

A European Health and Environment Information System for Exposure and Disease Mapping and Risk Assessment

FINAL REPORT

SAHSU, Dept. Epidemiology & Public Health, Imperial College
Dept. Statistics & Operation Research, University of Valencia
WHO European Centre for Environment & Health, Rome
National Board of Health
National Public Health Institute
Dept. Epidemiology, Stockholm Centre of Public Health
Dept. Public Health, UCD
INSERM

UK
Spain
Italy
Denmark
Finland
Sweden
Ireland
France

22nd February 2001

CONTENTS

EXECUTIVE SUMMARY	3
1 INTRODUCTION.....	5
1.1 PROJECT AIMS.....	5
1.2 OBJECTIVES.....	5
2 OVERVIEW OF GEOGRAPHICALLY REFERENCED HEALTH, DENOMINATOR, SOCIO-ECONOMIC AND ENVIRONMENTAL DATASETS WITHIN THE PARTNER COUNTRIES.....	7
2.1 DATASET REQUIREMENTS.....	7
2.1.1 <i>Health datasets</i>	7
2.1.2 <i>Denominator datasets</i>	7
2.1.3 <i>Socio-economic datasets</i>	8
2.1.4 <i>Environmental datasets</i>	9
2.1.5 <i>Geographical datasets</i>	9
2.2 DATASET AVAILABILITY , QUALITY AND SUITABILITY IN EACH COUNTRY	10
2.2.1 <i>Denmark</i>	11
2.2.2 <i>Finland</i>	12
2.2.3 <i>Italy</i>	12
2.2.4 <i>Spain</i>	13
2.2.5 <i>Sweden</i>	13
2.2.6 <i>UK</i>	14
2.3 SUMMARY	15
3 STATISTICAL METHODOLOGY	16
3.1 RECOMMENDATIONS FOR STATISTICAL ANALYSIS OF DISEASE/EXPOSURE MAPPING STUDIES AND POINT SOURCE INVESTIGATIONS.....	16
3.1.1 <i>General framework</i>	16
3.1.2 <i>Checking and adjusting for data quality</i>	16
3.1.3 <i>Preliminary analysis and checking model assumptions</i>	17
3.1.4 <i>Statistical modelling for disease/exposure mapping studies</i>	17
3.1.5 <i>Specification of prior distributions</i>	17
3.1.6 <i>Summarising results</i>	18
3.1.7 <i>Specific recommendations for point source investigations using small-area data</i>	18
3.2 APPLICATION OF STATISTICAL METHODOLOGY TO DATA IN EACH PARTNER COUNTRY.....	18
3.3 TRAINING COURSES IN STATISTICAL METHODOLOGY.....	19
4 DEVELOPMENT OF THE UK RAPID INQUIRY FACILITY.....	20
4.1 DESCRIPTION OF MODULES.....	20
4.2 SYSTEM DEVELOPMENT.....	20
5 FEASIBILITY OF IMPLEMENTATION OF THE SYSTEMS WITHIN THE PARTNER COUNTRIES	22
5.1 GENERIC FEASIBILITY ISSUES.....	22
5.1.1 <i>Political and organisational feasibility</i>	22
5.1.2 <i>Technical feasibility</i>	24
5.1.3 <i>Economic feasibility</i>	25
5.1.4 <i>Methodological feasibility</i>	25
5.2 FEASIBILITY IN EACH COUNTRY	25
5.2.1 <i>Denmark</i>	25
5.2.2 <i>Finland</i>	27
5.2.3 <i>Italy</i>	27
5.2.4 <i>Spain</i>	28
5.2.5 <i>Sweden</i>	29
5.2.6 <i>UK</i>	29
5.3 SUMMARY	29

EXECUTIVE SUMMARY

The EUROHEIS (European Health and Environment Information System) project aims to improve understanding of the links between environmental exposures, health outcome and risk through the development of integrated information systems for rapid assessment of relationships between the environment and health at a geographical level.

A one-year feasibility study was undertaken to assess the possibilities of implementing systems for point source investigations and disease and exposure mapping within the participating countries, modelled on a system (the Rapid Inquiry Facility (RIF)) being developed within the UK. The project also aimed: to develop the UK system to include more generalised modules for disease and exposure mapping; to explore techniques for taking into account inconsistencies in the data; and to assess the availability of indices of socio-economic status derived from routinely collected data sources. This report documents the findings of the project and makes recommendations for the implementation of the systems and methodologies within the various partner countries.

The successful implementation of systems for point source investigations and disease and exposure mapping is greatly dependent on suitable data being available for use within the systems. Not only must the required datasets be available, they must also be of sufficient quality to enable meaningful and interpretable analyses. This feasibility study concludes that datasets of suitable quality are available and accessible in some partner countries to enable the successful implementation of point source investigations and/or disease and exposure mapping. In some countries, the potential for very detailed small area studies is immense; in others there are dataset issues to be solved, which pertain mainly to the size of the geographical base units and to the sparseness of data. These issues must be investigated in more detail before a decision can be made regarding the type of implementation that is most appropriate. In these cases though, the potential of statistical methodologies at least partially to overcome such problems, has been explored as part of the feasibility study. In general, a Bayesian hierarchical modelling approach is recommended for statistical analysis of disease/exposure mapping studies and point source investigations. These methods allow raw disease rates to be 'smoothed' to overcome problems of sparse data, and provide a natural framework for incorporating common features of the data such as over-dispersion, spatial correlation, missing data, exposure measurement error and ecological bias.

As part of the project, the functionality of the RIF has been enhanced and expanded. This has been achieved through the development of a series of modules to enable small area disease mapping and enhanced point source analyses. The modules are complementary and were identified based on discussions with partners in the EUROHEIS project, and discussions in-house within SAHSU. The modules provide the mechanisms by which value can easily be added to the RIF: for example when new health datasets are acquired, when new geographies are needed, or when different methods of defining the study area are appropriate. The concepts and methodologies inherent in the modules can be generalised and because they are written using proprietary software, they are thus transferable to the other partner countries in EUROHEIS.

This feasibility study concludes that the implementation of point source investigation and disease/exposure mapping systems is feasible in all partner countries. Whilst there are generic issues to be addressed in all countries (including political, technical, economic and methodological issues), the precise form of the implementation will vary from country to country. This is dependent on the specific public health questions of importance in the country together with the different practical constraints and opportunities within that country.

1 INTRODUCTION

1.1 Project aims

The project aims to improve understanding of the links between environmental exposures, health outcome and risk through the development of integrated information systems for rapid assessment of relationships between the environment and health at a geographical level.

A one-year feasibility study was undertaken to assess the possibilities of implementing systems for point source investigations and disease and exposure mapping within the participating countries, modelled on a system being developed within the UK. The feasibility study also aimed: to develop the UK system to include more generalised modules for disease and exposure mapping; to explore techniques for taking into account inconsistencies in the data; and to assess the availability of indices of socio-economic status derived from routinely collected data sources. This report documents the findings of the feasibility study and makes recommendations for the implementation of the systems and methodologies within the various partner countries.

1.2 Objectives

The following work packages were identified and formed the specific objectives for the feasibility study:

Work package Objective

- 1 Overview of geographically referenced datasets currently available to each of the participating partners, and what is known of their accessibility, cost, completeness and quality, including sources of health, population, environmental and socio-economic data
- 2 Visits by the UK to partner countries to facilitate exploration of the feasibility of developing computerised systems in each country
- 3 Exploration of the feasibility of developing systems for point/area/line source investigations and disease and exposure mapping within partner countries as well as the possibilities of assessing socio-economic status from routinely collected data sources to be included in the systems
- 4 Exploration of mechanisms for elucidating inequalities in environmental exposure and health
- 5 Exploration of the application of previously developed statistical methodologies in the participating countries.
- 6 Generation of specifications for core data requirements for such small area analysis systems
- 7 Exploration of the impact of data quality and resolution on estimates of

disease occurrence

- 8 Development of the UK system to include a more generalised module for disease and exposure mapping
- 9 Dissemination to the European community

The rest of this report is arranged into five chapters, which broadly follow the work packages detailed above (although some work packages have been combined to form one chapter where appropriate).

Chapter 2 contains an overview of the geographically referenced health, denominator, socio-economic and environmental datasets available within each of the partner countries for use in the development of systems for point source investigations and disease/exposure mapping in the partner countries. It considers the suitability of these datasets for such studies in terms of their accessibility, cost, completeness and quality.

Statistical methodologies previously developed in the partner countries are explored in Chapter 3 and the impact of data quality and resolution on estimates of disease occurrence are considered. Guidelines on statistical analysis, ecological bias and epidemiological interpretation of point source investigations and disease/exposure mapping are provided.

Chapter 4 outlines improvements that have been made to the UK Rapid Inquiry Facility throughout the period of the feasibility study.

The feasibility of implementing systems for point source investigations and disease/exposure mapping in each of the partner countries is discussed in Chapter 5.

Appendix 1 considers the possibilities for assessing socio-economic status from routinely collected data sources within the partner countries and explores mechanisms for elucidating inequalities in environmental exposure and health.

Appendix 2 summarises the datasets available for use in the project within each of the partner countries.

Note also that further details regarding the project can be obtained from the EUROHEIS web site (http://www.med.ic.ac.uk/divisions/60/euroheis/copy_of_myweb/index.htm), which was developed during the period of the feasibility study. Additionally, during this period, a paper was presented at the International Society for Environmental Epidemiology Conference¹.

¹ Cockings S, Jarup L, Aylin P, Elliott P, Poulstrup A, Reuterwall C, Pekkanen J, Martuzzi M, Ferrandiz J, Staines A and Richardson S (2000). A European Health and Environment Information System for disease and exposure mapping and risk assessment. *Epidemiology*, 11(4), 343.

2 OVERVIEW OF GEOGRAPHICALLY REFERENCED HEALTH, DENOMINATOR, SOCIO-ECONOMIC AND ENVIRONMENTAL DATASETS WITHIN THE PARTNER COUNTRIES

The successful implementation of systems for point source investigations and disease and exposure mapping is greatly dependent on suitable data being available for use within the systems. Not only must the required datasets be available, they must also be of sufficient quality to enable meaningful and interpretable analyses.

2.1 Dataset requirements

One of the first programmes of work within the feasibility study was to identify the key types of datasets required for the implementation of such systems. To this end, five main types of dataset were identified: health, denominator, socio-economic, environmental and geographical datasets. The next step was then to consider the required contents of these datasets. Key fields were identified for each type of dataset and these are discussed below.

2.1.1 Health datasets

Health datasets are clearly a necessary requirement of such systems. The type of health event will, though, vary considerably; typical examples include mortality registrations, cancer registrations, hospital admissions, congenital malformations and perinatal mortality. The type of health data used in such systems is usually (but not necessarily) routinely collected datasets. Essentially, any dataset that fulfils the requirements of the system, in terms of its key fields and quality, can be used.

The key requirements for health datasets are some form of diagnostic code, date of event, age and sex of the person (or mother in the case of birth-related events) (usually at time of diagnosis, birth or death) and some form of geographical referencing.

Diagnostic coding will vary through time, but generally routinely collected datasets are coded according to the International Classification of Diseases (ICD) coding system, which ensures comparability at an international level. It is essential though, that the version of ICD being used at any particular time is known, as codes and their contents do change through time.

The health data must be geographically referenced in some way to enable them to be linked spatially to population, socio-economic and environmental data. The type of geographical referencing employed varies considerably, not only from dataset to dataset, but also between countries. Health data are frequently not available for the same geographical units as those for which denominator, socio-economic and environmental data are available. Some form of interpolation is therefore usually required to produce a set of coterminous units that are suitable for the analysis.

2.1.2 Denominator datasets

Denominator datasets are required to calculate rates of the diseases or health events within geographical areas. In many cases the denominator data will be obtained from population data, although for some health events an alternative denominator is more

appropriate. For example, for a study of congenital malformations the denominator is usually births, whilst for hospital admissions it is often more meaningful to use total admissions as a denominator rather than total population. The appropriate denominator to be employed will depend on the particular health event being investigated and the specific question being addressed.

In terms of required fields, denominator datasets must contain details of the date for which the data apply (or when they were collected); they must be broken down by age and sex (to enable age-sex specific rates to be applied); and they must be geographically referenced. In some instances, individual level denominator data may be available; in others only aggregated data will be available. As mentioned previously, denominator datasets are frequently not geo-referenced to the same set of geographical units as the health and other datasets. If this is the case, then some form of interpolation must be undertaken to derive a common set of units on which analysis can take place.

2.1.3 Socio-economic datasets

In studies of the distribution of diseases, especially those investigating links between the environment and health, factors which potentially confound the relationships under investigation must be taken into account. Variations in socio-economic characteristics at the individual and group level are particularly important in this respect as they frequently show strong correlations with both health and environmental factors. The precise socio-economic variables of relevance to such investigations varies considerably between countries. For example, in the UK, in the context of point source investigations and disease and exposure mapping, the most commonly used measure of socio-economic status of individuals and/or groups is deprivation, which is assessed by a score (e.g. Carstairs, Townsend) that combines different potential indicators of deprivation, such as social class and overcrowding. However, in Finland, people's occupation has been shown to be a useful measurement of socio-economic status relative to health. By contrast, in Denmark, education is considered to be one of the most important variables for such studies.

The required fields for socio-economic datasets are difficult to define due to the variety of datasets that could be employed. However, it is essential that there is some measure of the socio-economic variable being considered, together with the date that the data are relevant for and some form of geographical referencing. If the data are also to be used for the age-sex specific adjustment of rates, then the data also need to be broken down by age and sex.

The definition, measurement and assessment of both the absolute and relative socio-economic characteristics of individuals and populations is far from trivial. Appendix 1 contains a review of the evidence for socio-economic inequalities in health and the link with deprivation, together with a consideration of how deprivation may be measured at small-area level. It also identifies various implications for spatial modelling of disease patterns.

2.1.4 Environmental datasets

The environmental exposures of greatest concern tend to vary from one country to another, and over time, so the environmental datasets required to support the systems are also likely to vary. Current concerns in the UK, for example, include traffic-related air pollution, point-source emissions (e.g. from industrial activities, incineration), electro-magnetic fields, landfill and waste sites, and drinking water quality (e.g. disinfection by-products). In the wake of the BSE crisis, food-borne exposures are also a major focus of concern. In Spain, for example, environmental concerns focus on drinking water quality and air pollution. Intensive agricultural activities tend to disseminate chemical contaminants, whilst industrial and tourist activities are producing an excess of water consumption that tends to impoverish water resources. In main Spanish cities, air pollutants such as black smoke, total suspended particles, NO₂, SO₂ and CO are of concern for their effects on health.

The assessment of an individual's exposure to environmental factors is difficult. Studies using group level (routinely collected) data to assess exposure as well as health outcomes are usually referred to as "ecological". Such studies cannot reveal any causal relationships between exposure and disease. However, they can be very useful to suggest associations between (environmental) levels of a pollutant and health outcome. In order to obtain the best possible exposure estimates, the (modelled) group level data should be as close to actual individual exposures as possible, which necessitates a good spatial resolution. Exposure models should be validated using measured data in a sample of individuals. Most of the important public health problems in Western societies concern chronic diseases with long latency times (time between first exposure and diagnosis), in the order of years to decades. Thus, exposure should ideally be assessed up to 10-20 years before the health outcome, which is rarely feasible (although in some circumstances historical data may be available). Therefore, information on migration will help to interpret analysis, as well as information on previous levels of pollution in the study area. However, using current data on exposure and health outcome will usually lead to underestimation, rather than inflation, of the true risks.

In terms of the required fields for environmental datasets, it is essential that the data be geographically referenced and that they have some form of date indicating to what period or point in time the data pertain. Environmental datasets are usually not available for the same geographical units as health and other datasets. Interpolation must therefore usually be undertaken to derive measurements of exposure for an area, and at a resolution and scale, suitable for analysis.

2.1.5 Geographical datasets

As mentioned in previous sections, all the datasets to be employed in point source investigations and disease and exposure mapping must be geographically referenced to some common geographical referencing system. Most studies of this type are carried out at some small area, aggregated, level as individual level data are generally not available. Usually though, the different datasets employed in such studies are not available for the same sets of geographical units. A decision must therefore be made as to what constitutes the most suitable set of geographical units (or zonal system) for the analysis. This decision is by no means trivial and can have an important bearing on the validity and interpretation of results from the analysis. The units are normally selected depending on the datasets available and on the existence of look-up tables or suitable techniques which allow the datasets to be interpolated onto the selected set of units. It

is vital when making such decisions (and when interpreting the results of any subsequent analyses) that the resolution, scale and accuracy of the datasets, and the effects of any interpolation, are known and taken into account.

