

The Small Area Health Statistics Unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom

Paul Elliott, Andrew J Westlake, Michael Hills, Immo Kleinschmidt, Laura Rodrigues, Paul McGale, Kay Marshall, Geoffrey Rose

Abstract

Study objective—The Small Area Health Statistics Unit (SAHSU) was established at the London School of Hygiene and Tropical Medicine in response to a recommendation of the enquiry into the increased incidence of childhood leukaemia near Sellafield, the nuclear reprocessing plant in West Cumbria. The aim of this paper was to describe the Unit's methods for the investigation of health around point sources of environmental pollution in the United Kingdom.

Design—Routine data currently including deaths and cancer registrations are held in a large national database which uses a post code based retrieval system to locate cases geographically and link them to the underlying census enumeration districts, and hence to their populations at risk. Main outcome measures were comparison of observed/expected ratios (based on national rates) within bands delineated by concentric circles around point sources of environmental pollution located anywhere in Britain.

Main results—The system is illustrated by a study of mortality from mesothelioma and asbestosis near the Plymouth naval dockyards during 1981-87. Within a 3 km radius of the docks the mortality rate for mesothelioma was higher than the national rate by a factor of 8.4, and that for asbestosis was higher by a factor of 13.6.

Conclusions—SAHSU is a new national facility which is rapidly able to provide rates of mortality and cancer incidence for arbitrary circles drawn around any point in Britain. The example around Plymouth of mesothelioma and asbestosis demonstrates the ability of the system to detect an unusual excess of disease in a small locality, although in this case the findings are likely to be related to occupational rather than environmental exposure.

J Epidemiol Community Health 1992; 46: 345-349

In 1983 a television programme alleged that there was a 10-fold greater than expected incidence of childhood leukaemia at Seascale near the Sellafield nuclear reprocessing plant in West Cumbria.¹ An independent enquiry confirmed this finding, based on five cases accumulated over a 28 year period, and it called for a national organisation

“... to coordinate centrally the monitoring of small area statistics around major installations producing discharges that might present a carcinogenic or mutagenic hazard to the public. In this way early warning of any untoward health effect could be obtained”.² As a result, the Small Area Health Statistics Unit (SAHSU) was established in 1987 at the London School of Hygiene and Tropical Medicine, funded by government departments of health and the environment and the Health and Safety Executive, and with a remit covering the whole of the United Kingdom. The Unit has now completed its initial programme of system development and data acquisition. We describe here the background, purpose and methods of the Unit, and give an example of a first enquiry.

Background and methods

The Sellafield enquiry was constrained by factors which have limited the use of routine health statistics in small area enquiries, particularly that health data are aggregated for reporting purposes to administrative areas (eg, county district). These are usually inappropriate for detailed statistical analysis related to a point source: not only can the area of concern cross administrative boundaries (as with the Sellafield enquiry) but the size of the administrative units is likely to swamp truly local effects.

Subsequent developments, particularly in the use of the post code of residence, have made routine statistics in the United Kingdom available at a lower level of aggregation than was previously possible. A simplified version of the data structure and geography, in which the post code is a key element, is summarised in fig 1; a more detailed description is available elsewhere.^{3,4} The system of post codes, which covers the whole of the United Kingdom, was developed by the Post Office in the 1970s to facilitate delivery of mail. There are 1.6 million unit post codes covering 22 million postal addresses, so that each post code contains on average 14 households. Post codes are now added to routinely collected event data including births, deaths, and cancer registrations.

Small area population data (from the 1971 and 1981 national censuses) are used to give information on the populations at risk (fig 1). These data are published as computerised tables at the level of ward (1971) and enumeration district (1981), with a system of census tracts based on 1981 geography linking areas between the two cen-

Small Area Health Statistics Unit in the Environmental Epidemiology Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom
P Elliott
A J Westlake
M Hills
I Kleinschmidt
L Rodrigues
P McGale
K Marshall
G Rose

Correspondence to:
Dr Elliott

Accepted for publication
September 1991

suses. Similar data for enumeration districts (as well as some aggregate post code data) will be published from the 1991 census. There are approximately 10 000 wards and 130 000 enumeration districts with each of the latter corresponding on average to about 400 persons.

An approximate link between post codes and enumeration districts (which are not coterminous except in Scotland) is achieved via commercial directories. Map coordinates based on the national grid locate "centroids" of census enumeration districts, and similarly coordinates for post codes to 100 m accuracy are available through the *Central postcode directory*. These coordinates and geographical links allow mapping of cases, and computing of their relations to local populations and to "point sources" of interest.

