

A national facility for small area disease mapping and rapid initial assessment of apparent disease clusters around a point source: the UK Small Area Health Statistics Unit

Paul Aylin, Ravi Maheswaran, Jon Wakefield, Samantha Cockings, Lars Jarup, Richard Arnold, Gideon Wheeler and Paul Elliott

Abstract

Background Reports of disease clusters are often received by district health authorities and are, in some cases, associated with concerns about a pollution source. The Small Area Health Statistics Unit (SAHSU) has developed a Rapid Inquiry Facility, which will produce an estimated relative risk for any given condition for the population within defined areas around a point source, relative to the population in a local reference region. The system can also facilitate the production of annual reports and other health studies for Departments of Public Health Medicine through the creation of ward-level maps to illustrate disease variation across small areas.

Methods The facility uses routinely collected morbidity, mortality and population data at a small area scale, together with the computing facilities and expertise necessary to run such analyses quickly and efficiently. Using this facility SAHSU can supply a report within three working days. To aid interpretation, smoothed small area maps that account for sampling variability in the observed data can also be produced.

Results The paper reports on two case studies where the pilot system has been utilized by health authorities for both point source analyses and small area disease mapping.

Conclusions We believe that this facility would be of considerable use to districts. The local knowledge and expertise of the local public health specialist is essential in the interpretation and presentation of the facility's output. Feedback from public health specialists is helping SAHSU refine the output of the facility, so as to make the information presented as comprehensive and as useful as possible.

Keywords: small areas, disease mapping, disease clusters, geographical information systems

Introduction

District public health physicians are usually well placed to respond to reports of disease clusters. The response may include a variety of measures including establishing good rapport with the concerned parties, defining the cluster precisely both

clinically and in time and space, identifying potential sources of environmental pollution causing concern and liaising with other statutory agencies. Apparent disease clusters can cause substantial public anxiety and media interest, and need to be handled effectively.^{1,2}

A component of the response^{3,4} is to establish if the observed numbers for the apparent cluster are greater than would be expected based on the population at risk and on a reference set of disease probabilities. Determining the number of observed cases involves obtaining data on disease events over the study period, which may be several years. The calculation of the expected number of events must account for known risk factors that may include age, sex and socio-economic deprivation. A relative risk (see Methods section) can then be calculated by dividing the observed number by the expected. If a raised risk is found, it will need to be placed in context; e.g. the risk of a specific disease in a small area may be higher than expected when compared with the reference region, but may be lower than that in many other similar areas in the region.

The Small Area Health Statistics Unit (SAHSU) was established to assess the risk to the health of the population from environmental factors with an emphasis on the use and interpretation of routine health statistics.⁵ SAHSU is part of Imperial College and has strong ties with the Centre for

The Small Area Health Statistics Unit, Department of Epidemiology and Public Health, Imperial College School of Medicine, St Mary's Hospital, Norfolk Place, London W2 1PG.

Paul Aylin, Senior Clinical Lecturer in Epidemiology and Public Health

Ravi Maheswaran, Clinical Lecturer in Epidemiology and Public Health

Jon Wakefield, Senior Lecturer in Statistics

Samantha Cockings, Research Associate, Geographical Information Systems

Lars Jarup, Senior Clinical Lecturer

Richard Arnold, Researcher Associate, Statistics

Gideon Wheeler, Software Applications Developer

Paul Elliott, Head, Department of Epidemiology and Public Health

Address correspondence to Dr Aylin.

Environmental Technology within the T. H. Huxley School of Environment. It also has close links with bodies such as the Environment Agency and the Water Research Centre, from which data on specific exposures are obtained. This paper describes two examples of the output of a Rapid Inquiry Facility based at SAHSU, which was initially developed for the government funding departments to rapidly assess the risk to health around a point source. In addition, the facility will produce small area maps of disease distributions in user-specified spatial and temporal regions, which can be incorporated into annual health reports and other health studies.

Methods

SAHSU holds national cause-specific data on deaths (currently 1981–1997), on births (1981–1997), on cancers from the national cancer registry⁶ (1974–1992), on hospital admissions (1992–1997), and on congenital anomalies (1983–1997), using the postcode of residence to locate cases to within 10–100 m. In 1996, there were around 1.4 million residential postcodes in use in the UK containing, on average, 17 households each. SAHSU also holds a range of geographical, socio-economic and environmental data, all of which are geographically referenced. Using in-house database, statistics and geographical information systems technology and expertise, these datasets are integrated, analysed and displayed.

The system consists of a network of Sun⁷ Sparc servers. There are two Sun Sparc 20s each with four 200 MHz Hypersparc CPUs, 320 MB RAM and 100 GB disk arrays. These machines support two Oracle⁸ databases containing the principle datasets. A third Sun Sparc server runs the ARC/INFO⁹ geographical information system (GIS).

In general, each of the small areas of interest will have an associated set of disease risks for each age and sex stratum. The aim then is to summarize how these risks differ from those of a comparison region. A simple method of summarization is the standardized mortality or morbidity ratio (SMR), which is a measure of the quantity by which each of the reference risks is

multiplied to obtain the small area risks of interest. Consequently, the SMR is measuring the relative risk of disease and may be calculated by dividing the observed number of health events by the expected number based on the reference risks (see the Appendix for more details). Within the SAHSU Rapid Inquiry Facility these risks are calculated currently using the UK Standard Region that contains the study area as the reference. Other reference regions may be defined and used. The population data that are required for both the study and reference regions are available at the 1991 Census Enumeration Districts (ED) level, with an average of around 440 people per ED. The SMR follows from 'indirect' standardization.¹⁰ There is also the facility to calculate 'directly' standardized mortality ratios, although because of instability from small numbers, they have some practical limitations.¹¹

As well as the age and sex distribution of the area it is also important to consider the distribution of relative deprivation, as this has been shown to be a powerful predictor of ill-health.¹² The Carstairs index¹³ is a small area deprivation measure that has been shown to be strongly predictive of mortality and cancer incidence. The index is derived from Census statistics on overcrowding, access to a car, unemployment and social class of head of household, and is calculated at the ED level. Within the Rapid Inquiry Facility the Carstairs quintile of each ED is used to adjust disease risks for this possible confounder.