The requirements for such studies will therefore vary depending on the units selected and on the degree of interpolation required. Commonly though, the following are required: digital boundaries of the different geographical units being employed; centroids of the geographical units (preferably population-weighted); and look-up tables linking the various geographical units together. If look-up tables are not available, Geographical Information Systems (GIS) techniques can be employed to transfer data between the different geographical units. For example, grid-referenced address-based health data can be aggregated into grid squares using what are termed 'point-in-polygon' techniques; or exposure data for water supply zones can be interpolated onto another set of geographical units according to the proportion of each water supply zone's area falling within each unit (areal weighting).

2.2 Dataset availability, quality and suitability in each country

Having determined the datasets and fields required for the successful implementation of systems for point source investigations and disease and exposure mapping, the next stage of the project was to assess the availability, quality and suitability of such datasets within each of the partner countries.

In order to implement systems to undertake meaningful and useful investigations, it is essential that the key datasets are not only available and accessible (in terms of cost and time to obtain), but also of sufficient quality to be suitable for analysis. In this respect, the datasets need to be available for suitable geographical units, they must cover the appropriate study area(s) and they must be available for the years to be studied. High levels of case ascertainment and completeness are desirable, although unfortunately not always possible; what is essential though in this respect is that the levels of case ascertainment and completeness are known (so that they can be taken into account in interpretation of the results) and that they are not correlated with the exposure under study. Any costs involved in acquiring new datasets must be recognised, and an approximate timescale for obtaining the datasets must be identified. Any potential organisational barriers to acquiring datasets and to implementing such systems must be identified and any confidentiality and/or copyright issues highlighted.

The availability, quality and suitability of datasets within each of the partner countries were assessed by means of questionnaires and a series of discussions with the partners. Questionnaires were sent out, during January 2000, by the UK partner, to all participating countries, with the exception of France and Ireland (the French partner was acting as a statistical advisor and the Irish partner was leading the work on inequalities in health). Five different questionnaires were employed: one for each type of dataset (health, denominator, socio-economic, environmental and geographical). Samples of the questionnaires are available from the authors and were included in the EUROHEIS Interim Report (SAHSU, February 2000). The questionnaires were completed by the partners and returned to the UK partner prior to the first EUROHEIS meeting, which took place at SAHSU, Imperial College, on 24th-25th January 2000. The questionnaires formed the basis for discussions at this meeting.

Subsequently, during the period February to May 2000, representatives from the UK partner visited each of the other partner countries. One of the key aims of these visits was to discuss the availability and suitability of datasets within each country, building on the information that had been provided in the questionnaires.

Tables 1(a)-(e) to 6(a)-(e), in Appendix 2, summarise the findings of the questionnaires and the outcome of discussions with the partner countries. Table 7 provides details of EU-level environmental and socio-economic datasets, which are available to all partner countries.

2.2.1 Denmark

In Denmark (Tables 2(a) to (e)), the main health datasets of interest are cancer registrations, mortality and hospital admissions. These datasets all contain the required key fields and have high levels of case ascertainment and completeness. They are available at the individual level and can be geo-referenced at the address level. This geo-referencing is possible due to the existence of a national address register, which contains the grid references of all properties within the country, to an extremely high degree of precision and accuracy. Denominator data comes from the national population registry. All individuals in Denmark have a unique Central Population Register (CPR) number, which enables the cross-referencing of a wide range of datasets. Consequently, for example, health data can be cross-referenced with individual level socio-economic data, to obtain data on an individual's education, employment and income. Additionally, as all the data is obtainable from registers which are updated daily, the data are extremely timely. There is also the ability to track individuals through time and hence account for migration, an issue which is usually very difficult to address in epidemiological studies. Unfortunately (but quite typically), environmental datasets are not nearly as timely nor as complete. The main environmental datasets contain information on the location (and sometimes characteristics) of environmental features such as powerlines, industries, waste deposits, water sources, and so on. In some instances, data are available for modelling of, for instance, air pollution and traffic density.

In terms of geographical units for analysis, as the data are all available at the individual level, they can, in principle at least, be aggregated to virtually any set of geographical units deemed appropriate.

The datasets currently available to the Danish partner are for Vejle County. This is because a prototype system has been under development in this county, with a view to being a precursor for subsequent development on a national scale (see Section 6.2.1 for more details). The proposed development of a national system would obviously require that the relevant datasets be obtained for the whole country. In some cases payment would be necessary to obtain the required datasets. In addition, particularly for the National Prevention Registry, this would also set an important precedent as the data have never before been used outside the Central Statistical Bureau.

In summary then, in terms of the availability and quality of suitable datasets, the Danish partner is in an excellent position to develop a system for point source investigations and disease and exposure mapping.

2.2.2 Finland

Finland has extensive health registers, all of which contain the individual's social security number, which can be used to obtain coordinates of the place of residence through record linkage with Statistics Finland. The most important registers are those containing information on deaths, hospital discharges, births, asthma and infectious diseases.

The main health dataset of interest and currently available to the Finnish partner is cancer registrations. This dataset is available from 1981 to 1997 and contains all the required key fields. It has a high level of case ascertainment and a reasonable level of completeness. The data are available at the individual level and are geo-referenced by place of residence, at the time of diagnosis or registration, to an accuracy of 10 metres. The dataset also contains a socio-economic classification and level of education for every 5 years from 1970 to 1995 inclusive, together with the municipality of residence in 1980, 1990 and in the year before diagnosis, which enables some accounting for migration in the analyses. Population data are available from the Census Small Area Statistics, for 0.5km by 0.5km grid squares, broken down by 5-year age categories and sex, covering the whole of the country. These data are available for the years 1980, 1990 and 1997. The census data also contains a measure of socio-economic class, which is mainly based on occupation. Currently, there are no national environmental datasets available to the Finnish partner.

In terms of cancer registration data, therefore, there would be no problems implementing a system for point source investigations and disease and exposure mapping. Indeed, the Finnish partners have already started implementing such a system, which uses grid squares as the base units for analysis. A specific problem faced by the Finnish partners is the sparseness of population in large parts of the country. This leads to uncertainty in the estimates and difficulties in the implementation of smoothing techniques. Therefore, better statistical methods are needed to deal with these issues.

2.2.3 Italy

The main dataset available to the Italian partner is mortality data from 1980 to 1994. Due to confidentiality constraints, the data (collected by ISTAT, Italy's Bureau of Statistics) cannot be freely held and accessed at the individual level. The Italian partner, therefore, accesses the data through a query program, which extracts aggregated data above certain specified thresholds (minimum 10,000 people and 3 years of data). The highest, readily available, geographical resolution is the municipality level (around 8,000 municipalities in the country). Finer resolution is presently possible on an ad-hoc basis, by linking the mortality records with local registers of residents, held by municipalities. This is especially important as large cities consist of one municipality. Again, the record linkage necessary to obtain finer spatial resolution of deaths is subject to confidentiality constraints. This lack of direct access to the raw data may prove problematic for the development of a system for point source investigations that can provide rapid response to public health concern.

Population data, broken down by age and sex, are available from the Census, which is held every 10 years. Socio-economic data are available in the form of a newly developed index of deprivation (similar to the UK Carstairs Index of Deprivation), which uses various Census variables to derive scores and then quintiles.

In terms of the quality and suitability of data available for use in the development of systems, there are clearly some challenging issues that need addressing. The lack of direct access to the data and the large geographical base units pose significant problems for the development of systems for small area point source investigations. However, there is scope for small area disease and possibly exposure, mapping, and indeed prototype software has been developed for this means. In addition, for those selected cases where individual or census tract level data are available – notably Florence, Turin, Sicily and Puglia, the implementation of a facility for point source data analysis would be beneficial. Access to individual level health and socio-economic data for these areas will provide an opportunity to undertake some trial point source and hierarchical modelling analysis.

2.2.4 Spain

A wide range of health datasets has been identified by the Spanish partner for possible inclusion in such systems. The datasets are for two regions of Spain – the autonomous communities of Valencia and Andalusia. The health datasets include mortality, cancer registrations, hospital admissions, congenital malformations and notifiable communicable diseases, amongst others. These datasets are available for a range of years; most have a high level of case ascertainment and completeness; all contain the required key fields for use in a system for point source investigations and disease and exposure mapping. The main problem with the datasets available to the Spanish partner is that, as with Italy, the smallest geographical units for which they are available are municipalities. These vary considerably in size and population; for example, in Valencia, a municipality can contain anywhere between 23 and 746,683 people. This may be problematic for the implementation of small area analysis systems.

Population data, broken down by 5-year age categories and sex, are available from national and municipal censuses. The national Spanish Census and Municipal Census are each every 10 years but they are interleaved so that fresh data are available every 5 years. The last Spanish census was 1991 and the last municipal census 1996. From 1998 the municipal census is continuous and published every year. Some data on migration is available from the Census question 'Where did you live 5 years ago?', which reports movements into and out of the region. From this data, the Spanish partner can retrieve migration input-output tables between municipalities.

Socio-economic data are available at municipality level and are mainly concerned with employment levels. There is no measure of deprivation available. A range of environmental information is available, again at the municipality level, including data on atmospheric contamination, waste, water quality and traffic levels.

The Spanish partner is, therefore, in a good position to implement systems for disease and exposure mapping at the municipality level, and indeed has started doing so. Nevertheless, as in Italy, there are issues involving the interpretation of disease-exposure relationships at the small area level due to the high level of aggregation of the data.

2.2.5 Sweden

A range of health datasets are available to the Swedish partner, including cancer registrations, mortality, myocardial infarction incidence and hospital admissions. These datasets are available for various years and all contain the required key fields. Most

have a high level of case ascertainment and completeness. Specific permission would be required to use the datasets for this project, but no problems are anticipated in obtaining this permission. The datasets are available for what are termed base units, which are small geographical areas of reasonably homogenous characteristics, for which population data (by age and sex) can be obtained from the population register. The datasets are all linked via the 'civic registration number' (CRN), which is allocated to all Swedes upon birth or upon immigrating to Sweden. The CRN has been in existence since 1947.

In Sweden then, the potential for implementing both point source investigation systems and disease and exposure mapping facilities is great. Currently the dataset availability and feasibility has only been explored for Stockholm County, and it is for this area that the development of such systems is anticipated initially.

2.2.6 UK

It is the UK Rapid Inquiry Facility (RIF) (for point source investigations and disease and exposure mapping) that has been employed as the prototype system in fuelling this project. In this respect the question of dataset availability and suitability is not relevant as the datasets have already been proven to be appropriate. It is useful though, for completeness, to summarise the key datasets and their characteristics here.

The major datasets available for use in the RIF are cancer registrations, mortality, hospital admissions, congenital malformations and perinatal mortality. All the data are geo-referenced to the postcode level. A postcode typically contains around 14 households. The geographical base units currently employed in the RIF are Census Enumeration Districts (EDs), which are the smallest geographical unit for which population data (by age and sex), from the 10-yearly census, are available. Postcodes can be matched to EDs through look-up tables. Further denominator data are available through annual population estimates (calculated in-house) and, for birth-related events, from the births and stillbirths registers. Socio-economic data are available from the census, with the most commonly used measure being the Carstairs Index of Deprivation, which is derived from four census variables: overcrowding, unemployment, car ownership and social class of the head of the household.

Environmental and exposure data are less widely available and vary considerably in their timeliness, periodicity, geographical resolution and accuracy. Routine data are readily available for a number of air pollutants, through the national monitoring network, though the number of monitoring sites is not large and the sites do not give a representative picture of exposures across the population. Modelled emissions and concentrations data are available at 1km² level for some years. The UK is also well-provided with land cover data, derived from satellite imagery.

Most other environmental data are sparse, and/or not in the public domain. Access, therefore, is by negotiation with data providers, and often involves a significant cost. Many of these datasets are of variable quality and completeness (including their spatial and temporal coverage and accuracy) and thus require considerable checking and interpolation to make them suitable for analysis. Such data preparation may take considerable time, which limits the rapidity of response that can be achieved.

Digital geographical data (e.g. topography, urban areas, road lines) are available nationally at a variety of scales, but again – especially for large-scale (high resolution) data – may involve substantial costs.

2.3 Summary

In summary, datasets of suitable quality are available and accessible in all partner countries to enable the successful implementation of point source investigations and/or disease and exposure mapping. In some countries, the potential for very detailed small area studies is immense; in others there are dataset issues pertaining mainly to the size of the geographical base units and to sparseness of data. These issues must be investigated in more detail before a decision can be made regarding the type of implementation that is most appropriate. In these cases though, the potential of statistical methodologies at least partially to overcome such problems is being explored in another work package of the project (see Chapter 3).

3 STATISTICAL METHODOLOGY

One of the objectives of this project was to develop guidelines for the statistical analysis of point source investigations and disease/exposure mapping studies based on statistical methodologies previously developed by some of the partners in an EU-funded BIOMED project. The feasibility of applying these methods to the routine data available in each partner country has also been explored, taking into account the differences in data quality and resolution between countries. Training courses in the various methods have taken place in some of the partner countries.

3.1 Recommendations for statistical analysis of disease/exposure mapping studies and point source investigations

The Rapid Inquiry Facility being developed in the UK as part of this project, and similar systems proposed for the other partner countries, are intended to provide a rapid *preliminary* assessment of relationships between the environment and health at a geographical level. Any potentially important relationships identified should then be further investigated by carrying out more detailed studies. The recommendations below summarise some broad principles that should underpin such follow-up analyses, and more generally, any statistical analysis of routine disease/exposure data for the purpose of small-area mapping and point-source investigations. They are not intended to be followed rigidly for every study, since the available data and aim of the study should also guide the specific details of the methods used. The recommendations are based on a synthesis of the methodological and applied work carried out by the partner countries.

3.1.1 General framework

In general, a Bayesian hierarchical modelling approach is recommended for statistical analysis of disease/exposure mapping studies and point source investigations. These methods allow raw disease rates to be 'smoothed' to overcome problems of sparse data, and provide a natural framework for incorporating common features of the data such as over-dispersion, spatial correlation, missing data, exposure measurement error and ecological bias. It is assumed that data are available in the form of observed and expected disease counts in each area (the latter stratified by age and sex and possibly other known confounders), with possibly area-level exposure averages also available.

3.1.2 Checking and adjusting for data quality

Possible duplication of cases should be checked for by matching (e.g. on sex, date of birth and geographic location); any clear matches should be excluded. Outlying data points should be checked for by tabulating or plotting SMRs against area index to check for unusually high-risk estimates. Extreme SMRs may arise due to high sampling variability associated with sparse data, but may also indicate data anomalies. Possible solutions include: (i) check case records in suspicious areas and exclude erroneous cases (if any); (ii) consider using a robust model for the area-specific risks; and (iii) model the true number of cases in the area(s) as unknown.

Potential biases due to differences in registration procedures or case under-ascertainment between regions/disease registries may be adjusted for by using region/registry-specific reference rates to standardise expected counts.

Inaccuracies in population denominator data may also bias disease risk estimates. An initial approach is to treat the populations as known quantities, but if potential errors in the population data are of particular concern, possible approaches are to: (i) assume the observed population (or expected counts) are measured with error and carry out sensitivity analysis of inference to plausible estimates of the size of the measurement error; or (ii) model population change over time using, for example, a Markov chain or other temporally correlated model. The latter approach requires suitable data on population age and sex distribution by small area at frequent time intervals.

3.1.3 Preliminary analysis and checking model assumptions

For rare, non-infectious diseases it is usual to assume a Poisson model for the case count in each area. The assumptions of this model should be checked by, firstly, testing for lack of interaction between area and strata to check the proportional hazards assumption, and, secondly, by testing for Poisson over-dispersion.

(Informal) exploration of the spatial dependence between disease counts in each area should be carried out, for example using the variogram.

For point source studies, epidemiological knowledge should be used to hypothesise a plausible range for the putative exposure effect that will be used in exploratory analyses of the exposure-risk relationship (see section 3.1.7)

3.1.4 Statistical modelling for disease/exposure mapping studies

Smoothed estimates of disease risk should be obtained by fitting a Bayesian Poisson hierarchical model with exchangeable random effects for the area-specific log relative risks. Smoothed area-specific risk estimates can be plotted against area-level covariates and easting/northing to look for potentially important risk factors and large-scale spatial trend. The posterior mean or median risk estimates should also be mapped to look for spatial patterns. The model can then be re-fitted including spatially-structured area-specific random effects to capture spatial dependence in risk. Possible models for the spatial effects include conditional autoregressive models or multivariate normal models with spatially structured covariance. The latter model is only feasible for a moderate number of areas (up to a few hundred). If appropriate, the model can be re-fitted including covariates and spatial trend. The sensitivity of covariate effects to models with and without spatially-structured random effects should be checked. If exposure data are available at a finer resolution than the areas used for analysis, these may be introduced to model the within-area exposure distribution and help alleviate the problem of ecological bias.

Further details of the recommended models and methods can be found in, for example, Elliott et al (2000)², Pascutto et al (2000)³.

