

## **COLLATION AND COMPARISON OF MORTALITY, HOSPITAL ADMISSION, GENERAL PRACTICE AND SURVEY DATA ON RESPIRATORY DISEASE (STRACHAN/SURVEILLANCE/96/12)**

### **Executive Summary and recommendations**

#### **Summary of key findings**

- Variations in quality of data and frequent health authority boundary changes complicate analyses of routine data sources.
- Small numbers limited our ability to evaluate data on rarer diseases such as cystic fibrosis, idiopathic fibrosing alveolitis, sarcoidosis and pneumothorax.
- Each disease showed different epidemiological patterns and variable degrees of consistency between data sources. It was not possible to extrapolate from one disease to another.
- Asthma showed inconsistent disease patterns in different data sources and weak correlations for regional rankings. This suggests that high levels of emergency hospital admissions for asthma in a district would not necessarily correspond with a high prevalence of asthma
- The geographical patterns of COPD, acute bronchitis or bronchiolitis and pneumonia were more consistent in different data sources with higher rates in the North of England. COPD and pneumonia rates were also consistently higher in urban areas.
- Tuberculosis showed consistent patterns in different data sources including notifications, with highest rates in North Thames and West Midlands.
- Adjustment of the prevalences of COPD symptoms for social class and smoking habits using individual level data attenuated, but did not remove the regional and urban pattern for COPD

## **Aims**

This report describes work commissioned by the Department of Health under project reference STRACHAN/SURVEILLANCE/96/12

The aims of this work, were:

- (a) to investigate whether consistent patterns emerge from nationally available sources of data on respiratory disease when analysed by time, place and person and
- (b) to test the validity and feasibility of using routinely available data to explore environmental influences of respiratory disease.

## **Methods**

Different data sources giving information on 10 respiratory diseases in 1991-1995 were compared. Three routine data sources were used: mortality statistics, Hospital Episode Statistics (HES) and the General Practice Research Database (GPRD). Comparisons also included a national survey, the Health Survey of England in 1995 (HSE95) which gave information on symptoms for asthma, COPD and hayfever and on the social class and smoking status of individuals. The types of respiratory diseases studied were allergic diseases (asthma and allergic rhinitis), obstructive airways diseases (COPD and asthma), infectious conditions (pneumonia, acute bronchitis & bronchiolitis and tuberculosis) and rarer conditions (cystic fibrosis, fibrosing alveolitis, sarcoidosis, pneumothorax). Cancers were not included.

Four types of analysis were performed for each disease in order to assess the degree of consistency between data sources:

- Descriptive: within data source by age and sex, year on year time trends and by birth cohort
- Seasonal: within data source by week of the year (not possible for HSE95)
- Geographical: within data source by region and urban-rural classification
- Comparative: correlations of regional rankings across data sources using Spearman rank correlations as a descriptive measure of association for single years of data from 1991 and 1994

## Results

Each disease showed different patterns and it was not possible to extrapolate from one disease to another. The annual numbers of events for each disease and each data source is illustrated in Table 1. Division of the number of events by 100 gives an approximation of numbers expected in an average district health authority (DHA).

**Table 1 Total observed number of events in England (1994 data) for patient consultations in the GPRD, emergency hospital admissions in HES, deaths and Health Survey for England 1995**

| Condition                         | GPRD<br>(~6% popn)<br>Age 0-84 | HES<br>(100% popn)<br>Age 0-84 | Deaths<br>(100% popn)<br>Age 0-84 | HSE 95<br>(~0.04% popn)<br>Age 2-84 |
|-----------------------------------|--------------------------------|--------------------------------|-----------------------------------|-------------------------------------|
| Asthma                            | 81,905*                        | 78,921                         | 1,215                             | 2,003 †                             |
| COPD                              | 15,953*                        | 52,898                         | 18,388                            | 1,222                               |
| Pneumonia                         | 3,260*                         | 43,784                         | 22,436                            | -                                   |
| Acute bronchitis or bronchiolitis | 84,147                         | 25,913                         | 294                               | -                                   |
| Hayfever                          | 55,596                         | 71                             | 0                                 | 2,832                               |
| Tuberculosis                      | 174                            | 1,552                          | 260                               | -                                   |
| Cystic fibrosis                   | 100                            | 2,954                          | 101                               | -                                   |
| Sarcoidosis                       | 159                            | 427                            | 75                                | -                                   |
| Idiopathic fibrosing alveolitis   | 211                            | 1,075                          | 815                               | -                                   |
| Pneumothorax                      | 213                            | 4,937                          | 41                                | -                                   |

\*based on patient prescriptions not patient consultations †used an inhaler in the past year

Table 2 summarises age and sex, seasonal and geographical distribution for the 10 respiratory diseases including asthma, COPD and previously unpublished information on rarer conditions such as sarcoidosis and idiopathic fibrosing alveolitis. Year on year trends were not consistent across data sources for any disease over this five year period except for acute bronchitis or bronchiolitis and fibrosing alveolitis which were partially consistent. Pneumonia mortality rates showed an artefactual rise between 1992 and 1993 due to changes in coding rules for death certificates.

