The Rapid Inquiry Facility (RIF) Version 3.2

How to use the RIF

Authors (2010):
Beale L, Hodgson S, Abellan JJ, Andersson M, Fabbri, F, Jarup L

Small Area Health Statistics Unit (SAHSU)
Department of Epidemiology and Public Health
Imperial College London
Norfolk Place
London
W2 1PG

Website www.sahsu.org
Webboard http://rif.forum.cdc.gov
e-mail rif@imperial.ac.uk

Contents

1. Introduction to the RIF

1.1 Purpose

The Rapid Inquiry Facility (RIF) is an automated tool that provides an extension to ESRI® ArcGIS functions. The purpose of this facility is to rapidly address epidemiological and public health questions using routinely collected health and population data.

The RIF can perform risk analysis around putative hazardous sources, and can be used for disease mapping. It generates standardised rates and relative risks for any given health outcome, for specified age and year ranges, for any given geographical area.

The RIF has been developed by the Small Area Health Statistics Unit (SAHSU) at the Department of Epidemiology and Public Health, Imperial College London. This facility was initially designed as a tool for SAHSU staff to analyse routinely collected health data in relation to environmental exposures in the UK. The UK RIF was subsequently transformed for use by several European countries as part of the European Health and Environment Information System (EUROHEIS) project (http://www.euroheis.org). The Centers for Disease Control and Prevention (CDC) and SAHSU have collaborated to adapt and enhance the UK RIF software for use in CDC's National Environmental Public Health Tracking (EPHT) Network (http://www.cdc.gov/nceh/tracking).

This manual describes version **3.2** of the RIF (last updated 29th July 2010).
Note. For users of version 3.1 please note that this version requires three new database tables. For ACCESS users a new 3.2 database is provided but also the three new tables are included with the installer that can be added to your current database if preferred, to upgrade it to RIF 3.2 compliant database. For ORACLE users, one SQL script is provided to create a new 3.2 database and another SQL script is provided that will create the required three new tables.

1.2 Principal Features

- The software is designed to function with geo-referenced data stored in ODBC compliant databases (e.g.MS-Access).
- In addition to the point source 'risk analysis' and disease mapping options, it is also possible to import detailed exposure data, such as output from dispersion modelling.
- RIF provides a tool that allows users with skills in epidemiology to take advantage of the many functions that a GIS offers without requiring an in-depth knowledge of GIS.
- Since the application is embedded in ArcGIS, those with GIS skills will be able to use all the additional functionality that ArcGIS offers.
- Within the risk analysis tool the RIF performs tests on the relative risks to assess for homogeneity and linear trend with exposure.
- Within the disease mapping tool, the RIF performs empirical Bayes smoothing of the relative risks.
- The RIF can export data for further analysis in other (statistical) software

packages such as WinBUGS and SaTScan.

- The RIF can link directly to WinBUGS and SaTScan (where licences are held) and results from these software packages can be displayed concurrently in ArcGIS, with output from the RIF. This includes full Bayesian smoothing of the relative risks.

1.3 Input Facilities

As the RIF is embedded in ArcGIS, geo-referenced data can easily be incorporated into the RIF, for example the output from a dispersion model can be imported and used to define the study population in a risk analysis. In addition, contextual information which might aid interpretation can also be viewed alongside the RIF output.

1.4 Export Capability

The RIF is a versatile tool for generating smoothed disease maps and for calculating relative risks in populations living around putative sources of exposure. There are, however, additional software packages that can also be used to explore spatial and temporal trends in data, and to detect statistically significant clusters of disease that many users will wish to employ to aid their investigations. The RIF has been designed to work alongside these programmes, and can currently export data to WinBUGS, SaTScan, as well as to Microsoft Excel for further processing.

1.5 Scope of the Manual

Chapter 2 of this manual covers the basic background considerations the user should be aware of before undertaking an analysis of health risks at the small area level; including some of the geographic, dataset and statistical issues involved.

Chapter 3 indicates how to install and start up the RIF.

Chapters 4 and 5 provide all the necessary technical details to enable the user to develop databases for use in the RIF and to configure the RIF to run using those databases.

Chapters 6, 7 and 8 detail the RIF-specific menus, showing how to set-up and run a disease mapping or risk analysis query, and how to retrieve and delete previously run studies.
There are also two appendices: Appendix A lists the information, question, warning and critical error messages that might occur in the RIF, and gives details on how to optimise the performance of the RIF in MS Access.

Appendix B contains technical information, including details of the statistics applied in the RIF, details of geographic and population weighted centroids, information on covariates, and the basic use of SQL statements and clauses.

As the RIF is embedded in ArcGIS, there is access to all the functionality of ArcGIS, however this RIF user manual does not provide detail on these functions, as these can be accessed using the help files in ArcGIS, the ArcGIS user guides or from www.esri.com.

Additional relevant documentation can be found in an upcoming Environmental Health Perspectives Mini-monograph (Beale et al. 2008 116: 8: 1105:1110) and the Environmental Health Perspectives 2004 Mini-monograph on Information systems (EHP 2004, 112:995-1044).

The source code owners do not warrant that this Software and User Guide will meet your requirements or operate free from error, and disclaims all other warranties and conditions either expressed or implied, including the warranty of merchantability and fitness for a particular purpose.

Under no circumstances and under no legal theory, whether in tort, contract, or otherwise, shall the source code owners or its contractors be liable to you or to any other person for any indirect, special, incidental, or consequential damages of any character including, without limitation, damages for loss or goodwill, work stoppage, computer failure or malfunction, or for any and all other damages or losses.

2. Background Considerations

This chapter will give a very brief overview of some of the considerations that should be made when planning, undertaking, and interpreting a RIF study. These considerations are not unique to studies undertaken using the RIF, and although the RIF will help to speed up point source and mapping studies, users are cautioned to plan RIF studies as carefully as they would any other epidemiological investigation they would undertake. More details on these issues can be found in the following papers:

Beale L, Hodgson S, Abellan JJ, LeFevre S, Jarup L, 2010, Evaluation of spatial relationships between health and the environment: The Rapid Inquiry Facility, *Environmental Health Perspectives.* doi:10.1289/ehp.0901849

Beale L., Abellan J., Hodgson S., Jarup, L, 2008, Risk assessment using spatial epidemiological methods, *Environmental Health Perspectives*, Volume 116, Number 8

Ball W., LeFevre S, Jarup L., Beale L., 2008, Comparison of Different Methods for Spatial Analysis of Cancer Data in Utah, *Environmental Health Perpectives,* Volume 116, Number 8

2.1 Disease mapping or risk analysis?

There are two types of study that the RIF can undertake, disease mapping and risk analysis.

The risk analysis approach can be used to explore whether a source or some particular exposure (risk factor) is having an impact on health in a local population. To carry out a risk analysis study the geographical position of the putative risk factor will need to be known (as a point or a plume for example), and some consideration should be given to what distance the exposure of interest might be expected to have an impact. Thought should also be given to whether the exposure is likely to have a short or long term effect, as this will determine which years of health data will be most appropriate to study.

The disease mapping approach can be used to visualise mortality/morbidity rates and risks across an area. Disease mapping can provide an invaluable tool to explore spatial patterns of health outcomes; identify potential issues regarding data quality by geographical area; and identify areas which need additional resources or remediation.

Careful consideration should always be given to the most appropriate scale of investigation, which will depend on local circumstances (i.e. population density), and on the outcome of interest (i.e. whether this is a very rare outcome or not). The most appropriate geographical resolution to be used in any particular study will depend on individual circumstances and is often a compromise between having a high enough resolution to allow differences in disease risk to be assessed by small area, and having a large enough area (or population) to ensure that disease rates are sufficiently stable to permit interpretation. When mapping a rare disease across a sparsely populated area, thought should be given to the value of mapping at the smallest units available; if these units lead to very unstable risk estimates due to small populations, it may be preferable to lose some of the geographical resolution to gain more stable disease rates. While there may be basis for investigating the population living in very close proximity to a putative pollution source, thought should be given to whether the size of this ‘exposed’ population is sufficient to provide a meaningful risk estimate.

When assessing potential disease clusters post hoc, special care must be taken to avoid the ‘Texas sharpshooter’ effect, where the cluster is tightly defined in space and time, thus minimising the population at risk, and maximising the excess risk.

2.2 Geographical data issues

There are many different types of enumeration areas (e.g. administrative, health, electoral, postcode etc) and frequently their boundaries do not align. To use the RIF, however, the geographical data for any study must be hierarchical, with the boundaries at higher resolution areas being subdivisions of the larger areal units. In most countries census data are hierarchical. Since these boundaries tend to be defined administrative boundaries rather than physical boundaries, the boundary locations can, and do, change over time. Area names and codes can also change, which can be further complicated by the fact that different government departments can develop different coding systems for administrative geographies, or use slightly different names for the same area.

Inconsistent geography is problematic for any temporal studies that span time periods when boundary changes have occurred and are a major problem when trying to produce and compare meaningful statistics over time. The **Modifiable Areal Unit Problem (MAUP)**, as it is known, can affect any spatial study that utilises aggregate data sources (Openshaw, 1984). Since enumeration areas are often arbitrary and can change spatially and temporally, they are said to be 'modifiable’. Many spatial datasets are collected at a fine resolution (i.e. a large number of small spatial units) but are released only after being spatially aggregated to a coarser resolution (i.e. a smaller number of larger spatial units). This is usual for census data which are collected from every household, but released as aggregated data for an enumeration area. When values are averaged during the process of aggregation, variability in the dataset is lost and values of statistics computed at different levels of spatial resolution will be different. This change is called the **scale effect**. The **aggregation** or **zonation effect** must also be considered, which occurs due to the variation in numerical results that can occur due to the grouping of smaller areas into larger units (e.g., enumeration areas into census tracts). If we grouped EAs into zones of similar size to census tracts, but in a different spatial arrangement, we would likely find the statistical results are different between the two groupings of data.

Problems related to the **ecological fallacy** should also be considered. Users should be wary of interpreting results solely from aggregate statistics and making assumptions about the nature of individuals from data that relates to groups.

2.3 Health and population database issues

The appropriate statistical techniques and tools are available to calculate and map small area risks, but meaningful results can only be achieved if the underlying health and population data are accurate and complete. Local variations in ascertainment of health data, changes in health event recording over time (e.g. adoption of a new ICD revision), errors in the denominator (population) data (e.g. due to migration), or incomplete/inaccurate geocoding of either health or population data (e.g. greater positional errors for rural than for urban addresses) may introduce spurious temporal or spatial patterns in risk.

Any underlying data problems are not corrected merely by running the analysis through the RIF. It is vital that any data quality issues are known about, dealt with where possible, and where issues remain, that these are considered fully when interpreting the results.

2.4 Exposure data

GIS methods are available to improve pollution modelling and exposure assessment. While many ‘exposed’ populations have been classified based on proximity to a point source of pollution, these populations are being increasingly classified based on more realistic modelled exposure levels. Being embedded in a GIS, the RIF can easily handle the output from dispersion models or can attribute exposures based on available monitoring data. However, it should always be remembered that no matter how well modelled, or how appropriately monitored, environmental levels do not equate to actual exposure (Briggs, 2003).

Whereas it is often reasonable to assume that areas closest to a point source will have higher exposure than areas further away, using radii around a point source is a rather crude way of estimating exposure, as radii take no account of prevailing wind, topography, emissions, etc. If more detailed exposure information is available (e.g. dispersion modelling or monitoring data), it would be preferable to use these data to determine the ‘exposed’ populations, rather than using the simple radii. Where such data are not available, and radii are to be used, give careful consideration to what distances are appropriate with respect to the putative risk factor (is it likely to have only very local impact, or a wider effect?). Also consider the population density of the area of interest; if the area is very sparsely populated, even a wide band around a point source will return very unstable risk estimates. If the area is densely populated, a smaller band radius might still return stable risk estimates. Where little is known about whether the impact is likely to be very local or widespread, it might be prudent to define ‘default’ radii a priori (e.g. 0-2km and 2-7.5km have often been used by SAHSU in the UK (Aylin et al., 1999)). Where radii are employed, remember that there is likely to be significant exposure misclassification, and that this will normally bias the resultant risk estimates towards the null.

While the RIF can easily handle exposure data, thought should always be given to the potential for exposure misclassification. When investigating chronic disease, latency periods between exposure and disease onset must also be taken into account (though these are often not well characterised), and in such situations, migration of the population into and out of the area under study (which in many countries is poorly measured) must also be considered. Misclassification of exposure (differential or non-differential), either due to environmental levels not appropriately representing actual exposure or due to population migration, will reduce the study power and can potentially lead to biased study results.