3.1.5 Specification of prior distributions

These models can be sensitive to the choice of prior distribution for the random effects variance parameters, particularly if the data are sparse (area-specific expected counts

² Elliott P, Wakefield J, Best N and Briggs D, eds. (2000). *Spatial Epidemiology: Methods and Applications*. Oxford University Press, Oxford.

³ Pascutto C, Wakefield J, Best N, Richardson S, Bernardinelli L, Staines A and Elliott P (2000). Statistical issues in the analysis of disease mapping data. *Statistics in Medicine*, 19, 2493-2519.

< 5 on average). Moderately informative priors should be used and sensitivity to choice of prior should be checked. There is no default prior to recommend. Suitable families of distributions include the gamma or chi-squared distributions for the random effects precision (inverse variance) or log-normal distributions for the variance. Guidelines for choosing appropriate parameter values for such priors are given by Bernardinelli et al (1995)⁴ and Pascutto et al (2000)².

3.1.6 Summarising results

Summaries of both the posterior mean and posterior uncertainty of the area-specific risk estimates should be reported. Posterior uncertainty may be summarised by quoting/mapping posterior quantiles, standard deviations and/or posterior exceedance probabilities (e.g. probability that the relative risk exceeds 1.0).

3.1.7 Specific recommendations for point source investigations using small-area data

The approach has some common features with disease/exposure mapping, but the exposure data will typically be some measure of the distance between each area and the source, or a specific measure of exposure that may be attached to each area. If exposure is represented by distance from source, the form of the distance-risk model should be explored by plotting smoothed risk estimates obtained as in section 3.1.4 against distance, using prior opinion to inform this process (see section 3.1.3). For further discussion of suitable models, see Morris and Wakefield (2000)⁵.

Unstructured or spatially-structured random effects may be needed to account for background variation in disease risk not associated with the point source. However, the random effects (particularly spatially-structured effects) may be confounded with the estimated risk associated with exposure to the point source. The definition of the background risk therefore needs careful consideration. It might be useful to include in the study “control regions” of similar spatial range to the region of interest that will provide information on the level of background variation in disease risk, excluding the putative exposure effect of the point source.

More work is needed to test how best to disentangle “real effect” of point source versus “typical cluster pattern” for the disease and the spatial scale of the analysis

3.2 Application of statistical methodology to data in each partner country

The general methodological framework outlined in section 3.1 is applicable to model area-level disease and exposure data in all the partner countries. However, differences between countries in data quality and resolution will require a different focus for the modelling effort.

⁴ Bernardinelli L, Clayton D and Montomoli C. (1995). Bayesian estimates of disease maps: how important are priors?. *Statistics in Medicine*, 14, 2411-2431.

⁵ Morris S and Wakefield J (2000). Assessment of disease risk in relation to a pre-specified point source. In: *Spatial Epidemiology: Methods and Applications* (Elliott P, Wakefield J, Best N and Briggs D, eds.), Oxford University Press, Chapter 9, p.153-184.

For countries with detailed, fine-resolution data the sparseness of the data in each area presents the main challenge. Results are likely to be sensitive to data inaccuracies, the spatial dependence structure assumed between areas and the choice of priors for the random effects variances. Since data quality is generally good for these countries, data anomalies should not present a major problem. Sensitivity analysis to model assumptions is important though, and it may be appropriate to aggregate data to larger geographical units to facilitate some analyses.

For countries with data available on larger geographical units, analyses are likely to be more robust to modelling assumptions since data are less sparse (although sensitivity should still be explored). None the less, interpretation of the modelled disease-exposure relationships requires care due to the possibility of ecological bias. The problem can be partially addressed by modelling the within-area distribution of exposures, provided some exposure data are available at a finer resolution.

3.3 Training courses in statistical methodology

The UK partner has organised training courses in methods and software for Bayesian modelling in spatial epidemiology in the UK (December 1999 and July 2000) and also visited the Finnish partner in February 2000 to present a similar course.

The Spanish partner organised a course on Bayesian methods and geographical risk assessment as a satellite activity of the First Joint Meeting of the Spanish Society of Epidemiology and the Spanish Society of Biometry (Spanish Region of the International Biometric Society). The event was supported by the Valencian Health Authority and took place in Valencia on June 21st, 2000. The practical session was based in WinBUGS and used tutorial material supplied by the UK partner.

4 DEVELOPMENT OF THE UK RAPID INQUIRY FACILITY

The methodologies and concepts underlying the UK Rapid Inquiry Facility (RIF) have been reported previously elsewhere (Aylin et al, 1999⁶, and SAHSU, 2000⁷), and so will not be repeated here. Instead, this chapter will focus on the developments undertaken during the period of this feasibility project. The aim of these developments was to improve the RIF for use both within the UK and for implementation within the EUROHEIS partner countries.

In order to enhance and expand the functionality of the RIF, a series of modules was devised. This type of modular development was advantageous as the modules could be developed independently and integrated into the core RIF as and when appropriate. The modules are complementary and were identified based on discussions with partners in the EUROHEIS project, and discussions in-house within SAHSU. The modules provide the mechanisms by which value can now easily be added to the RIF: for example when new health datasets are acquired, when new geographies are needed, or when different methods of defining the study area are appropriate. The concepts and methodologies inherent in the modules can be generalised and are therefore transferable to the other partner countries in EUROHEIS.

4.1 Description of modules

The first module is an essential building block for the RIF, for both point source investigations and disease mapping. It enables a list of EDs and/or wards to be read into the system to define either or both of the study area and the comparison region. This module forms the interface not only for more flexible point source queries but also for disease mapping and for extensions to the RIF such as importing a study area defined by air pollution modelling.

The second module provides the mechanism by which observed and expected counts and rates (directly and indirectly standardised, smoothed and unsmoothed) can be calculated 'on-the-fly' for the specific inquiry submitted. These can then be mapped using the techniques developed in the third module.

The third module provides a flexible interface for disease and exposure mapping. It produces a range of disease, exposure, socio-economic, demographic and geographical feature maps, which can be printed and exported in various formats.

4.2 System development

Major system development has been undertaken during the period of this feasibility project. The previous version of the RIF (Aylin et al, 1999) was developed within a Unix environment, using a range of software packages and development tools (Tcl/Tk, ORACLE PL/SQL, ArcInfo AML, C, HTML). As such, the system was restricted to

⁶ Aylin P, Maheswaran R, Wakefield J, Cockings S, Jarup L, Arnold R, Wheeler G, Elliott P (1999). 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. *Journal of Public Health Medicine*, 21 (3), 289-298.

⁷ Small Area Health Statistics Unit (2000). *A European Environment And Health Information System For Exposure And Disease Mapping And Risk Assessment: First Interim Report (Proceedings of first EUROHEIS meeting)*. Interim Project Report to the EU, pp.25.

running within a Unix platform and required that potential users obtain all the above packages to enable them to run the system. The majority of the EUROHEIS partners were operating within a Windows environment, and did not have the ORACLE DBMS, nor the ArcInfo GIS, nor the staff to use either of these packages. To implement the old RIF in the other countries would therefore be, at the least, very costly and time consuming.

It was decided, therefore, that the UK system would be ported across to a Windows environment and that it would be re-written using software that was readily obtainable and usable by the partner countries. The main aim for the feasibility project was to develop the modules outlined in Section 5.1, within the new development environment. This transition was not a trivial one, but it was made possible within the timescale of the feasibility project due to the experience gained from implementing the previous version of the RIF. The new system now employs two relatively common and affordable software packages: ArcView and Oracle (note that partners need not obtain the full version of Oracle to run the RIF: a copy of Oracle Personal will suffice). It is now platform independent, making it much more exportable to the other partner countries. Furthermore, where possible, the code has been written in a generic format so that it will operate on any datasets which are in the appropriate format.

Figure 1 gives a schematic overview of the new system, showing the software employed for the various components.

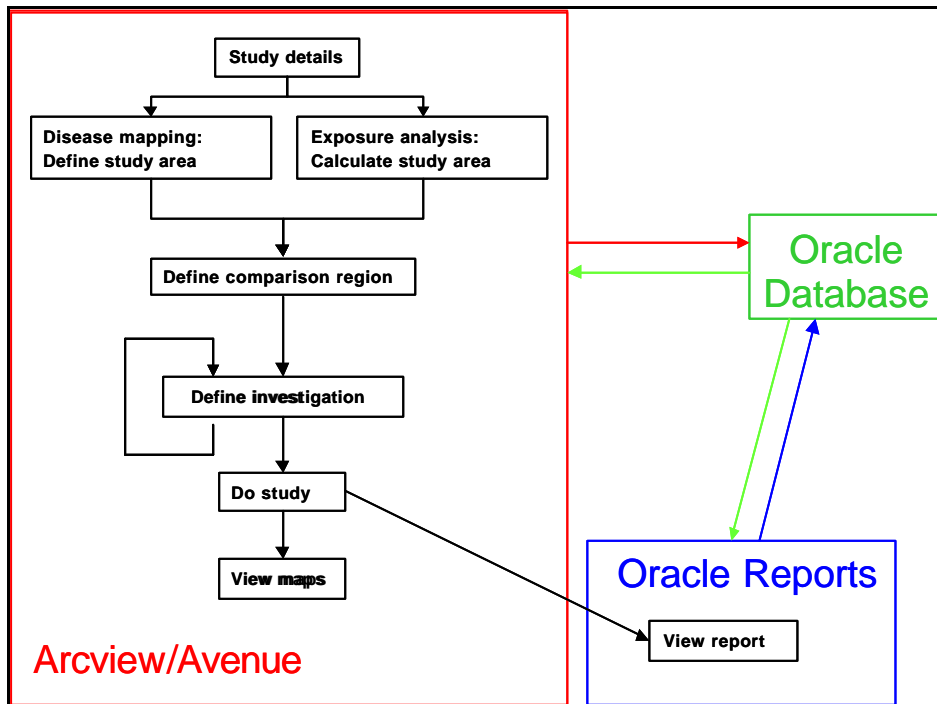


Figure 1 Schematic diagram of UK Rapid Inquiry Facility

5 FEASIBILITY OF IMPLEMENTATION OF THE SYSTEMS WITHIN THE PARTNER COUNTRIES

The purpose of this chapter is to explore the feasibility of implementing systems for point source investigations and disease/exposure mapping in each of the partner countries. Issues concerning the availability and accessibility of data suitable for such rapid inquiry systems have already been explored in Chapter 2, which concluded that suitable data are available within each of the partner countries. In this Chapter, we consider the wider issues concerning the development and implementation of such systems. We firstly outline the generic political, technical, economic and methodological issues which must be considered in assessing the feasibility of implementing a system in each country. We then go on to explore these issues with respect to each partner country, summarising any existing systems or methodologies within the country and assessing the overall feasibility of implementation.

5.1 Generic feasibility issues

There are issues regarding the feasibility of implementation of systems for point source investigations and disease/exposure mapping which must be considered in all countries. These issues may be classified as: political and organisational feasibility, technical feasibility, economic feasibility and methodological feasibility.

5.1.1 Political and organisational feasibility

Political support and organisations involved

The degree of political support and organisations involved in implementing and employing systems for point source investigations and disease/exposure mapping depends largely on the scale of implementation and the proposed usage of the system. The scale of implementation could potentially range from a local system through a regional facility to a national system. Clearly a national system requires the collaboration and cooperation of far more organisations than a local one, but all implementations, whatever the scale, require political support to be successful. In terms of the usage of the system, there is a range of potential scenarios, from dealing with infrequent inquiries on an ad hoc basis, through to using the system for some sort of regular surveillance.

Clearly, the degree of political support and organisational involvement required will vary depending on the precise use of the system within each country; some generic requirements can, nonetheless, be identified.

Any implementation will need the support of data providers, including those responsible for the maintenance and supply of health, denominator, socio-economic, environmental and geographical datasets, together with the national mapping and cadastral agencies. For the successful development of such systems, not only do these organisations need to provide the actual datasets but they also need to be integrated into the development and support of the system so that issues regarding the format, quality and appropriate use of the data can be discussed as an on-going process. This sort of dialogue is essential if meaningful results are to be gained from analyses, especially if the results are to be used to inform the formulation of policy.

Whatever the scale of implementation, the successful use of such systems requires the development of a network of key agencies involved in the investigation of relationships between the environment and health. Organisations likely to be involved include: the national bodies responsible for health and the environment; regional and local health authorities; local authorities; census agencies; national statistics offices; national, regional and local health registries; national mapping agencies; private data suppliers; and academic, research and quasi-public institutions.

The issue of where, or within what sort of institutional setting, such systems should be developed and administered is one which is clearly country-specific. It could be based within an academic institution, such as in the Small Area Health Statistics Unit (SAHSU) UK (note though that SAHSU is funded by various organisations within the UK government). In other countries though, there may be significant advantages to be gained by developing such a system within a governmental setting. Clearly the aims of the system, the agencies involved and the resources available will play a large part in this decision.

Data access and confidentiality issues

In order to carry out meaningful small area analyses, data must be available at a suitably low level of aggregation. In most countries there is some form of constraint on the level of aggregated data available, which seeks to preserve the confidentiality of individuals. If the minimum threshold is too high, analyses may not be able to produce meaningful information to inform policy decisions.

If such systems are to be used to provide rapid initial assessment of potential links between the environment and health, the infrastructure must be in place to enable rapid querying of the required datasets. In essence, this means that either the required datasets (see Chapter 2) must be obtained from the various dataset suppliers, compiled and then integrated into one geo-referenced system in one organisation, or, if the datasets are to be held by different organisations, then extreme close-coupling must exist between the systems holding the datasets in the different organisations (with rapid access available over some form of internet connection).

Clearly, confidentiality concerns will also play an important role in deciding whether the datasets are held centrally or by individual organisations; a central database offers the advantage of security in that data can be held on a secure, private network (as in the UK example), but it does mean that the dataset providers have to release the data to the central organisation, which may not have been carried out before. By contrast, if the datasets are retained by the individual dataset providers and data are transferred as and when necessary on an ad hoc basis, over some form of internet network, then there are clear security concerns, even if the internet link is purported to be secure.

The decision over the institutional setting of the system is also likely to be heavily influenced by the size of the datasets involved. For example, in the UK, because the Rapid Inquiry Facility is a national system, the datasets involved are very large, making efficient indexing and cross-referencing of the datasets essential for rapid data retrieval – such database operations would not be possible across an internet on datasets held on different systems; as such, analyses could not be carried out in a timely manner.

Experience in the UK at SAHSU has shown the very definite advantages to be gained by compiling all the datasets within one organisational setting. The various health, demographic, socio-economic, environmental and geographical datasets have been collated, integrated and interpolated onto one, secure, private network. This enables rapid data access, together with assurances that confidentiality can be maintained. Perhaps most important of all though, is that it has enabled the cross-checking of the various datasets and the detection of a range of errors ranging from case ascertainment and completeness checks through to the identification of systematic problems with the geo-referencing of the datasets. This type of 'added-value' quality control is only possible due to the integration of the various datasets into one common system. It is useful to note at this point though, that the time and effort required to compile, integrate and maintain such a database is not insignificant. It requires dedicated, skilled and experienced staff and resources and is an aspect of the development of such systems that should not be under-estimated.

Copyright and legal issues

It is important, at an early stage in the planning of the implementation of such systems, that an assessment be made of any likely costs and restrictions on the distribution and publication of results of analysis from the system.

5.1.2 Technical feasibility

Hardware requirements

The hardware requirements of such systems are mainly dependent on the scale and infrastructural setting of the implementation, which will be country-specific. If the system is to be a national one, then it will clearly require more hardware (processing power, memory, disk space, data transfer, backup and storage media) than a local one, due to the dataset size and processing requirements. The type of implementation will determine the type of network requirement - as described in section 6.1.1.2 there are various ways that the system could be configured: a private network may be required, or some sort of secure internet connection may be necessary.

Software requirements

The essential software requirements for the development of systems for point source investigation and disease/exposure mapping are some form of database management system (DBMS) and a geographical information system (GIS).

The DBMS provides the functionality to compile, cross-reference, index, search and query the various datasets required for the analysis. Additionally, in the form of SQL (structured query language), it provides the means of actually calculating the statistical counts and rates for the analysis. The GIS provides the ability to compile, integrate, interpolate, spatially query, interactively explore and visually display the various geo-referenced datasets. Importantly, it provides the means of defining areas of exposure and identifying populations at risk, and therefore the ability to define study areas and comparison regions. The DBMS and the GIS must be able to communicate seamlessly – i.e. closely coupled systems – so that not only can information be passed between the two systems, but ideally, they can also execute programs within one another.

Note that most GIS do incorporate significant database management functions, so the complete system could be implemented using just a GIS. However, if the datasets are

large, and if the system is to be used in a multi-user environment, then the database functions of the GIS are likely to be found wanting. In these instances, a proper DBMS is usually preferential, as it provides functionality for sophisticated indexing, security, user logging and so on. Additionally, reporting software is required to display the tabular results of the analysis. Depending on the DBMS and GIS software employed, this functionality could be available within those packages, or an additional software package may be required.

