Accessing and using hospital activity data

Anna Hansell, Alex Bottle, Linda Shurlock and Paul Aylin

Summary

Hospital activity data can be accessed from a variety of sources ranging from hospitals to the Department of Health. These data provide valuable and widely used information, but care is needed in their use and interpretation. Hospital activity rates reflect not only the underlying prevalence and severity of disease, individual factors and referral practices, but also variations related to provider-specific factors: the ‘provider effect’. This includes completeness in the data, supply of hospital beds, admission policies, hospital access and distance from hospital. The provider effect can be controlled to a certain extent in statistical analyses. Although data quality has improved considerably in the last decade, this should still be investigated where trusts are being compared and in small area studies because missing data may lead to artefactual differences in rates. ‘Dump’ postcodes, where missing or unknown postcodes are assigned to a local postcode, may affect small area analyses and linkage if a proxy patient identifier is constructed that includes postcode.

Keywords: hospital records, patient admission, hospitals, data collection, data interpretation, statistical

Introduction

Hospital activity data are currently used for a wide range of purposes, including policy development; monitoring and evaluation; performance management; resource allocation; public health enquiries such as needs assessments; research; and public and parliamentary accountability. Advantages of using these datasets are that they are relatively comprehensive, comparatively cheap to use and ostensibly similar throughout the National Health Service (NHS).\(^1\) This paper discusses methodological and practical issues in their use and interpretation. It focuses on data originating from hospital-based information systems such as English Hospital Episode Statistics (HES), but findings will apply to similar UK hospital datasets.

How to access hospital activity data

Information on hospital activity can be obtained in a number of formats from a variety of sources:

(1) Published annual reference volumes (ARVs) of HES data. These data use a 100 per cent sample from 1994/1995 and a 25 per cent sample before this, and are grossed (adjusted to take account of variations in coverage and completeness of diagnostic coding).

(2) Within other publications and outputs such as the Hospital Activity Bulletin, Health and Personal Social Statistics, and Social and Regional Trends.

(3) Within various indicator sets, e.g. clinical indicators, clinical effective indicators, high-level performance indicators and the Public Health Common Dataset.

(4) Electronic media: CD-ROMs from 1994/1995 onwards. These are based on the ARVs containing tabulated grossed data at national (England), regional office and Health Authority (HA) level and are free to NHS users. Internet access via the Department of Health Web site to supporting information and ARV tables from April 1998 onwards is available on-line at http://www.doh.gov.uk/public/stats3.htm (accessed September 2000).

(5) *Ad hoc* tabulations and extracts and information on data quality. Possible routes include:

(a) Directly from providers’\(^2\) Patient Administration Systems (PAS) or from HAs – only practical if data from only one or a few providers or purchasers are being analysed.

(b) From the HES Enquiry Point at the Department of Health [external enquiries are forwarded to IBM (HES)]. Extracts can be taken from the cleaned annual data held from 1989/1990 to the most recent completed financial year and, if more up-to-date data are required, from the quarterly data (available approximately 10 weeks in arrears), which have been only partially cleaned. Release of extracts is covered by strict conditions protecting patient confidentiality – all requests have to be approved by the Security and Confidentiality Advisory Group (SCAG) in England (or equivalent in other UK countries).

(c) From the NHS-Wide Clearing Service (NCWS) database, maintained by a commercial company on behalf of the NHS.\(^3\)

---

\(^1\)Department of Epidemiology and Public Health, Imperial College School of Medicine at St Mary’s, London W2 1PG.

\(^2\)SD2, HES, Department of Health, Skipton House, 80 London Road, London SE1 6LH.

\(^3\)From the NHS-Wide Clearing Service (NCWS) database, maintained by a commercial company on behalf of the NHS.

© Faculty of Public Health Medicine 2001
NHS. The NCWS is a clearing house for the hospital activity data from which the HES dataset is derived. Data extracts can be supplied for specified periods and on a regular basis if required. A full dataset is available from 1997/1998, with a partial dataset for 1996/1997. These data have not been cleaned, although cleaning algorithms can be used as required. Data requests have to be agreed by the NCWS Management Board.