For estimating the risk surrounding a point source, concentric circles (usually of radius 2 km and 7.5 km) are drawn around the source (specified either as a postcode or an Ordnance Survey National Grid Reference) and EDs with their population weighted centroid falling within the bands are included in the study area. Two or more point sources may be combined in a single study area by selecting EDs that fall within a specified distance from any of the sources. For large industrial areas or for irregular boundary definitions, study areas can also be defined from a list of EDs or wards. The risk is then calculated for the EDs contained within the bands. To account for sampling variability 95 per cent confidence intervals for these risks are also calculated. A description of the SAHSU approach to analysis of data around point sources and more

Table 1 Populations, observed and expected counts and standardized mortality and morbidity ratios after standardization for age, sex and deprivation from Rapid Inquiry Facility output

| Area | Population | Observed cases | Expected cases | Relative risk (standardized mortality and morbidity ratios) | 95% CI |
|---|------------|----------------|----------------|---|-----------|
| <i>Mortality from respiratory diseases</i> | | | | | |
| Within 2 km | 71495 | 1196 | 1120.6 | 1.07 | 1.01–1.13 |
| 2–7.5 km | 783943 | 13970 | 12765.8 | 1.09 | 1.08–1.11 |
| <i>Hospital admissions for respiratory diseases</i> | | | | | |
| Within 2 km | 71495 | 2834 | 2867.2 | 0.99 | 0.95–1.03 |
| 2–7.5 km | 783943 | 30827 | 29239.7 | 1.05 | 1.04–1.07 |

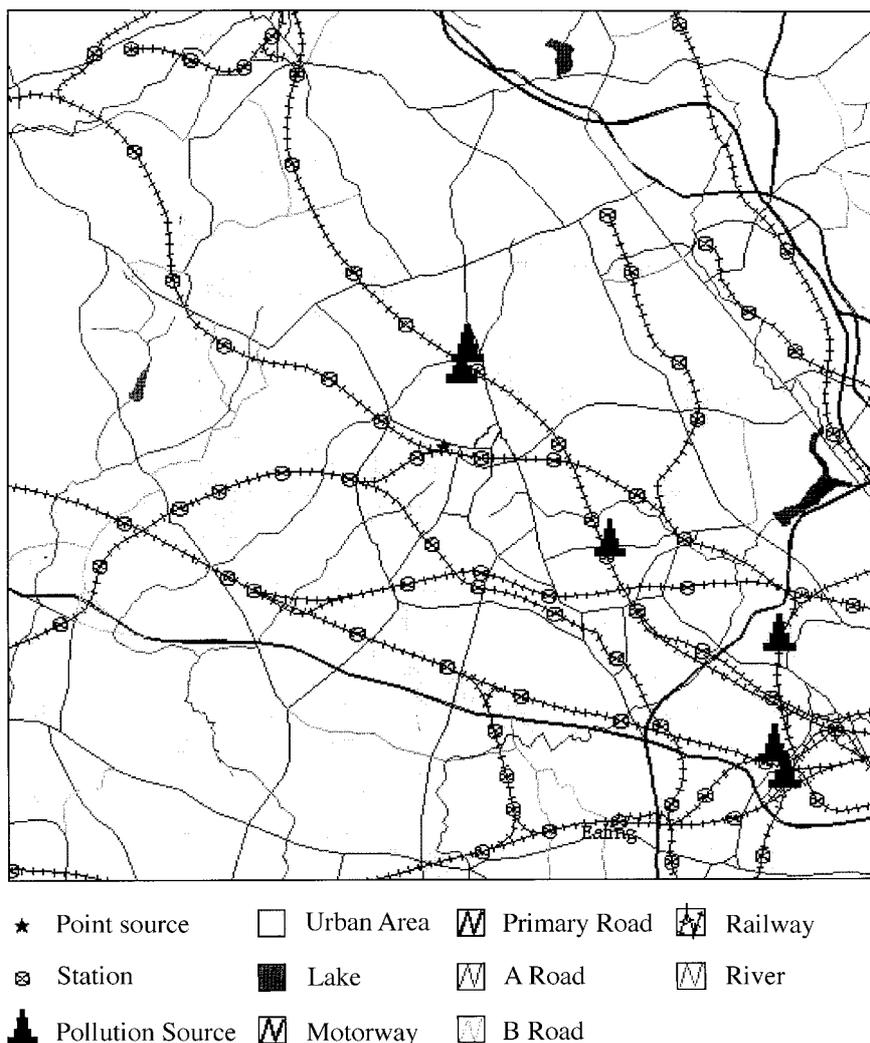


Figure 1 Example of map to illustrate location of the point source.

general geographical epidemiology can be found in the Appendix and elsewhere.^{14,15}

For small area disease mapping, large differences in health risk between small areas may arise simply by chance, even when several years of data are used. This is particularly true when the numbers of cases are very small (for example, typically, an electoral ward will have fewer than ten deaths from heart disease in the under-75s per year). In the Appendix a statistical smoothing technique that may be used to stabilize the ward rates is described. The smoothed estimates can provide more stable estimates of the 'true' ward relative risks than the raw SMRs.