Staff requirements

Studies of links between the environment and health are inherently multi-disciplinary in nature. To develop and employ a health and environment system that offers rapid initial assessment of potential disease clusters, together with disease and exposure mapping, requires the development of a network of skilled experts from a range of disciplines and professions. The complex processes of study design, data collation, integration and interpolation, spatial statistical analysis, the visualisation of results, the interpretation of results and recommendations for actions, requires the input (to varying degrees and at varying times) of: epidemiologists; public health professionals; statisticians; geographers and GIS specialists; environmental scientists; and database, programming and IT specialists.

5.1.3 Economic feasibility

Many of the datasets required for use in the systems may have a cost associated with their acquisition and use for such studies. In some instances, these datasets may already be held in-house, in others, they may have to be obtained. In some cases, even if the datasets are already held in-house, there may be a charge associated with using them for this specific project. The staff costs for developing and implementing the system must be calculated, together with the staff costs for administering the system once it is up and running, and for any future and on-going developments to the system. The costs of purchasing additional hardware and specific software packages must also be identified.

As the majority of countries are usually building on existing networks of experts in the field, there are frequently already existing, well-established, programmes of research and system development. As such, the EUROHEIS programme has already benefited from, and will continue to benefit from, substantial levels of non-EU funding, notably in the form of the salaries of the majority of partners involved in the project.

5.1.4 Methodological feasibility

As the range of questions and problems being investigated varies considerably from country to country, there are specific methodological issues associated with each particular country, and these are discussed briefly in Section 6.2.

5.2 Feasibility in each country

5.2.1 Denmark

There is political support at the national level for the development of a health and environment system for Denmark. A prototype system has already been developed for Vejle County, which has proved, at a local level at least, that the required datasets are both available and suitable for use in such a system. Additionally, the prototype has

gone some way towards developing some of the necessary techniques and methodologies for preparing and linking the datasets for analysis. Crucially, the project has the support of key organisations and individuals, including the Chief Medical Officer and the National Board of Health, the Department of Statistics, the National Institute of Public Health and the National Survey and Cadastre. The National Board of Health will provide the necessary data for the implementation of such a system. The National Survey and Cadastre has undertaken to provide the necessary geo-referenced datasets and to play a central role in the development of methods to link the various health, socio-economic, demographic and geographical datasets. There are no confidentiality constraints on the use of the various datasets within the system. It is most likely that the system would be based at the National Board of Health or National Institute of Public Health, Copenhagen. In terms of political and organisational feasibility then, the development of a Danish rapid inquiry facility is definitely feasible.

Technically, Denmark is well placed to develop such a system. Because of the range of organisations supporting the development of the system, there is a wide range of skills expertise available within the developing network. The National Survey and Cadastre has already invested significant resources in creating a national geo-referenced address register, which can be linked to the various health and socio-economic datasets. The Vejle County prototype demonstrated the usefulness of this address-based geo-referencing in giving extremely high resolution data about present and past residence of populations and hence the identification of very precise populations for study.

In terms of hardware and software, the current Vejle County prototype system is a customised application, written in the GIS package MapInfo. It is possible that for a national system, a specific DBMS may prove useful. However, if the datasets involved are not too large, then it may be possible to retain the whole system in a GIS. The UK system has been developed in the GIS ArcView, which has very similar functionality to MapInfo. The National Survey and Cadastre has a range of GIS packages, such as ArcInfo, ArcView and MapInfo, available so there are a number of options for the future development of the Danish system.

There will be charges associated with obtaining the Central Population Register and the National Prevention Register. It is impossible to say for definite exactly how much this will cost, but there does seem to be a political move towards letting register related and health surveillance research be free of charge when dealing with the national registries. EU funds are therefore not anticipated to be required for this purpose. The major costs in the Danish system will be for dedicated staff to develop methodologies for the linking of the various geo-referenced datasets. As stated previously, this is a considerable undertaking, but with the support of the various organisations within Denmark, and with the sharing of experience amongst the EUROHEIS partners, this should be attainable. A post has already been funded by the National Board of Health, dedicated to link their health registries to the GIS.

The key methodological issues to be resolved in Denmark involve the linking of the various datasets. One particular methodological issue concerns the design of base geographical units for the aggregation of the datasets onto some common geography. Whereas many countries are constrained to pre-defined administrative units for their base units, Denmark is in the enviable position of being able to design purpose-specific units. The National Survey and Cadastra has already been investigating methodologies for this purpose and has assigned a full time person to deal with this aspect.

In summary, Denmark is very favourably placed to develop systems for point source investigations and disease/exposure mapping. In terms of political, organisational, technical, economic and methodological issues the implementation of such a system is very definitely feasible.

5.2.2 Finland

A prototype point source investigation facility, SMASH (Small Area Statistics on Health), has been developed at the National Public Health Institute (KLT) in Kuopio. This system provides the ability to carry out point source investigations for cancer registry data for Finland. All the required data are already held by KTL. There are no plans at the moment to extend the system to incorporate other datasets. Rather, the main area of development is likely to be in addressing some significant methodological issues concerned with sparse data and the interpretation of results (see section 3.2.2.4). There are no significant organisational or political barriers to the further development of the system.

The current system has been developed within the GIS package ArcView, with customised programs written in Avenue (ArcView's programming language). This is the same GIS software as that used in the UK, meaning that any programs and methodologies should be easily transferable between the two partners. The Finnish partner does not have GIS or database expertise in-house and is therefore likely to look to gain this sort of expertise from other partner countries.

The main costs of the implementation in Finland will therefore be staff costs for the development of methodologies for dealing with sparse data. There are no significant additional data costs envisaged, and additional hardware and software requirements should be minimal.

The main issue to be addressed under the next stage of the EUROHEIS project in Finland is the development of methodologies for dealing with sparse data. In particular, the use of smoothing techniques is problematic due to the presence of small, densely populated areas within large expanses of unpopulated areas. The UK, French and Finnish partners are expected to work very closely in developing techniques to address such problems.

In summary, the main challenge to the further development of the prototype rapid inquiry facility in Finland is methodological. The political, organisational, technical and economic feasibility of the project is largely without problems.

5.2.3 Italy

The main health dataset currently available to the Italian partner is the national mortality dataset. However, the Italian partner is not able to access the mortality database directly on their computers. They, and most other organizations, are only able to access data according to strict confidentiality constraints. Data can be accessed for a minimum period of three years, and for a minimum population threshold of 10,000 people. The smallest geographical area for which data are available is a municipality, of which there are over 8000 in Italy. The population within the municipalities varies considerably, with major cities comprising one municipality. The University of Milan has developed a Mortality and Population GIS 'Atlas', which enables the querying of the national mortality

dataset. The parameters of the query have to be entered into the system interactively; it is not possible to run batch programs to automate the process. The system can calculate various statistics, it can carry out statistical tests and it can produce disease maps. However, the Italian partner uses it mainly to create tables with counts and populations for the area of interest, which they then import into their own programs to run statistical analysis and to produce maps. They have written macros, which automate the process of producing maps within the GIS software package MapInfo. Most statistical analyses are carried out using S-plus. Despite having a range of statistical functionality, the usefulness of the system is constrained by the confidentiality constraints imposed on the data.

As noted in Section 2.2.3, there are some areas (Florence, Turin, Sicily and Puglia) for which individual level address-based mortality data are, or soon will be, available. For these areas, it is possible that a system could be developed to carry out point source investigations and disease/exposure mapping at a local level. As there is little in-house expertise available for the development of such a system, the most likely process of implementation would be through the adoption of the UK-developed system, with the Italian data being formatted appropriately for use in the system.

There are significant political and methodological issues to be addressed in Italy due to the high level of aggregation. In some areas, complete cities are contained within one municipality, which poses difficulties for the interpretation of results from small area analyses. These methodological issues are likely to form a significant part of the next stage of the EUROHEIS project.

5.2.4 Spain

The Spanish partner is supported by a wide range of experts from academia and from the health authorities. The focus of the Spanish implementation will be in the two autonomous communities of Andalusia and Valencia. The Spanish partners do not have an existing system for the rapid initial assessment of potential disease clusters, but they do have experience in disease mapping and in the application of spatial statistical smoothing techniques (using the WinBUGS software). As in Italy, the smallest geographical unit for which data are available is the municipality, which poses constraints on the degree to which small area analyses can be carried out. However, a key focus of the Spanish implementation will be to investigate to what degree such aggregated data can support small area analyses. The majority of datasets required for the implementation of such systems are already held by the Spanish partners, and most do not have costs associated with their use in the project. The implementation is therefore feasible in political, organisational and economic terms.

Technically, the Spanish partner is well set up to implement such a system. They already have the GIS package ArcView, which would enable them to take modules of the UK-developed system and apply them to appropriately formatted data. There is, though, a lack of suitably trained in-house GIS staff, so they would be looking to other countries to share expertise in this respect.

Methodologically, as mentioned previously, one of the key focuses of the Spanish implementation would be to explore the extent to which meaningful interpretation can be obtained from small area analyses of the municipality-level data.

5.2.5 Sweden

There is political support for the development of a system for disease and exposure mapping within Stockholm County, and potentially for point source investigations. The Stockholm Centre of Public Health (SCPH) is committed to the project and collaborates with the Centre for Epidemiology at the Swedish National Board of Health and Welfare and with the Office of Regional Planning and Urban Transportation within the Stockholm County Council.

A prototype mapping system already exists for Stockholm County, but it is limited in its flexibility and functionality. The main focus of the Swedish programme will therefore be to adapt and implement the UK-developed rapid inquiry facility for use with the Stockholm data and to prepare and link the appropriate datasets for use in the system. Population counts, disease and mortality data are already available, whilst data on socio-economic indicators will need to be obtained and prepared. The cost of the socio-economic data will be dependent on the variables required.

Technically, there should be few problems with implementation as the datasets are of a manageable size. In terms of staff, there is a lack of sufficient GIS expertise in-house, therefore the Swedish partner will be looking to gain experience from the other EUROHEIS partners, especially the UK, in this respect.

5.2.6 UK

The implementation of a national prototype rapid inquiry facility has already been successfully carried out within Great Britain (England, Wales and Scotland). It is this prototype system that has formed the basis for the EUROHEIS project. Improvements to the system that have been carried out during the period of this feasibility project have already been described in Chapter 4.

5.3 Summary

In summary, the implementation of point source investigation and disease/exposure mapping systems are feasible in all partner countries. Whilst there are generic issues to be addressed in all countries, the precise form of the implementation varies from country to country, and is dependent on the specific public health questions of importance in the country and on the different practical constraints and opportunities within the country.

ACKNOWLEDGEMENTS

UK

The Small Area Health Statistics Unit is funded by a grant from the Department of Health, Department of the Environment, Transport and The Regions, Health and Safety Executive, Scottish Executive, National Assembly for Wales and Northern Ireland Assembly. The views expressed in this publication are those of the authors and not necessarily of the funding departments. This work is based on data provided with the support of the ESRC and JISC and uses census and boundary data, which are copyright of the Crown, the Post Office and the ED-LINE consortium.

Sweden

The disease mapping system currently in use in the Stockholm County was developed from a system designed by the Center for Epidemiology (EpC) at the Swedish National Board of Health and Welfare. The development was carried out in close collaboration with the EpC and was funded by grants from the Stockholm County Council.

Spain

The Department of Statistics and Operation Research in the University of Valencia has been funded by a collaborative agreement with the Valencia Regional Health Authority. Data has been provided by the Valencia Regional Health Authority, Valencia Institute of Statistics, Valencia School of Health Studies and Andalusian School of Public Health as partner institutions jointly with the support of the Valencian Environment Department.

APPENDIX 1

**A perspective on health inequalities and
area-based socio-economic deprivation**

A perspective on health inequalities and area-based socio-economic deprivation

Dr Alan Kelly, Small Area Health Research Unit (SAHRU), Department of Community Health & General Practice, Trinity College, Dublin

Abstract

There is a general consensus - based on the substantial body of international and national research, that differentials in socio-economic status lead to differentials in experienced morbidity, premature mortality and health services demand. These differentials will not only exist as a direct consequence of economic factors - they will reflect inequities by gender, age, social marginalisation and geographic location. Analysis of health and exposure data by geographic area, particularly by small-area, must recognise the serious potential for confounding that deprivation - however measured - represents. This report briefly reviews the evidence for socio-economic inequalities in health, the link with deprivation, how deprivation may be measured at small-area level and identifies various implications for spatial modelling of disease patterns.

Key words

Inequalities in health, small-area analysis, socio-economic deprivation, area deprivation, Bayesian spatial modelling.

1. Overview

This report begins with a review of the evidence of the links between health inequalities and socio-economic status. The importance of *context*, i.e. a geographic perspective – rather than simply a cultural perspective - in the investigation of health experience and inequalities is stressed. The report then addresses the issues of what is deprivation and how has it been measured in recent decades. At this point the important conceptually important distinction between social inequality and material deprivation is indicated. Considerations in the creation of a material deprivation index by small area are set out. The possibility of a standard small area Euro deprivation index based on the existing NUTS regions is considered but regarded as impractical at present. An alternative approach, that relies on existing national administrative areas – and these will differ by country - is considered practical. In addition, structurally equivalent indices of material deprivation should be capable of being developed and some recent efforts in continental European countries are cited. The question of whether to formally adjust, or not, and in what manner, for area-level deprivation during the modelling of disease patterns or exposure is briefly addressed. The report concludes with a short set of recommendations. A technical annex provides – by way of illustration – the steps in the development of a national small-area deprivation index.

2. Socio-economic inequalities in Health - a review

In 1980 the DHSS in the United Kingdom published a seminal on Inequalities in Health which has become anonymously known as the 'Black Report', after Sir Douglas Black, the Chairman of the working group who drew up the report (DHSS, 1980). In essence, this report compiled evidence of marked differentials in health outcomes - "*...from birth to old age, those at the bottom of the social scale have much poorer health and quality of life than those at the top... gender, area of residence and ethnic origin also have a deep impact*" (Townsend et al. 1988). The impetus provided by this report has galvanised inquiry in the United Kingdom with the consequence that a wide diversity of health outcomes have

been researched in detail with respect to social class membership. Davey Smith et al. (1990) revisited the findings of the 'Black Report' 10 years on and concluded that socio-economic inequalities in health outcome had widened and indicated that the observed trends "... suggest that further widening of mortality differentials may be expected". This was recently confirmed by Raleigh et al. (1997) who reported that "*Differences in life expectancy had widened over the decade ... and prosperous areas with greatest longevity had seen the largest gains*".

There is now considerable evidence that material deprivation - whether indirectly assessed in terms of social class, socio-economic status or by an area-based index of deprivation - is strongly linked with many common diseases (physical and psychiatric) (Eachus et al. 1996; Weich and Lewis, 1998) and with premature mortality (Eames et al., 1993). These findings have been replicated internationally as reported by Reijneveld and Schene (1998), Borrell et al. (1997) and Mcisaac and Wilkenson (1997) and with an overview in a BMJ editorial (1998). In Northern Ireland, Kee et al. (1996) confirmed socio-economic differentials in cause-specific mortality. Further afield, Ross et al. (2000) show the link between income and mortality in the US. However, contrary to the vast majority of publication on this topic, Ross et al. challenge the assumption that the generally held relationship between income and health inequalities holds universally with their analysis of data from Canada. They conclude (given the lack of such an association in Canada): "... *that the effects of income inequality on health are not automatic and may be blunted by the different ways in which social and economic resources are distributed in Canada and in the United States.*" A number of researchers have shown that deprivation has a direct impact on demands placed on primary care provision (Worrall et al., 1997; Ben-Shlomo et al., 1992; Scaife et al., 2000, Gunning-Schepers, 1998). In a recent study in Scotland, Pell et al. (2000) set out to determine whether the priority given to patients referred for cardiac surgery is associated with socio-economic status. They found that socio-economically deprived patients are more likely to

develop coronary heart disease but are less likely to be investigated and offered surgery once it has developed. Such patients may be further disadvantaged by having to wait longer for surgery because of being given lower priority.

Hart et al. (2000) review the influence of socio-economic circumstances from early to late in life in Scotland, conclude that poorer socio-economic circumstance is associated with greater stroke risk, with adverse early-life circumstances of particular importance. In Spain, Vazquez et al. (2000) show socio-demographic differentials in asthma in the population. Also, Alvarez-Dardet and Ruiz (2000) offer evidence for health inequalities in Spain. In Sweden, Hagquist (2000) explores the issue of socio-economic differences in adolescents' smoking behaviour, using academic orientation as an indication of social position.