2.5 Statistics

One problem associated with investigating health risks in small areas is that small populations have a small number of expected and observed events, which can lead to unstable risk estimates. This can result in misleading risk maps, especially if the areas with the smallest populations are quite large (rural areas for instance), as these areas with the least stable risk estimates can dominate a map. In an attempt to overcome this problem and to aid interpretation of the disease mapping output, the RIF also performs empirical Bayes smoothing of the raw relative risks to account for sampling variability in the observed data. These methods can allow more meaningful risks to be calculated at the small area level; however these statistical techniques need to be applied with due consideration and caution. While raw risks can produce noisy maps that are difficult to interpret, over-smoothed maps may produce a homogenous risk surface. Obviously there is a trade off between high sensitivity (where true high risk areas can be identified), and high specificity (where areas of no excess risk are correctly identified) (Richardson et al. 2004).

The RIF calculates standardised mortality (or incidence/morbidity) ratios (SMRs), however these measures are not directly comparable between different exposure groups as they are not based on the same standard population (i.e. the age, gender and socio-economic make up between the populations being compared are not exactly the same). This should only result in misleading comparisons where the population structure is significantly different between the groups being compared (Goldman & Brender 2000). An alternative to using indirectly standardised measures would be to use directly standardised rates and assess comparative mortality figures (CMFs) (or incidence/admissions figures) (Julious, Nicholl, & George 2001). The use of CMFs is advised for studies in which there are substantial numbers of cases in each study area or exposure category; however at the small geographical level, the number of cases is usually so few that directly standardised rates are unstable and the imprecision of this measure makes comparisons very difficult. In such situations it is appropriate to use SMRs instead of CMFs, provided the stratum specific death rates for each exposure class are proportional to the standard population rates, and bearing in mind that the rates in each exposure group may not be directly comparable with each other (Jarup & Best, 2003).

In a risk analysis investigation, the RIF tests the global null hypothesis (that the risks for each level of exposure are simultaneously equal to unity) (see appendix B.1). One disadvantage of this test is the lack of power against the specific alternative hypothesis of a trend in risks with increasing exposure. Even if the global null hypothesis cannot be rejected, substantial evidence of an exposure response trend may still be generated if the risks are in the hypothesised order (Breslow & Day, 1987). The RIF therefore also implements the Poisson trend statistic to detect a monotonic linear exposure response relationship (see appendix B.1).

Currently the RIF does not perform any type of temporal analysis. If users are interested in time trends in rates or relative risks, they might use the RIF to explore trends by running several annual (or other time length) periods and then plotting the rates/risks obtained. This would be in spirit similar to a moving average analysis. Although this could be valid for explorative purposes, users should be aware that it is not a proper moving average analysis, and therefore it lacks their properties, hence results should be interpreted carefully.

2.6 Interpretation and Limitations

Crucial to effective communication of spatial information is the use of suitable mapping techniques that convey results objectively. Effective mapping requires both an understanding of the mapped phenomena as well as the mechanisms to present the data appropriately. This is particularly true for maps that display data related to epidemiological risk in order to avoid misinterpretation or to over or under-emphasise particular results. Data symbolization has been chosen that should be suitable to effectively display results from RIF studies, however, the user can easily alter the map display if required.

The main advantages of undertaking spatial epidemiology at the small rather than large area level is increased interpretability - small-area studies are less susceptible to ecological bias created by within-area heterogeneity; they also allow local effects (such as impacts of point sources of pollution) to be investigated (Elliott & Wartenberg 2004). While analysis at the small area can help reduce components of ecological bias, unless the analysis is carried out at the individual level it is impossible to rule out this bias entirely. Factors associated with national or regional disease rates may not necessarily be associated with disease in individuals (Morgenstern 1998), and while the RIF can help assess whether a reported cluster is statistically significant or can demonstrate spatial trends in disease risk, the RIF cannot infer a causal relationship between an environmental factor and a disease. If cause for concern around a particular site is confirmed, data should be checked and validated (for completeness, diagnostic accuracy, etc). Replication around other or multiple sites with similar discharges (if they can be found) can be carried out or indeed etiologic studies at the individual level can be designed and carried out.

It should always be remembered that that RIF-type studies are subject to the limitations outlined above, and the user should therefore always consider what impact inconsistent geography, health and population data, exposure misclassification, ecological bias, and so on, will have on the study output. The RIF output therefore needs to be interpreted with caution and with expert local knowledge.

2.7 References

Aylin P, Maheswaran R, Wakefield J et al. 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.

Breslow NE & Day NE. 1987. Statistical Methods in Cancer Research Volume 2 – The design and analysis of cohort studies. IARC Scientific publications no 82.

Briggs D. 2003. Environmental measurement and modelling: geographical information systems, in Exposure Assessment in Occupational and Environmental Epidemiology, M. J. Nieuwenhuijsen, ed.

Elliott P & Wartenberg D. 2004. Spatial epidemiology: current approaches and future challenges. Environmental Health Perspectives 112(9):998-1006.

Goldman DA & Brender JD. 2000. Are standardized mortality ratios valid for public health data analysis? Statistics in Medicine 19(8):1081-1088.

Jarup L & Best N. 2003. Editorial comment on Geographical differences in cancer incidence in the Belgian Province of Limburg by Bruntinx and colleagues. European Journal of Cancer 39(14):1973-1975.

Jarup L. 2004. Health and Environment Information Systems for Exposure and Disease Mapping, and Risk Assessment. Environmental Health Perspectives 112(9):995-997.

Julious SA, Nicholl J, & George S. 2001. Why do we continue to use standardized mortality ratios for small area comparisons? Journal of Public Health Medicine 23(1):40-46.

Openshaw S. 1984. The Modifiable Areal Unit Problem, CATMOG, Concepts and Techniques in Modern Geography, No. 38, Norwich, GeoAbstrats

Morgenstern H. 1998. Ecologic Studies, in Modern Epidemiology, Second Edition, KJ. Rothman & S Greenland, eds, Lippincott Williams & Wilkins, pp. 459-480.

Nuckols JR, Ward MH, & Jarup L. 2004. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environmental Health Perspectives 112(9):1007-1015.

Richardson S, Thomson A, Best N et al. 2004. Interpreting posterior relative risk estimates in disease-mapping studies. Environmental Health Perspectives 112(9):1016-1025.

3. Starting up

3.1 Installing RIF

The RIF is freeware and no charge will be made for its use, however, this software is still under development. This current version of the RIF package is available only to eligible classes of users. Distribution is strictly through application to CDC's National Environmental Public Health Tracking (EPHT) Network or through the Small Area Health Statistics Unit (SASHU), Imperial College London. An SFTP site address and password will be supplied to successful applicants from where a copy of the RIF software can be downloaded.

The RIF is supplied as a self-extracting software which once copied from the SFTP can be easily loaded onto a PC.

The RIF requires the following:

- Windows XP (service pack 2 or higher), Windows 2003 (service pack 1 or higher) or Windows Vista.
- Microsoft Office 2003, including MS Access, MS Word and EXCEL 2003
- ArcGIS v9.0 or higher (ArcView licence level)
- Optional: ORACLE 10g release 2 (10.2 or higher)

Note: this version of the RIF requires that decimal numbers are separated by points and not commas.

The destination folder for the software can be specified on setup.

Note: A number of Microsoft security patches prevent Microsoft ActiveX controls from working (used for menu components). An effective solution to dealing with these, which does not require the removal of any patches, is the use of a virtual machine. A separate virtual machine setup document is available.

3.2 Test data

Before using your own data, we recommend using the sample health, population and geography data sets provided with the RIF software. These data give an idea of how the RIF works and help indicate what format data need to be in before they can be used in the RIF. The test data are automatically installed with RIF software, and in this version of the RIF/manual these data relate to a fictitious area known as Sahsuland.
Note: all these datasets are fictitious datasets and may not reflect patterns observed in reality.

The data consist of:

- Population data (by five year age group[1] by gender), for the period 1989-1996.
- Cancer incidence data for the period 1989-1996.

[1] Age groups are actually by one year age group for ages 0 to 4, then by five year age groups from ages 5 to 85, e.g. age groups 0, 1, 2, 3, 4, 5-9, 10-14, …, 80-84, 85+ (see section 4.2.3)

- Covariate data[2] on socio-economic status, ethnicity, and proximity to TRI (Toxic Release Inventory) sites.

The example dataset ‘Sahsuland', supplied with the RIF software, can be used to test the software setup and as a template for database construction.

Sahsuland is approximately 32860 km^2. The area of Sahsuland (Figure 1) uses four different hierarchical enumeration areas or *levels of geography*. Each area can be identified by a unique ID value. This also follows a hierarchical form, so that LEVEL2 areas are unique by LEVEL1 area, LEVEL3 areas are unique by LEVEL2, and so on. A unique ID at the highest resolution of level 4 is a combination of the level 1 ID, the level 2 ID and the level 3 ID, and follows the system used by many countries for their census data (e.g. FIPS in the USA, Output areas in the UK – see Table 1).

Administrative area			
Sahsuland	***USA***	***UK***	***Canada***
Level 1	State	District	Province (PR)
Level 2	County	Standard table Wards (ST Wards)	Census Division (CD)
Level 3	Tract	Super Output Areas (SOAs)	Census-subdivision (CSD)
Level 4	Census Block Group	Census Output areas (OAs)	Dissemination Area (DA)

Table 1. The census areas in Sahsuland

All level IDs are stored as text values with LEVEL1 areas using two characters, LEVEL2 uses 3 characters, which is joined with the LEVEL1 unique ID to make a LEVEL2 unique ID of 2 and 3 characters separated by a dot. The LEVEL3 units use a 6 character value and the LEVEL4 is a single character. Again, unique IDs for each region are achieved by concatenating each lower resolution area such that the proceeding level falls within each separated by a single dot.
Note. The data formats described in this section refer to Sahsuland data only. These data formats are not a requirement by the RIF. Data requirements are covered in section 4.

Screen shots and examples in this manual are based on this Sahsuland data.

[2] The RIF can handle ecological level covariate data (see section 4.2.2 and Appendix B.3).

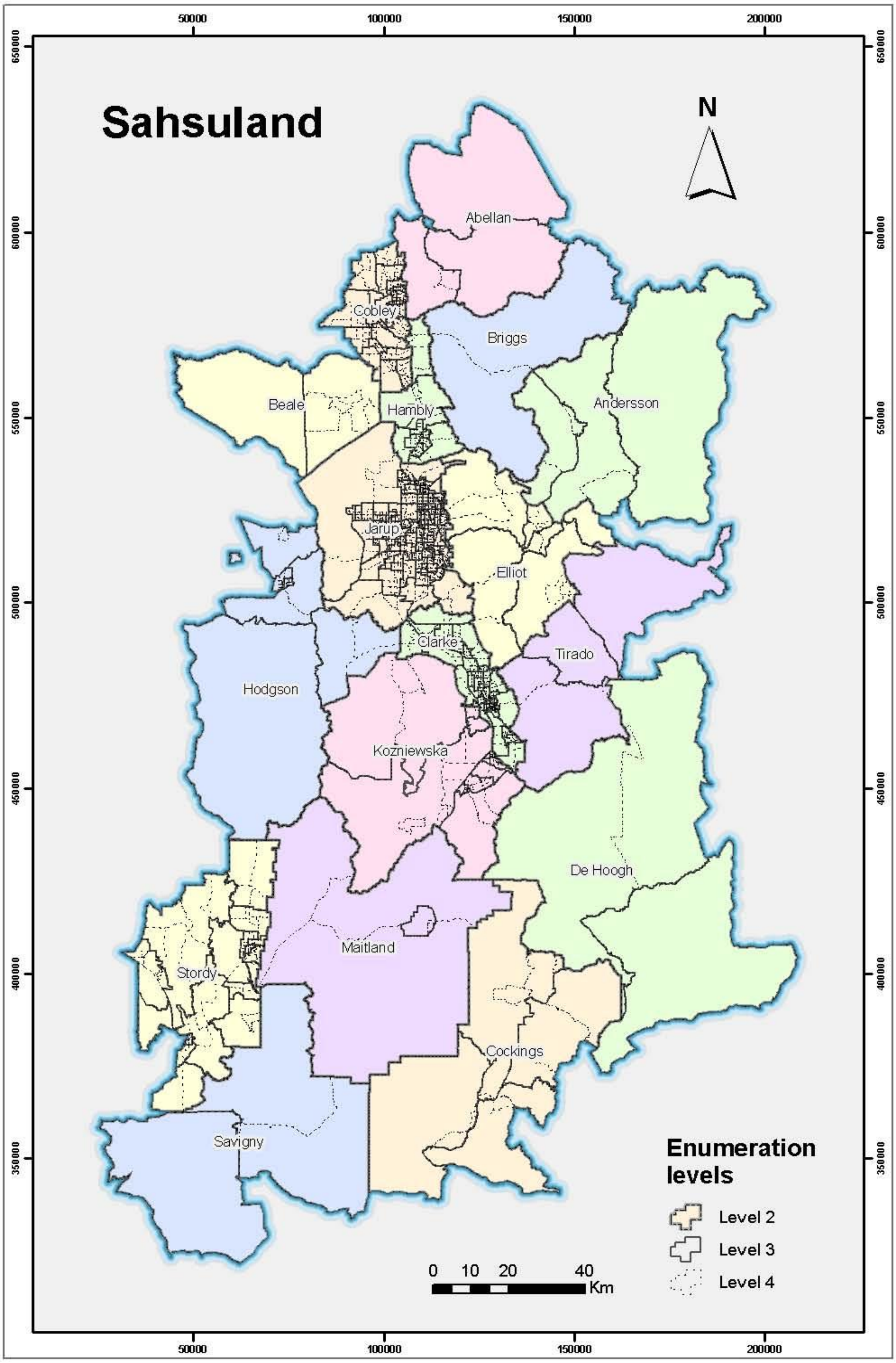
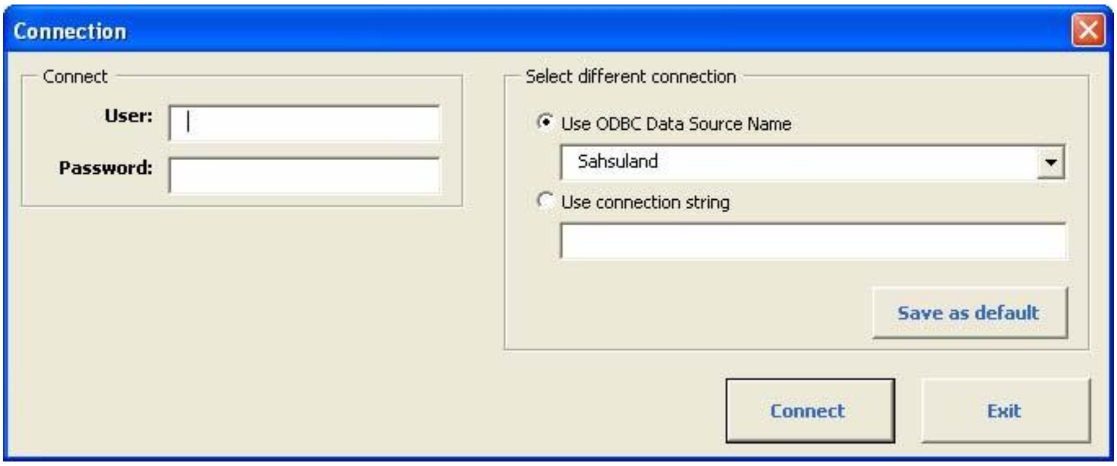