Consistency across data sources varied by condition (Tables 2 & 3). Asthma showed inconsistent disease patterns and weak geographical correlations across data sources, but COPD and tuberculosis were fully consistent. Hayfever, acute bronchitis and bronchiolitis and pneumonia were consistent only for some analyses. COPD, acute bronchitis or bronchiolitis and pneumonia all showed higher (age-sex standardised) rates in Northern areas of England and COPD and pneumonia showed higher rates in urban areas. Adjustment of the prevalences of COPD symptoms for social class and smoking habits using individual data from the HSE95 attenuated the regional and urban patterns but did not remove them.

**Table 2 Summary of age-sex, seasonal and geographical analyses for 1991-5**

| Disease                           | Age-sex, seasonal, geographical   |
|-----------------------------------|---|
| Asthma                            | <ul style="list-style-type: none"> <li>- Inconsistent age-sex patterns: ↑ deaths in elderly, ↑ emergency hospital admissions in ages 0-4, ↑ inhalers for asthma in ages 5-15</li> <li>- Inconsistent seasonal pattern: ↑ deaths in winter, ↑ hospital admissions in September, ↑ first ever GP consultations in early summer</li> <li>- Inconsistent regional and urban rural patterns: ↑urban ↓rural gradient in HES &amp; mortality, no gradient in GPRD or HSE95</li> </ul>  |
| Acute bronchitis or bronchiolitis | <ul style="list-style-type: none"> <li>- Partially consistent age-sex patterns. Rates highest at extremes of age but relative magnitudes varied. M&gt;F except ages 15-50 in HES and GPRD</li> <li>- Consistently highest in December and January</li> <li>- North &amp; Midlands &gt; South. No consistent urban rural pattern</li> </ul>  |
| COPD                              | <ul style="list-style-type: none"> <li>- Rates highest in elderly and M&gt;F in all data sources</li> <li>- Winter consistently higher than summer</li> <li>- North&gt;south, urban&gt;rural areas in all data sources</li> </ul>   |
| Hayfever                          | <ul style="list-style-type: none"> <li>- Comparisons limited to HSE95 and GPRD because of small numbers</li> <li>- Boys&gt;girls but F&gt;M in adults in all data sources with highest rates in children or young adults</li> <li>- Rates consistently highest in June and July</li> <li>- ↑SW Thames and Oxford, ↓Yorkshire region in all data sources.</li> <li>- No consistent urban rural pattern</li> </ul>  |
| Pneumonia                         | <ul style="list-style-type: none"> <li>- Consistently highest in elderly, M&gt;F</li> <li>- Winter&gt;summer in all data sources</li> <li>- Consistent regional and urban rural patterns. North&gt;south, urban&gt;rural areas</li> </ul>   |
| Cystic fibrosis                   | <ul style="list-style-type: none"> <li>- Highest in adolescence, but sex distribution varied by data source</li> <li>- Small numbers limited ability to interpret seasonal patterns</li> <li>- ↑Yorkshire, Mersey and Wessex, ↓Northern, SW Thames and NE Thames. Small numbers limited ability to interpret urban rural patterns</li> </ul>  |
| Idiopathic fibrosing alveolitis   | <ul style="list-style-type: none"> <li>- Increased with age, male rates ≈ 2x female rates</li> <li>- Partially consistent for seasonal pattern. Higher in winter in mortality and HES, no seasonal pattern in GPRD</li> <li>- Not consistent for urban rural. Urban&gt;rural for mortality. No urban rural pattern for HES and GPRD</li> <li>- Small numbers limited ability to interpret regional patterns</li> </ul>  |
| Pneumothorax                      | <ul style="list-style-type: none"> <li>- Male rates ≈ 5x female rates. Partially consistent for age: GP consultations and emergency hospital admissions ↑ in teenagers and ↑ in elderly, but single peak in the elderly for deaths.</li> <li>- No seasonal patterns seen in any data source</li> <li>- Small numbers ability to interpret regional and urban rural patterns</li> </ul>  |
| Sarcoidosis                       | <ul style="list-style-type: none"> <li>- Male rates&gt;female rates to age 50 then females&gt;males. Partially consistent age distributions: ↑ GP consultations and emergency hospital admissions in ages 40-60, ↑ deaths in ages 55-85.</li> <li>- Small numbers limited interpretation of seasonality but GP consultations ↑June and July while hospital admissions and deaths ↓July to early October.</li> <li>- ↑ North Thames, East Anglia, South Western, ↓ North of England. Higher SERs in rural and conurbation than mixed and urban areas.</li> </ul> |
| Tuberculosis                      | <ul style="list-style-type: none"> <li>- Highest rates in elderly, M&gt;F in all data sources</li> <li>- Consistent lack of seasonal pattern</li> <li>- Highest regions North Thames &amp; West Midlands, conurbation &gt; rural</li> </ul>   |

