdoor Air Pollution.[4] This assigned Health Authorities in 1990 to one of four categories: rural, mixed, urban and conurbation. Some of the mergers and splits of district health authorities between 1991-1995 involved districts in different urban-rural categories and these were coded as 'indeterminate' - in total, 19 of 89 non-conglomerate districts and five of six conglomerate districts existing in 1995 were assigned to an 'indeterminate' urban rural code (Table 4.6).

Table 4.6 Number of districts at end 1995 resulting from mergers of districts with different urban-rural codes in the period 1991-5 by region

Region	Number of 'non-conglomerate' districts resulting from mergers of districts with different urban-rural codes in 1991-1995	Total number of 'non-conglomerate' districts at end 1995	Number of 'conglomerate' districts resulting from mergers of districts with different urban-rural codes in 1991-1995	Total number of 'conglomerate' districts at end 1995
A	0	8	0	0
B	3	5	1	1
C	0	10	0	0
D	2	4	0	1
E	2	4	0	0
F	0	5	1	1
G	1	5	1	1
H	2	6	0	0
J	1	3	1	1
K	1	5	0	0
L	2	6	0	0
M	3	15	0	0
N	0	3	1	1
P	2	10	0	0
Total	19	89	5	6

Key to Region: A= Northern, B = Yorkshire, C = Trent, D = East Anglia, E = North West Thames, F = North East Thames, G = South East Thames, H = South West Thames, J = Wessex, K = Oxford, L = South Western, M = West Midlands, N = Mersey, P = North Western

More detailed tables showing urban rural codes for each district in 1995 by region are available in Appendix A4.

Geographical comparisons across data sources

Years chosen for geographical comparisons across data sources

Health Authority was the only geographical identifier common to all data sources used, but because of the large number of mergers and splits of Health Authorities over the five years of the analysis, some districts could not be assigned to a single region or urban rural code for analyses of combined years (and were labelled as 'indeterminate'). This limited our ability to make comparisons across different data sources using all five years of data. We therefore conducted analyses using two different approaches:

1. Statistics for 1991 from GPRD, HES and mortality were compared across the 14 regions existing in that year, which allowed urban-rural comparisons and minimised the number of indeterminate urban rural codes. There were three potential problems with this:
 - (i) Mortality data was coded on year of registration not year of occurrence prior to 1993. Therefore some 1991 deaths appear in 1992 files but were be coded to April 1992 boundaries. To overcome this we had to use 1992 boundaries but 1991 figures, which led to a small amount of data with indeterminate urban rural codes.
 - (ii) HES is recorded by financial year. Jan-March 1991 used 1990 boundaries and April-December 1991 used 1991 boundaries. However, this did not seem to be a particular problem as boundary changes between 1990/1 and 1991/2 were restricted to a small number of mergers.
 - (iii) It was not possible to compare this data with Health Survey for England 1995 (HSE95) data because the survey was conducted four years later making any comparisons of limited validity and because of subsequent boundary changes.
2. GPRD, HES, mortality and HSE95 were compared across 14 regions without subdividing by urban rural for 1994 only. We did not use urban rural codes because of the large number of indeterminate codes. 1994 was chosen because this was the most compatible with HSE95 data as it asked questions about symptoms over the past year (i.e. in 1994). Also, the majority of the observed Health Authority mergers and splits had occurred by then. The potential problems identified with this approach were:
 - (i) GPRD was coded to 1991 regions and four districts or parts of districts had changed regions by 1994 (Table 4.7). We do not know if any of the GPRD practices were in these districts. We therefore decided to ignore these small discrepancies between dataset boundaries.
 - (ii) In the HSE95, the original codes for the 14 districts had been recoded to the then current 8 regions. In order to analyse data by 14 regions, we had to specially request these codes from the Data Archive.

Table 4.7 Discrepancies between the GPRD regions and the HES/mortality regions in the 1994 graphs and scatterplots

	GRPD region	HES/mortality region
South Cumbria	A	P
Bedfordshire	E	D
50% North Hants	H	J
30% Hounslow/Spelthorne	E	H

Spearman rank correlations

A statistical assessment of the consistency between data sources for the selected ten respiratory conditions was conducted by using correlations of rankings of standardised event ratios by region and by region plus urban rural group between different data sources (Table 4.8) using Spearman's rank correlation coefficient. This test was chosen so that no assumptions needed to be made about the distribution of the underlying data. Since it uses the rank of the data, rather than the actual data value, it should be more robust to outliers than the product-moment correlation coefficient.