Figure 1. Sahsuland

3.3 Starting up

Double click on RIF desktop icon RIF to open up RIF.

This opens up an ArcGIS screen, which has the usual ArcGIS menus along the top. There is also a pop-up window, the RIF **Connection** screen (Figure 2).



Connection

Connect

User:

Password:

Select different connection

● Use ODBC Data Source Name

Sahsuland

○ Use connection string

Figure 2. Connection screen

This screen is used to connect to any compatible existing data source. The available ODBC data sources will depend on what data have been defined. New connections to RIF databases can be created with ODBC connectivity through Windows Control Panel - Admin Tools - Data sources (ODBC) Utility (see section 5.1). Password protection can be set-up for user-defined databases where required. If password protection has not been set-up then both user and password text boxes can be left blank.

Select the radio button that relates to the required connection type e.g. using an ODBC data source name or by using a connection string (this contains the information that the provider needs to know to be able to establish a connection to the database or the data file). If an ODBC data source name has been selected, the required data source should then be selected using the drop down list.

Click on the **Connect** button to connect to RIF.

Having connected to the RIF, an additional menu will be available in ArcGIS (Figure 3). The last menu (after **Help**) is now **RIF**. Clicking on this menu reveals the following options:

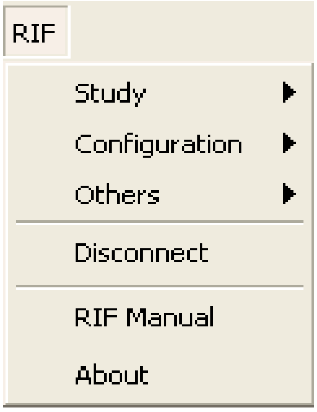


Figure 3. The RIF menu

The **Study** submenu is used to start a new RIF study (see chapters 6 and 7) or to retrieve or delete a previous study (see chapter 8).

The **Configuration** submenu is used to define the study geography, databases, numerator and denominators, covariates etc (for details see chapters 4 and 5).

The **Others** submenu allows the user to personalise the **Appearance** of the RIF windows and forms.

The **Disconnect** submenu allows the user to disconnect from the RIF; replacing the RIF menu with a **RIF [disconnected]** menu.

The **About** submenu identifies what version of the RIF is being used, and provides contact details of SAHSU.

4. Data specifications for RIF studies

4.1 Introduction

This chapter outlines how a database should be set up to be used with the RIF. Examples are given of the tables required.

There are 48 generic tables that must exist in any RIF database for studies to be carried out. These tables are supplied as a database ‘rif_empty.mdb’, for use with an ACCESS database. This ACCESS database should be copied and renamed to any chosen name. The original ‘rif_empty.mdb’ should be kept as a template for future database development. A number of additional tables will need to be added to the database and are outlined in this database development chapter.

In addition to storing the input data, the tables in the database will be used to store all output from any RIF study that has been carried out. All results can, therefore, be accessed in this database.

Note. Since the RIF stores all results, users should remember that database size will continue to expand. Unwanted studies can be deleted at any time (see Chapter 8)

4.2 Concepts

The key concepts for each data type are detailed further in the following section. A full understanding of the RIF requirements should help users to set-up their database. This section gives more details about the geography, covariates, age-sex groups and health event coding.

4.2.1 Geography

The geographical levels are assumed to be hierarchical so that the boundaries of the smaller areal units are nested in the larger unit boundaries (Figure 4). Using the example data for Sahsuland each consecutive level is a further subdivision of the previous level, so that LEVEL1 represents the whole of Sahsuland (e.g. State) and LEVEL2 shows 17 divisions (e.g. County). LEVEL3 shows the 17 LEVEL2 areas further sub-divided so that there are 200 unique areas (e.g. Tract). These 200 LEVEL3 areas are then further divided to give 1230 unique areas at LEVEL4 (e.g. Census Block Group). This is the highest resolution data for Sahsuland.

For each of the 1230 areas in LEVEL4 there will be a corresponding LEVEL3, LEVEL2 and LEVEL1 area in which it falls. Knowing which higher resolution areas are contained in the next lowest resolution dataset enables the spatial distribution of each different level or hierarchy of geography to be understood. This ‘link’ between the different levels of geography is used in the database (outlined in detail in chapter 5).

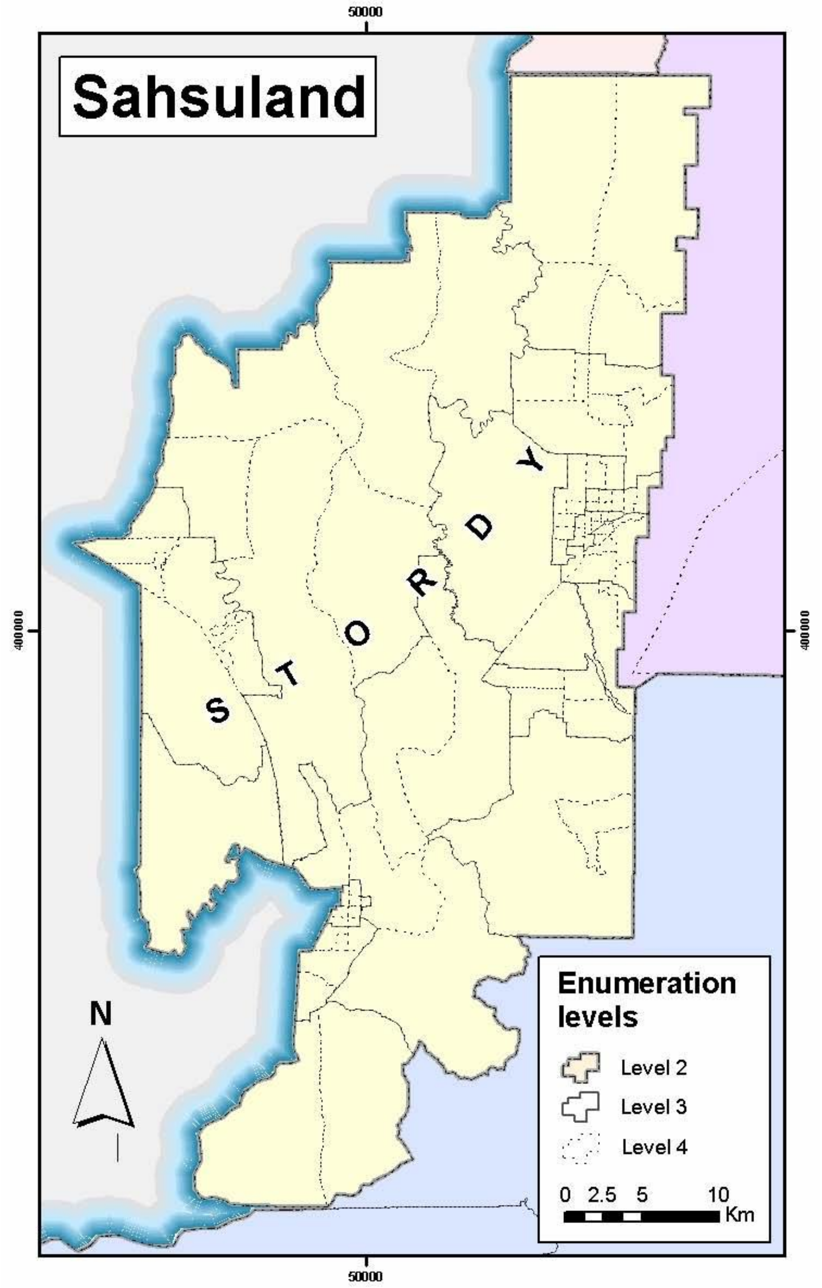


Figure 4. Hierarchical boundaries

In many countries administrative boundaries are clearly established and well-known e.g. census boundaries (see section 2.2). These often follow this hierarchical nature and, therefore, can easily be implemented in the RIF.

Geographical data should use an appropriate map projection and all shapefiles **must** use the same projection (explain). Map projections are used so as to portray the surface of the earth (or a portion of the earth) on a flat surface. All projections will distort reality in some way, for example in distance, direction, scale and/or area. The RIF will assume that all geographical data follows the same map projection as the first shapefile that is used. It this is not true the locations of further geographical data may be incorrect and will lead to erroneous results or problems in running the RIF.

4.2.2 Covariate data

Covariate data can be included in the RIF for both adjustment (‘*adjusting covariates’*) and as a means of defining study areas (‘*exposure covariates’*). The RIF uses ecological level covariate data. Exposure covariates cannot vary with time, and where covariate values do change over the period of the investigation the user must create an appropriate exposure metric (such as the mean value, or maximum value) to apply throughout. The RIF supports adjustment for zero or more ‘adjusting covariates’. Each investigation within a study may include adjustment for a different set of covariates. Finally, the covariate resolution must always be the same as the resolution of the study area.

Covariate data must be stored in a specific table, separate from the health or population data. The name of this covariate table must be defined in the Geographical levels dialog.

When choosing study areas with respect to a numeric (continuous) covariate, the RIF can suggest cut-points to create evenly distributed categories. No less than two categories (one cut-point) and no more than 7 categories can be made. A numeric (continuous) covariate can only be used to select study areas and cannot be used for adjustment.

Results from the homogeneity test can be performed regardless of the type of exposure covariate; however, the linear trend test will not be performed where the covariate data are categorical, nominal data.

See Appendix B.3 for more details of the use of covariates in the RIF.

4.2.3 Age Groups

The age groups can be defined by the user although default values are set in the RIF database (‘rif_empty.mdb’) in the table ‘RIF_AGE_ GROUPS’. An age group table must exist in any database to be used with RIF. Table 2 shows the 22 different categories used in the default table but of course this table can be replaced in preference of a user defined table. For example, a user may prefer to combine ages 0-4 years into a single age group. Thus, in this example, there would now be only 18 age groups spanning the range 00 to 17.

Age Group	Age
00	0
01	1
02	2
03	3
04	4
05	5-9
06	10-14
07	15-19
08	20-24
09	25-29
10	30-34
11	35-39
12	40-44
13	45-49

14	50-54
15	55-59
16	60-64
17	65-69
18	70-74
19	75-79
20	80-84
21	85+

Table 2: Age-sex group categories

The same age categories that are defined in this table must also be present in the health data (see section 4.3.3 Tables). The same age groups will be used for both male and female sex groups.