(d) Through a regional ‘HES Safe Haven’ in England. Currently, two pilot sites are in operation: the South East Institute of Public Health and the University of Birmingham. This facility is likely to be extended to other regions via the new Public Health Observatories in future.

The choice between NWCS clearing house data and HES data will be affected by data requirements and the type of study. Studies needing cleaned data that are consistent with national figures or data before 1996 should use HES data; studies needing the most up-to-date data available should use clearing house data.

Types of analyses

Many analyses conducted using HES will include basic measures such as age- and sex-specific admission rates, cohort analyses and simple time trends. The focus of the analysis may be on specific conditions or treatments, or on administrative information (e.g. length of stay or waiting times), whereas the purpose of the analysis may be to explore one of the following areas:

1. Inequalities (e.g. differences by ethnic group, age or sex);
2. Geographical variations (ranging from small area level studies to national);
3. Organizational variations (such as NHS Trusts, Local Authorities or HAs).

More complex analyses are possible (e.g. the impact of air pollution on health), using a range of explanatory variables and an HES-derived variable as the outcome.

Patient admissions, readmissions and episodes

Rates of ‘finished consultant episodes’ (FCEs) are only an indirect measure for patient admission rates. A good approximation for admissions is to select episodes where the field ‘EPI-ORDER’ = 1 and choose the primary diagnosis field (DIAG_1) as the main reason for admission. Counts of admissions will be increased if patients are discharged early only for them to be readmitted later (discharging patients ‘sicker and quicker’), or if patients go home for the weekend and are readmitted on Monday.

The calculation of the rates of patients admitted in a certain time period or of readmission rates (however defined) requires a patient identifier such as the NHS number (assuming it has been entered correctly) in PAS. This was introduced into HES from April 1997, but to date this field has not been complete enough to use in routine analysis. There was no such personal identifier in previous HES years, but it is possible to approximate a unique patient identifier using date of birth, sex and postcode with at least 90 per cent accuracy (L. Gill, personal communication). This method requires correct data entry and invariance of postcode (less likely in more mobile urban populations). These special analyses can be run (at a cost) for specific data requests (see above).

Activity measures

Options for measures of hospital activity include hospital episodes and hospital admissions (which can be subdivided into emergency, elective or ‘other’). Many analyses using HES select emergency admissions, but for some enquiries and/or diagnoses all admissions may be more appropriate, particularly if admissions are being used as a proxy for underlying disease prevalence. For example, an analysis of hospital admissions for respiratory diseases (Table 1) suggested that almost all asthma admissions were emergencies, but only half those for idiopathic fibrosing alveolitis were. It is possible to look at admission rates by specialty of admitting consultant, but specialty codes are not consistently defined between Trusts.

### Table 1: Method of admission for all and selected respiratory diseases for admissions with a start date in 1991–1994 (numbers, with percentages given in parentheses)

<table>
<thead>
<tr>
<th>Disease (ICD9 code)</th>
<th>Elective admissions</th>
<th>Emergency admissions</th>
<th>Other admissions</th>
<th>All admissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>All respiratory diseases (460–519)</td>
<td>702165 (32)</td>
<td>1463195 (66)</td>
<td>35334 (2)</td>
<td>2200694 (100)</td>
</tr>
<tr>
<td>Asthma (493)</td>
<td>10480 (3)</td>
<td>330306 (96)</td>
<td>3166 (1)</td>
<td>343952 (100)</td>
</tr>
<tr>
<td>COPD (not asthma) (490–492, 494–496)</td>
<td>27852 (10)</td>
<td>244760 (88)</td>
<td>5283 (2)</td>
<td>277895 (100)</td>
</tr>
<tr>
<td>Hayfever (477)</td>
<td>70423 (96)</td>
<td>295 (4)</td>
<td>27 (0)</td>
<td>7365 (100)</td>
</tr>
<tr>
<td>Pneumonia (480–486)</td>
<td>9084 (4)</td>
<td>219585 (92)</td>
<td>9421 (4)</td>
<td>238090 (100)</td>
</tr>
<tr>
<td>Acute bronchitis or bronchiolitis (466)</td>
<td>989 (1)</td>
<td>91803 (97)</td>
<td>1468 (2)</td>
<td>94260 (100)</td>
</tr>
<tr>
<td>Idiopathic fibrosing alveolitis (516.3)</td>
<td>3566 (44)</td>
<td>4322 (53)</td>
<td>212 (3)</td>
<td>8100 (100)</td>
</tr>
<tr>
<td>Pneumothorax (512)</td>
<td>2340 (9)</td>
<td>21600 (82)</td>
<td>2278 (9)</td>
<td>26218 (100)</td>
</tr>
</tbody>
</table>