**Table 3 Suggested routine data sources with sufficient numbers to permit annual rankings at district and regional health authority level and degree of consistency between data sources for regional rankings**

| <b>Disease</b>                    | <b>Sufficient nos* for district rankings</b> | <b>Sufficient nos† for regional rankings</b> | <b>Consistency of regional rankings‡</b>   |
|-----------------------------------|--|--|--|
| <b>Common diseases</b>            |  |  |  |
| Asthma                            | HES<br>GPRD                                  | Mortality<br>HES<br>GPRD<br>HSE95            | Weak geographical correlations across data sources   |
| Acute bronchitis or bronchiolitis | HES<br>GPRD                                  | HES<br>GPRD                                  | Moderately good geographical correlation between GPRD and HES  |
| COPD                              | Mortality<br>HES<br>GPRD                     | Mortality<br>HES<br>GPRD<br>HSE95            | Good geographical correlations between data sources  |
| Hayfever                          | GPRD   | GPRD<br>HSE95                                | Weak geographical correlation between symptoms and GP prescriptions for hayfever   |
| Pneumonia                         | Mortality<br>HES                             | Mortality<br>HES                             | Moderately good positive correlations between HES and mortality  |
| <b>Rarer diseases</b>             |  |  |  |
| Cystic fibrosis                   | -  | HES  | Good consistency of regional rankings across data sources  |
| Idiopathic fibrosing alveolitis   | -  | Mortality<br>HES                             | Moderate consistency across data sources of regional rankings  |
| Pneumothorax                      | -  | HES  | Could not be assessed due to small numbers even in combined years  |
| Sarcoidosis                       | -  | -  | Moderate consistency across data sources of regional rankings  |
| Tuberculosis                      | -  | HES<br><br>Notifications                     | Good consistency of rankings between HES and mortality. Moderate consistency of GPRD with HES and mortality<br><br>Good consistency of notifications with HES and mortality. Moderate consistency with GPRD. |

\* at least 100 events per average district, based on observed number of events in 1994

† total of at least 800 events, based on observed number of events in 1994

‡ based on one year of data for common diseases, based on several years data for rarer diseases

## Discussion

### **Practical problems in using routine data to explore the geographical distribution of respiratory disease**

- (i) Boundary changes. Information is often required by administrative boundaries, but these are prone to frequent change. Postcoded or small area data are needed to aggregate data to the required boundary. Facilities need to be in place to allow researchers access to data aggregated to specified boundaries (for example, five years of data to 1999 boundaries).
- (ii) Data quality. This is a particular problem in using hospital admissions data as this varies between trusts which may affect certain types of studies.
- (iii) Diagnosis and coding. Regional differences in clinical practice and clinical coding may not be captured on routine quality reports.
- (iv) Comparability of coding systems. Different data sources use different clinical coding system or versions: mortality currently uses ICD9, HES uses ICD10, GPRD uses OXMIS codes, while surveys use text-based questionnaires.
- (v) Small numbers. These may limit precision, particularly with age-specific rates using single years of data or in small geographical areas such as DHA.

### **Investigation of environmental influences on respiratory disease**

The format of this will be determined by the data sources with sufficient numbers to permit meaningful statistical analysis and by the level of consistency between them (Table 3). Generally, routine data might be sought to investigate environmental influences on respiratory disease at a district health authority in two circumstances:

#### **(i) Rates from a routine data source are reported to be higher than average**

For example, hospital admission rates for a particular disease in the past year were significantly higher than average. Districts may wish to compare rates with previous years and with other routine data sources, taking into account the practical problems as above, and to refer to Table 3 or make their own assessment of consistency.