Case studies

To pilot the facility, both a point source investigation and a disease mapping exercise were carried out in collaboration with two health authorities.

Case 1: a point source investigation

SAHSU responded to a request from the Department of Public Health Medicine at Barking and Havering Health Authority. The local MP had expressed concerns regarding a complaint from a constituent about chemical air pollution near two factories on the same site in a deprived area of the district. Local GPs had been contacted and no changes in illness patterns had been noticed in the area. Environmental health officers had inspected the factories and produced a detailed report. The Environment Agency had also been involved. Both factories had been operating for 16 years. Local mortality data had been analysed but no unusually high rates had been apparent. However, respiratory mortality rates were generally high in that part of the district. As the complaint was specifically about respiratory illness, the Department of Public Health decided to focus on respiratory admissions and mortality in the vicinity of the two factories. The co-ordinates for the location of the factories were provided to SAHSU.

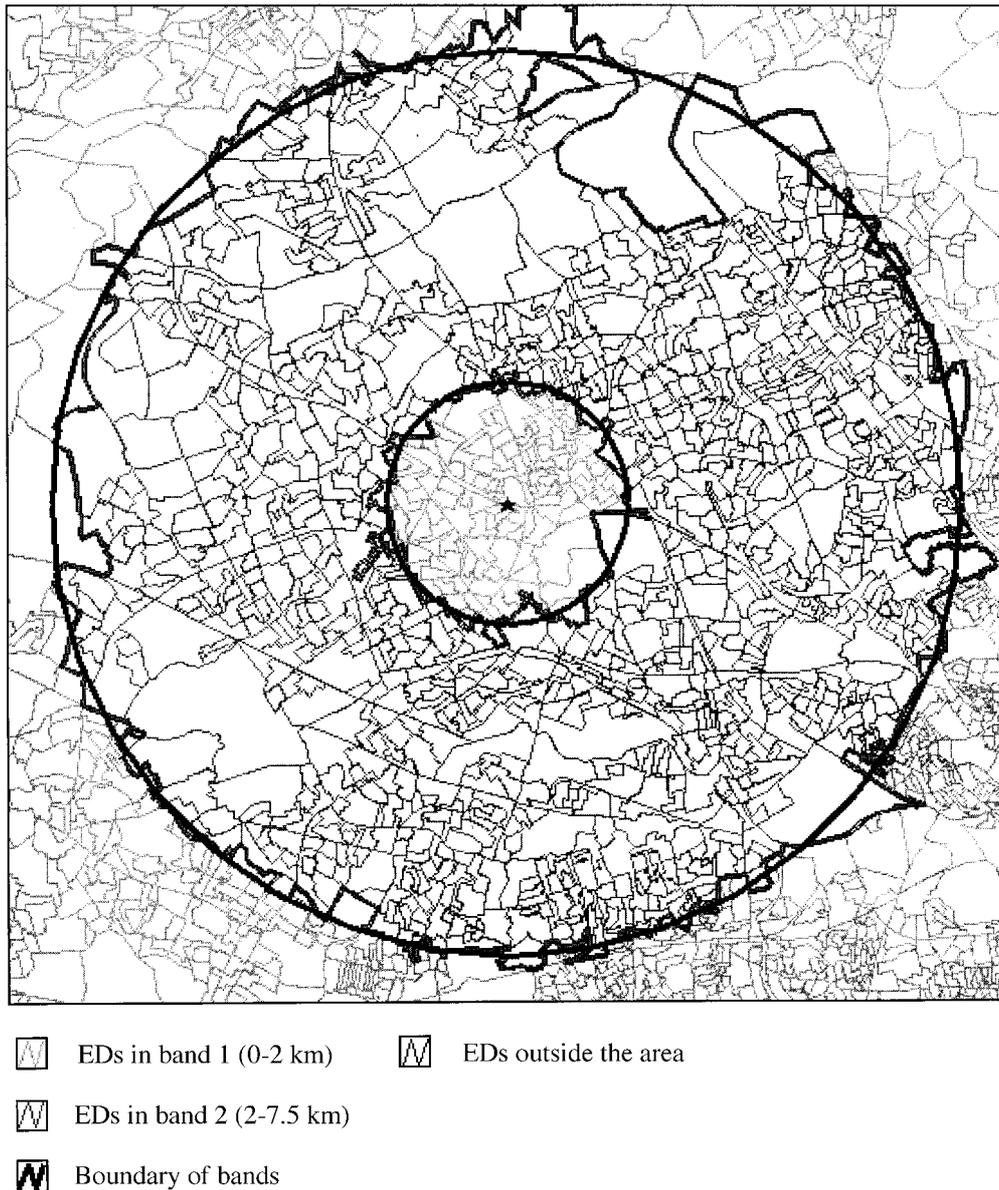


Figure 2 Example of map to illustrate the concentric bands used to calculate relative risk.

After standardizing for age, sex and deprivation, mortality from respiratory diseases appeared to be marginally raised in the immediate vicinity (i.e. within a 2 km radius) of either of the two factories (relative risk 1.07, 95 per cent confidence interval (CI) 1.01–1.13). However, the excess seemed to be similar to that seen in the area 2–7.5 km from the factories (relative risk 1.09, 95 per cent CI 1.08–1.11). This suggested that the risk within 2 km might be a reflection of excess risk in the whole area, relative to areas with similar populations and deprivation in the South East Standard Region, rather than being specifically related to the two factories. There did not appear to be any increase in hospital admissions for respiratory diseases in the immediate vicinity of the two factories (relative

risk 0.99, 95 per cent CI 0.95–1.03). However, there was a small but significant raised risk of hospital admission for respiratory diseases 2–7.5 km from the two factories (relative risk 1.05, 95 per cent CI 1.04–1.07). None of the excess risks observed was greater than 10 per cent. Table 1 shows the populations and the number of observed and expected cases within each concentric band.

