Investigation of disease risks in small areas

Paul Elliott

Abstract
The investigation of disease risks in small areas is complicated by many issues including data quality, the retrospective nature of the statistical testing, the problems of boundary definitions in time and space around a putative disease cluster, and the lack of generally accepted definitions of the key terminology. Routine data systems have disease surveillance in an initial investigation of disease risks near sources of environmental pollution, although problems of data analysis and interpretation remain. This is especially true of unmeasured socioeconomic confounding, which could generate apparent positive results near a pollution source.

Keywords: small areas; disease risks

Public health physicians are becoming increasingly involved in environmental health issues beyond the traditional and well defined role of outbreak investigation of communicable disease. The focus has shifted to the non-communicable diseases, especially cancers, in response to immense public concern about environmental issues, and, in particular, the possible health effects of industrial pollution. Thus public health authorities may be faced with an unsuspected and previously unreported “cluster” of disease, or the need to respond to reports of unacceptable levels of pollution from a local industrial source.

Allegations of a putative cluster may have been made by concerned members of the public, or have been reported first by an ever alert media. How should the public health physician respond? Is there a duty to be proactive—that is, undertake disease surveillance in an attempt to identify and quantify potential environmental health problems locally?

Investigation of apparent disease clusters
In response to an alleged cluster of disease, often understanding and simple reassurance is all that is required, as many putative clusters do not stand detailed scrutiny. For rare diseases in small areas, an apparent disease cluster may depend crucially on only one or two cases. What at first sight seems to be cause for concern may evaporate once case histories and diagnoses are checked, duplicates—for example, of cancer registrations—are removed, recent migrants, possibly with a previous history of the disease in question, are identified, and other anomalies detected during an initial case by case review.

Such an investigation of course deals only with the numerator (cases), whereas a proper epidemiological enquiry is concerned equally with the denominator (population). The next steps will therefore usually involve identification of a population at risk and specification of a time frame so that disease rates can be calculated. Some difficulties now arise—what should be the geographical extent of the enquiry? Are relevant population data available, especially at the small area level? Are the population data accurate and up to date? Has there been recent migration into or out of the area that could distort population estimates? What standard rates should be used and are they readily available—in an accessible form, and in timely manner? What time period should be studied?

These questions are crucial as apparent clusters can be made to disappear or reappear simply by loosening or tightening the boundaries around the putative cluster (in time, space, or both time and space), by over or underenumeration of the population at risk, or by choice of different sets of standard rates. Often decisions are implied rather than openly acknowledged as they are dependent on existing data; only a limited data window may be available for analysis. Matters are complicated further if an apparent cluster of cases is then linked to a putative pollution source. Formal statistical testing in such circumstances is rendered invalid given the retrospective nature of the observation, but in any case, testing for departure from some Poisson process in small area analyses is fraught with difficulty. For example, the area surrounding the initial observation can theoretically be thought of as being drawn from a much larger set of small areas—for example, covering the whole country—some of which will also have high disease rates (while others will have low rates).

One of the problems that has beset work of this type is that there have been no generally acceptable definitions of the terms disease cluster or the more general concept of clustering. Knox has attempted a biological as opposed to a statistical definition of a disease cluster as “... a geographically and/or temporally bounded group of occurrences related to...
each other through some social or biological mechanism, or having a common relationship with some other event or circumstance". Knox himself recognised that this definition provided more questions than answers to the crucial underlying query in a cluster investigation: "Is it real?".

Rothman has warned that reports of disease clusters rarely lead to advances in our understanding of the aetiology of disease. As cluster investigation invariably implies considerable use of (limited) resources, both in financial and human terms, then, according to Alexander and Cuzick, "... it can be justified from a scientific point of view only if it leads to identification of an aetiological factor which is either (a) a substantial cause of the total burden of the particular disease in society, or (b) an overwhelming cause of a disease in a particular area, though rarely causative elsewhere." These conditions are likely to be only rarely satisfied. For further discussion, including a critique of the statistical methods available for the spatial analysis of disease occurrence, see Elliott et al.

Disease mapping
As well as the investigation of disease clusters, public health physicians are increasingly involved in the closely related activity of disease mapping, with the intention of using maps as a surveillance tool either to identify areas at high risk of disease requiring further investigation, or to aid resource allocation. In the United Kingdom, such maps appear as standard in many district public health reports, not least because of the ease with which they can now be produced with current computer technology. Their intuitive visual appeal, however, masks very real problems of interpretation, especially for small areas where the map can be dominated by random variation.

Methods based on Bayesian statistics can be used to remove the random component from the map, with the aim of showing any true underlying systematic variation. There are currently few examples of their use in public health or environmental health applications.

Small area studies
The term small area is itself only loosely defined. One rather unhelpful definition in the context of environmental health studies is that it relates to an analysis carried out at sub-national or perhaps sub-regional level—that is, at a geographical scale below that of the standard (published) reporting of disease rates. As this definition does not take into account population size—which may differ substantially between areas—disease frequency, or suitability of the available geography to study the problem in question, its use is limited to the pragmatic—data on an inappropriately broad geographical scale may be the only data available! Putative disease clusters are unlikely to respect the administrative boundaries that traditionally have determined the reporting of population data or the availability of health statistics; any disease excess may be occurring on a much finer geographical scale and cross arbitrary geographical boundaries. Under such conditions, a purely local excess is unlikely to be detected in the routine statistics.