4.2.4 Health event coding

Every health record stored in the database must have, at least, one field containing a health end point. The RIF can use numerator data coded in any format (including user-defined codes, SEER codes, ICD codes, etc). Users select the specific end point of interest from the **Investigation Details** screen (see section 6.4 and 7.4) using one of three approaches:

- *Choose a predefined group*: Users can develop a list of frequently used end points by adding the appropriate SQL clause and description to the table RIF_PREDEFINED_GROUPS (supplied as part of the ‘rif_empty.mdb’). Elements stored in this table can then be selected from a drop down list to be used as the end point of interest in subsequent studies. This table can be particularly useful where health data are not coded using ICD 9 or 10 codes (when the chapter lists can be used).
- *Enter SQL clause***:** Users can type out an SQL clause directly into the **Investigation Details** screen (although this will not be stored for subsequent use).
- *Pick ICDs*: Users can, if their health data are coded using the International Classification of Diseases (ICD) revision 9 or 10, pick ICD codes from a pop-up box listing the ICD9 and ICD10[1] codes in a treeview. Any chapter, group, or specific 3 or 4 digit code can be selected simply by clicking in the box adjacent to the outcome of interest. Where health data span several years and include a change-over in ICD coding, the user may also specify in the **Configuration menu** (see chapter 5) which year the coding changed from one classification to another in that dataset. When the year of change is specified the text on the Investigation details menu reflect which ICD codes are required for the study (e.g. ‘Only ICD10 is required’, or ‘ICD9 and ICD10 are required’).

The option to select ICD codes from the drop down list has been added for the convenience of users with ICD coded data who may not be confident using SQL clauses; however these lists should always be used in conjunction with

[1] ICD-10 codes, terms and text used by permission of WHO, from: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10). Second Edition Vols 1-3. Geneva, World Health Organization, 2004.

the full details of inclusions/exclusions that apply for each ICD code, group or chapter, which are available from the WHO ICD books (http://www.who.int/classifications/icd/en/).

The advantage of using SQL clauses to identify the health end point of interest is to allow very specific end points to be returned, for example renal disease in diabetics, or a specific cancer type by site (provided this level of detail is present in the database itself). More details on using SQL clauses can be found in appendix B.3.

Where the numerator data do not include a health end point, perhaps in a register for a specific disease where all individuals share the same health end point (for instance in a diabetes register), a health end point field will still need to be present. It is possible to add this field and populate it with a single character. For example, a field could be added called ‘end point’, and populated with ‘1’ for every record. With this example, the data would be queried in the RIF using a predefined query or SQL clause such as:
end point = ‘1’

4.3 Tables and Shapefiles

4.3.1 RIF data formats

Data for RIF studies are primarily stored in database tables. Spatial data are also required in the form of shapefiles. The RIF links the spatial data and the database data so that spatial functions and queries can be easily be carried out, whilst - large datasets and intensive calculations are efficiently handled.

The data stored in both the database and the shapefiles must cover the same geographical areas and have the same column names. The column names are not fixed, and can be user specified, however, some of the naming must be consistent between shapefiles and the database. Shapefile column names should not exceed 10 characters in length.

When a database is being built for the first time it is good practice to include all known hierarchies of geographical data, both in the database and as shapefile data as this will ensure that future studies can be carried out at any geographical level, depending upon the available health data available. The resolution of the data that can be used for RIF studies can be easily controlled in the RIF configuration menu (see chapter 5). Users should be aware that administrative boundaries tend to change over time and, therefore, different boundary data may be required for certain studies (see section 2.2).

It is important that the assumptions outlined in section 4.2.1 are true such that the geography is hierarchical and the areas in the database are unique.

4.3.2 Shapefiles

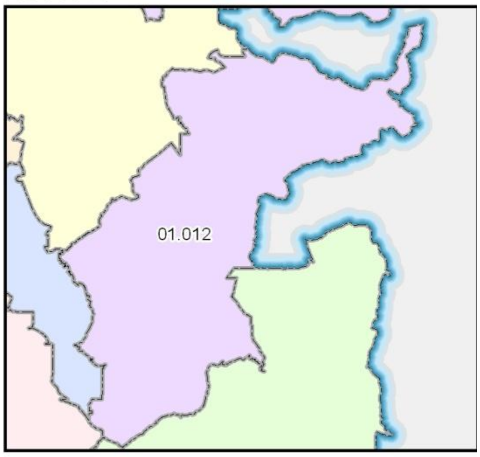
Shapefiles are used by the RIF to define the geographical extent and resolution of any study. At minimum geographical boundary data must be held for the same geographical resolution as the health data. All shapefiles must be held in the same map projection. If more than one resolution of spatial data is to be used the geographical boundaries between levels must be hierarchical so that the boundaries of the higher resolution data are subdivisions of the larger areas (Figure 5).

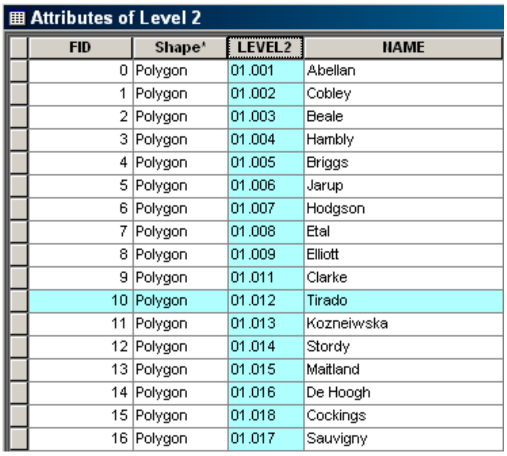
The shapefiles for each geographical boundary dataset to be included must be defined at setup. The RIF does not require any specific naming convention for shapefiles. A column must exist that provides a link to the database. This column must have the same name as the corresponding geographical data table, stored in the database. Again, a column name can have a maximum of 10 characters. For example, if a shapefile Sahsu_LEVEL4.shp contains an ID field called *LEVEL4* this will require a population table with the suffix *_LEVEL4* in the database.

In addition to geographical boundary data, users can use point shapefiles to represent the ‘centre’ of each geographical area. In the case of population, the ‘centre’ is defined from the spatial distribution of the population rather than the spatial boundaries and can, therefore, can be termed population-weighted centroid data. Population weighted centroid data are not compulsory as geographical centroids are automatically calculated for any area selected for study. Including population weighted centroid data where available will, however, significantly reduce exposure misclassification by selecting the most appropriate populations in any study (see appendix B.2).

Note. Users should note that the study area is selected spatially (and not on ID), therefore, errors will occur if any population weighted centroid does not lie inside its relevant spatial boundary (this may arise if any polygon areas have large concavities).

Level 2

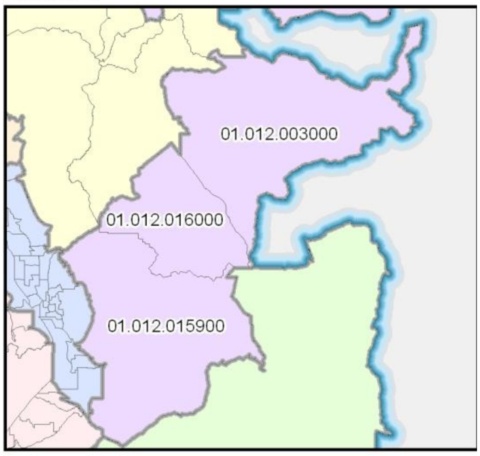


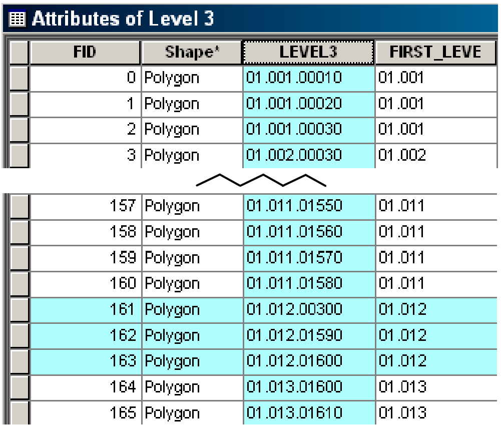


Attributes of Level 2

FID	Shape*	LEVEL2	NAME
0	Polygon	01.001	Abellan
1	Polygon	01.002	Cobley
2	Polygon	01.003	Beale
3	Polygon	01.004	Hambly
4	Polygon	01.005	Briggs
5	Polygon	01.006	Jarup
6	Polygon	01.007	Hodgson
7	Polygon	01.008	Etal
8	Polygon	01.009	Elliott
9	Polygon	01.011	Clarke
10	Polygon	01.012	Tirado
11	Polygon	01.013	Kozneiwska
12	Polygon	01.014	Stordy
13	Polygon	01.015	Maitland
14	Polygon	01.016	De Hoogh
15	Polygon	01.018	Cockings
16	Polygon	01.017	Sauvigny

Level 3

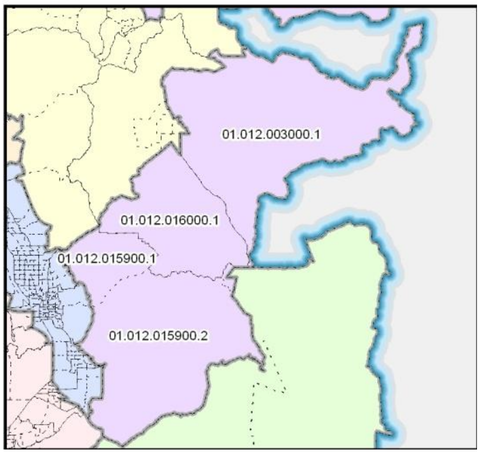


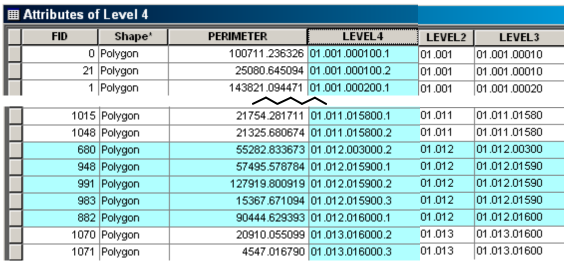


Attributes of Level 3

FID	Shape*	LEVEL3	FIRST_LEVE
0	Polygon	01.001.00010	01.001
1	Polygon	01.001.00020	01.001
2	Polygon	01.001.00030	01.001
3	Polygon	01.002.00030	01.002
157	Polygon	01.011.01550	01.011
158	Polygon	01.011.01560	01.011
159	Polygon	01.011.01570	01.011
160	Polygon	01.011.01580	01.011
161	Polygon	01.012.00300	01.012
162	Polygon	01.012.01590	01.012
163	Polygon	01.012.01600	01.012
164	Polygon	01.013.01600	01.013
165	Polygon	01.013.01610	01.013

Level 4





Attributes of Level 4

FID	Shape*	PERIMETER	LEVEL4	LEVEL2	LEVEL3
0	Polygon	100711.236326	01.001.000100.1	01.001	01.001.00010
21	Polygon	25080.645094	01.001.000100.2	01.001	01.001.00010
1	Polygon	143821.094471	01.001.000200.1	01.001	01.001.00020
1015	Polygon	21754.281711	01.011.015800.1	01.011	01.011.01580
1048	Polygon	21325.680674	01.011.015800.2	01.011	01.011.01580
680	Polygon	55282.833673	01.012.003000.2	01.012	01.012.00300
948	Polygon	57495.578784	01.012.015900.1	01.012	01.012.01590
991	Polygon	127919.800919	01.012.015900.2	01.012	01.012.01590
983	Polygon	15367.671094	01.012.015900.3	01.012	01.012.01590
882	Polygon	90444.629393	01.012.016000.1	01.012	01.012.01600
1070	Polygon	20910.055099	01.013.016000.2	01.013	01.013.01600
1071	Polygon	4547.016790	01.013.016000.3	01.013	01.013.01600

Figure 5. Unique area IDs in shapefiles and database tables

4.3.3 Tables

Data for all RIF studies are selected from a number of database tables. These tables must be present in the database for the RIF to carry out any study. Initially the RIF will be set-up to work with the provided example dataset of Sahsuland - however, for valid studies using other data a database must be built. The tables in a database for use with RIF must include the 48 provided in the ‘rif_empty.mdb’ *plus* numerator, denominator and geographical data (and covariate data where available).

The following describes some key concepts for the storage of data in RIF tables.

- **Numerator and denominator tables**
  The denominator data represent the population data while the numerator data contain the health data. All health records and population data must have an associated age group. Most commonly, numerator data should be stored in a single table (with all geographical resolutions contained in that table). Denominator data are normally more efficiently stored in one table per geographical resolution. This is the adopted method for the Sahsuland database.