Following the recommendations of the Sellafeld enquiry, one of the Unit's most important tasks is to analyse the routinely available health statistics around specific industrial sites in the United Kingdom, selected *either* on a priori grounds *or* because of a local report of a possible health problem. Using the data structure summarised above, arbitrary small areas can be constructed around the source; then, by linking health events to their populations at risk (via the post code), disease rates can be calculated and compared with expected values derived from national, regional, or more local experience. Statistical power can be increased by aggregating results from a number of similar sites.

Other functions of the Unit include the examination of background local variations of disease rates (hitherto largely unexplored), and the development and application of appropriate database and statistical methodologies. As a national unit, SAHSU also coordinates work with other interested groups within the United Kingdom.

External requests for investigations by the Unit, from governmental or other sources, are directed to a Steering Committee, representing the funding departments and their independent scientific advisers. Once the programme of work is agreed, the Unit proceeds independently and is fully free to publish its findings.

DATA

The Unit is currently able to analyse mortality and cancer incidence data for a range of arbitrarily selected circles around any point of Britain. Post-coded mortality data are available from 1981 in

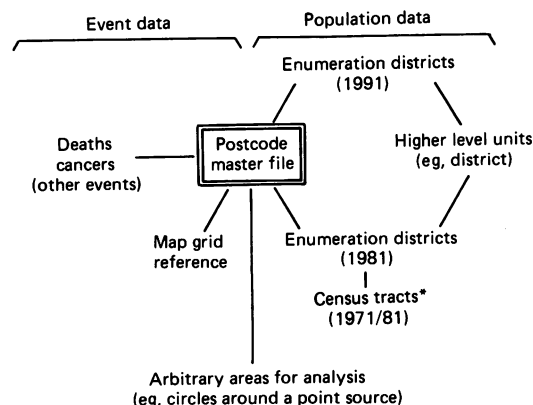


Figure 1 Diagrammatic representation of post code and census geography giving links between health events and populations at risk, based on decennial census

* Based on 1981 enumeration districts

England and Wales, 1974 in Scotland, and 1986 in Northern Ireland. Following a retrospective postcoding exercise by the Medical Statistics Division, Office of Population, Censuses and Surveys (OPCS), postcoded cancer incidence data are also available from 1974 for England and Wales. Similar data from 1975 are currently available for Scotland. A separate file of incidence of childhood leukaemia and non-Hodgkin lymphomas, postcoded from 1966, has been received from the Childhood Cancer Research Group in Oxford. Files of live births and stillbirths for England and Wales from 1981, and congenital malformations from 1983, are also available. (The postcoded file of births will permit accurate year by year denominator data at the level of post code for perinatal and early childhood events.)

As shown in fig 1, denominators for rates are derived from small area statistics from the decennial census. Currently tables are held giving age and sex counts in five year groups for each enumeration district. In addition, SAHSU holds year by year population estimates at district level (from OPCS), estimates of population for 1986 at enumeration district level, and tables giving sociodemographic information derived from census for enumeration districts (eg, car ownership, per cent economically active), which have been obtained from the small area statistics. A file giving a neighbourhood classification for each enumeration district based on census characteristics has also been acquired (ACORN, available from CACI Ltd, 89 Kingsway, London WC2B 6RH). This type of information goes some way to enabling control for demographic and social confounding variables.

SYSTEM DEVELOPMENT

Details of the SAHSU computing system and its development are available elsewhere.^{4 5} Briefly, SAHSU needs to hold a large amount of health, census, and other data, to retrieve relevant data rapidly for any enquiry in the United Kingdom, to analyse the local data, and to compare results with national or more local rates. The volume of data is already about two gigabytes (2000 megabytes) and storage requirements will approximately double when all the currently available data are in place. Indexing of the data (to allow rapid and efficient access to the database) accounts for much of these storage requirements. Thus the computing system requires large storage capacity as well as, for any particular enquiry, rapid and efficient access to small subsets of the data.

The Unit's strategy has been to work within the microcomputer environment, on grounds of cost, accessibility, and the possibility of a system which can be copied by other users. A close working relationship has been established with the environmental monitoring project at OPCS, which is establishing a parallel system. This has involved joint staff appointments and the exchange of data, programs, and computer expertise. The recent development of RISC (Reduced Instruction Set Computer) technology has provided the appropriate mix of computing power and data storage facilities. The SAHSU system currently operates with a DEC 5500 super micro-

computer running under Unix with about 10 gigabytes of disk storage.