A more useful definition of small area would take into account the number of cases observed, which itself would depend on population size, disease frequency, and time period of analysis. As a rough guide, any region containing fewer than about 20 cases of disease can be considered a small area. As many cancers have annual incidence rates of around 5/100 000, over a five year period a small area might comprise a population of around 100 000 or less. For rare diseases or small populations in remote areas, the population size might be much less, but usually populations of at least 10 000 or so are required to form an aggregation of minimal size.

Cuzick and Elliott attempted to characterise the purpose and methods of small area enquiries; in so doing seven types of study were identified.

STUDIES OF REPORTS OF DISEASE EXCESS (CLUSTERS) IN SPECIFIC LOCALITIES WITHOUT A PUTATIVE SOURCE.
This is probably the most common problem faced by the public health authorities, as discussed above. In this type of study, problems of interpretation are severe as many apparent clusters are certain to arise by chance.

STUDIES OF POINT SOURCES OF INDUSTRIAL POLLUTION
Public or media concern about an industrial source of pollution may lead to a review of the available health statistics for the area in the vicinity of the plant. A major problem in interpretation is how to deal with retrospective reports of disease excess near an industrial plant. One approach, if the data systems are in place, is to replicate the study around other similar industrial sites (if such can be found).

STUDIES OF CLUSTERING AS A GENERAL PHENOMENON
The purpose here is to examine for some general tendency of a disease to show patterns of clustering. One example is that of Hodgkin's disease where the tendency for local clustering has led to suggestions of an infectious aetiology, although anomalies in the data, especially the population estimates, offer an alternative explanation.

ECOLOGICAL CORRELATION STUDIES OF HEALTH AND THE ENVIRONMENT
Ecological studies at the small area level are rare, as often the health, environmental, and population data are not available at the same level of geographical resolution. As an example, the EU is funding a small area methodological study (small area variations in air quality and health—SAVIAH) based in the United Kingdom, The Netherlands, the Czech Republic, and Poland. The study is examining the use of passive samplers for the
Investigation of disease risks in small areas

measurement of NO₂ and SO₂ in small areas, geographical information systems (GIS) and pollution modelling, and Bayesian map smoothing techniques in the analysis of outdoor air pollution and the geographical distribution of childhood wheeze and asthma.

DESCRIPTIVE STUDIES AT THE SMALL AREA LEVEL

Little is known of the natural variation of disease rates over small areas (related, for example, to the underlying sociodemography), although such information is necessary to place specific disease clusters in proper context. An example is the geographical epidemiology of childhood leukaemia and non-Hodgkin’s lymphomas in Great Britain where small area variations in incidence are related to both regional and socioeconomic factors.

GEOGRAPHICAL SURVEILLANCE

Modern computer systems allow automatic surveillance of routine health data to detect areas of high disease incidence although many false positives are bound to occur with currently available methods. These endeavours need to be viewed as experimental in nature, and results treated with extreme caution until more about the sensitivity and specificity of the methods is known.

ACUTE ACCIDENTS

Studies of chemical or nuclear accidents have much in common with studies of point sources of pollution, with the added dimension of time. An example is the explosion at the chemical plant in Seveso, where geographically defined cohorts were identified according to presumed exposure to dioxins.

The Small Area Health Statistics Unit

In the United Kingdom, one media report, of an alleged increase of leukaemia in children and young adults near the Sellafield nuclear reprocessing plant, led to a series of epidemiological studies which continue some 10 years later. These studies have suggested a possible new aetiology for childhood leukaemia related to occupational exposure of the fathers to ionising radiation around the time of conception. More generally, the Sellafield enquiry generated considerable scientific interest in the statistical and epidemiological methods available for small area studies, and led directly to the setting up of the Small Area Health Statistics Unit (SAHSU) for the United Kingdom.

The SAHSU is an independent national facility, funded by the United Kingdom Government, for the investigation of disease near point sources of industrial pollution. Its terms of reference currently are as follows.

(1) To examine reports of unusual clusters of disease, particularly in the neighbourhood of industrial installations and advise authoritatively as soon as possible.

(2) In collaboration with other scientific groups, to build up reliable background information on the distribution of disease in small areas so that specific clusters can be placed in proper context.

(3) To study the available statistics to detect any unusual incidence of disease as early as possible and, where appropriate, to investigate.

(4) To develop the methodology for analysing and interpreting statistics relating to small areas.

This remit includes examination of the health data in an area where there has been public or media concern. Examples include a study of cancer incidence and mortality near Baglan Bay petrochemical works, South Wales and near a pesticides factory. We have been particularly concerned with the study of disease around multiple industrial sites, either to replicate an enquiry conducted retrospectively around one site (by studying other sites in Britain producing similar discharges) or to test hypotheses related to a particular industrial process. Such studies include the investigation of cancer incidence near waste incinerators, radio transmitters, and the incidence of angiosarcoma of the liver near vinyl chloride plants.