(a) *Inconsistent routine data sources.* For example, asthma. One cannot infer that asthma prevalence or mortality rates are high in an area with high hospital admission rates. Factors such as the threshold for hospital admission, geographical proximity to hospital and quality of and access to primary care are more likely than environmental influences to explain the inconsistency between data sources.

(b) *Consistent routine data sources.* For example, COPD. High hospital admission rates suggest high underlying prevalence of disease. Before investigation of environmental influences, known confounders such as smoking need to be adjusted for. Adjusting for social class can be performed, but may partially adjust for environmental influences or for lifestyle factors such as diet. If differences remain, it would be reasonable to suggest that environmental influences may be responsible. However, it may not be clear whether current environmental exposures or prior exposures (such as in childhood or previous years) are more important and further work may be needed to clarify this.

**(ii) There is a known or suspected environmental hazard locally**

Where data are clearly inconsistent, such as asthma, the data source most clearly related to the problem needs to be used. For example, asthma severity might be better assessed using hospital admissions, while prevalence might be better assessed by survey data on symptoms. Where data are clearly consistent, such as COPD, any data source could be used to estimate the impact of environmental influences.

## **Recommendations**

### **(a) To the Department of Health**

#### **Use of routine data to investigate environmental influences on respiratory disease**

1. Routine data can be used to give information about the patterns of disease, but they should be interpreted with care. In particular, asthma shows striking inconsistency between routine data sources and high rates of hospital admissions for this disease cannot be interpreted as an indicator of an adverse environmental effect.

#### **Improvements to PACT and HES**

2. Inclusion of age (even if limited to a simple distinction between child and adult) would improve the epidemiological usefulness of PACT as a routine data source.
3. Dual coding should be performed for HES in the year before ICD changes as is already performed for mortality data. This could be performed on a 1% national sample. It is recommended that this is performed retrospectively on hospital admissions for 1995/6 to assess the impact of the ICD9 to ICD10 changes.

### **(b) To the Office for National Statistics**

#### **Decennial supplement**

4. Comparative analyses across data sources provide useful information about the burden of disease. A comparison using ten years of data could usefully be included with the decennial supplement, concentrated on diseases with particular public health relevance such as asthma and COPD.

#### **Improvements to the GPRD**

5. The value of the GPRD for epidemiological analysis would be enhanced if postcode-based socio-economic indicators for each patient were linked to the clinical records. This could be done at the practice level to avoid compromising patient confidentiality by centralised compilation of postcodes.
6. The validity of the smoking data contained in the GPRD needs to be evaluated by analysing it in relation to outcomes such as lung cancer and COPD which are known to be strongly related to smoking..

### **(c) To Health authorities, Primary Care Groups and Trusts**

#### **Quality of data**

7. Health Authorities should routinely monitor the quality of hospital data.
8. The implementation of the government's IT strategy "Information for Health", clinical governance and the use of routine data for performance monitoring as specified in the White Paper should be used as opportunities to improve the quality of HES.

#### **Conducting epidemiological analyses**

9. Researchers undertaking detailed epidemiological analyses using HES should use the full (100% sample) unadjusted data rather than published data (which is adjusted and uses a 25% sample).
10. *Systematic* variations in coverage and missing diagnostic codes in HES should generally be investigated in the following circumstances:
  - (i) Comparison of trusts
  - (ii) Investigation of time trends using aggregated data from a small number of trusts, for example investigation of a disease cluster within a Health Authority or of an environmental hazard.

If levels of missing data are high, trusts may have to be excluded from studies. Where this is not possible (for example, performance management of local trusts) local knowledge, trust level data quality reports and liaison with the trust concerned may be required. In epidemiological studies, statistical adjustment may be needed.

*Systematic* variations in coverage and missing diagnostic codes may *not* need investigation in the following circumstances:

- (iii) Daily or weekly time series analyses such as short-term fluctuations in air pollution levels (variations unlikely to vary with exposures)
- (iv) Large aggregations of data – national and probably at regional level (variations in quality likely to even out over larger areas)
- (v) Using "cross-sectional" data from a single trust (variations likely to be internally consistent, at least over short periods of time)
- (vi) Using all admissions irrespective of cause (completeness of coverage varies less than missing diagnostic codes)