A report was provided to the district department, which included details of the request, types of data used, the time frame, the conditions investigated, age groups studied, details of the standardization, the geographical areas, the results, a brief commentary and an outline of the limitations of the analysis. Maps were also included to show the location of the

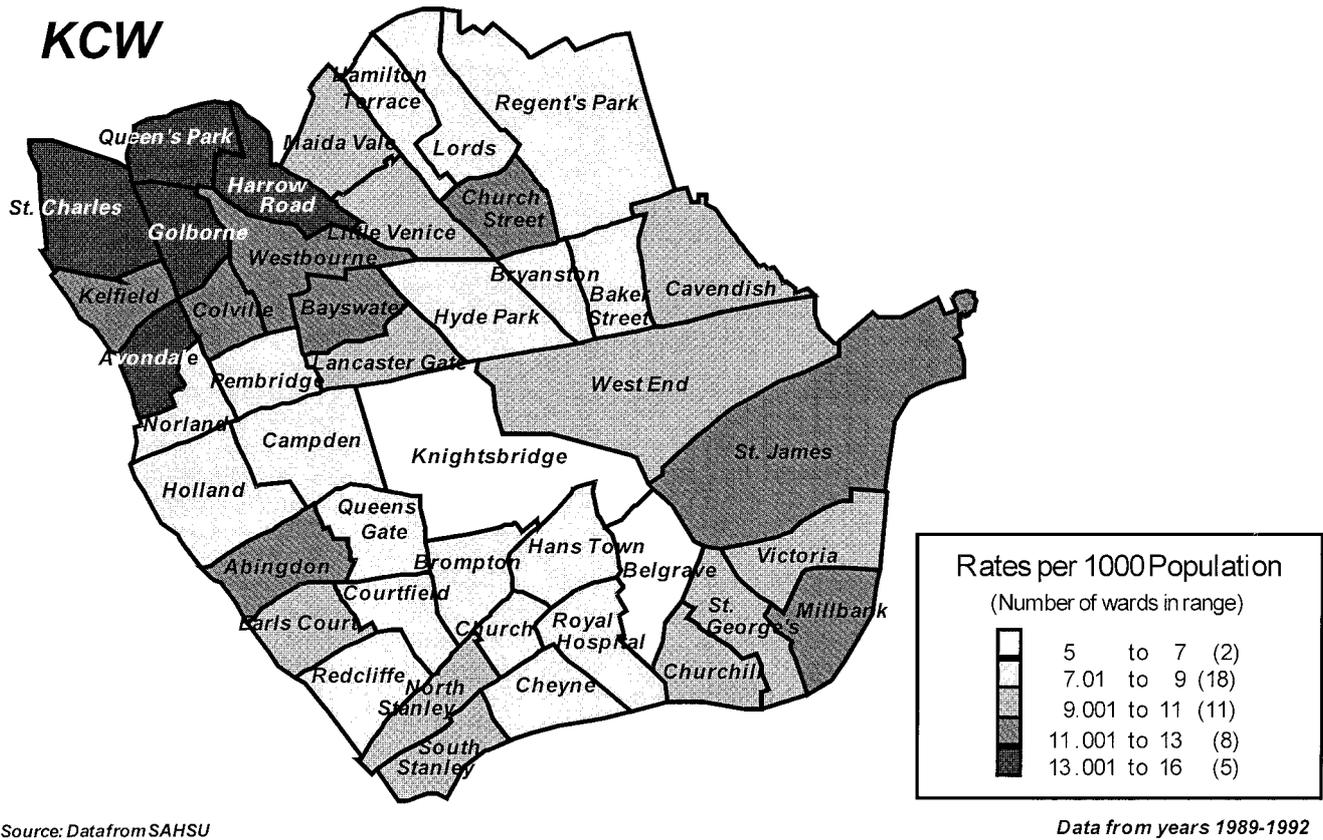


Figure 3 Deaths from all causes by ward – males and females all ages (directly standardized).

point source, the concentric bands used to calculate relative risk and the pattern of deprivation in the area. Examples of the kind of maps that may be produced routinely by SAHSU are given in Figs 1 and 2.

Case 2: small area disease mapping

SAHSU responded to a request from the local Department of Public Health Medicine at Kensington, Chelsea and Westminster (KCW) Health Authority to provide ward-level maps of disease variation across the district. One of the main topics of the Annual Public Health Report was to address health inequalities within the district. Maps for specific conditions were required. Because of the newly developed Rapid Inquiry Facility, SAHSU was in a unique position to provide quickly the required tables and maps.