Table 4.8 Data sources used for geographical correlations by condition

Condition	Mortality	HES	GPRD	HSE95
Asthma	✓	✓	✓	✓
COPD	✓	✓	✓	✓
Hay fever	x	x	✓	✓
Pneumonia	✓	✓	✓	x
Acute bronchitis and bronchiolitis	x	✓	✓	x
Tuberculosis	x	x	x	x
Cystic fibrosis	x	x	x	x
Sarcoidosis	x	x	x	x
Fibrosing alveolitis	✓	✓	x	x
Pneumothorax	x	x	x	x

For some of the diseases, the numbers of events was considered to be too small to give a reliable estimate of ranking. The 1991 correlations involved subdivisions by region and urban rural, giving 50 'cells' in the comparison between HES and mortality and 33 for GPRD comparisons, while the 1994 correlations involved subdivisions by 14 regions. We considered that rankings and therefore correlations based on ranking would be unreliable if any of the cells contained less than 10 events. This affected correlations for tuberculosis, cystic fibrosis, sarcoidosis and pneumothorax. It limited meaningful comparisons for hayfever to comparisons

between HSE95 and the GPRD, for acute bronchitis or bronchiolitis to comparisons between GPRD and HES and for idiopathic fibrosing alveolitis to regional level comparisons between HES and mortality.

It was not possible to use all data sources for all conditions. The HSE95 only contained information on asthma, COPD and hayfever symptoms so it was not possible to include HSE95 data in correlations of other diseases. In contrast, the GPRD offered more than one way of examining both asthma and COPD morbidity (diagnosis only, inhaler therapy, inhaler therapy plus diagnosis).

Analyses were conducted on ages 0-84 for all respiratory conditions with additional analyses by broad age groups (0-14, 15-84) for asthma, pneumonia and acute bronchitis or bronchiolitis. However, small numbers affected comparisons with all three asthma outcomes in the HSE95 in children aged 2-4 and with GP consultations for pneumonia in children aged 0-14.

Within database comparisons

In order to assess the consistency of geographical patterns for different disease outcomes from the same dataset, the regional patterns within HSE95 and the GPRD were examined and Spearman rank correlation coefficients were calculated:

- (i) HSE95. Within database comparisons were made for asthma and COPD within the where information on symptoms, use of inhalers and self reported asthma was available.
- (ii) GPRD. Within database comparisons were also made for the three conditions for which we had extracted prescribing information – asthma, COPD and hayfever.

Adjustments for confounders such as smoking and social class

In order to assess the extent of geographical confounding by smoking and socio-economic status standardised event ratios by region from the HSE95 for hayfever, COPD and different measures of asthma were adjusted for smoking status and social class variables as well as urban rural category, age and sex. These SERs were compared with those used in previous analyses (only adjusted for urban-rural category, age and sex). The extent to which the urban rural patterns were confounded by smoking and socio-economic status were examined in the same way.

Format of standard output graphs

Standard output graphs and tables of number of events, crude rates and SERs were produced for each respiratory condition and can be found in Appendices A6 to A15. The general format of the appendices is listed below.

Age and sex distributions, yearly trends and cohort analysis (crude rates)

HSE95 (for asthma, COPD, hayfever)

GPRD, England 1991-5

HES: Emergency hospital admissions, England 1991-4

Mortality, England 1991-5

Seasonality (crude rates by week of year)

GPRD, England 1991-5

HES: Emergency hospital admissions, England 1991-4

Mortality, England 1991-5

Geographical comparisons and yearly trends (standardised event ratios)

HSE95 (for asthma, COPD, hayfever)

GPRD, England 1991-5

HES: Emergency hospital admissions, England 1991-4

Mortality, England 1991-5

Geographical comparisons and yearly trends (standardised event ratios) further adjusted for smoking and social class

HSE95 only (for asthma, COPD, hayfever)

Scatterplots and correlation coefficients (standardised event ratios standardised for age and sex). Tables of observed number of events, denominators for rates, crude rates, expected number of events, SERs standardised for age in five year bands and sex on which scatterplots and coefficients were based

- 1991 region and urban-rural comparisons between GPRD, HES and mortality, England, ages 0-84 (all conditions) and age-specific for children (ages 0-14) and adults (ages 15-84) and children vs adults (for asthma, acute bronchitis and bronchiolitis, pneumonia)
- 1994 regional comparisons between GPRD, HES and mortality. Also HSE95 for asthma, COPD and hayfever. England ages 0-84 and age-specific for children (ages 0-14) and adults (ages 15-84) and children vs adults (for asthma, acute bronchitis and bronchiolitis, pneumonia)

Within data source comparisons

HSE95 Different measures for asthma including inhaler prescription

- Age-sex distributions, yearly trends and cohort analysis (crude rates)
- Geographical comparisons (standardised event ratios)

GPRD Different measures for asthma including inhaler prescription

- Age-sex distributions, yearly trends and cohort analysis (crude rates)
- Seasonality (crude rates by week of year)

References

1. Intercooled Stata [computer program]. 5.0 for Windows 95. USA: Stata Corporation, Texas; 1997;
2. 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.
3. Department of Health Statistics Section SD2 HES. HES The Book. London, UK: Department of Health; 1998.
4. Committee on the Medical Aspects of Air Pollution. Asthma and outdoor air pollution. London, UK: HMSO; 1995.