The hierarchical structure of the data (fig 1) lends itself to the relational database model, and the system has been implemented using Oracle Version 6.0. Data retrievals are based, for example, on post code or grid reference of the point source, sex, age range, calendar period, ICD codes of interest, and a range of circle sizes. Specific applications routines have been written in the programming language C and they access the database using the standard database language SQL.

To gain rapid access to a particular area for enquiry, the United Kingdom has been divided into a series of one km grid squares which are numbered according to an x-y lattice (eastings and northings). Post codes and enumeration districts are indexed according to the grid square to which they belong. Simple arithmetic computation retrieves all squares needed to cover the circle of interest. The database is then accessed (via the index) to obtain all the post codes or enumeration districts within these squares and to examine those at the boundary to test whether they lie inside or outside the circle. A potentially large data retrieval is thus quickly reduced to a relatively small problem.

At present, the system can produce events, populations at risk, and disease rates for arbitrary circles drawn anywhere in Britain. Some modification is necessary for studies in Northern Ireland which has different census geography. Extension is planned to more complex models (eg, to take account of prevailing winds) and to extended structures such as rivers and coastal areas. This might involve further improvements to the system, using for example links to the geographical retrieval component of a geographical information system.

STATISTICAL METHODS

Statistical analysis is based on the numbers of cases within the bands between adjacent circles, counting the smallest circle as the first band. Expected values are based on age and sex specific national rates. The analysis proceeds in two phases, the first uses a choice of just two bands, made a priori, and the second uses data on all bands.

The two bands chosen a priori in this case were less than 3 km, and between 3 and 10 km. The results for each band are analysed separately, making the assumption that the observed number of cases in the band has a Poisson distribution

with mean λE where E is the expected number for the band, and λ is the true rate ratio.

For the analysis of the data from all bands we make the assumption that the observed number of cases in the i 'th band has a Poisson distribution with a mean $\lambda_i E_i$. The null hypothesis of no increase of relative risk compared to national rates for k bands corresponds to $\lambda_1 = \lambda_2 = \dots = \lambda_k = 1$. Following Stone⁸ this hypothesis is tested against the class of alternatives in which $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$, that is, the risk decreases with distance. The test is based on the ratio of the likelihood when all values of λ are estimated from the data, subject to the restriction that the estimated values do not increase with distance, to the likelihood when all the λ values are equal to 1. The p value for this ratio is obtained by generating sets of observations from the Poisson distribution with means E_1, E_2 , etc, and counting how often the resulting likelihood ratio exceeds the one observed.

The class of alternatives to the null hypothesis includes situations where the rate ratio (λ) declines with distance from the source, and also those where the rate ratio shows no trend but is uniformly greater than (or less than) 1. For this reason we refer to the test as an overall test of departures from 1. In order to test specifically for trend the null hypothesis, $\lambda_1 = \lambda_2 = \dots = \lambda_k = \lambda$ is tested, where λ , the common value of the rate ratio for all bands, is estimated from the data. The p value for the likelihood ratio is now obtained by generating observations as before, but subject to the restraint that the total number of cases is the same as that actually observed.

An example: mesothelioma and asbestosis mortality near the Plymouth dockyards

As a test of this new system, mortality from pleural mesothelioma and asbestosis for 1981-87 was examined around the naval dockyards at Plymouth. In an analysis of mortality from mesothelioma in local authority areas in England and Wales during 1968-78, Plymouth was identified as having one of the highest standardised mortality ratios in the country.⁹ Mesothelioma and asbestosis are known to be related to industrial exposure, and it is also known that dockyard workers were formerly exposed to asbestos. Thus it was likely that any increase in rates seen in the current analysis would represent occupational rather than environmental exposure; nonetheless, the study afforded a useful test of the methods and computing system.

Circles ranging from 1 to 10 km were drawn, according to the preset rules, around the mid-point of the Plymouth docks (Ordnance Survey SX44805610). Deaths from pleural mesothelioma (ICD 163.9, 9th revision) and asbestosis (ICD 501) were accumulated in the circles of interest for the period of study.