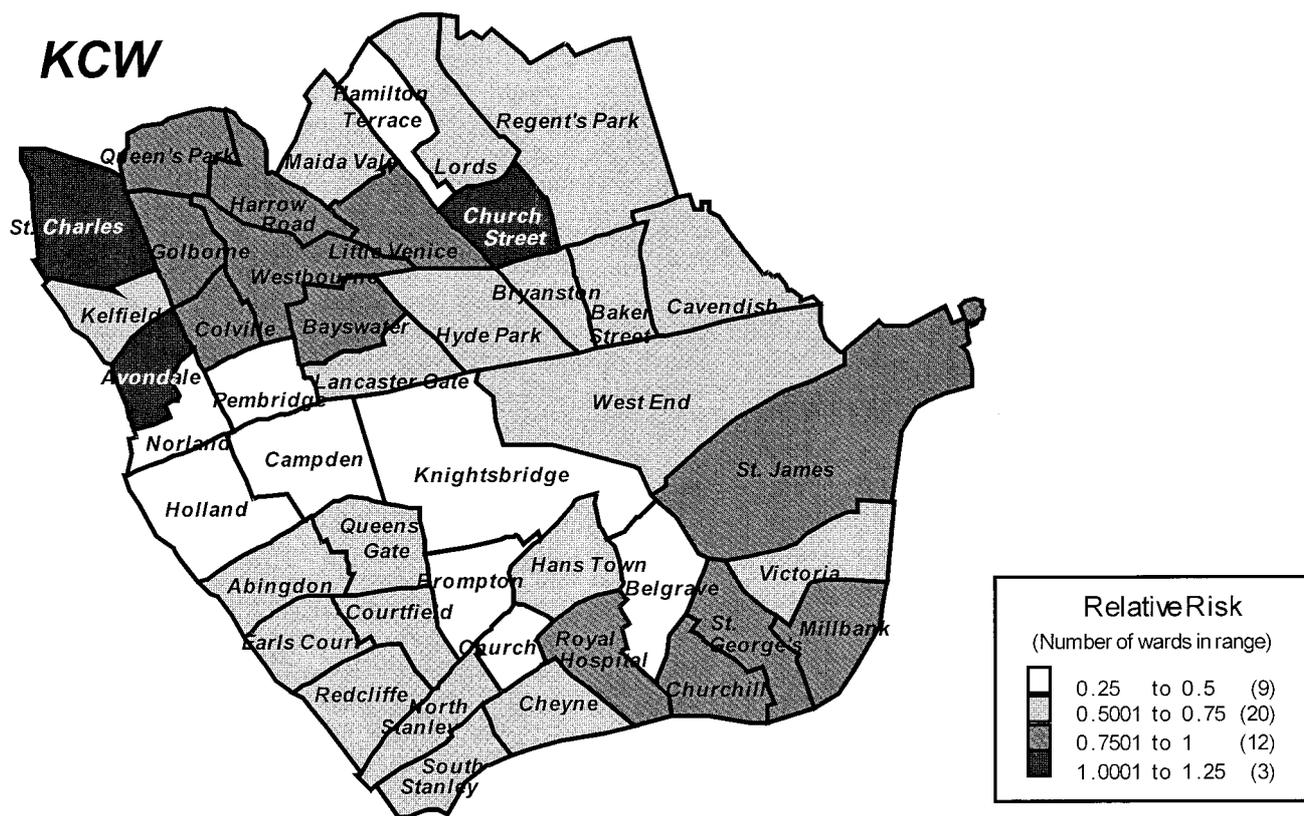
The maps showed the distribution of all-cause death rates to be variable across the health district. Figure 3 displays directly age-standardized mortality rate (see Appendix) calculated for all causes and in all ages for the four years 1989–1992. The data were standardized to the district population. The figure shows that, for the particular standardization used, the rates in the Queens Park, St Charles, Golborne, Avondale and Harrow Road area are around three times those in Knightsbridge and Belgravia.

Figure 4 displays estimates of the relative risks of death from heart disease by ward, for males under the age of 75 only. These estimates have been smoothed using the techniques described in the Appendix.

Figure 5a shows the five wards with the highest levels of deprivation as measured by the Carstairs index. Alongside this, Fig. 5b–d shows the wards with the highest death rates from all causes, heart disease and lung cancer, respectively. There is a striking similarity between these figures. These findings were incorporated into the Public Health Annual Report and informed its recommendations.¹⁶

Discussion

The two examples above illustrate the potential utility of the Rapid Inquiry Facility for district Departments of Public Health Medicine. In general, health districts only routinely have access to health data across the district as a whole or, in some instances, at the level of electoral ward within the district. Even data at the ward level is often at too coarse a resolution to investigate localized pollution sources, for example, incinerators or factories. The production of ward maps, with rates adjusted for age, sex and deprivation and also statistically



Source: Data from SAHSU

Data from years 1989-1992

Figure 4 Deaths from heart disease by ward – males under 75 years (smoothed).

smoothed is another service that most districts would find hard to replicate in-house.

Although the potential outputs of this facility would appear to be valuable to public health doctors, there are limitations to be considered.

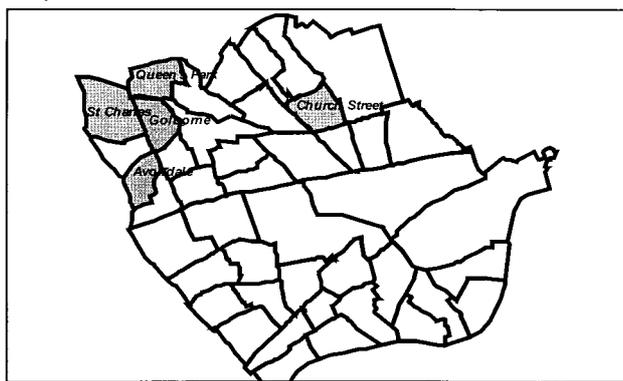
The source data for the facility are supplied from routinely collected national data, therefore factors such as diagnostic and geographical completeness, accuracy and timeliness need to be considered. Mortality data are known to suffer from inaccuracy of cause-specific diagnosis, particularly in the elderly.¹² The completeness of cancer registration may vary from registry to registry, from cause to cause and over time. A small proportion of cases are diagnosed but not registered. In addition, some people develop cancer but are never diagnosed. Under-registration may be non-random, especially at the local level. Difficulties can arise where patients cross regional boundaries to receive treatment. It is important that these factors are considered, especially if the disease is rare.⁴ At least two sources of potential bias are present in using hospital data for this type of epidemiological study. First, there are differences between admissions policies and coding practices between different provider units.¹⁷ Second, there are differing referral policies between general practitioners.¹⁸ The local expertise of the Department of Public Health will be valuable in determin-

ing the extent to which these factors are likely to be operating in a particular area. Further work is needed to investigate the extent to which such local factors affect estimates of risk.