In light of the increasing proportions of children living in poverty in the UK, Reading (1997) shows that wide inequalities continue to occur in most types of infectious disease in childhood and the consequences stretch into adulthood. He suggests that *“Environmental and material factors are likely to have a stronger influence than social differences in behaviour, attitudes and lifestyle, and they can affect susceptibility to infection at the level of individuals, family, and community, and in access to health care.”* And concludes that *“More emphasis needs to be placed on structural and community wide interventions than those directed at changing individual behaviour if a reduction in inequalities in childhood infection is to be achieved.”*

2.1 Recent key overviews of the international evidence

Davey Smith et al. (1998) concluded, following a major longitudinal study, that:

- adverse socio-economic circumstances in childhood have a specific influence on mortality from stroke and stomach cancer in adulthood;
- deprivation in childhood influences risk of mortality from CHD and respiratory disease in adulthood;

- and mortality from lung cancer, other cancer, and accidents and violence is predominantly influenced by risk factors that are related to social circumstances in adulthood.

Knust, A.E., et al. (EU Working Group on Socio-economic Inequalities in Health)

North-south gradient: mortality from IHD was strongly related to occupational class in England & Wales, Ireland, Finland, Sweden, Norway and Denmark, but not in France, Switzerland and the Mediterranean countries. They find that

- in the latter countries, cancers other than lung cancer and gastrointestinal diseases made a large contribution to class differences in total mortality;
- inequalities in lung cancer, cerebrovascular disease, and external causes of death also varied greatly between countries; and
- the mortality advantage of people in higher occupational classes is independent of the precise diseases and risk factors involved.

Mackenbach et al. (2000) undertook a comparative analysis of socio-economic inequalities in relation to reported cardiovascular disease in the United States and 11 western European countries found that:

- in all countries mortality from cardiovascular diseases is higher among persons with lower occupational class or lower educational level;
- within western Europe, a north-south gradient is apparent, with relative and absolute inequalities being larger in the north than in the south;
- and for ischaemic heart disease, but not for cerebrovascular disease, an even more striking north-south gradient is seen, with some 'reverse' inequalities in southern Europe;

This group concluded that:

- socio-economic inequalities in cardiovascular disease mortality are a major public health problem in most industrialized countries.
- closing the gap between low and high socio-economic groups offers great potential for reducing cardiovascular disease mortality.

- developing effective methods of behavioural risk factor reduction in the lower socio-economic groups should be a top priority in cardiovascular disease prevention.

2.2 From Evidence to policy and policy to action

Whitehead (Ed., 1995) with colleagues, in *Tackling Inequalities in Health*, provided a timely reminder that the evidence was in, and that there was now an urgent need to progress to policy initiatives to address health inequalities. In response, the DHSS commissioned a further major review in 1998 under the chairmanship of Sir Donald Acheson which was published as *The Independent inquiry into inequalities in health*, (inevitably known as the “Acheson Report”), The focus of this report is on mechanisms for tackling inequalities in health in the settings of schools, the workplace and neighbourhoods. They propose a socio-economic model of health and its inequalities:-

Socio-economic inequalities in health reflect differential exposure – from birth and across the life span – to risks associated with socio-economic position. These differential exposures are also important in explaining health inequalities which exist by ethnicity and gender. ... the research task is to trace the paths from social structure, represented by socio-economic status, through to inequalities in health. This can be done in stages, for example showing that work is related to pathophysiological changes such as raised blood pressure or biochemical disturbances what are in turn related to disease risk; or showing that the social environment in which people live is related to their health behaviour, such as patterns of eating, drinking, smoking and physical activity.

For a critical appraisal of this report one year on see Black et al. (1999).

For a detailed review of the evidence presented to the Acheson Committee and an interpretation of the policy implications, the report edited by Gordon, Shaw, Dorling and Davy Smith (1999) is both comprehensive and critical.

The WHO strategy 'Health 21' provides 21 targets which outline the needs of the whole European Region and provide suggestions for action to improve health. It offer a framework for action at all levels and benchmarks against which progress can be measured. Key strategies for action ensure that scientific, economic, social and political sustainability drives its implementation:

- multisectorial strategies to tackle the determinants of health, taking into account psychosocial, economic, social, cultural and gender perspectives, and ensuring the use of health impact assessment;
- a participatory health development process that involves relevant partners for health at home, school and worksite, local community and country levels, and which promotes joint decision-making, implementation and accountability.

In the Foreword to a recent WHO report, Dr. Agis Tsouros, Head, Centre for Urban Health, WHO, Copenhagen argues that *“The good news is that decision-makers at all levels increasingly recognise the need to invest in health and sustainable development. To do this, they need clear facts as much as they need strategic guidance and policy tools.”* and *“A call to decision-makers and public health professionals to address the social determinants of health should rest on clear evidence. ... The provision of up-to-date information on the key areas of social determinants, in a concise, clear and authoritative form.”*

Wilkinson and Marmot's 1998 report (published by the WHO), entitled *Social Determinants of Health*, summarises in a very accessible manner, the societal determinants of inequalities that must be recognised nationally and addressed structurally.

3. A Geographic dimension – the importance of place

Geographic differences in health experience and outcome have long been noted. The 'Black Report' quantified the historical North-South divide in England; and the WHO has for many years signalled the importance of diet in relation to risk of cardiovascular disease on the basis of the observed differences in CHD outcome in those countries bordering the Mediterranean as compared to the more northerly states within Europe.

Analysis by geographic area, i.e. areas, rather than individuals, are the units of analysis, is both appealing and often more practical and cost effective in epidemiological research into population exposure and disease patterns. This is evidenced by the significant number of articles retrieved - as an exercise - from a search of just two major international journals (The Journal of Epidemiology and Community Health - JECH, and the British Medical Journal – the BMJ) for the last couple of years (see Bibliography following References). The search involved the following key words: area + deprivation + disease + exposure. The diversity of the material retrieved confirms the level of activity and interest by researchers in many European countries in area-based or ecological studies into health patterns and inequalities.

To the extent that such research is likely to be very influential in raising awareness concerning national problems and thus engaging the interest of policy makers in developing new initiatives (with the experience of the UK providing an leading example), it is reasonable to be concerned about the implications for research conclusions and subsequent recommendations of the so-called *ecological fallacy*. Robinson, as long ago as 1950, described this problem: as set out more recently by MacRae (1994) “ ... *the problem is that the correlation between two variables when the group is used as the unit of analysis may be quite different from the correlation between those two variables when individual*

people are used as the unit of analysis .” MacRae points to the existence of substantial evidence [in socio-economic inequalities in health] using individuals as the unit of observation to support the conclusion that ecological correlations between socio-economic deprivation and health arise from associations among the relevant variables in individuals. He concludes: “... those who would seek to criticise or ignore research on socio-economic deprivation and health using ecological correlation studies in which appropriate indices of deprivation have been used cannot legitimately seek support from the ecological fallacy.”

3.1 Treat the area or the individual?

Macintyre (1999) presents an additional highly important implication of area-based, as distinct from individual-based, research. She states the issue as follows: (I quote *in extenso*): “... whether these observed geographical inequalities are simply a reflection of the composition of the population in different areas (poor people die earlier, so areas with lots of poor people will have high deaths rates; and poor people will have the same death rates wherever they live), or additionally reflect something to do with the physical and social context (most poor people may have no choice but to live in areas which have health damaging characteristics, and the death rates of poor or affluent individuals will vary depending on what sort of area they live in).” Macintyre points out that a consideration of this question may well lead to research and policy initiatives that are directed more toward the individual (or the group to which he/she belongs) rather than addressing the health damaging and health promoting features of areas. This is fundamentally about the importance of *context*—by this I mean the wider social and environmental circumstances in which one (affluent or poor) lives. Clearly, this is undisputed when we are concerned with the issue of *equity of access* to services – where geographic location may represent a barrier to both rich and poor alike (although usually not to the same degree!). Macintyre cites many studies that offer evidence both in favour of the importance of context and where context did not appear to

contribute over-and-above the sum of individual effects. But she does offer the following in conclusion: “*Thus, the relationship between socio-economic status and health at both an individual and small-area level seems to vary by latitude, by type of ward, district or region, and by rurality.*”

4. What is deprivation and how is it measured?

Deprivation is a concept that has taken a variety of forms and has had many different meanings that have evolved over time. It is generally recognised as a composite concept, in that there is no single variable that can be said to measure it but rather a number of variables must be combined in some way. While poverty, as measured by household income, is usually recognised as an important component of deprivation it is only one of many variables affecting quality of life. Moreover the measurement of poverty remains controversial, governments tend to dislike the political implications and the poor dislike the stigma of being labelled poor.

The measurement of deprivation has been pursued energetically in the UK since the early 1980s and a number of deprivation indices have been put forward (for a discussion of these, see Annex H of PAT18 Report (2000), Working Paper: *Measuring Deprivation – A review of Indices in Common Use*, and also – *Indices of Deprivation*, DETR, 2000).

The terms ‘material deprivation’ and ‘social deprivation’ are often and inappropriately used as though these are equivalent conceptually and practically. This is not the case. Townsend explicitly defined material deprivation as entailing “*the lack of goods, resources, amenities and physical environment which are customary, or at least widely approved in the society under consideration*”. The idea has come to be applied to conditions (i.e. physical and social circumstances) rather than resources or income and can therefore be distinguished from the concept of poverty, though the two are closely related.

This conceptualisation can explain why people can experience deprivation but do not necessarily live in poverty. Implicit in the definition is the notion of *relativity* (Wilkinson, 1997), i.e. material deprivation is specific to a given time and place – it is not absolute in nature. By way of contrast, Townsend defined social deprivation as “*non participation in the roles, relationships, customs, functions, rights and responsibilities implied by members of a society and its sub-groups. Such deprivation may be attributed to the affects of racism, sexism and ageism.*”

While the Townsend Index and the Carstairs Index relate to material deprivation, other indices are effectively a combination of indicators of both material and social status – the so-called Jarman Index (an underprivileged area score), for instance. This mixing is conceptually and practically problematic, but its appeal may have to do with the ‘big is better’ school of index construction. Townsend warned against this, yet we see - even within the health sector in the UK – a tendency to use both the Townsend Index (or equivalently, the Carstairs Index in Scotland) and the Jarman Index. However, it is interesting to note that Public Health clearly favours the uniform use of the Townsend Index. (Although this choice is not without its difficulties in that the Index is census-based, and in the UK the census is only conducted every 10 years. This problem with the currency of the index may have contributed, in part, to the imperative to develop a more timely and locally-based set of indices as in the effort of the Department of the Environment and regions, see DETR 2000.)

In Ireland we reviewed the various approaches to creating deprivation indices since the early 1980s in the UK, before deciding (SAHRU, 1997) – following detailed simulation and an appreciation of the clear conceptual rationale - on a model that closely resembles the Townsend Index. This index has been adopted by the regional health authorities and other state agencies in the Republic of

Ireland.⁸ Some of the practical issues that arise in terms of choices to be made during the construction are illustrated in the Annex.

4.1 A generic recipe for Index construction

What follows is a set of key questions and issues that require consideration prior to, or during the construction of a deprivation index.

Pre-requisites:

- conceptual validity - what are you trying to capture?
- interpretability – not too many component indicators please!
- data quality
- statistical sensitivity – quantifying variation
- geographic level - how 'local' do we need to get?
- data update frequency - monthly/annual/5-yearly/every 10 years?
- comparability/consistency with existing indexes

Irrespective of where applied, the conceptualisation of material deprivation as developed by Townsend remains the most compelling in the context of this exercise. Note, this does not imply that the choice of indicators must follow his original proposal. National levels of poverty and deprivation differ, so too should an index designed to be sensitive to geographic variation.

Key question - are suitable data available?

- data must be available at the same geographic level
- covering the same time period
- updated with the same regularity
- of comparable quality

⁸ However, as already noted, as the index is census-based (which is conducted every 5 years in Ireland) there is a strong incentive to develop locally-relevant community indicators of deprivation. This is work in-hand and is being tested in several Local Authority areas presently.)

General technical issues:

- choice of indicator expression – given interest in ‘unemployment’, how to define this
- denominator is important – know your reference population
- is variable rescaling required – do you need to Log transform?
- how to combine individual indicators - adding indicators *versus* model-based weighting
- PCA *versus* Factor Analysis for weighting
- qualities of resulting Index
- Finally, do not mix ‘apples’ and ‘oranges’ – examine correlation structure of indicators

Avoid the ‘Big is Better’ Model of index creation

The Jarman Index and that proposed by the DETR in the UK, might well qualify under this rubric. Such indices effectively combine indicators of *material deprivation* with socio-demographic indicators of *potential risk*, and this is not only conceptually flawed, but may also be seriously misleading in practice.

Negative correlations exist between selected indicators *within* the set of demographic indicators and *between* this group and the material deprivation group - more indicators, more problems!

Inclusion of more indicators may add little to explanatory power (as these are highly correlated with existing indicators) hence little to be gained in defining an index on a more complex basis. It is not clear how one interprets a high score in terms of the underlying indicators in this situation - a requirement for any practical index.

Before you begin - Do you really need an index?

Assuming yes, then apply The KISS principle-

- keep it simple and interpretable
- consider how to combine indicators - simple addition will not (usually) do!
 - weight contributions for individual indicators – PCA is one choice.
- consider how to present the index - ranking followed by mapping is very compelling!

Generally, tailor the choice of indicators being considered for the index to solve specific questions/problems - avoid generic solutions – an index developed in one country is unlikely to fit the requirements of another with the same degree of sensitivity.

4.2 Existing approaches

Cairstairs (2000) compares and contrasts the construction of four of the most popular indices in use in the UK and further afield, *viz.* the Townsend, Carstairs, Jarman and DETR indices. In addition to documenting the choice of indicators comprising these indices, she discusses the calculation of the respective deprivation scores at the operational area level.

Key differences – apart, that is, from content of the indices – relate to the statistical treatment of the underlying variables. For example,

- the prior standardisation of the variables using *z-scores* (Carstairs, Jarman, Townsend) or *signed chi-square* (DETR);
- use of variable weighting to allow for intrinsic or empirical differences in importance (Jarman used weights derived from self-reported degree of importance attached to a list of questions by surveyed General Practitioners – this approach has been criticised and alternative multivariate statistical methods have been used, see below);
- use of variable transformations – e.g. the \log_e or arc-sin transform (Townsend, DETR, Jarman)

In addition, the choice of cut-off points of the resulting score – often employed, to define ‘deprived areas’ or ‘affluent areas’ – may differ from index to index, but is most often based on the deciles or quantiles of the computed score variable.

This latter approach seems counter-intuitive.

4.3 European experience – use of area-based Deprivation Indices

At the beginning of the last decade, Mackenbach et al. (1991) reported few published instances of research on the link between small-area level deprivation and mortality at a national level within Europe (outside of the UK).

A web-based search of key journals publishing European research in English turned up very few articles in which any form of small-area deprivation indices had been developed. The exceptions are noted next. Of course, many more instances might well appear in national journals in languages other than English.

In the Netherlands, Reijneveld (1996) has developed a Jarman index to study GP workload in Amsterdam. In Sweden, Malmstrom, et al. (1998) contrasted the Swedish Care Need Index with a local version of the Jarman Index and the Townsend Index and found that the correlations between scores was quite high. Bajekal, M., et al. (1996) also report on a similar exercise in Sweden in the use of the Jarman Index. Benach, and Yasui, (1999) analysed the geographic patterns of mortality in the 2220 small-areas in Spain using an index similar to the Townsend Index. As already noted above, in Ireland we have a Townsend-style national small-area (3421 District Electoral Divisions) deprivation index and this has been used in the analysis of mortality and morbidity data (SAHRU, 1997, Kelly et al., 1998, Kelly, 1999) and with a particular focus on health inequalities (Kelly, Sinclair, 1997). Further details of our experience in its construction appear in the Annex.

4.4 Potential for a common definition of ‘small-area’

In this report, I am working from the premise that relevant questions concerning health status, population needs, health care delivery and uptake, are sensitive to both scale and location. High levels of aggregation mask significant geographic variation in population characteristics of concern to public health. Analysis on a sub-county level must provide a more sensitive basis for the identification of need and delivery of care. Hence the importance of small-area analysis. But what constitutes a ‘small-area’?

Reijneveld, et al. (2000) confirm the importance of the careful choice of small-area classification. MacRae (*op. cit.*) has provided ample warning regarding the *ecologic fallacy* – and geographers have long struggled with the so-called Modifiable Area Unit Problem (MAUP for short). The capacity for defining a ‘small-area’ varies significantly throughout Europe, with the use of a 1 kilometre square grid employed in parts of Scandinavia, to rather more artificial and irregular boundaries derived for census enumeration in the UK and Ireland and on to whole municipalities in some other European states. Pragmatically, the definition of ‘small-area’ is often a direct function of the availability of census (or other regularly collected administrative data) being available periodically (and the frequency of this also varies across the continent).

If a common definition of ‘small-area’ was to be employed in a majority of EU countries, then the existing NUTS classification may provide an operational basis.

The acronym NUTS stands for Nomenclature des Unités Territoriales Statistiques or translated as: Nomenclature of territorial units of statistics. (See <http://irpud.raumplanung.uni-dortmund.de/irpud/pro/sasi/sasid7.htm> for the following review of NUTS.)