- **Denominator data rotation**
  Denominator data can be stored in two different formats depending on whether the data are rotated or not.
  The data *rotation* is with respect to the age-sex grouping. In simple terms, the age-sex group information is either held by row (record) or by column (field):
  - Non-rotated denominator data contains rows for every age, sex, and, optionally, covariate-specific group, for each geographical area.
  - Rotated denominator data contain rows for each covariate specific group, for each geographical area, and contain a column for each age-sex group (e.g. 44 columns with default settings) (see table 5).

- **Summing and counting data**
  In the special case when data are not aggregated (shown in table 4) and each row refers to an individual, the total population will be the total number of data rows, and the data can hence be ‘*counted’*. Conversely, if the data are aggregated, an additional column is required in the database. This should store the total population per area, age-sex-group and year. The total population is then calculated by the RIF by *summing* this column. The name of this column must be specified by the user at configuration (see section 5.8).

- **Numerator data**
  Numerator data must be stored in a *non-rotated* form with only one column specifying age-sex-group. In most cases (e.g. cancer incidence) storing the numerator data in a rotated form would be inefficient since many cells (especially those representing younger age groups) would contain zeros.

  Numerator data are more likely than population data to be available at the individual level and so may well be countable (in that the total number of data rows equals the total number of individuals).

The structure of the numerator table is shown below (table 5).

- **Data formats in the RIF**

  Both non-rotated and rotated methods of denominator data storage are acceptable for use with the RIF. Choice of method will be driven by the user data format, although users are encouraged to consider data redundancy and storage efficiency. Details on how to optimise the performance of the RIF in MS Access are provided in Appendix A.6.

  The naming convention for denominator data that is specified at different geographical levels is

  <name root>_<geographical level>

  e.g. SAHSU_POPULATION_LEVEL4

  This naming convention can also be used for the numerator table but please note that the _<geographical level> extension is for user information only and the RIF expects a single numerator table.

  The suffix for the geographical level will be automatically added if the checkbox “Add suffix for geography” is checked (see section 5.4). The table name should then be entered as <name_root> only.
  If you only have numerator and denominator data for one level, adding a suffix is optional.

Only required with grouped data

Only required with grouped data

Column	Year	<Level1>	<Level2>	…	<Level n>	<Covariate1>	…	<Covariate n>	AGE_SEX_GROUP	<Total>
Data Type	Number (integer)	Text	Text	Text	Text	Number	Number	Number	Number	Number

Table 4: Denominator data: Age-sex group non-rotated

Column	Year	<Level1>	<Level2>	…	<Level n>	<Covariate1>	…	<Covariate n>	<ASG1>	<ASG2>	…	<ASG44>
Data Type	Number (integer)	Text	Text	Text	Text	Number	Number	Number	Number	Number	Number	Number

Table 5: Denominator data: Age sex group rotated

Only required with grouped data

Column	Year	<Level1>	<Level2>	…	<Level n>	<ICD>	AGE_SEX_GROUP	<Total>
Data Type	Number (integer)	Text	Text	Text	Text	Text	Number	Number

Table 6: Numerator data

Where[*]:

Column shown with tags (e.g. <Column>) do not have fixed column names and the name can be replaced by any of the user's choice.

Column names shown without surrounding tags must use the exact name shown in the example table.

The data type required for each field is shown below the Column name. These are mandatory as the database will not work with the RIF with different formats.

Numbers can be assumed to be decimal data unless specified as integer.

[*] This convention is used for all tables is this manual

- **Covariate tables**
  The structure of the covariate table can be seen in table 6. It does not need to contain any other information than year, geographical information and the covariate values. The RIF can handle ecological covariates *only*, i.e. one value per covariate per area per year. There can never be two different values of the same covariate for the same area and year for a given geographical level.

  The covariate table name must be entered when configuring the geographical levels.

Field	Year	<Level1>	<Level2>	…	<Level n>	<Covariate1>	…	<Covariate n>
Data Type	Number (integer)	Text	Text	Text	Text	Number	Number	Number

Table 6: Covariate table structure

If ecological covariates exist at two different levels (e.g. Socio-economic status [SES] at LEVEL4 and INCOME at LEVEL3) two different covariate tables must be created. Both tables need to have information about all lower resolutions of geography.

Example: A covariate table containing SES values at LEVEL4 has to contain details for all lower resolutions, such as LEVEL3, LEVEL2 and level 1 (table 7). A covariate table of INCOME values at LEVEL3 would require IDs of LEVEL2 and LEVEL1 to be stored in the same (see table 8).

Field	YEAR	LEVEL1	LEVEL2	LEVEL3	LEVEL4	SES
Data Type	Number (integer)	Text	Text	Text	Text	Number

Table 7: Covariate data for LEVEL4

Field	YEAR	LEVEL1	LEVEL2	LEVEL3	INCOME
Data Type	Number (integer)	Text	Text	Text	Number

Table 8: Covariate data for LEVEL3

Note: the covariate level must always be the same as the study level.

- **Look-up tables for geographical levels**
  For each geographical level it is necessary to define a *look-up table*. This table is used to check the validity of the areal units selected for the study in the RIF. The same data that is selected from a shapefile in the GIS must exist in the database for the study to proceed.
  The database tables can be built using the '.dbf' file from any shapefile, however, in the database each area must be unique and this may not be the case in the geographical data. An area may be spatially disparate (e.g. islands) but they belong to the same administrative region. In this case a new

record will be held in the shapefile for each geographical area but the same attribute value will be stored that defines the unique level value. The shapefile data must be '*dissolved*' using the field that defines the area values before it can be used to build the geographical level lookup table. The geographical data, however, are appropriate with multiple spatial areas and a single identifier (e.g. an enumeration area that is physically split such as Islands).

The table should have two columns, one which defines the geographical unit and is a unique value for each area (this column is compulsory) and a column to which some descriptive information can be added, such as the name of the geographical area. The column that holds the unique value should be defined as the 'primary key' as this will provide a check that each row is unique.

Two additional columns can also be added to the table that the RIF can use if desired. If the geographical unit has X/Y coordinate data it can be used for risk analysis studies in the RIF, e.g.

Column	<zone_id>	<name>	<X coordinate>	<Y coordinate>
DataType	Text	Text	Number(decimal)	Number(decimal)

compulsory optional

- **Look-up tables for geographical hierarchy**
  The RIF needs a table that defines the relationship between the geographical units. This table will by definition have the same total number of records as that of the highest resolution dataset that is to be used e.g. LEVEL4 (1230 records).

  For each LEVEL4 areal unit there will be a corresponding LEVEL3, LEVEL2 and LEVEL1 value. A *single table* should be defined with a total number of records from the highest resolution data (LEVEL4) and a column for each level of geography that is to be included (e.g. LEVEL3, LEVEL2 and LEVEL1). This allows the RIF to understand the hierarchical relationship of the geographical data.

  Sahsuland example*:* If you wish to use all the areas in LEVEL2 area of 'Tirado' ('value 01.012') then the RIF will automatically select the three LEVEL3 areas that are contained in this area of '01.012.0030000', '01.012.016000' and '01.012.015900'.
  The name of this table can be any user specified name. The table has the structure shown below.

Column Name	<Level1>	<Level2>	<Level3>	<Level4>
DataType	Text	Text	Text	Text

Table 9: Geographical table structure

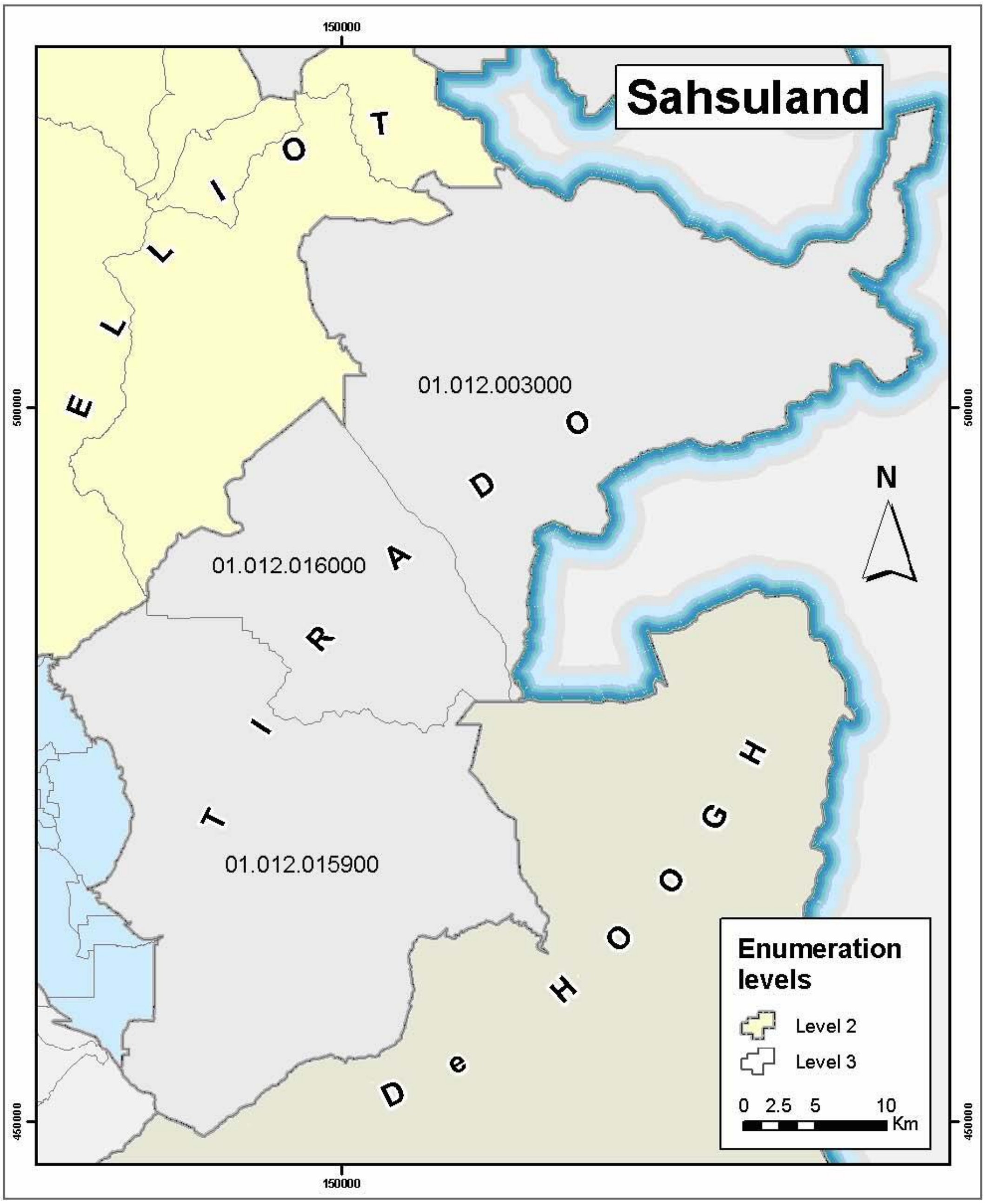


Figure 6. Link between hierarchical enumeration data

Sahsuland example: Data tables (with column names) and shapefiles required for Sahsuland using enumeration areas where data for four different census levels are held (LEVEL1, LEVEL2, LEVEL3, LEVEL4)
Denominator data (grouped to each geographical level):

SAHSU_POP_LEVEL4: year, level4, level3, level2, level1, M0, M1,…M85PLUS, F0..F85PLUS
SAHSU_POP_LEVEL3: year, level3, level2, level1, SES, ETHNICITY, AreaTRI1km, TRI_1km M0, M1,…M85PLUS, F0..F85PLUS
SAHSU_POP_LEVEL2: year, level2, level1 SES, ETHNICITY, AreaTRI1km, TRI_1km M0, M1,…M85PLUS, F0..F85PLUS
SAHSU_POP_LEVEL1: year, level1, SES, ETHNICITY, AreaTRI1km, TRI_1km M0, M1,…M85PLUS, F0..F85PLUS

Numerator data (available only at level4):

SAHSU_CANCER_LEVEL4: year, level4, level3, level2, level1, AGE_SEX_GROUP, ICD, TOTAL

Covariate table:

SAHSU_COVARIATES_LEVEL4: Year, level4, level3, level2, level1, SES, ETHNICITY, AreaTRI1km, NEAR_DIST, TRI_1km

Look-up tables for geographical units:

SAHSU_LEVEL1: zone_id, name
SAHSU_LEVEL2: zone_id, name
SAHSU_LEVEL3: zone_id, name
SAHSU_LEVEL4: zone_id, name

Look-up table for hierarchy:

SAHSU_GEOGRAPHY: level4, level3, level2, level1

Shapefiles:

Boundary files:
SAHSU_GRD_level1
SAHSU_GRD_level2
SAHSU_GRD_level3
SAHSU_GRD_level4
Centroids:
SAHSU_CEN_ level4

4.4 Database development: Quick check

The following should be checked as part of database setup:

- Geographical data are hierarchical
- Geographical areas are unique in the database
- The GIS data are in shapefile format
- The key for the geographical areas in the database and the shapefiles identify the same areas
- A single lookup table exists for the different levels of geography
- Shapefile column names match relevant column names in the database
- All 48 tables as defined in ‘rif_empty.mdb’ exist (i.e. make a copy of this mdb database and data to this)
- Numerator and denominator tables have been added
- Column names and data types follow RIF specifications

5. Configuring the RIF

5.1 Introduction

This chapter describes how to define the environment variables in the RIF. It therefore assumes that a suitable database for use with RIF exists (see chapter 4). Initially, the example dataset can be used but for other datasets many of the environment variables described herein will need specifying. Indeed, many variables must be defined for each new database that is to be used with the RIF.