The SAHSU incorporates a national database which includes mortality (by specific cause), cancer registrations, births, and congenital malformation data. Data retrieval is based on the postcode of residence at the time of the registered event (death or diagnosis). The postcode relates to only 14 households on average, and can be located as a point on a map to 100 m accuracy. Population data for small areas (enumeration districts) are available from the national census. A typical enumeration district gives population counts and socioeconomic data for about 400 people. By use of a database retrieval system with postcode as a key, rates of mortality and cancer incidence can rapidly be analysed for arbitrary circles located anywhere in Britain. Recent addition of a geographical information system has meant that occurrence of disease in relation to irregularly shaped areas (wind pattern) or linear structures (coast line, roads, power lines) can also be studied.

Postcoded data sets held by SAHSU include:


(2) Live and stillbirths for England and Wales from 1981. The birth data provide accurate year by year postcoded denominator counts for perinatal and childhood events.


For each individual event, SAHSU holds the postcode of residence, diagnosis (ICD code, and histology code for cancer registrations), age in years and months, and an identi-
fier allowing linkage to individual records through the Office of Population, Censuses, and Surveys (OPCS). A similar system is in place for Scottish data.

The SAHSU system runs under Unix with RISC (reduced instruction set computer) technology including a Digital (DEC) 5500 super-microcomputer and two Sun Sparc-servers 10 model 41. Currently around 15 000 Mbytes of data storage are required. We use the Oracle relational database management system and implement algorithms in C with “embedded” SQL for access to the database. Geographical data retrieval and mapping is done by Arc-Info, with links to the Oracle database. Statistical analysis is done mainly in the statistical package Splus.

For the analyses around a point source, currently only a simple radial dispersion model is used. For example, a range of circle sizes around a source is chosen first, and the numbers of events observed, and the numbers expected, are obtained for the bands between adjacent circles. Expected numbers are calculated from national rates standardised for age and sex. Adjustment is also made for a measure of socioeconomic deprivation in small areas, and for region—to allow for regional differences in disease incidence and (for cancers) for variation in the quality and completeness of registration.

Statistical testing is currently based on an adaptation of Stone’s method for data over a range of circles, to test whether there is evidence of decline in risk with distance from the source. Data from several installations can readily be pooled and tested in the same way.

An example: cancer of the larynx and lung near incinerators of waste solvents and oils in Great Britain

A report of cancer incidence near a defunct incinerator of waste solvents and oils at Charnock Richard, Coppull, Lancashire, north west England, suggested that there was an apparent cluster of five cases of cancer of the larynx nearby. This post hoc finding was examined with the SAHSU database, and the enquiry extended to all 10 eligible incinerator sites in Great Britain that could be identified as burning a similar type of waste. With Stone’s method, the excess of cancer of the larynx near Charnock Richard was found to be within chance limits, based on a small number of cases. In an analysis pooled over all 10 sites (which ensured adequate statistical power) no significant excess of either cancer of the larynx or lung was found. It was concluded that the apparent cluster of cancer of the larynx previously observed at Charnock Richard was unlikely to be due to its former incinerator.

Problems of interpretation in small area studies

A major problem in interpretation is the issue of socioeconomic confounding. People living close to an industrial site located in a deprived area are not a random sample of the population, but rather tend to be disadvantaged in society. It is well known that deprivation in itself is associated with ill health (reflecting, for example, differences in smoking rates) so that studies of disease risks in the vicinity of industrial plants are likely to be confounded. The size of the deprivation effect in many cases is large enough to swamp possible health effects related to low level environmental pollution, at least for some cancers such as stomach and lung.

Data availability and confidentiality

Few countries except for the United Kingdom and the Scandinavian countries have the capability of carrying out small area analyses with routine data. A major constraint is the availability of appropriate health data at the small area level, such as cancer registrations, often because of reasons of data confidentiality. A further constraint in many countries is the lack of a system of geocoding of health and population data analogous to the use of the postcode in the United Kingdom. A particular problem in the United Kingdom has been the lack of reliable morbidity data—for example, hospital discharges—other than cancer registrations, although recent changes in the health service have led to improvements in the database. For example, postcoding of hospital discharges is now almost 100%. Further improvements to the database, including possible use of general practice data can be expected.

Summary

The study of the effects of environmental pollution on health is complicated by problems of data quality, and the geographical resolution at which relevant data are available. Recent advances in methods for small area studies have meant that the initial investigation of disease near point sources of pollution can largely be automated, although care is needed in the interpretation of a positive result, both to check and validate cases and to allow for possible socioeconomic confounding. More widespread use of small area methods, to enable the rapid replication of findings internationally as well as nationally, would enhance our ability to quantify the effects of environmental pollution on human health.

The Small Area Health Statistics Unit (SAHSU) is funded by grants from the Department of Health, Department of the Environment, Health and Safety Executive, Home and Health Department of the Scottish Office, Welsh Office, and the Northern Ireland Department of Health and Social Services. An earlier version of this paper was presented as the opening address to the 10th annual meeting of the Spanish Society of Epidemiology, Alicante, September 1994.


Investigation of disease risks in small areas