The Nomenclature of Statistical Territorial Units (NUTS) subdivides the European territory into interrelated levels. It is essentially a hierarchical classification in which each higher level region involves a grouping of a set of lower level regions. Beyond NUTS level 0 which corresponds to countries the classification consists of three regional levels (NUTS 1-3) and two local levels (NUTS 4 and 5). NUTS level 3 is however generally the lowest level at which Eurostat data are published. The current NUTS nomenclature subdivides the territory of the European Union into 15 NUTS 0 regions (countries), 77 NUTS 1 regions, 206 NUTS 2 regions and 1031 NUTS 3 regions. The NUTS classification tries to integrate the different national administrative units for which statistics are collected. However, the classification does not provide a harmonised set of regional units. At each level the number and size of regions vary greatly from country to country and not all countries and regions have a lower level subdivision. Generally, the variations in the size and population of the units increase with the level of disaggregation.

Thus, NUTS Levels 4 and 5 are evidently not defined for most EU countries. This would preclude, for now, any equivalent operational definition of small-areas (NUTS Level 5) across partner countries. However, this does not preclude the development of regionally similar structures. For example, in the UK and the Republic of Ireland – for historical reasons – a similar structure applies at a sufficiently disaggregated level to clearly qualify as ‘small-area’ for these purposes, viz. NUTS Level 5 which corresponds to electoral wards in the UK and in the Republic of Ireland. Similarly, in several Scandinavian countries, a 1 kilometre or 10 kilometre square grid exists for administrative purposes. And, as we have seen above, analogues of NUTS Level 5 have been developed for the Netherlands and Spain, and these point to the possibilities for other partner countries. The availability of population economic and demographic data for similar small areas – at some point during the last decade – suggests that nationally relevant deprivation indices can be developed. These need not strictly follow the same recipe as employed in the UK. Because the chosen indicators

comprising the Townsend Index (as a case in point) are appropriate and sensitive to deprivation in the UK, it does not follow that the same indicators are thereby the best choice for measuring deprivation in Denmark or France, for instance. Indeed, it is likely that other indicators will prove more appropriate in the different socio-economic circumstances pertaining in different countries.

5. Modelling issues

For this section, I am assuming that the modelling entails small area analysis of, for example, a disease pattern and that we are typically concerned with providing a valid estimate of the underlying standardised incidence or prevalence ratio. The methodology is a variant on the Bayesian spatial model and the estimation engine is Markov chain Monte Carlo as implemented, for example, in WinBugs. Many examples of such applications are now readily available in both journal form (various special issues of *Statistics in Medicine*) and in chapters in books (see contributions by Mollie and Best and others in Lawson et al., 1999, and by Diggle, by Wakefield, Best and Waller and by Mollie, in Elliot et al., 2000, and Best et al., 2000).

The issue I wish to consider here is not the mechanisms for incorporating a deprivation measure into a Bayesian spatial model, but rather to raise some questions concerning the desirability of so doing and what alternative(s) may be possible.

The key issue is: should one formally adjust for area-based deprivation as a potential confounder, i.e. as part of the modelling process, or not? What are the alternatives?

1. Adjust for area-level deprivation during the standardisation process, prior to modelling. Routinely we adjust for different age and sex profiles by area in the determination of a risk ratio. There is no reason, in principle, to confine the standardisation to these traditional

confounders alone – area-level deprivation could also be taken into account at this stage. This approach has been advocated by some, but it is not commonly applied.

2. Adjust for deprivation during the modelling. This approach, by contrast, is often seen. The model incorporates an index of deprivation (or other confounder) directly. This results in an adjusted (for deprivation) estimate of the risk ratio.
3. Leave the estimate of the risk ratio unadjusted (for deprivation, but adjusted for age and sex as is normal) and undertake a *post-hoc* exploration of the relationship between the model estimates and the deprivation score or level.

The first choice seems to me to be too radical. This approach explicitly places *deprivation* on a par with age and sex - while long experience in epidemiological research provides a clear justification for the treatment of area differences by age and sex, this does not extend to a similar treatment of deprivation. Additionally, it is far from clear that would serve our purposes in understanding the role of deprivation.

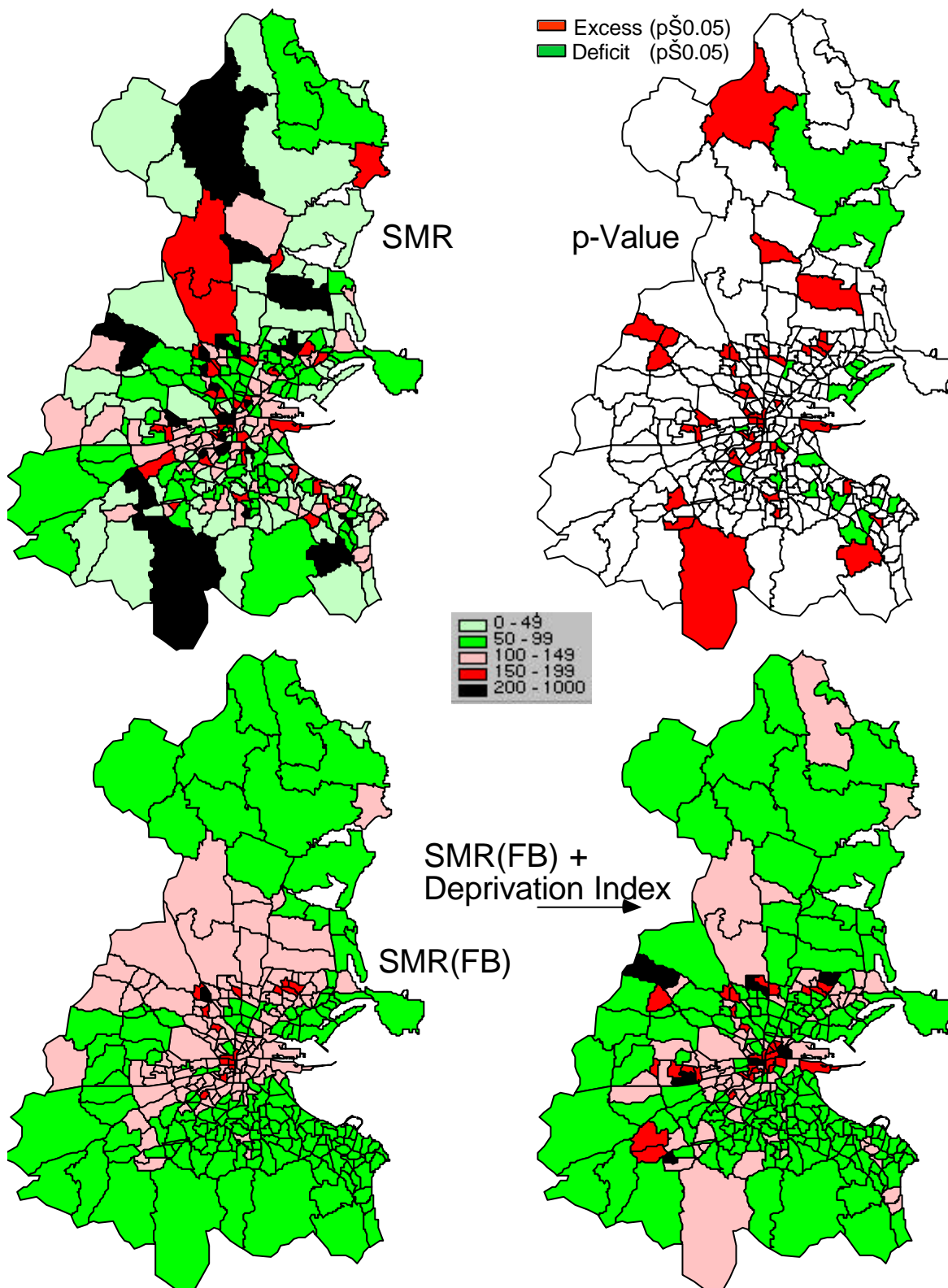
The second choice represents a fairly standard modelling approach. In the context of a Bayesian spatial model, in addition to 'borrowing strength from neighbouring areas' the small-area estimate will take into account any relationship between the dependant variable and deprivation, should a relationship exist – as with any routine regression model. Assuming that a substantive relationship exists, then the resulting estimate will be made more similar to areas that share a similar level of deprivation, whether they are neighbours or not.

This approach is based upon the premise that in comparing disease rates or exposure levels by small-area, we must compare like with like, and thus we need

to remove the potential complicating effect of different levels of deprivation. This choice would clearly appeal to the statistician.

The last option has more relevance to the Public Health officer or advocate. It permits an inspection of how the risk ratio varies by level of deprivation. In this instance, no prior adjustment for deprivation takes place thereby 'hiding' any relationship. This allows for a visualisation of the differential mortality (or other health outcome) as a function of deprivation level. In my experience, this approach is more intuitive to the end-user (public health specialist or service planner). This addresses the question: is there spatial variation in disease and does it vary with level of deprivation? This is an important requirement for presentation to decision-makers who may need empirical evidence that deprivation counts! For a very recent example of this approach (in relation to a Bayesian spatial analysis of mortality rates for chronic obstructive pulmonary disease and the links with potential explanatory factors such as smoking, rainfall, population density and air quality, see Nandram et al., 2000.)

Of course, there is no reason why choices 2 and 3 could not go hand-in-hand. With efficient means to routinely estimate full Bayesian models now to hand, it is entirely practical to compute and present the results of the modelling in both formats, perhaps also including a map of the raw SMRs and the Poisson p-values as examples of incorrect analyses and presentation – see example below for mortality from lung cancer for males (15-64 years) for small areas in Dublin County and County Borough.



Lung cancer (males 15-64 years): raw SMR, p-value, SMR(full Bayesian) and SMR(Full Bayesian) adjusting for Deprivation Index. Dublin county and County Borough.

6. Conclusion

The dominant factor in the spatial variation of disease incidence, after taking population size, age and sex into account, is usually socio-economic (i.e. material) deprivation. The particular importance of estimating the effect of deprivation on the spatial pattern of disease arises from three facts. Most human settlements display a strong degree of stratification by socio-economic status. Most diseases are more frequent, more severe, and have worse survival amongst deprived people and communities. There is frequently a strong correlation between being poor, and living in or near potentially contaminated areas.

The evidence linking area-based deprivation and inequalities in health outcomes is compelling. The rationale for developing national or regional (within country) measures of deprivation is clear. As noted, serious problems arise when considering the development of a single EU-wide deprivation score. A more feasible, and more valuable goal, would be the construction of deprivation measures with a similar structure in each of the partner countries, and a systematic investigation of candidate variables for inclusion in such a system.

This will achieve three objectives:

- By control of socio-economic confounding it will permit optimal assessment of the health effects of environmental exposures in European populations.
- By measurement of the environmental exposure of different socio-economic groups it will contribute to work on social justice and environmental equity in Europe.
- It will enhance our understanding of the common features, and the national and regional differences, in the origins of social inequalities in health in Europe.

During the next phase of the study, partners will contribute candidate socio-economic variables. These will include a small set of standard variables from each country, for example unemployment data, housing quality data and household composition data. Other data items will be selected to reflect specific aspects of the social structure of partner countries. Partners will work jointly on the interpretation of these variables, and how they differ from place to place within each country, and on the effect of these variables, as factors explaining variations in mortality between small areas in each country.

References

- Alvarez-Dardet C, Ruiz MT., *Rethinking the map for health inequalities*, Lancet 356: S36-S36, Suppl. S DEC 2000.
- Bajekal, M., Jan, S., Jarman, B., *The Swedish UPA scores: an administrative tool for identification of underprivileged areas*, Scandinavian Journal of Social Medicine, 24(3):177-84, 1996.
- Benach, J., Yasui, Y., *Geographic patterns of excess mortality in Spain explained by two indices of deprivation*, JECH, 3:423-31, 1999.
- Ben-Shlomo, Y, White, I., McKeigue, P.M., *Prediction of general practice workload from census based social deprivation scores*, Journal of Epidemiology and Community Health, 46:532-536, 1992.
- Best, N. G., Katja Ickstadt, and Robert L. Wolpert, *Spatial poisson regression for health and exposure data measures at disparate resolutions*, J. American Statistical Association, 95: 1110-1118, Dec, 2000
- Black, D., Morris, J.N., Smith, C., Townsend, P., *Better benefits for health: plan to implement the central recommendations of the Acheson report*. BMJ 318: 724-727, 1999.
- BMJ Editorial, *The Health of Adult Europe*, BMJ 1998; 316: 1620-21
- Borrell, C., et al., *Widening social inequalities in mortality: the case of Barcelona, a southern European city*, Journal of Epidemiology and Community Health, 51:659-667, 1997.
- Cairstairs, V., *Socio-economic factors at areal level and their relationship with health*, In: Elliot, P., Wakefield, J.C., Best, N.G., and D.J. Briggs (Eds.) *Spatial Epidemiology: Methods and Applications*, OUP, Oxform, 2000.
- Davey Smith, G., Bartley, M., Blane, D., *The Black report on socio-economic inequalities in health 10 years on*, BMJ, 301:373:377, 1990.
- DETR (Department of the Environment and the Regions, UK), *Indices of Deprivation 2000*, DETR, London, 2000.
- DHSS, *Report on inequalities in health ("The Black Report")*, 1980.
- DHSS, *Independent inquiry into inequalities in health ("The Acheson Report")*, 1998.

Eachus, J., et al., Deprivation and cause specific morbidity: evidence from the Somerset and Avon survey of health, *BMJ*, 312:287-312, 1996.

Eames, M., Ben-Shlomo, Y., Marmot, M., Social deprivation and premature mortality: a regional comparison across England, *BMJ*, 307:1097-1102, 1993.

Elliot, P., Wakefield, J.C., Best, N.G., and D.J. Briggs (Eds.) *Spatial Epidemiology: Methods and Applications*, OUP, Oxford, 2000.

Gordon, D., Shaw, M., Dorling, D. and Davy Smith, G., *Inequalities in Health*, The Policy Press, University of Bristol, United Kingdom.

Gunning-Schepers, L. *Equity on both the scientific and the policy agenda*, *BMJ* 316: 1035-1036, 1998.

Hart CL. Hole DJ. Smith GD., *Influence of socio-economic circumstances in early and later life on stroke risk among men in a Scottish cohort study*. *Stroke*. 31(9):2093-7, 2000

Hagquist C. *Socio-economic differences in smoking behaviour among adolescents - The role of academic orientation*, *J. of Child Research*, 7: (4) 467-478 NOV 2000

Kee, F., et al., *Socio-economic circumstances and the risk of bowel cancer in Northern Ireland*, *Journal of Epidemiology and Community Health*, 50:640-644, 1996.

Kelly, A., et al., *Use of Bayesian adjustment to SMRs based on small-area data for health and health service planning*, *Irish Journal of Medical Science*; 167, Suppl 9, 28-9, 1998.

Kelly, A., *Case studies in Bayesian disease mapping for health and health service research in Ireland*, In, eds, Lawson et al. (Eds.) *Disease mapping and risk assessment for Public Health*, John Wiley & Sons, UK, p349-63, 1999.

Kelly, A., Sinclair, H., *Deprivation and Health: Identifying the Black Spots*, *Journal of Health Gain*, 1/ 2, p13-14. 1997.

Lawson et al. (Eds.) *Disease mapping and risk assessment for Public Health*, John Wiley & Sons, UK, p349-63, 1999.

Macintyre, S., *Geographical inequalities in mortality, morbidity and health-related behaviour in England*, In: Gordon, D., Shaw, M., Dorling, D. and Davy Smith, G., *Inequalities in Health*, The Policy Press, University of Bristol, United Kingdom.

Mackenbach JP., Kunst, AE, Looman, CWN., *Cultural and economic determinants of geographical mortality patterns in the Netherlands*, JECH, 45:231-7, 1991.

Mackenbach JP., Cavelaars AE. Kunst AE. Groenhouf F., Socio-economic inequalities in cardiovascular disease mortality; an international study. *European Heart Journal*. 21(14):1141-51, 2000.

MacRae, K , Commentary: Socio-economic deprivation and health and the ecological fallacy, *BMJ* 309:1478-1479, 1994.

Malmstrom, M., Sundjuist, J., Bajekal, M., Johansson, SE., *Indices of need and social deprivation for primary health care*, *Scandinavian Journal of Social Medicine*, 26(2):124-30, 1998.

Mcisaac, S., Wilkenson, R., Income distribution and cause-specific mortality, *European Journal of Public Health*, 7: 45-53, 1997.

Nandram, B., Sedransk, J. and Linda Williams Pickle, *Bayesian analysis and mapping of mortality rates for chronic obstructive pulmonary disease*, *J. American Statistical Association*, 95: 1110-1118, Dec, 2000.