Upon launch, the RIF will require connection to any database other than Sahsuland. This should be done before starting the RIF by setting up an **ODBC connection.** This must be done with the Windows environment by:

start > control panel

Administrative tools > Data Sources (ODBC) > Add > Microsoft Access Driver (*.mdb)

Type in the desired database name (and a description if required)

Select > Browse to the database

Set Options > Buffer size to 0

5.2 Configuration menus

The configuration menus are only required for the setup of RIF to define the environment settings for any RIF study. The configuration submenu is accessed from the RIF drop down menu which appears upon installation of the RIF (Figure 6).

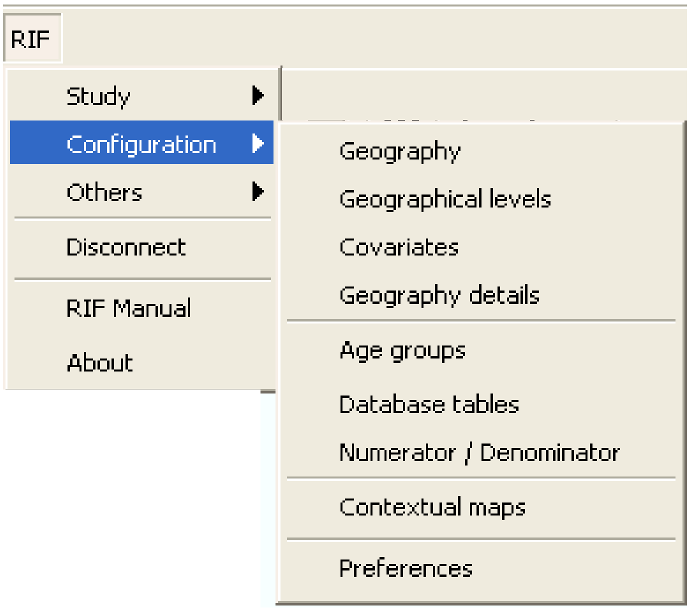
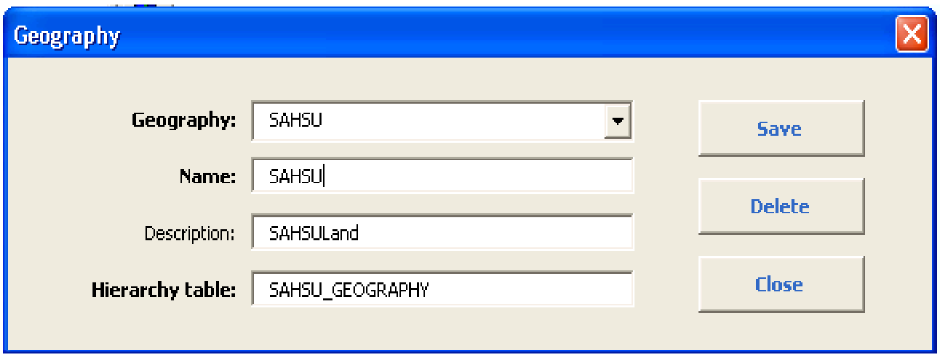


Figure 6. Configuration menus

Each of these submenus will be described in turn.

5.3 Geography

A name should be given to a dataset that can be used for any number of subsequent RIF studies in that area. Conventionally, a name that represents the geographical area that the shapefiles and health, population and covariate data cover is used. Naming datasets by geographical area enables several separate databases to be stored for use with the RIF. A description of the area can also be included.



Geography: SAHSU
Name: SAHSU
Description: SAHSULand
Hierarchy table: SAHSU_GEOGRAPHY

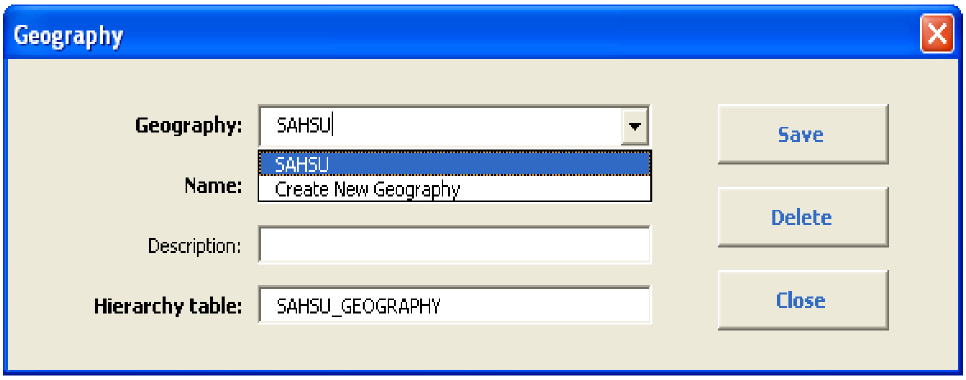
Figure 7. Geography configuration menu

A name of the geographical hierarchy table must also be entered on this menu. This table should contain information showing the links between the different geographical hierarchical levels of data i.e. which high resolution areas fall inside which lower resolution areas. This table is crucial as it outlines the spatial link between different resolutions of data (section 4.3.3. Tables)

For each 'geography database' **all** the relevant shapefiles and database tables must then be specified.

To allow the RIF to use a new dataset it must be added in configuration menus:

- Use the drop-down menu and select 'Create New Geography'
- The Name, Description (optional) and Hierarchy table name must be added
- 'Save' before closing the menu

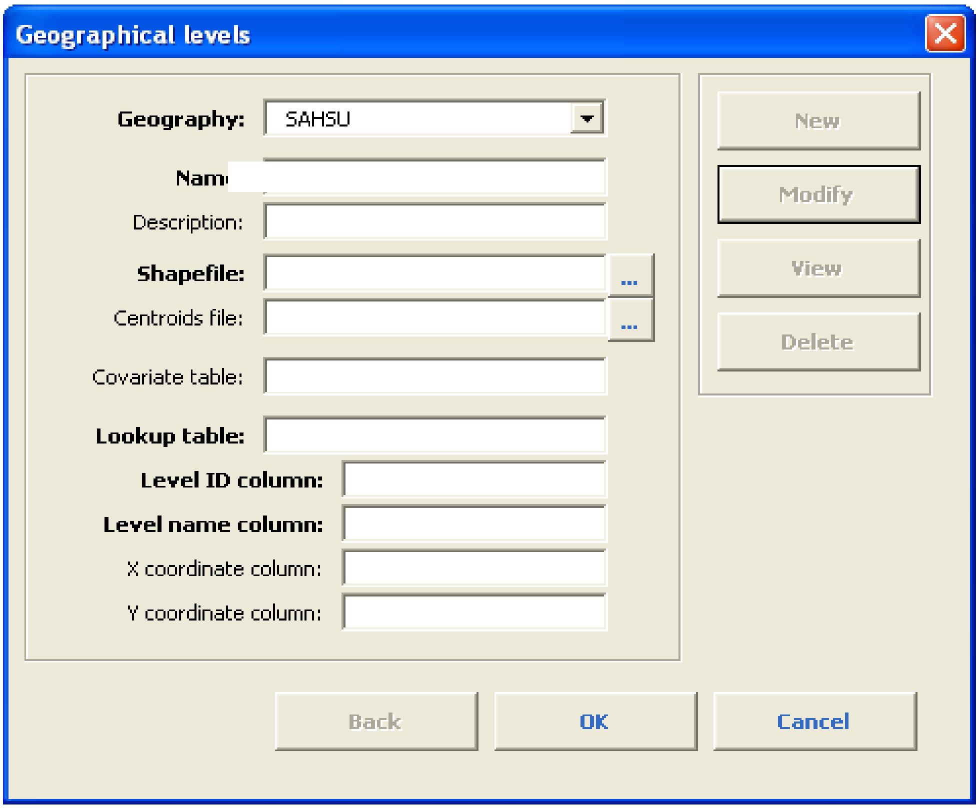


Geography: SAHSU
Name:
Description:
Hierarchy table: SAHSU_GEOGRAPHY

Figure 8. Adding New Geography

5.4 Geographical levels

This menu allows all the information that relates to the spatial information to be specified. The details of each geographical level should be entered. This step should be repeated for every geographical level of data that are to be included in the specified dataset.



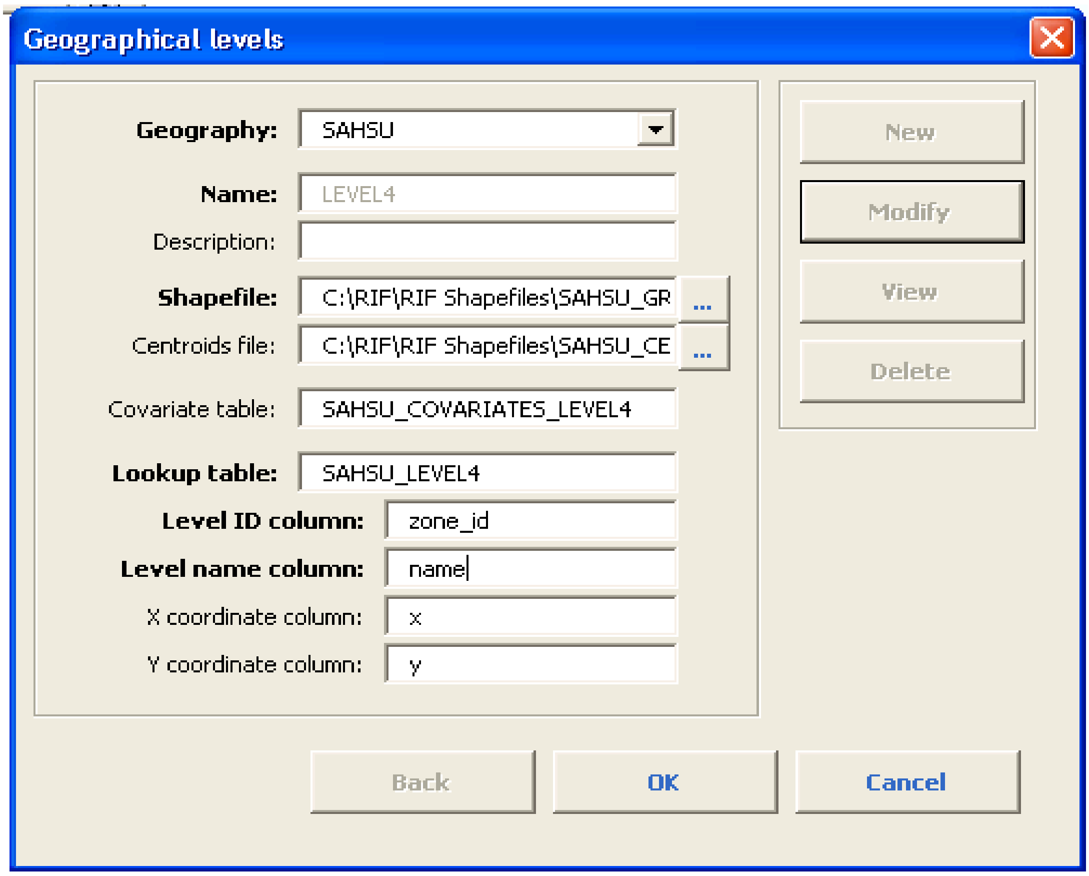
Geography: SAHSU

Figure 9. Geographical levels menu

Compulsory fields as shown in **bold** on the menu. The features that can be defined are:

- *Geography*: the name of the geography for which you want to define the details
- *Name*: the geographical level name. This name must be the same as that used for the column name in the **shapefiles**
- *Description*: a brief description of the geographical level
- *Shapefile*: the **full** pathname and name of the related shapefile for the defined geographical level (shown by Name)
- *Centroids file*: the **full** pathname and name of the related point shapefile for any population weighted centroids at the defined geographical resolution
- *Covariate table*: the name of the covariate table that exists for the selected geographical level data (if covariates are to be made available for use in any RIF study).
- *Look up table*: name of the lookup table in the database
- *Level ID column*: name of the field that holds the unique ID (which should match the ID in the relevant shapefile)
- *Level name column:* additional descriptor for each area e.g. the administrative name of the area
- *X coordinate column*: the field in the lookup table that stores the X coordinate
- *Y coordinate column*: the field in the lookup table that stores the Y coordinate

Example: In our Sahsuland example below (Figure 10), the LEVEL4 geography has a boundary shapefile, a weighted centroids file and a covariate table. The corresponding lookup table called SAHSU_LEVEL4 has 'zone_id' as the level ID column and 'name' as the level name column. For LEVEL4 areas there are X, Y coordinates available in columns called 'x' and 'y' respectively.