Source: analysis of unpublished, ungrossed HES data.
Outcome measures

Information can be obtained on outcomes of care such as deaths in hospital, length of stay, readmissions (using a patient identifier), complications as a result of surgery, or method of delivery in maternity care. However, there are some limitations to this information. Complications as a result of surgery can be coded using ICD codes, but are likely to be underreported, as providers may be reluctant to volunteer these. Whereas deaths occurring in hospital are likely to be well recorded, and could be used in investigations of complications of care after allowing for ‘case mix’ (the health of the patients before admission), some, if not most, of these deaths will occur after the patient has left hospital. HES and mortality data are not currently linked, but feasibility studies to see if this could be done in future are being carried out. Information on case mix currently available within hospital activity data is likely to be limited to age and co-morbidity (which is variably recorded and may be impossible to distinguish from iatrogenic events).

Coding changes

ICD code changes and the changes in coding rules accompanying them have been recognized to lead to artefactual changes in rates of diseases. As with previous changes of ICD codes, dual coding for mortality (using both ICD9 and ICD10) was carried out in 2000, the year before the change to ICD10, allowing the calculation of conversion factors. However, routine dual coding was not performed for hospital admissions data in the change from ICD9 to ICD10 in 1995. This may make comparisons within HES across the change years of 1994/1995 and 1995/1996 problematic for some conditions. Potential problems also exist in comparing hospital (using ICD10) and mortality data (using ICD9) between 1995 and 2000. Mapping one coding version to the other would be required for such comparisons.

Geographical analyses

From 1991/1992 the HES dataset has been collected by provider (hospital or community trust), but the dataset contains a field for postcode of residence and for both HA of residence and of treatment. For most epidemiological analyses the area of residence is the most relevant, but for some the HA of treatment may be preferred. For example, in analyses of time trends in air pollution the HA of treatment may correspond better to exposure.

Analysis over time by administrative region can be difficult because of boundary changes. English HAs have undergone

| Table 2 | English health administrative regions, 1990–1999 |
|-----------------------------------------------|
| **14 Regional Health Authorities (RHAs)** | **8 Regional Offices** | **January 1999‡** |
| April 1990–March 1994* | April 1994–March 1999† | |
| Northern | Northern | Northern |
| Yorkshire | Trent | Trent |
| Trent | East Anglia | Anglia & Oxford |
| East Anglia | Oxford | Eastern |
| Oxford | North West Thames | North Thames |
| North West Thames | North East Thames | South Thames |
| North East Thames | South East Thames | South Thames |
| South East Thames | South West Thames | South Thames |
| South West Thames | Wessex | South and West |
| Wessex | South Western | South East |
| South Western | West Midlands | South West |
| West Midlands | Mersey | West Midlands |
| Mersey | North Western | North West |
| North Western | | |

*Parts of two HAs changed region in April 1993: (i) Spelthorne (part of Hounslow & Spelthorne) moved from North West Thames to South West Thames; (ii) North Hampshire (part of West Surrey & NE Hants) moved from Wessex to South West Thames.
†The change from 14 RHAs to eight regional offices in April 1994 involved merging 12 RHAs into six regional offices; the boundaries of the remaining two RHAs were unchanged. Region fields in HES continued to use the 14 RHAs until March 1996, although the administrative boundaries changed in April 1994. Two HAs changed region in April 1994: South Cumbria moved from Northern to North West region, Bedfordshire moved to Anglia & Oxford region. Seventeen parts of 13 HAs changed region in April 1996 as DHA boundaries were aligned with local government boundaries.
‡Changes in 1999 were related to the creation of London as a region and involved changes to Anglia & Oxford, North Thames, South Thames, and South and West regions.
major changes in the past decade (Table 2). Between 1991 and 1995, 144 district health authorities (DHAs) merged and 10 were split (the majority of changes occurred in 1993 and 1994), with the most complicated realignments occurring in Thames regions. Further large-scale changes occurred in April 1996, when HA boundaries were aligned with local government boundaries.\(^9\) This involved (mainly minor) boundary changes to 80 of 105 HAs.