Results for men are summarised in table I and fig 2. As anticipated there is a clear increase in risk in the band within 3 km of the docks. The increase is by a factor of 8.41 compared to national rates for mesothelioma (95% confidence interval 5.5-12.3) and by a factor of 13.64 for asbestosis (5.0-29.7). This factor is still appreciably greater than 1 for the band from 3-10 km: 5.35 for mesothelioma

Table I Observed (O) and expected† (E) deaths and observed/expected ratios (O/E) for deaths from mesothelioma and asbestosis by band from Plymouth docks, men, 1981-87

Band (km)	Mesothelioma			Asbestosis		
	O	E†	O/E	O	E†	O/E
0.0-1.0	4	0.36	11.11	2	0.05	40.00
1.0-2.0	14	1.30	10.77	1	0.19	5.26
2.0-3.0	8	1.43	5.59	3	0.20	15.00
3.0-4.9	20	3.78	5.29	4	0.55	7.27
4.9-6.3	5	1.30	3.85	2	0.18	11.11
6.3-7.4	5	0.77	6.49	3	0.10	30.00
7.4-8.3	4	0.58	6.90	1	0.08	12.50
8.3-9.2	3	0.43	6.98	0	0.07	0.00
9.2-10.0	1	0.24	4.17	0	0.03	0.00
0.0-3.0	26	3.09	8.41	6	0.44	13.64
3.0-10.0	38	7.10	5.35	10	1.01	9.90

†Expected values based on England and Wales rates

and 9.90 for asbestosis. Results for the two likelihood ratio tests are shown in table II. The test overall confirms that at least some of the bands have rate ratios greater than 1. This is of little interest in this example because we have already demonstrated that the bands selected a priori have rate ratios considerably in excess of 1, but would be more useful when the results for these bands are equivocal. The second likelihood ratio test (for trend) shows that, although the rate ratio for at least some of the bands is appreciably greater than 1, there is no trend with distance from the source, either for mesothelioma or for asbestosis.

Review of all death certificates with either mesothelioma or asbestosis as underlying cause confirmed a close association with work on the dockyards. In men, of the 64 certificates coded to mesothelioma, 43 (67%) mentioned the naval dockyards; the jobs of several of the remainder were very possibly associated with the dockyards (eg welder, electrical fitter, labourer, lagger). Similarly, 12 out of the 16 certificates (75%) coded as asbestosis mentioned the dockyards.

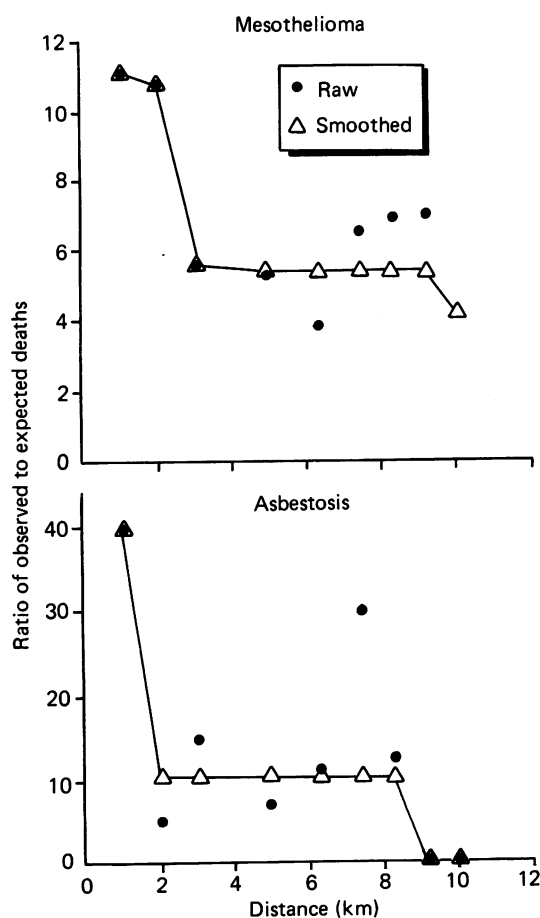


Figure 2 Plot of the ratio of observed to expected deaths, for different bands, by distance from Plymouth docks (men). The smoothed line joins the estimated values of $\lambda_1, \dots, \lambda_9$ subject to the constraint that λ is a decreasing function of distance. Numbers of cases are shown in table I

Table II Likelihood ratio tests for an overall departure from a rate ratio of 1, and a decreasing trend in rate ratio with distance from source

		Mesothelioma	Asbestosis
Overall	likelihood ratio	8.82×10^{28}	2.16×10^{11}
	p value†	0.0000	0.0000
Trend	likelihood ratio	17.46	9.35
	p value†	0.1040	0.1578

†based on 10 000 simulated sets of data

Discussion

The need for appropriate and timely investigation of disease clusters has been stressed.¹⁰ Following the Sellafield enquiry, a major concern was that the increase in childhood leukaemia had not been detected using routinely available data. The advent of postcoding of health events and the ability to build arbitrary areas for analysis using the post code as key have meant that analyses in small areas can be conducted rapidly and with much greater geographical resolution than was hitherto possible. SAHSU is now able to provide rates of mortality and cancer incidence for arbitrary circles drawn around any point in Britain. The example around Plymouth of mesothelioma and asbestosis (an extremely rare cause of death) demonstrates the ability of the system to detect an unusual excess of disease in a relatively small locality.