Geographical levels

Geography: SAHSU

Name: LEVEL4

Description:

Shapefile: C:\RIF\RIF Shapefiles\SAHSU_GR

Centroids file: C:\RIF\RIF Shapefiles\SAHSU_CE

Covariate table: SAHSU_COVARIATES_LEVEL4

Lookup table: SAHSU_LEVEL4

Level ID column: zone_id

Level name column: name

X coordinate column: x

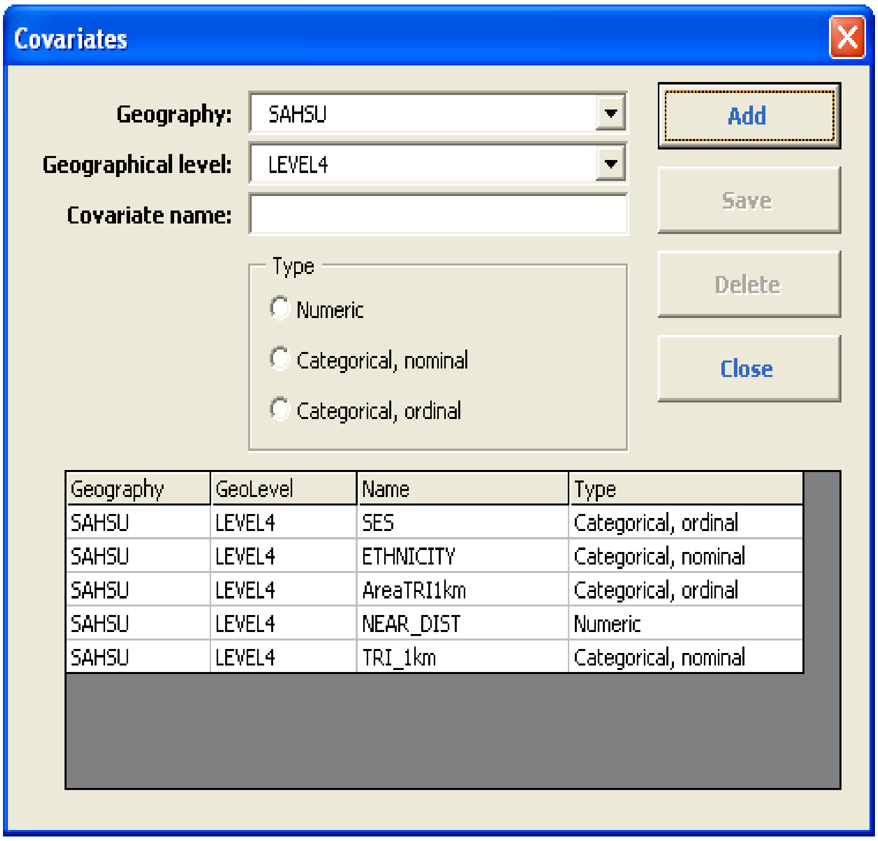
Y coordinate column: y

Figure 10. Geographical levels menu with centroid data

The details of the relationships between these different geographical levels, in terms of resolution i.e. which areas nest inside others, are specified by the geographical hierarchy table (section 5.3).

5.5 Covariates configuration menu

This menu allows the available covariates for a specified dataset to be defined.



Geography: SAHSU

Geographical level: LEVEL4

Covariate name:

Geography	GeoLevel	Name	Type
SAHSU	LEVEL4	SES	Categorical, ordinal
SAHSU	LEVEL4	ETHNICITY	Categorical, nominal
SAHSU	LEVEL4	AreaTRI1km	Categorical, ordinal
SAHSU	LEVEL4	NEAR_DIST	Numeric
SAHSU	LEVEL4	TRI_1km	Categorical, nominal

Figure 11. Covariates menu

Information that is required:

- *Geography*: use the drop down list to select the correct database for which you wish to enter/view the covariate data.
- *Geographical level*: the geographical data to which this covariate data relates (in terms of resolution/ level).
- *Covariate Name*: the name of the column that holds the covariate information. *Type*:
    - *Numeric* :The covariate data are in numeric (continuous)
    - *Categorical, nominal:* The covariate data is categorical data
    - *Categorical, ordinal:*

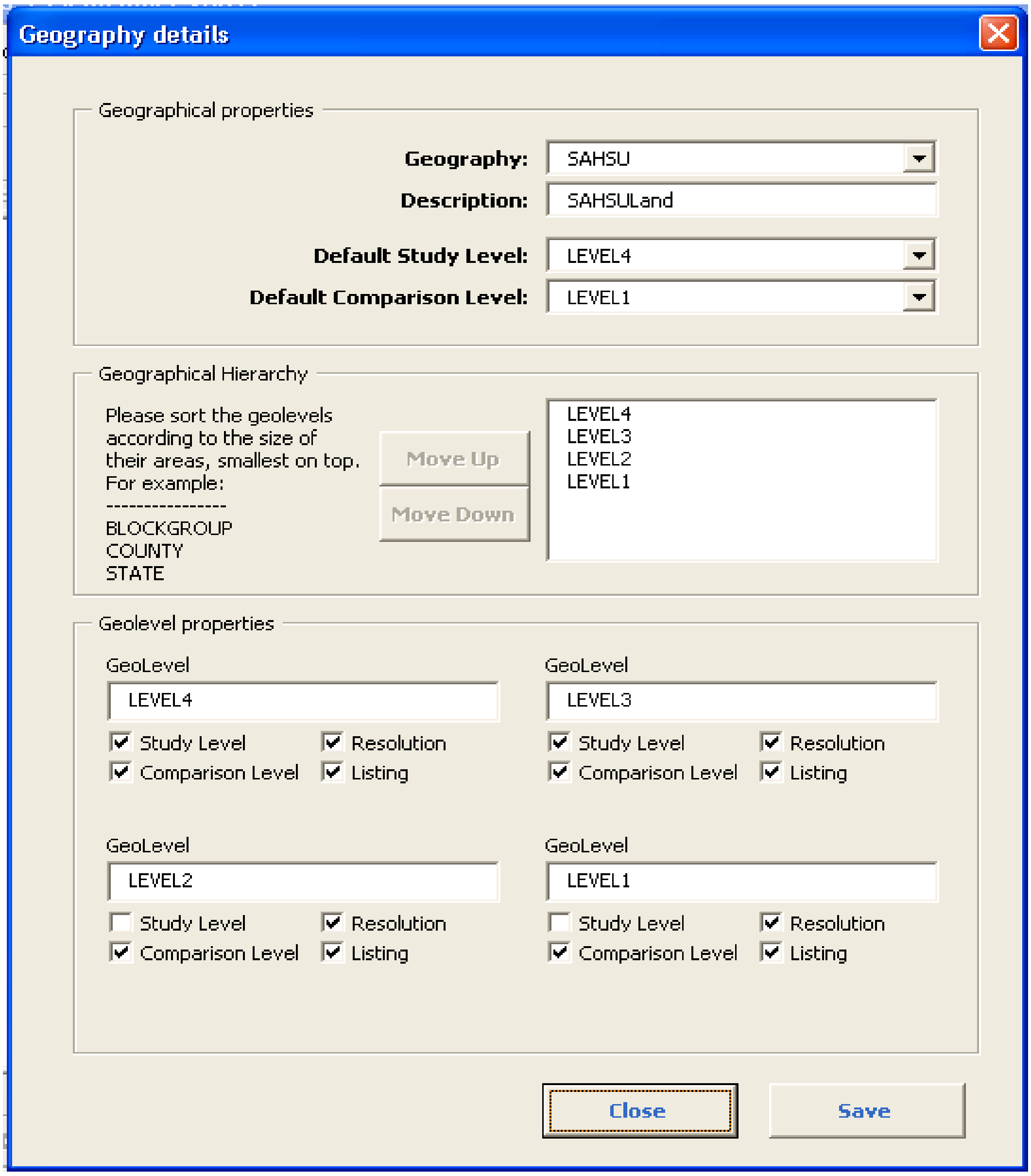
Covariates can be removed by clicking the row and selecting delete. This does not delete the data but means that (until added again) they can not be used in RIF studies.

Note. Although all types of covariate can be used, users should note that numeric covariates can **only** be used as exposure covariates and cannot be used for adjustment in studies.

5.6 Geography details

This menu must be used to define the details for the different levels of geographical boundary data that will be available for studies together with the specified dataset.

The properties for each of the geographical levels must be specified here. The geographical levels **must** have already been defined in the Geographical levels menu prior to defining the details for a geography using this menu (section 5.4). Although the different resolutions of geographical data will automatically populate this box, the list **must** be correctly ordered before this menu is closed to avoid errors in studies.



Geographical properties

Geography: SAHSU

Description: SAHSULand

Default Study Level: LEVEL4

Default Comparison Level: LEVEL1

Geographical Hierarchy

Please sort the geolevels according to the size of their areas, smallest on top.
For example:

BLOCKGROUP
COUNTY
STATE

LEVEL4
LEVEL3
LEVEL2
LEVEL1

Geolevel properties

GeoLevel: LEVEL4
☑ Study Level ☑ Resolution
☑ Comparison Level ☑ Listing

GeoLevel: LEVEL3
☑ Study Level ☑ Resolution
☑ Comparison Level ☑ Listing

GeoLevel: LEVEL2
☐ Study Level ☑ Resolution
☑ Comparison Level ☑ Listing

GeoLevel: LEVEL1
☐ Study Level ☑ Resolution
☑ Comparison Level ☑ Listing

Figure 12. Geography details menu

The features that should be defined here are:
Geographical properties:

- *Geography name*: Select the name of the geography to configure
- *Description*: Populated automatically. This is read only.
- *Default Study Level*: The level or resolution of the geographical data can be selected that will appear as the default setting. Any number of different geographical data resolutions can be used but the default setting will appear at the top of the list.
- *Default Comparison Level*: In the same way as for the study level, default

geographical resolution can be defined.

- *Geographical hierarchy*: If different resolutions of geographical adapt are to be used in the RIF, these datasets must be hierarchical. The order of this hierarchy **must** be defined by the user i.e. the order in which these datasets nest inside one another must be outlined. The highest resolution dataset (that with the smallest areas) should be placed at the top of the list and descend to the geographical level of lowest resolution (largest areas).

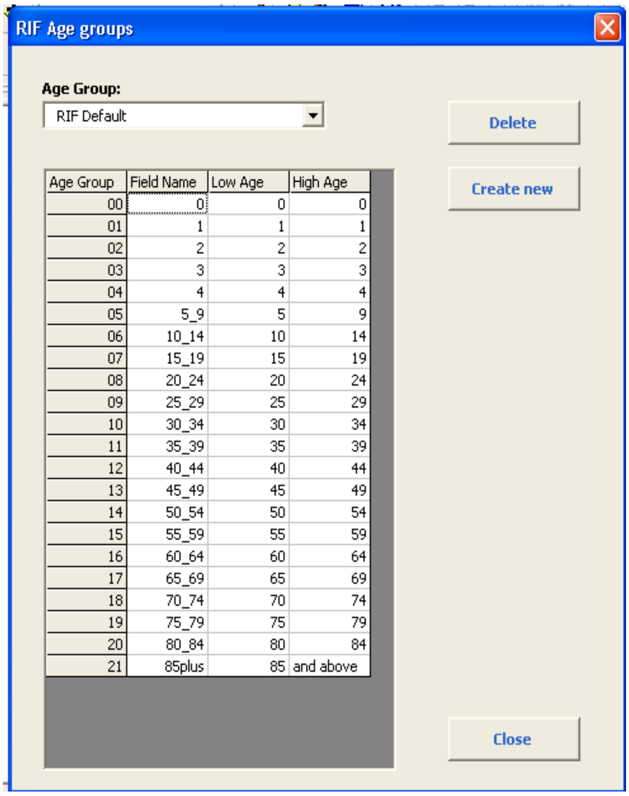
Geolevel properties:

Here details that relate to every geographical level of data should be entered. A box showing the geographic resolution plus a tick box for each geographical level that has been defined in the geography will be visible. These specify the following:

- *Study level*: select the tick box if the shown geographic data can be used as a study area.
- *Comparison level*: select the tick box if the specified geographical level should be available for use as comparison area data.
- *Resolution*: ticking this box allows users to select large areas for a disease mapping study using a simple method. The geographical unit can be used to select all lower level geographical units within the higher level unit. An associated boundary shape file must exist at this geographical resolution.
- *Listing*: ticking this box will mean that the values of the geographical unit can be listed in list boxes (used as a selection method).
  Note. for high resolution areas, where this list may contain thousands of elements, memory problems can occur when the RIF tries to populate the list. If a geographical level has more than 2000 elements, the listing capability should be deactivated (see section 5.8).