Changes in HA boundaries can be identified by consulting the Health Services Yearbook for each year and with help from the Department of Health Statistics Division 2 (SD2). An easier alternative for data requests is to specify the boundaries required so that postcodes can be used to allocate admissions to uniform boundaries.

Postcoded HES data are necessary to define the boundaries of smaller geographical areas.\(^6\) The postcode field is rarely released to individual researchers, to protect patient confidentiality, but special analyses for specified postcodes can be run (at a cost). Two particular areas of concern in using postcodes are postcode changes over time and ‘dump postcodes’ (where missing or unknown postcodes are assigned to a local postcode such as that of the hospital), which will not be detected using routine quality checks. In geographical studies of small areas, counts by postcode may need to check for unexpected gaps or high numbers of observed cases. Problems with postcodes may also affect linkage where patient identifiers are constructed using postcodes.

**Hospital admission rates as a measure of community prevalence**

Hospital admission rates may not be a good measure of underlying community prevalence at a small area level for all but a limited range of conditions that are easily diagnosed and where there is a consensus on treatment (e.g. myocardial infarction).\(^11\) For example, variations in self-reported prevalence of asthma or chronic bronchitis explained only half (54 per cent) of the variation in hospital episodes for these conditions in a study of electoral wards in northern England in 1991.\(^11\) Another study of 90 HAs in England found that socio-economic characteristics, health status and secondary care supply factors explained 45 per cent of the variation in HA admission rates between 1989/1990 and 1994/1995 for asthma and 33 per cent of the variations for diabetes.\(^12\)

**Adjusting for the ‘provider effect’**

The provider effect can be considered as the part of the variation in admission rates between providers that can be explained by provider-specific factors. Part of this may be artefactual and related to variations in the completeness of the data and coding differences. Other factors include supply of hospital beds,\(^15\) variations in admission policies, hospital access (which tends to be less for less-privileged groups),\(^14\) and distance from hospital.\(^15,16\) The provider effect may also partly encompass general practitioner (GP) referral rates, for example, if these are influenced by admission policies. Other factors influencing admission rates are individual patient decision thresholds to seek medical advice, which are influenced by many psychosocial factors,\(^17–19\) the quality of primary care,\(^20\) and the prevalence and severity of disease. Observed variations in admission rates may be large, even where providers are adjacent.

Adjustment for completeness of the data using grossing factors is generally not advised for epidemiological analyses, as it may introduce unrecognizable distortions. For example, the quality report for Northern RHA for 1992/1993 (available on request from the Department of Health) stated that one of the most common invalid diagnoses in paediatrics 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 respiratory disease in children in the Northern region in 1992/1993 by multiplying by the grossing factor (the inverse of the percentage of missing diagnoses).

We have developed a relatively simple method to incorporate the provider effect into Poisson regression small area analyses. A more complex Bayesian method has been developed to incorporate fixed provider effects with other factors (e.g. general practice factors, distance from hospital and ward level deprivation) to model small area variations in hospital admissions to a number of adjacent providers.\(^21\)

**Interpreting provider performance measures**

Performance measures such as surgical mortality rates at different centres will vary for at least four reasons: (1) random variation; (2) the nature and quality of care provided; (3) data quality; and (4) variation in the ‘case mix’. A full investigation of these measures therefore requires proper adjustment for the role of chance and case mix in determining performance in each of the centres carrying out similar procedures. To address the question of whether a centre is a genuine outlier or simply appears extreme as a result of chance, the mortality rate has to be estimated for a typical or average centre based on the rates observed in the remaining centres. This estimate takes the form of a distribution giving the relative probability of possible ‘true’ values for the typical mortality rate. Bayesian hierarchical methods\(^22–24\) can be used to estimate this distribution for procedure and age. The extent to which the distribution of mortality rates in a centre differs from the distribution of mortality rates expected in a typical or average centre reflects the extent to which it may be viewed as a genuine outlier.\(^25\)