The use of the database as outlined above will be only the first line in an enquiry. If the initial findings appear to be valid, several important issues of interpretation then need to be considered:

(1) Estimates of statistical significance will have been exaggerated if the reason for the enquiry was a report of a local excess (as at Sellafield), or if the excess came to light as a result of systematic dredging of the database. No formal correction is possible, but some judgement may be made by considering the local results in relation to the size of background variation in disease rates, which the SAHSU system is able to examine. Unless the local excess is extreme, it will need to be confirmed or refuted by extending the enquiry to other similar sites (if such can be found).¹¹

(2) Unless the local effect is large, the study of a single site will lack the power to demonstrate it clearly, leading to either an inconclusive or false negative result. Again, one solution is to aggregate results from a number of sites producing similar environmental discharges.¹¹

(3) The residents around an industrial site are not a random sample of the population, especially if the area suffers from pollution. A true excess of disease may then reflect sociodemographic factors rather than the effects of pollution. Judgements on causality must be cautious; but they may be assisted either by controlling for the confounding effects of local demographic characteristics (which the SAHSU system can undertake), or by testing whether the observed excess is specific rather than shared by other diseases that can be linked with those demographic characteristics.

(4) As in the example set out here, an excess of disease may be primarily occupational rather than environmental in origin.

(5) Local field studies (including environmental measurements) may be necessary in order to test specific hypotheses.

The present system uses death and cancer incidence as measures of health outcome. These reflect the exposure experience of many years previously, and associations may be concealed by industrial changes and population mobility. Recognition of a wider range of health effects, more closely related to current exposures, would require access to postcoded morbidity data such

as hospital discharges or general practice consultations.

In summary, we have described a new system for the conduct of small area epidemiological studies using routinely available data. The system makes possible for the first time at national level a rapid display of disease rates around point sources of environmental pollution, with evaluation of their statistical significance and an ability to control for some demographic confounding factors. Further studies now in hand will help to assess the power of the system to detect relatively small effects, and the extent to which the findings will support causal inferences.

The Small Area Health Statistics Unit is funded by grants from the Department of Health, the Department of the Environment, the Health and Safety Executive, the Scottish Home and Health Department, the Welsh Office, and the Northern Ireland Department of Health and Social Services.

1 Yorkshire Television. "Windscale—the Nuclear Laundry". Broadcast on Independent Television on November 1, 1983.

- 2 Black D. *Investigation of the possible increased incidence of cancer in West Cumbria*. London: HMSO, 1984.
- 3 Carstairs V, Lowe M. Small area analysis: creating a base for environmental monitoring and epidemiological analysis. *Community Med* 1986; 8: 15–28.
- 4 Westlake AJ, Kleinschmidt I. The implementation of area and membership retrievals in point geography using SQL. In: Michaelawicz Z, ed. *Proceedings of the 5th International Conference on Statistical and Scientific Database Management*. Berlin: Springer Verlag, 1990.
- 5 Kleinschmidt I. A database structure and analysis system for small area health statistics. In: Westlake AJ, ed. *Geographical methods in small area health studies*. London: Small Area Health Statistics Unit, London School of Hygiene and Tropical Medicine, 1992.
- 6 Hills M, Alexander F. Statistical methods used in assessing the risk of disease near a source of possible environmental pollution: a review. *J R Stat Soc A* 1989; 152: 353–63.
- 7 Elliott P, ed. *Proceedings of a meeting on the methodology of enquiries into disease clustering*. London: Small Area Health Statistics Unit, London School of Hygiene and Tropical Medicine, 1989.
- 8 Stone RA. Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Stat Med* 1988; 7: 649–50.
- 9 Gardner MJ, Acheson ED, Winter PD. Mortality from mesothelioma of the pleura during 1968–78 in England and Wales. *Br J Cancer* 1982; 46: 81–7.
- 10 Cuzick J, Elliott P. Small area studies: purpose and methods. In: Elliott P, Cuzick J, English D, Stern R, eds. *Geographical and environmental epidemiology: methods for small area analysis*. Oxford: Oxford University Press, 1992 (in press).
- 11 Elliott P, Hills M, Beresford J, et al. Incidence of cancers of the larynx and lung near incinerators of waste solvents and oils in Great Britain. *Lancet* 1992; 339: 854–8.