PAT18 report of Policy Action Team 18: Better Information, Annex H, Working Paper: *Measuring Deprivation – A review of Indices in Commun Use*, The Stationary Office, London. April 2000. (Annex available at: <http://www.cabinet-office.gov.uk/seu/2000/pat18/Depindices.htm>)

Pell J.P., *et al.*, *Effect of socio-economic deprivation on waiting time for cardiac surgery: retrospective cohort study*, *BMJ* ; 320:15-19, 2000.

Raleigh, V.S., Kiri, V.A., Life expectancy in England: variations and trends by gender, health authority, and level of deprivation, *Journal of Epidemiology and Community Health*, 51:649-658, 1997.

Reading R *Social disadvantage and infection in childhood*, *Sociology of Health and Illness*, 19: (4) 395-414, 1997

Reijneveld, SA., *Predicting the workload in urban general practice in The Netherlands from Jarman's indicators of deprivation at patient level*, JECH, 0(5): 541-4, 1996.

Reijneveld, SA., Verheij, R., de Bakker, DH., *The impact of area deprivation on differences in health: does the choice of the geographical classification matter?*, JECH 54:306-13, 2000.

Reijneveld, S., Schene, A., *Higher prevalence of mental disorders in socio-economically deprived urban areas in the Netherlands: community or personal disadvantage?* Journal of Epidemiology and Community Health, 52:2-7, 1998.

Robinson WS. Ecological correlations and the behavior of individuals. *American Sociological Reviews* 15:351-7, 1950.

Ross, N.A., et al., *Relation between income inequality and mortality in Canada and in the United States: cross sectional assessment using census data and vital statistics*, BMJ, 320:898-902, 2000.

SAHRU, *Small Area Analysis in Health & Health Service Research: Principles & Application*, Technical Report No. 1, Small Area Health Research Unit, Trinity College Dublin. May 1997.

SAHRU, *A National Deprivation Index for Health & Health Services Research*, Technical Report No. 2, Small Area Health Research Unit, Trinity College Dublin. August 1997.

Scaife B. Gill P. Heywood P. Neal R. *Socio-economic characteristics of adult frequent attenders in general practice: secondary analysis of data*. Family Practice. 17(4):298-304, 2000.

Townsend, P., Davidson, N., Whitehead, M., (Eds.) *Inequalities in Health: The Black Report & The Health Divide*, Pelican Books, 1988.

Vazquez, Juncal S., Mi. Seoane G. Blanco-Aparicio M. Vereá H., *Applicability of the Asthma Opinion Survey in the Spanish population: distribution and relationship with sociodemographic and clinical variables*. J. of Asthma, 37(6):469-79, 2000.

Weich, S., Lewis, G., *Material standard of living, social class, and the prevalence of the common mental disorders in Great Britain*, Journal of Epidemiology and Community Health, 52:8-14, 1998.

Wilkinson, R, Marmot, M., *Social Determinants of Health: The Solid Facts*, WHO, Copenhagen, 1998.

Wilkinson, R.G., *Socio-economic determinants of health: Health inequalities – relative or absolute material standards?* BMJ 314, 1997.

Whitehead, M., (Ed.) *Tackling Inequalities in Health*, King's Fund, United Kingdom, 1995.

Worrall, A., Rea, J.N., Ben-Shlomo, Y., *Counting the cost of social disadvantage in primary care: retrospective analysis of patient data*, BMJ, 314:38:42, 1997.

Additional bibliography on socio-economic influences on health

(From recent issues of JECH and BMJ)

Baker D., and J Henderson, *Differences between infants and adults in the social aetiology of wheeze*. The ALSPAC Study Team. Avon Longitudinal Study of Pregnancy and Childhood, JECH 53: 636-642, 1999

Barnett, S., P Roderick, D Martin, and I Diamond, *A multilevel analysis of the effects of rurality and social deprivation on premature limiting long term illness*, JECH 55: 44-51, 2001

Benigni, R., Rosa Giaimo, Domenica Matranga, and Alessandro Giuliani, *The cultural heritage shapes the pattern of tumour profiles in Europe: a correlation study*, JECH 54: 262-268, 2000

Bowen, H J ., S R Palmer, H M P Fielder, G Coleman, P A Routledge, and D L Fone, *Community exposures to chemical incidents: development and evaluation of the first environmental public health surveillance system in Europe*, JECH 54: 870-873, 2000

Davey Smith, G., Carole Hart, Mark Upton, David Hole, Charles Gillis, Graham Watt, and Victor Hawthorne, *Height and risk of death among men and women: aetiological implications of associations with cardiorespiratory disease and cancer mortality*, JECH 54: 97-103, 2000

Dedman, D.J., D Gunnell, G Davey Smith, and S Frankel, *Childhood housing conditions and later mortality in the Boyd Orr cohort*, JECH 55: 10-15, 2001

Eachus, J., P Chan, N Pearson, C Propper, and G Davey Smith , *An additional dimension to health inequalities: disease severity and socio-economic position*, JECH 1 53: 603-611.

Gunnell, DJ,. G Davey Smith, S Frankel, K Nanchahal, FE Braddon, J Pemberton, and TJ Peter, *Childhood leg length and adult mortality: follow up of the Carnegie (Boyd Orr) Survey of Diet and Health in Pre-war Britain*, JECH 52: 142-152, 1998

Kinra, S., Robert P Nelder, and Gill J Lewendon, *Deprivation and childhood obesity: a cross sectional study of 20 973 children in Plymouth, United Kingdom*, JECH 54: 456-460, 2000.

Krantz, G., and Per-Olof Östergren, *Common symptoms in middle aged women: their relation to employment status, psychosocial work conditions and social support in a Swedish setting*, JECH 54: 192-199, 2000

Lamont, DW., FM Toal, and M Crawford, *Socio-economic deprivation and health in Glasgow and the west of Scotland--a study of cancer incidence among male residents of hostels for the single homeless*, JECH 51: 668-671, 1997

Lee PN., and BA Forey, *Trends in cigarette consumption cannot fully explain trends in British lung cancer rates*, JECH 52: 82-92, 1998

Loughlin, JE., KJ Rothman, and NA Dreyer, *Lymphatic and haematopoietic cancer mortality in a population attending school adjacent to styrene-butadiene facilities, 1963-1993*, JECH 53: 283-287, 1999

Maheswaran, R., P Elliott, and DP Strachan, *Socio-economic deprivation, ethnicity, and stroke mortality in Greater London and south east England*, JECH 51: 127-131, 1997

Manson-Siddle, CJ., and MB Robinson, *Super Profile analysis of socio-economic variations in coronary investigation and revascularisation rates*, JECH 52: 507-512, 1998

Neeleman, J., and G Lewis, *Suicide, religion, and socio-economic conditions. An ecological study in 26 countries*, JECH 53: 204-210, 1999

Niedhammer, I., M Goldberg, A Leclerc, S David, I Bugel, and MF Landre, *Psychosocial work environment and cardiovascular risk factors in an occupational cohort in France*, JECH 52: 93-100, 1998

Richards, H., Alex McConnachie, Caroline Morrison, Keith Murray, and Graham Watt, *Social and gender variation in the prevalence, presentation and general practitioner provisional diagnosis of chest pain*, JECH 54: 714-718, 2000

Stronks, K., HD van de Mheen, and JP Mackenbach, *A higher prevalence of health problems in low income groups: does it reflect relative deprivation?*, JECH 52: 548-557, 1998

Vahtera, J., P Virtanen, M Kivimaki, and J Pentti, *Workplace as an origin of health inequalities*, JECH 53: 399-407, 1999

Warnes, AM., GK Armstrong, and D Peters, *Population predictors of community health and social service use in Northern Ireland*, JECH 51: 722-730, 1997

White, IR., D Blane, JN Morris, and P Mourouga, *Educational attainment, deprivation-affluence and self reported health in Britain: a cross sectional study*, JECH 1999 53: 535-541, 999

Bibliography of area-based studies on inequalities in health

(from recent issues of JECH & BMJ)

Benach, J., and Y Yasui, *Geographical patterns of excess mortality in Spain explained by two indices of deprivation*, JECH 1999 53: 423-431.

Brugal, MT., A Domingo-Salvany, A Maguire, JA Cayla, JR Villalbi, and R Hartnoll, *A small area analysis estimating the prevalence of addiction to opioids in Barcelona*, 1993, JECH 1999 53: 488-494.

Connolly, V., N Unwin, P Sherriff, R Bilous, and W Kelly, *Diabetes prevalence and socio-economic status: a population based study showing increased prevalence of type 2 diabetes mellitus in deprived areas*, JECH 2000 54: 173-177.

Cubbin, C., Felicia B LeClere, and Gordon S Smith, *Socio-economic status and injury mortality: individual and neighbourhood determinants*, JECH 2000 54: 517-524.

Jones, C.M., G O Taylor, J G Whittle, D Evans, D P Trotter, *Water fluoridation, tooth decay in 5 year olds, and social deprivation measured by the Jarman score: analysis of data from British dental surveys*, *BMJ* 315:514-517, 1997

Davey Smith, G. Carole Hart, David Blane, David Hole, *Adverse socio-economic conditions in childhood and cause specific adult mortality: prospective observational study*, *BMJ* 316:1631-1635, 1998.

Davy Smith, G., C Hart, G Watt, D Hole, and V Hawthorne, *Individual social class, area-based deprivation, cardiovascular disease risk factors, and mortality: the Renfrew and Paisley Study*, JECH 1998 52: 399-405.

Pattenden, S., H Dolk, and M Vrijheid, *Inequalities in low birth weight: parental social class, area deprivation, and "lone mother" status*, JECH 1999 53: 355-358.

Pickett, K.E., and M Pearl, *Multilevel analyses of neighbourhood socio-economic context and health outcomes: a critical review*, JECH 2001 55: 111-122.

Reijneveld, S.A. and AH Schene, *Higher prevalence of mental disorders in socioeconomically deprived urban areas in The Netherlands: community or personal disadvantage?*, JECH 1998 52: 2-7.

Reijneveld, S.A., Robert A Verheij, and Dinny H de Bakker, *The impact of area deprivation on differences in health: does the choice of the geographical classification matter?*, JECH 2000 54: 306-313.

Salmond, C., P Crampton, S Hales, S Lewis, and N Pearce, *Asthma prevalence and deprivation: a small area analysis*, JECH 1999 53: 476-480.

Spencer, N., S Bambang, S Logan, and L Gill, *Socio-economic status and birth weight: comparison of an area-based measure with the Registrar General's social class*, JECH 1999 53: 495-498.

APPENDIX 2

Summary of health, denominator, socio-economic and environmental datasets available in each partner country

DENMARK

Table 1(a) Health datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Disease classification</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Death Registry	Ministry of Health	Denmark	Individual address (CPR No.) Municipality	1977-	Yes	ICD8 until 1993; ICD10 from 1993	No	No	Months	High case ascertainment and completeness
Cancer Registry	Ministry of Health	Denmark	Individual address (CPR No.) Municipality	1965-	Yes	ICD7 Coded to ICD10 (as well as ICD7) since 1994	No	No	Months	High case ascertainment and completeness
Hospital Admissions	Ministry of Health	Denmark	Individual address (CPR No.) Hospital	1977-	Yes		No	No	Months	In-patients and (recently) out-patients ICD8 until 1993; ICD10 from 1994 High case ascertainment and completeness Some diagnostic uncertainty

¹ Key fields are: diagnostic code, date of event, age, sex and geographical reference

Table 1(b) Denominator datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Type of dataset</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Central Population Registry	Ministry of Interior	Register	Denmark	Individuals	1970-	Yes	No	Yes	Days	All individuals (alive and deceased) have a personal identification number High level of completeness Place of birth, previous and present addresses are recorded Civil status is recorded Special permission required to obtain the dataset Probably need to negotiate special price for large amounts of data

¹ Key fields are: date, age, sex and geographical reference

Table 1(c) Socio-economic datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
National Prevention Registry	Central Statistical Bureau (CSB)	Denmark	Individuals (CPR No.)	Varies: Demographic: 1977- Income/employment: 1977- Education: 1980- Public assistance: 1984- Housing: 1980-	Income and employment Education Public assistance Housing	No	Yes	Unknown	Nobody has ever used this data outside of the CSB before

Table 1(d) Environmental and background feature datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
GIS on Environmental Issues	Counties	Denmark	High resolution and accuracy	Varies	Powerlines, waste deposits, industrial waste, industrial plants, water sources, roads and more	For Vejle county only	No	Months	Most of the datasets are not historical – i.e. only present day features

Table 1(e) Geographical referencing and linkage datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
National Address Registry	Counties Municipalities	Denmark	Metres	1999-	For Vejle and 3 other counties	No	Unknown	Data is for addresses which existed from 1999 High level of coverage – c.98% of properties in 2000; remaining ones being geo-coded in on-going project This dataset provides the geographical referencing for the health, population and socio-economic datasets through linkage with the CPR No. and the address

Notes

Municipalities vary in size from 3,500 to 400,000 inhabitants. Counties are made up of aggregations of municipalities. There are 16 counties and 275 municipalities. Individual level data can be linked via the CPR number – this links health, population, socio-economic data together, at the address level.

FINLAND

Table 2(a) Health datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Disease classification</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Cancer Registrations	Finnish Cancer Registry	Finland	Individual	1981-97	Yes	ICD7	Yes	Yes	NA	High level of case ascertainment Completeness >95% Contains socio-economic classification (in 1970, 75, 80, 85, 90, 95) Contains metric coordinates of residence in 1980, 1990 and year before diagnosis

¹ Key fields are: diagnostic code, date of event, age, sex and geographical reference

Table 2(b) Denominator datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Type of dataset</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Population data	Statistics Finland	Census	Finland	0.25km ² grid squares	1980, 90, 97	Yes	Yes	Yes	NA	Categorical age ranges High level of completeness

¹ Key fields are: date, age, sex and geographical reference

Table 2(c) Socio-economic datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Census data	Statistics Finland	Finland	0.25km ² grid squares	1980, 90, 97	Socio-economic class (mainly based on occupation)	Yes	Yes	NA	

Table 2(d) Environmental and background feature datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
None available to partner, but see EU-level table (Table 7)									

Table 2(e) Geographical referencing and linkage datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Datataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
No information provided								

Notes

Health data can be aggregated to grid squares for analysis within a GIS.

ITALY

Table 3(a) Health datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Disease classification</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Mortality Registry	ISTAT	Italy	Municipality	1980-94	Yes	ICD9	Yes	Yes	NA	High case ascertainment and completeness Confidentiality constraints restrict access (minimum period 3 years; minimum population threshold 10,000) Cannot access single records directly – query it through an interface

¹ Key fields are: diagnostic code, date of event, age, sex and geographical reference

Table 3(b) Denominator datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Type of dataset</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Pay for dataset</i>	<i>Notes</i>
Mortality/Census	ISTAT	Census	Italy	Municipality	1980-94	Yes	Yes	No	NA	Categorical age ranges (5 yrs groups)

¹ Key fields are: date, age, sex and geographical reference

Table 3(c) Socio-economic datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Socio-economic status (SES)	ISTAT	Italy	Municipality	1981, 1991	Several census socio-demographic variables and deprivation index at municipality level	Yes	No	NA	Developed in collaboration with Piedmont Regional Health Authority

Table 3(d) Environmental and background feature datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
None available to partner, but see EU-level table (Table 7)									

Table 3(e) Geographical referencing and linkage datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Datataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
No information provided								

Notes

Municipalities vary in size considerably. There are over 8000 municipalities, and cities are often one municipality. Municipalities nest within provinces (c. 200), which nest within regions (20).