**Effect of NHS reforms**

A number of structural changes occurred in the early 1990s, including the 1990 NHS and Community Care Act reforms, introduction of the internal market, the gradual introduction of GP fundholding schemes, hospitals forming provider trusts,
hospital mergers and the use of hospital data for contracting purposes. This undoubtedly had an impact on hospital admissions and may affect studies using these years. For example, some researchers reported an increase in the number of transfers between consultants (each transfer being recorded as a separate episode) from 1991. This in part represented improvements in data collection and changes in practice (e.g. the use of admission wards) but may also have included some ‘episode inflation’. The changes may also have affected GP referral behaviour but assessments of them were hampered by the almost simultaneous introduction of the 1990 performance-related GP contract and the evolutionary nature of the reforms. It is likely that structural changes introduced in 1997 such as the gradual transfer of purchasing to primary care groups will affect both GP referral behaviour and admission rates, but it is difficult to predict the nature and scale of these effects.

**Effect of variations in quality on analysis**

Although overall quality has improved since the early 1990s, some HA-level data show wide variations, particularly in missing diagnostic codes. These variations are even larger at the level of individual trusts. Data quality can be investigated through Data Quality Indicator reports, available from the Department of Health (and sent to all HAs and providers).

Variations in coverage and missing diagnostic codes will generally need investigation in the following circumstances:

1. **Comparison of trusts.** Epidemiological studies may choose to exclude those with high levels of missing data. For example, a study examining national differences in hospital death rates in 1991/1992 to 1994/1995 excluded hospital trusts where more than 30 per cent of primary diagnoses were missing. Exclusion may not be possible in certain circumstances; for example, if comparing trusts for performance management purposes when local knowledge, trust-level data quality reports and liaison with the trust concerned may be needed.

2. **Investigation of time trends using aggregated data from a small number of trusts;** for example, investigation of a disease cluster within an HA or of an environmental hazard. Little has been published on this, but it can be considered as one component of the ‘provider effect’.

Variations in coverage and missing diagnostic codes may not need investigation in the following circumstances:

1. Variations are usually ignored in daily or weekly time series analyses, as it can be assumed that they are unlikely to vary systematically with the exposure variable (for example, daily air pollution levels).

2. Variations are usually ignored when large aggregations of data are used; for example, at national and probably at regional level. In this case, the variations in completeness and in missing diagnostic codes are likely to occur randomly and therefore to cancel each other out.

3. Investigation of apparent disease clustering can be particularly sensitive to systematic variations in quality and completeness of the data. However, data from a single trust should be less prone to systematic differences in coding across time or across small geographical areas.

**Conclusion**

This paper has attempted to give some practical help in the use and interpretation of these valuable but complicated data sources. Hospital activity data always need careful interpretation and may not always reflect the underlying prevalence of disease. Small area studies and comparisons of trusts are particularly vulnerable to bias resulting from variations in quality of the data and the ‘provider effect’. The provider effect is made up of a number of components, but statistical methods have been developed to allow for this in analyses. Quality has definitely improved in recent years, but coverage and completeness should still be investigated where trusts are being compared and in small area spatial analyses using hospital data.

**Acknowledgements**

We thank the staff at SD2 Department of Health for providing information on HES fields, how to access the data, HA boundary changes and data quality reports. We also thank Professor David Strachan and Dr Pat Sturdy (Department of Public Health Sciences, St George’s Hospital Medical School) and Dr Stephen Oliver (Department of Social Medicine, University of Bristol) for comments on earlier drafts of this paper.

**References**


26 Newton J, Goldacre M. How many patients are admitted in districts other than their own and why? J Publ Hlth Med 1994; 16(2): 159–164.


Accepted on 18 August 2000