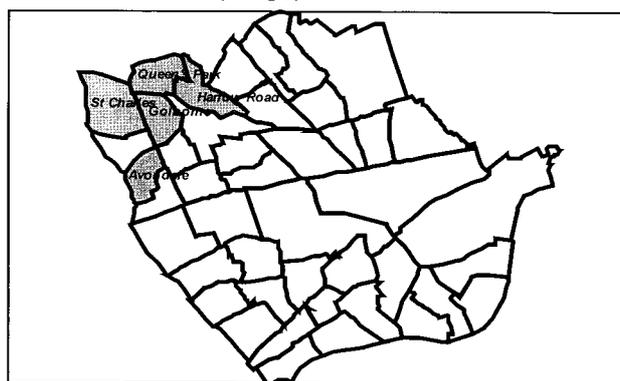
The way in which areas are grouped into concentric circles around point sources is oversimplified and may not be the best representation of exposure around emission sources. Further analysis incorporating prevailing winds or utilizing pollution monitoring data or atmospheric dispersion modelling¹⁹ would help to define groups of areas according to pollution concentrations. This may provide a better estimate of the geographical exposure patterns surrounding such point sources.

Many point sources of pollution are located in deprived areas. Although the effect of deprivation on health can be adjusted for using standardization, the possibility of residual confounding remains. Dolk *et al.*²⁰ concluded from their study of mortality around cokeworks that the effects of deprivation and region 'explained' 12 per cent of the observed 15 per cent excess of mortality, leaving only 3 per cent excess mortality related to residence near a cokeworks. The authors concluded that residual socio-economic confounding was a strong candidate to also explain the remaining 3 per cent. There may be additional sources of pollution close to the study area, which may also contribute to an increased relative risk of disease.

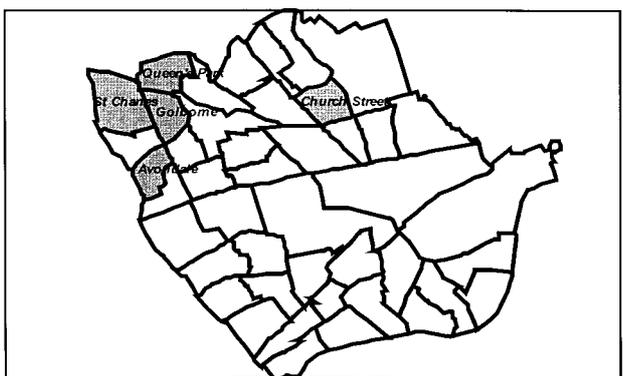
A. Deprivation - Carstairs Score



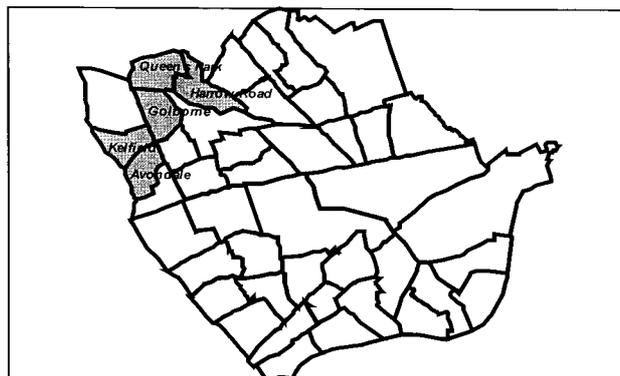
B. All Cause Death Rates (All Ages)



C. Death from Heart Disease (Under 75s - Smoothed)



D. Death from Lung Cancer (All Ages - Smoothed)



Source: Data from SAHSU

Data from Years 1989-1992

Figure 5 Wards with the highest levels of deprivation, death rates, death from heart disease and death from lung cancer.

Again, local knowledge can be useful in interpreting the estimates provided by the analysis. An additional problem is the accuracy to which postcodes are mapped to EDs. This may affect the deprivation score assigned to a health event and the calculation of the relative risk.²¹

A further problem is the use of ED centroids to represent the spatial distribution of the population at risk. It is assumed that all of the population within the ED lies at the geographical point location of the centroid. If that centroid falls within the band, the entire population of the ED is allocated to the band, when in reality, some of the population may lie outside the band. Equally, populations may be wrongly excluded if their centroid does not fall within the band. On average, these errors may be expected to cancel out, but for small area studies, the inclusion or exclusion of one or two EDs could potentially have a large effect on the final analysis.

One of the major problems with cluster investigation is the choice of boundaries used to define a cluster. Often a group of cases is identified before defining spatial or temporal boundaries. When boundaries are then drawn to include these cases, the denominator population is also included within these boundaries. The tighter the boundaries around the cluster, the higher the risk will be relative to a comparison population. This has been described as the 'Texas sharpshooter' effect, whereby

a sharpshooter first empties his gun into a barn door and then draws a target around the bullet holes.²² As a way to minimize the effect of boundary shrinkage the Rapid Inquiry Facility routinely uses *a priori* standard 'near' and 'far' bands of 0–2 km and 2–7.5 km surrounding a putative cluster of cases around a point source. Although arbitrary, the bands have been used in earlier SAHSU studies and, in general, achieve a useful compromise between population size and proximity to the point source.