SPAIN

Table 4(a) Health datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Disease classification</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Hospital Admissions (Minimum Basic Dataset (CMBD))	Regional Health Authority	Valencia region	Municipality	1993-98	Yes	ICD9???	Partially	No	Months	High case ascertainment Completeness not as good for first two years (1993-94)
Mortality Registry	Mortality Registry of Andalusia	Andalusia region	Municipality	1991-	Yes	ICD9	Yes	No	NA	
Mortality Registry	Mortality Registry of Valencia	Valencia region	Municipality	1987-	Yes	ICD9 until 1998; then ICD10	Yes	No	NA	High case ascertainment and completeness
Breast Cancer Incidence in the province of Granada	Cancer Registry of Granada	Granada province	Municipality	1985-	Yes	ICD9	Yes	No	NA	High level of completeness There are no other cancer registries in the Andalusian region – only in Granada province
Breast Cancer Incidence in the city of Granada	Cancer Registry of Granada	Wards of the city of Granada	Wards	1985-	Yes	ICD9	Yes	No	unknown	This is a pilot project of the Andalusian Scholl of Public Health. We do not know if data will be available this year
Childhood Cancer Registry	Valencian Regional Health Authority	Valencia region	Municipality	1983-	Yes	ICD0-2 Stands for ICD for Oncology 2nd. revision	Yes	No	NA	Childhood cancers < 15 years High level of case ascertainment and completeness
Stillbirths Registry	Valencian Institute of Statistics	Valencia region	Municipality	1996-	Yes	1996-1998: ICD9 1999-: ICD10	No	No	Months	Stillborn or born alive but dead within 24 hours
Congenital Malformations	ECEMC School of Medicine, Madrid	Valencia region	Hospitals	1976-97	Yes	ICD9	No	No	Unknown	Voluntary notification by hospitals
Induced Abortion Registry	National Health Authority	Valencia region	Municipality (Postcodes) ²	1986-	Yes	None	Yes	No	NA	Compulsory notification by hospitals and health centres Low case ascertainment (80%) and completeness (85%)

Table 4(a) Health datasets (continued)

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Disease classification</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Renal Diseases Registry	Valencian Regional Health Authority	Valencia region	Municipality (Postcodes) ²	1991-	Yes	EDTA code	Yes	No	NA	Compulsory notification by hospitals Case ascertainment high; completeness lower (85-100% for key fields)
Notifiable Communicable Diseases	Valencian Regional Health Authority	Valencia region	Municipality (Postcode) ²	1990-	No age or sex	ICD9	Yes	No	NA	High level of case ascertainment and completeness
Creutzfeldt-Jakob Disease (CJD)	Valencian Regional Health Authority	Valencia region	Municipality	1998-	Yes	ICD9 monographic registry, no disease code recorded	Yes	No	NA	

¹ Key fields are: diagnostic code, date of event, age, sex and geographical reference

² Postcode: a municipality is an administrative unit that can be a town, vilage or a city. Every town/village has a unique postcode, but cities are divided into districts, each of which has its own postcode. To perform an epidemiological analysis in a city those districts could be considered a partition of the city and would allow us to work at a finer level of aggregation.

Table 4(b) Denominator datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Type of dataset</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Municipal population	Valencian Institute of Statistics	Census	Valencia region	Municipality	1960(5) 1996 + 1998	Yes	Yes	No	NA	5-year age categories
Births Registry	Valencian Regional Health Authority	Register	Valencia region	Municipality	1990-	Yes	No	No	Months	Low case ascertainment (70%) Low completeness (85%)
Municipal Emigrations and Immigrations	Valencian Institute of Statistics	From Census	Valencia region	Municipality	1996	Only sex	Yes	No	NA	Migration since last Census (1991) – between municipalities within region

¹ Key fields are: date, age, sex and geographical reference

Table 4(c) Socio-economic datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Andalusian Economic Atlas	Society of Economic Studies	Andalusia	Municipality	1991	Population, employment, agriculture, industry, meteorological variables	No	No	Immediate	
SIMA	Andalusian Institute of Statistics	Andalusia	Municipality	1991	Population, employment, ... and more ...	No	No	Immediate	Some socioeconomic variables only since 1998?
% Unemployed	Valencian Institute of Statistics	Valencia region	Municipality	1990-1997	Registered as unemployed with National Institute – does not include retired or job-seekers	Yes	No	NA	

Table 4(d) Environmental and background feature datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Monitoring system of air quality	Department of Environment in Andalusia	Main Andalusia cities	Daily average air pollution in big 8 cities	1993-99	Black smoke, TSP, NO ₂ , SO ₂ , CO	Yes	No	NA	Automatic monitoring stations
Monitoring system of air quality	Department of Environment in Valencia	Main Valencian Cities plus some specific areas	Daily average air pollution	1995-98	Black smoke, TSP, NO ₂ , SO ₂ , CO, O ₃	NO	No	weeks	28 Automatic monitoring stations
Vegetation land-use	Valencian Institute of Statistics/ Regional Agricultural Ministry	Valencia region	Municipality	1996	Number of hectares of total land, crop, forest, pasture etc.	Yes	No	NA	Collected every 5 years Specific census
Water Quality Monitoring Programme	Department of Environment, Valencia	Valencia region	Municipality	1992-99	Concentrations of parameters included in the Water Quality Monitoring Programme (inc. nitrates, sulphates, calcium, microbiological)	No	No	weeks	

Table 4(d) Environmental and background feature datasets (continued)

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Datataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Number of motor vehicles	Valencian Institute of Statistics/	Valencia region	Municipality	1996-97	Number of motor vehicles – broken down into cars, vans, buses and tractors Excludes motorbikes	Yes	No	NA	Obtained from number registered per year, minus withdrawals and modifications. Some vehicles cannot be linked to a municipality

Table 4(e) Geographical referencing and linkage datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Datataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
MuniView	ESRI	Spain	Municipality	1996	Yes	No	NA	GIS with municipalities, provinces, regions and main cities boundaries plus rivers, roads, etc.

Notes

Municipalities vary in size considerably e.g. in Valencia they range from 23 to 746,683 people. There are 760 municipalities in Andalusia. Municipalities nest within provinces, which nest within regions.

SWEDEN

Table 5(a) Health datasets

<i>Dataset name</i>	<i>Dataset provider/ owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Disease classification</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Cal 1990-97 (Cancer Registrations)	Stockholm County Council	Stockholm County	Base unit	1990-97	Yes	ICD9	Yes	Yes (for geo-referencing)	c. 6 weeks	High case ascertainment and completeness Permission required to use the data for this specific project
MI 1990-95	Dept Epidemiology Stockholm	Stockholm County	Base unit	1990-95	Yes	ICD9	Yes	No	1 week	Data from linking hospital discharge register with cause of death register Identifies first time MI and recurrent MIs High case ascertainment and completeness Annual data but summed over period
SLV 1990-98 (Hospital discharges)	Stockholm County Council	Stockholm County	Base unit	1990-98	Yes	ICD9 and ICD10 (from 1997)	Yes	Yes (for data during 1990-95)	c. 6 weeks	High case ascertainment Completeness c. 95% Permission required to use the data for this specific project
DOS 1990-97 (Death Registry)	Dept Epidemiology Stockholm	Stockholm County	Base unit	1990-97	Yes	ICD9 and ICD10 (from 1997)	Yes	Yes	c. 4 weeks	High case ascertainment and completeness Permission required to use the data for this specific project

¹ Key fields are: diagnostic code, date of event, age, sex and geographical reference

Table 5(b) Denominator datasets

<i>Dataset name</i>	<i>Dataset provider/ owner</i>	<i>Type of dataset</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Population register	Stockholm County Council	Population register	Stockholm County	Base unit	1990-98		Yes	Yes (for geo-referencing 1990-95)	c. 6 weeks	

¹ Key fields are: date, age, sex and geographical reference

Table 5(c) Socio-economic datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Income, education, social assistance	Statistics Sweden	Stockholm County	Base unit	1990-	Income Education social assistance	No	Yes	c. 6 weeks	Detailed information over several years is there, but the information is expensive to buy
Income, education, social assistance	Stockholm County Council	Stockholm County	Base unit/parish	Varies from annually to single year/years depending on variable	Income, education, social assistance	No	No	Certain data can be downloaded = work time	Several changes of dataset owners affects periodicity, geo-information, and availability

Table 5(d) Environmental and background feature datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Environmental data	SLB-analys (Environment and Health Protection Administration)	City of Stockholm; the Stockholm County	Varies from ~municipality to some measurement point; grids	Varies (ranges from 1960 to only last year)	NO ₂ , ozone, particulate matter; traffic noise; traffic counts	No	Yes	??	Datasets are created on request. Basic info is certain measurements, used as indata for modelling of NO ₂ , particulate matter,

Table 5(e) Geographical referencing and linkage datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Base unit boundaries and look-up tables	Dept Epidemiology, Stockholm	Stockholm County	The resolution varies across base units with respect both to geogr area and to number of inhabitants (0-5000) included	1990- (boundaries according to 1999 revision for all years)	Yes	No	NA	

Notes

Base units are small geographical areas representing reasonably homogenous areas. There are 1248 base units within Stockholm County, with between 0 and 5000 persons per unit. These can be aggregated to two levels: 18 community sectors and 6 health administrative regions.

UK

Table 6(a) Health datasets

<i>Dataset name</i>	<i>Dataset provider /owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Disease classification</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Mortality Registry	ONS	England & Wales	Postcode	1981-98	Yes	ICD9	Yes	No	NA	
Mortality Registry	GROS	Scotland	Postcode	1981-98	Yes	ICD9	Yes	No	NA	
Cancer Registry	ONS	England & Wales	Postcode	1981-97	Yes	ICD9	Yes	No	NA	Welsh data only to 1994
Cancer Registry	GROS	Scotland	Postcode	1981-96	Yes	ICD9	Yes	No	NA	
Hospital Episode Statistics	DH	England	Postcode	1991-98	Yes	ICD9/10	Yes	No	NA	
Hospital Episode Statistics	WHIS	Wales	Postcode	1991-94	Yes	ICD9	Yes	No	NA	
Hospital Episode Statistics	GROS	Scotland	Postcode	1992-98	Yes	ICD9/10	Yes	No	NA	
Congenital Malformation Registry	ONS	England & Wales	Postcode	1983-98	Yes	ICD9	Yes	No	NA	
Congenital Malformation Registry	GROS	Scotland	Postcode	1988-94	Yes	ICD9	Yes	No	NA	
Perinatal mortality	SAHSU	GB	Postcode	1981-98	Yes	ICD9	Yes	No	NA	Stillbirths and deaths within first week

¹ Key fields are: diagnostic code, date of event, age, sex and geographical reference

Table 6(b) Denominator datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Type of dataset</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Key fields¹</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Census Small Area Statistics	ONS	Census	GB	Enumeration District	1981, 1991	Yes	Yes	No	NA	Under-enumeration of certain sections of society is a problem 1991 estimated 1 million missing
Registrar General Mid-Year Estimates	EwC Project	Statistical estimates	GB	Enumeration District	1991	Yes	Yes	No	NA	Population estimates with missing million imputed
Registrar General Mid-Year Estimates	Registrar General	Population registers/statistical estimates	GB	Local Authority District	1974-96	Yes	Yes	No	NA	Annual estimates at LAD level which can be used for calculating annual estimates at ED level
Births Registry	ONS	Registry	GB	Postcode	1981-98	Yes	Yes	No	NA	
Stillbirths	ONS	Registry	GB	Postcode	1981-98	Yes	Yes	No	NA	

¹ Key fields are: date, age, sex and geographical reference

Table 6(c) Socio-economic datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical base unit</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Carstairs Index of Deprivation	ONS	GB	Enumeration District	1991	Unemployment; overcrowding; social class of head of household; car ownership	Yes	No	NA	Derived from Census variables Scores are ranked and put into quintiles (for GB, England & Wales, Scotland, or by country)
ONS Ward Classification	ONS	GB	Ward	1991	Various Census variables	Yes	No	NA	Areas classified according to similarity of various Census variables Clusters, families, groups

Table 6(d) Environmental and background feature datasets

<i>Dataset name</i>	<i>Dataset provider/ owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Automobile Association digital data	Automobile Association	GB	From 1:200,000 source data	1995	Roads, rivers, railways, urban areas, placename gazetteer	Yes	Yes	NA	Quite generalised features
Topographic data	Ordnance Survey	UK	Variable: 1:10k – 1:250k	Current	Coastline, roads, contours, spot heights, urban boundaries etc	Yes	No	NA	Available via national CHEST agreement
Chemical Release Inventory	EA & SEPA & Industrial Pollution Inspectorate, Department of the Environment NI	UK	Point data	1992 - 1997	Over 500 pollutant releases	Yes	No	NA	
Air pollutant concentrations	NETCEN	UK	Point data (rounded to 100 metres)	1962-1998	Mean monthly/annual concentrations of SO ₂ , black smoke, NO ₂	Yes	No	NA	SO ₂ and black smoke data available for up to 600 sites from continuous monitoring stations; NO ₂ data available for ca 1200 sites from passive samplers (1993 onwards only)
Background air pollutant concentrations	NETCEN	GB	1km ²	1996	Mean annual concentration of SO ₂ , PM ₁₀ , NO ₂ , black smoke	Yes	No	NA	Modelled data for each 1km square; other pollutants available
Atmospheric emissions	NETCEN	GB	1km ²	Late 1990s	Total emission (by source) of particulates, SO _x , NO _x , VOCs	No	No	Downloadable from Internet	
Landfills	Environment Agency; Scottish Environmental Protection Agency	GB	Point data (rounded to 100 metres, but with +/- 500 metre accuracy)	1920-1998	Site location; site name; waste type; site area; opening date; closing date	Yes	No	NA	Data set is incomplete for many fields; data have been extensively edited and checked to obtain a workable data set
Water Quality	Water Companies	3 water companies	Water zones	1992-98	Trihalomethane concentrations	Yes	No	NA	Water quality zones do not nest with other geographical units Some problems with source data – inaccuracies due to digitising; missing data

Table 6(d) Environmental and background feature datasets (continued)

<i>Dataset name</i>	<i>Dataset provider/ owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Land cover	Centre for Ecology and Hydrology	GB	1 km ²	1988	Ca. 40 land cover types, including habitats, built-up land	Yes	No	NA	Available as part of Countryside Information System. Updated data due in 2001
Soils	Centre for Ecology and Hydrology	GB	1 km ²	1988	Dominant soil type in each grid cell	Yes	Yes	NA	Available as part of Countryside Information System. Updated data due in 2001
Meteorology	BADC	UK	Point data	To date	Hourly wind speed, wind direction, precipitation, cloudcover for UK land surface stations	No	No	Downloadable from Internet	
Solid geology	British Geological Survey	UK	1:625,000	Various	Major stratigraphic groups	No	Yes (annual fee)	Weeks	

Table 6(e) Geographical referencing and linkage datasets

<i>Dataset name</i>	<i>Dataset provider/owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Datset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
Digital Census boundaries	JISC/ESRC	GB	ED/Ward/District /County/Region	1991	Yes	No	NA	Boundaries of Eds, wards, districts, counties and regions Some inaccuracies due to labelling errors
Enumeration District centroids	JISC/ESRC	GB	+/-100m in England/Wales +/-10m in Scotland	1991	Yes	No	NA	Population-weighted centroids of EDs Some errors in grid-references – have been corrected by SAHSU
Postcode grid-references	Royal Mail/ONS	GB	+/-100m in England/Wales +/-10m in Scotland	1995	Yes	Yes	NA	Grid-references of first house in postcode Some inaccuracies in grid-references New, terminated and recycled postcodes could be missing or have incorrect grid-references
Postcode to ED links	Royal Mail/ONS	GB	NA	1995	Yes	No	NA	Links postcodes to EDs Some inaccuracies in allocation of postcodes to EDs
Census Ward to Health Authority link	ONS	GB	NA	1996	Yes	No	NA	Links 1991 Census wards to Health Authorities

Notes

Enumeration Districts (EDs) are the smallest geographical unit for which Census data are available. EDs vary considerably in size; average population size is approximately 300 to 400 people, with a minimum population threshold of 100 people – below this threshold, population is set to zero and population is exported to other neighbouring EDs. EDs nest within wards, which nest within districts, counties and then regions. There are 10 Standard Regions: Scotland and Wales are both individual regions. Postcodes vary considerably in size; median number of households is 14. Postcodes do not nest within EDs – they must be linked using look-up tables or GIS techniques.

Table 7 EU-level environmental and socio-economic datasets

<i>Dataset name</i>	<i>Dataset provider/ owner</i>	<i>Area available for</i>	<i>Geographical resolution and accuracy</i>	<i>Years available</i>	<i>Variables contained</i>	<i>Dataset held already</i>	<i>Pay for dataset</i>	<i>Time to obtain dataset</i>	<i>Notes</i>
CORINE land cover	EEA	EU	1:1 million	1985-90	Ca. 45 land cover classes, including residential land, industrial/commercial land	N/A	No	Days-weeks	National data sets also available, giving more detailed classification and higher spatial resolution
Healthy Cities Air Management Information System (AMIS)	WHO	World	By city	1997-2000	Mean annual (and 95%ile) SO ₂ , NO ₂ , CO, O ₃ , SPM, PM ₁₀ , black smoke, cadmium and lead concentrations	N/A	No	Days-weeks	Ca. 40 cities in Europe
Emissions	EEA	EU	NUTS level 3 and 50km ²	1985-present	Annual emissions of CO, CO ₂ , SO _x , NO _x , particulates, VOCs etc by sector (e.g. transport, industry, energy)	N/A	No	Days-weeks	Includes detailed breakdown of emission sources
Climate	EEA	EU	Point data	1960-present	Mean annual temperature, rainfall, relative humidity etc	N/A	No	Days-weeks	Long-term average data for several hundred stations in the EU
Urban centroids	Eurostat	EU	1:100,000	Current	Point locations and populations of all major towns	N/A	No	Days-weeks	
REGIO	Eurostat	EU	NUTS level 3	1980s-present	Socio-economic data (population, employment etc)	N/A	No	Days-weeks	Large data set including several hundred socio-economic variables