Further problems occur in estimating the population at risk. The population data upon which the summary risk estimates are based are subject to inaccuracies as they arise from the Census. The Census provides only a snapshot view of the population every 10 years and may not reflect population changes between Censuses. The 1991 Census was most notably subject to the problem of under-enumeration,²³ which could inflate risk estimates, especially at younger ages. We address this latter problem by using the adjusted 1991 Census counts from the 'Estimating with Confidence' project.²⁴ Small area level populations for non-Census years are then calculated as follows. An initial estimate of the population in each ED is made for the years between 1981 and 1991, by interpolating between the two Censuses, whereas for years after 1991, the 1991 populations are used. These initial estimates are then

rescaled in proportion to the annual local authority district populations.

All these limitations are highlighted in the standard report produced by the facility. It is important to recognize that the Rapid Inquiry Facility is only part of the scientific investigation of clusters and disease inequalities. The facility for rapidly producing maps and analyses will be offered to Departments of Public Health Medicine only where careful consideration by a local public health specialist will enhance the value of these outputs and indeed will be essential in interpretation and presentation of the findings. We believe that this facility would be of considerable use to districts. Feedback from public health specialists is helping SAHSU to refine the output of the facility, so as to make the information presented as comprehensive and as useful as possible.

Contributors

P.E., P.A. and L.J. were involved in the conception and management of the project. S.C., J.W., P.A., R.M., R.A., G.W. and L.J. were involved in the development of the project, with S.C. taking responsibility for the GIS aspect and J.W. and R.A. overseeing the statistics. P.A. and R.M. prepared the first draft of this paper, and L.J., S.C., J.W., R.A., G.W. and P.E. contributed to the final submitted version. J.W. also prepared the statistical appendix. P.A. is the guarantor.

Acknowledgements

We would like to thank Dr Kishor Padki, Consultant in Public Health Medicine at Barking and Havering Health Authority, and Dr Dorothy Gregson, Consultant in Public Health Medicine at Kensington, Chelsea and Westminster Health Authority, for their valuable comments in preparing this paper and for granting us permission to use their SAHSU enquiries as examples in this paper. The Small Area Health Statistics Unit is funded by a grant from the Department of Health, Department of the Environment, Transport and Regions, The Health and Safety Executive, the Scottish Office, the Welsh Office and the Department of Health and Social Services (Northern Ireland). We are grateful to the Office for National Statistics and the ESRC for provision of and permission to use their data. This work was supported, in part, by an equipment grant from the Wellcome Trust. The views expressed in this publication are those of the authors and not necessarily those of the funding departments.

References

- 1 Fiore BJ, Hanrahan LP, Anderson HA. State health department response to disease cluster reports: a protocol for investigations. *Am J Epidemiol* 1990; **132**(1 Suppl.): S14–S22.
- 2 Olsen SF, Martuzzi M, Elliott P. Cluster analysis and disease mapping—why, when, and how? A step by step guide. *Br Med J* 1996; **313**: 863–866.
- 3 Maheswaran R, Staines A. Cancer clusters and their origins. *Chem Ind* 1997; **7**: 254–256.
- 4 Leukaemia Research Fund. *Handbook and guide to the investigation of clusters of diseases*. Leeds: Centre for Clinical Epidemiology, University of Leeds, 1997.
- 5 Elliott P, Westlake AJ, Kleinschmidt Hills M, *et al*. The Small Area Health Statistics Unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom. *J Epidemiol Commun Hlth* 1992; **46**: 345–349.
- 6 National Cancer Registration Bureau. B6/02. London: Office for National Statistics (various years).
- 7 <http://www.sun.com/>
- 8 <http://www.oracle.com/>
- 9 <http://www.esri.com/>
- 10 Breslow N, Day NE. *Statistical methods in cancer research, volume II—the design and analysis of cohort studies*. Scientific Publications, 82. Lyon: International Agency for Research on Cancer, 1987.
- 11 MacMahon B, Trichopoulos D. *Epidemiology, principals and methods*, 2nd edn. Reading, MA: Little, Brown, 1996: 55.
- 12 Jolley D, Jarman B, Elliott P. Socio-economic confounding. In: Elliott P, Cuzick J, English D, Stern R, eds. *Geographical and environmental epidemiology: methods for small-area studies*. Oxford: Oxford University Press, 1982: 115–124.
- 13 Carstairs V, Morris R. *Deprivation and health in Scotland*. Aberdeen: Aberdeen University Press, 1991.
- 14 Elliott P, Martuzzi M, Shaddick G. Spatial statistical methods in environmental epidemiology: a critique. *Statist Meth Med Res* 1995; **4**: 149–161.
- 15 Wakefield J, Elliott P. Issues in the statistical analysis of small area health data. *Statist Med* (in press).
- 16 *Annual public health report 1997. Inequality in health*. London: Kensington, Chelsea and Westminster Health Authority, 1998.
- 17 Dixon J, Sanderson C, Elliott P, *et al*. Assessment of the reproducibility of clinical coding in routinely collected hospital activity data: a study in two hospitals. *J Publ Hlth Med* 1998; **20**: 63–69.
- 18 Hippisley-Cox J, Hardy C, Pringle M, *et al*. The effect of deprivation on variations in general practitioners' referral rates: a cross sectional study of computerised data on new medical and surgical outpatient referrals in Nottinghamshire. *Br Med J* 1997; **314**: 1458–1461.
- 19 ADMS 2 The Multiple Source Air Dispersion Model. Cambridge: Cambridge Environmental Research Consultants.
- 20 Dolk H, Thakrar B, Walls P, *et al*. Mortality among

- residents near cokeworks in Great Britain. *Occup Environ Med* 1999; **56**: 34–40.
- 21 Collins SE, Haining RP, Browns IR, *et al.* Errors in postcode to enumeration district and their effect on small area analyses of health data. *J Publ Hlth Med* 1998; **20**(3): 325–330.
- 22 Rothman KJ. A sobering start for the cluster busters' conference. *Am J Epidemiol* 1990; **132**(Suppl.): S6–S13.
- 23 Victor C. Underenumeration in 1991 census: forms not retrieved. *Br Med J* 1993; **307**: 1564.
- 24 Simpson S, Cossey R, Diamond I. 1991 population estimates for areas smaller than districts. *Population Trends* 1997; **90**: 31–39.
- 25 Clayton D, Kaldor J. Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics* 1987; **43**: 671–681.

Accepted on 30 March 1999

Statistical appendix

In what follows we assume that the health endpoint of interest is rare and that cases occur independently of each other. In this case the starting point for analysis is that the number of cases follow a Poisson distribution. In general, we suppose we have n areas in our study region of interest. These areas may be wards, EDs or artificially created areas around a location of interest (for example, a putative source of pollution). In area i , stratum j and in year t we have populations, N_{ijt} . We suppose there are J strata, and T years in the study period. These strata are defined by factors upon which it is known disease risk depends, for example, age and sex. We denote by p_{ijt} the risk of disease in area i , stratum j and in year t . Let us suppose we wish to compare the areas of interest with a reference region within which the risk of disease in stratum j and in year t are given by p_{jt} . The aim then is to estimate the relative risk in each of these areas, that is, the ratio of the risks in the area divided by the risks in a reference region p_{ijt}/p_{jt} . Hence a relative risk of two implies that each of the risks are doubled in the area, compared with the reference region. We denote the observed number of disease cases in area i , summed across strata and the study period, by O_i .

The expected number of cases in area i is calculated from the expression

$$E_i = \sum_{j=1}^J \sum_{t=1}^L N_{ijt} p_{jt}$$

Given the observed and expected numbers in area i we may simply calculate an estimate of the relative risk (SMR) via $R_i = O_i/E_i$.

As an alternative to producing maps of the relative risks of each area we may also obtain a map of a directly standardized

rate. We first emphasize that, as was noted when SMRs were introduced, the use of a single number to summarize the set of risks in a given area will rarely completely represent the true risks and may in some instances be misleading.¹⁰ First, we let M_{jt} denote the size of the population in stratum j and year t in a reference area, and let

$$M = \sum_{j=1}^J \sum_{t=1}^L M_{jt}$$

denote the total population. So the M_{jt} together provide a reference population. Again, we let p_{ijt} be the risk of disease in small area i , stratum j and in year t . These probabilities may be estimated by O_{ijt}/N_{ijt} , where O_{ijt} and N_{ijt} represent the number of cases and (as before) the population at risk in area i , stratum j and in year t , respectively. The directly standardized rate for area i , which we denote S_i , is then produced by applying the area-specific estimated risks to the reference population and then dividing by the total population, i.e.

$$S_i = \sum_{j=1}^J \sum_{t=1}^L M_{jt} p_{ijt} / M = E_i^* / M$$

The directly standardized rate is the number of cases that would be expected in the reference population E_i^* if the risks in this population were the same as in the small area of interest i . An important point to note is that the quantity S_i may be highly unstable, as the estimates of the area-specific probabilities p_{ijt} will themselves be unstable. Hence low and high directly standardized rates may simply reflect sampling uncertainty. The instability in the SMRs is likely to be lower, and for the SMR there is a relatively simple way, which we now describe, by which the estimates may be made more robust.

The variance of the SMR is proportional to $1/E$. Unfortunately, therefore, for small areas in particular, the relative risk estimates are likely to be highly unstable because they are based on small numbers. This manifests itself in the areas with the smallest populations producing the highest (and lowest) relative risk estimates by chance alone. As areas with small populations are often geographically large this can lead to maps that are misleading. Hence we would like to make our estimates more robust by the incorporation of additional information. This is achieved via what is known as multilevel modelling. This technique is the standard in disease mapping.²⁵

The basic rationale behind the method is the following. We acknowledge that although we expect regional rates within a region to vary, we do not expect them to be very dissimilar. One way of modelling this belief is to assume that the rates within the study region are a sample from a probability distribution. If this distribution has the bulk of its mass in a small range then we would observe that the rates did not vary greatly in our study region. We choose the gamma distribution because it is defined for positive quantities and is skewed to the right (which has empirically been found to mimic relative risks across small areas), and because it is computationally convenient.²⁵

We now turn to the explicit form of our study estimates. The gamma distribution has two parameters which we shall denote α and β . The mean of the distribution is given by α/β and the variance by α/β^2 . As $1/\beta$ is equal to the variance divided by the mean it is a measure of the scale of the distribution. That is, if $1/\beta$ is large then the distribution of rates has a large spread. Therefore different choices of α and β can reflect a range of distributional shapes and locations. With the assumption that the rates arise from a gamma distribution with parameters α and β the smoothed estimates of the relative risks are given by

$$R_i^* = w_i R_i + (1 - w_i) \alpha / \beta$$

where $R_i = O_i/E_i$ is the raw relative risk and $0 \leq w_i \leq 1$ is a weight that is given by

$$w_i = E_i / (E_i + \beta)$$

with

$$1 - w_i = \beta / (E_i + \beta)$$

Hence we see that the smoothed risk estimate is a compromise between the raw SMR and the mean of the risks across the region as a whole. The raw SMR dominates if E_i is large compared with β , whereas for areas with small populations (which result in small expected numbers) the overall mean dominates. The above discussion has assumed that the parameters of the gamma distribution, α and β , are known. In practice, this is not true, but we can estimate them using the data from all regions. We achieve this using a technique known as the Empirical Bayes method.²⁵ The Empirical Bayes relative risk estimates are less dispersed than the SMRs as they have been smoothed.