5.7 RIF age groups menu

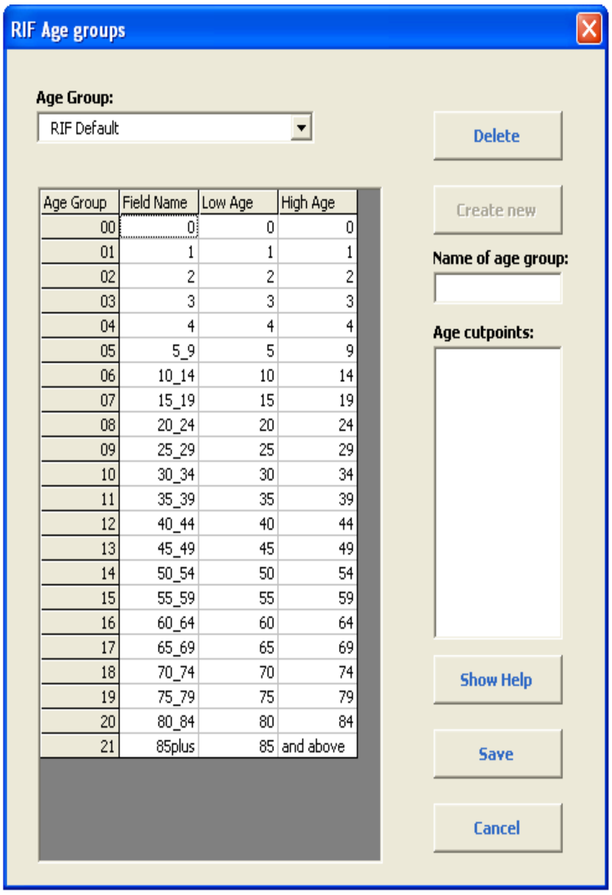
The user can specify the age groupings, depending upon the data and/or needs. Default age bands are provided but do not have to be used. To create a new age group, a name must be given to the age group and then cut points must be specified. Any age band is defined by the cut point which should be the upper value in a required range e.g. for age band 0-5 the cutpoint is 5. Any non-numeric character will be treated as a separator while whitespace will be ignored. Any number above 99 will be interpreted as an indefinite upper age limit. A maximum of one single age band is acceptable (e.g. for birth defects studies).



Age Group: RIF Default

Age Group	Field Name	Low Age	High Age
00	0	0	0
01	1	1	1
02	2	2	2
03	3	3	3
04	4	4	4
05	5_9	5	9
06	10_14	10	14
07	15_19	15	19
08	20_24	20	24
09	25_29	25	29
10	30_34	30	34
11	35_39	35	39
12	40_44	40	44
13	45_49	45	49
14	50_54	50	54
15	55_59	55	59
16	60_64	60	64
17	65_69	65	69
18	70_74	70	74
19	75_79	75	79
20	80_84	80	84
21	85plus	85	and above

Figure 13. RIF Age groups menu



Age Group: RIF Default

Age Group	Field Name	Low Age	High Age
00	0	0	0
01	1	1	1
02	2	2	2
03	3	3	3
04	4	4	4
05	5_9	5	9
06	10_14	10	14
07	15_19	15	19
08	20_24	20	24
09	25_29	25	29
10	30_34	30	34
11	35_39	35	39
12	40_44	40	44
13	45_49	45	49
14	50_54	50	54
15	55_59	55	59
16	60_64	60	64
17	65_69	65	69
18	70_74	70	74
19	75_79	75	79
20	80_84	80	84
21	85plus	85	and above

Figure 14. Create New Age group

5.8 RIF database tables

This menu is used to specify the data tables that should be used for RIF studies.

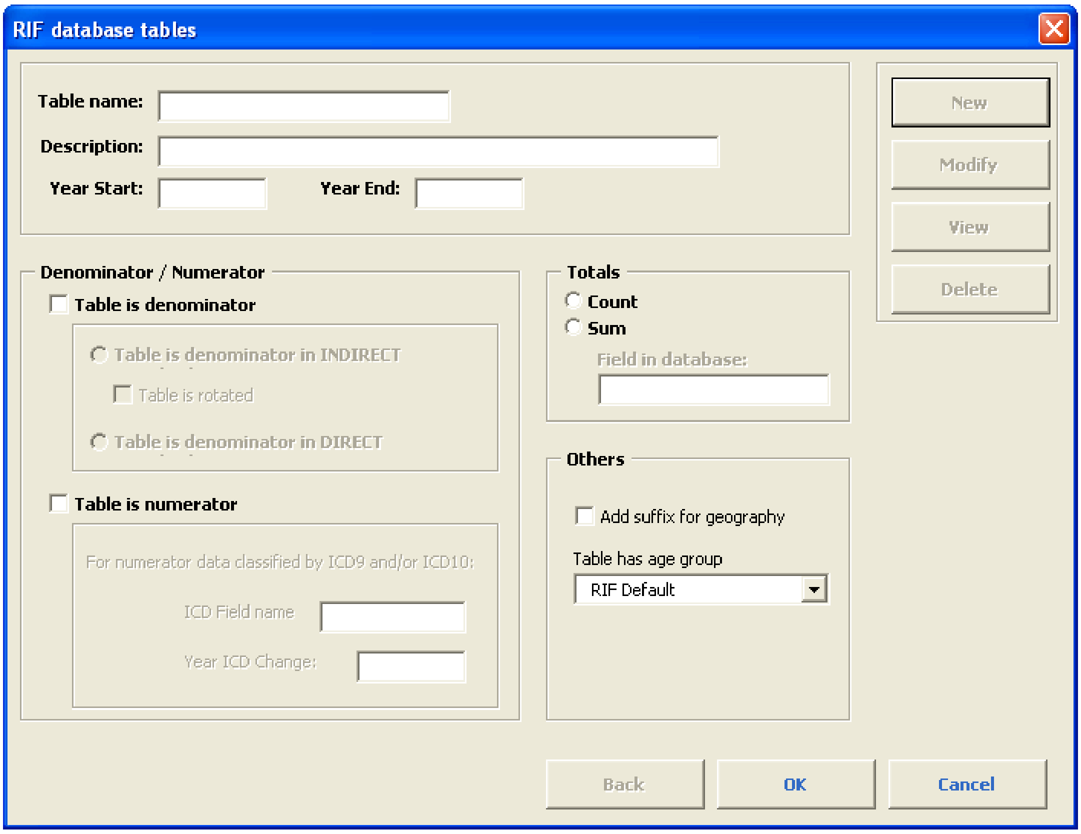


Figure 15. RIF study tables

A number of different features related to the data tables must be specified:

- *Table name*: name of the table in the database for which you are about to enter the details.
- *Description*: brief description about the table for information only.
- *Year start* and *Year end*: enter the range of years for which these data are available.
- *Table is denominator*: select if the table is to be used as denominator data.
  - *Table is denominator in INDIRECT* [Standardisation]: select if the table will be available to be used as comparison area data to calculate relative risks.
  - *Table is rotated*: the RIF uses age sex group values from only one column (non rotated). If the denominator table has one column per age sex group then check table is rotated and the RIF will automatically convert the data to a suitable data for runtime.
    Note. the data will remain in the original format in the database.
- *Table is numerator*: Select when the table can be used as numerator data.
  - *ICD Field Name*: If numerator data are coded using ICD9 and/or ICD10, the name of the ICD field in the numerator table should be added here. This is then the field that will be searched when using the **Pick ICDs** option to select ICD codes (see section 4.2.4).
  - *Year ICD change*: the World Health Organisation recommended that

the Tenth revision of the International Classification of Diseases (ICD10) should come into effect from the 1st January 1993, however users can specify the year in which this change happened in their country/State. This allows the RIF to indicate in the investigation details screen whether they should select ICD9, ICD10 or both depending on the range of years selected for the investigation.

- *Totals*: The data in the tables may need to be totalled during run-time. The RIF provides two different methods:
  - *Count*: count applies to individual level data, and involves 'counting' the number of records (rows) that match the conditions specified.
  - *Sum*: sum applies to aggregated data and involves summing the totals in a user specified field (column) in the box below (see section 4.3.3).
  - *Others*:
    - add suffix for geography: The RIF can take advantage of pre-grouped data stored in different tables for each geographical level to increase the speed of the calculations. The names of the tables will be identified as a common root name and a suffix depending on the geographical level to which the data has been grouped. The suffix for the geographical level will be automatically added if the checkbox "Add suffix for geography" is checked.
    - Table has age group: Select the age group to be used for the denominator/numerator table

Example: Using the Sahsuland example, we have four SAHSU_POP tables: _LEVEL4, _LEVEL3, _LEVEL2, and _LEVEL1 (Figure 16). The data is pre-aggregated at the corresponding levels. By activating the 'Add suffix for geography' the RIF will create the final table name combining the table name (in this case root name) and the geographical level used in each case. In Sahsuland the age sex groups are in 44 columns so the table is rotated.

There is one numerator table in the Sahsuland example; SAHSU_CANCER_LEVEL4 (Figure 16). The ICD column name is simply 'icd'; the year of ICD change is undefined. This table has the column 'total' to sum to give the number of cases.

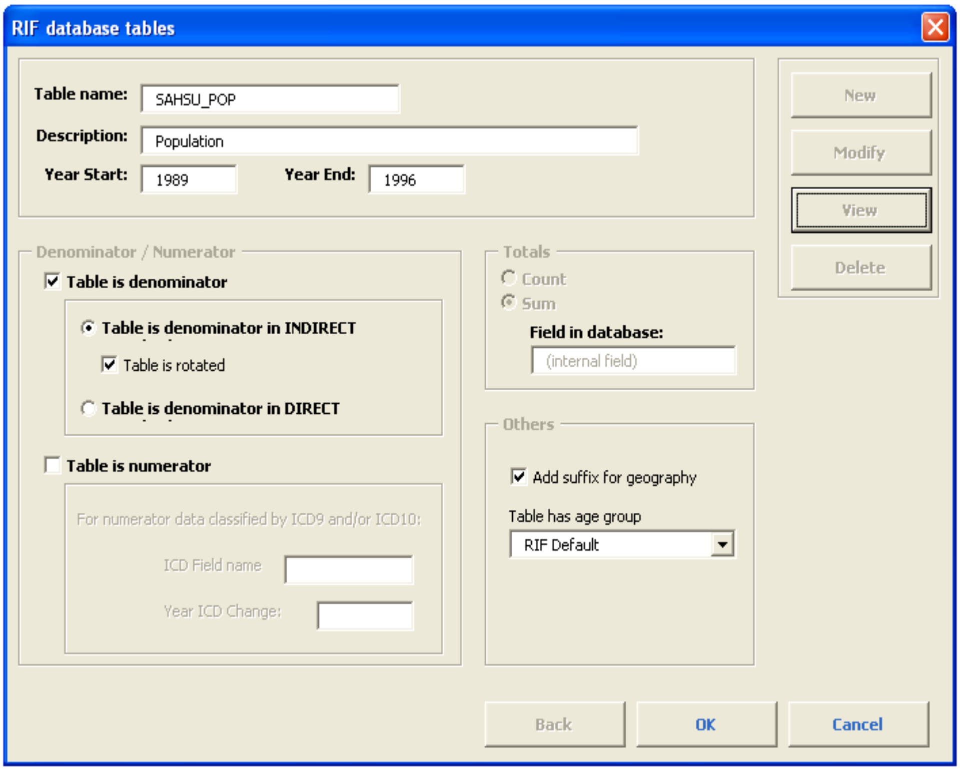


Table name: SAHSU_POP

Description: Population

Year Start: 1989 Year End: 1996

Denominator / Numerator

☑ Table is denominator

● Table is denominator in INDIRECT

☑ Table is rotated

○ Table is denominator in DIRECT

☐ Table is numerator

For numerator data classified by ICD9 and/or ICD10:

ICD Field name

Year ICD Change:

Totals

○ Count

● Sum

Field in database: (internal field)

Others

☑ Add suffix for geography

Table has age group: RIF Default

Figure 16. SAHSU_POP

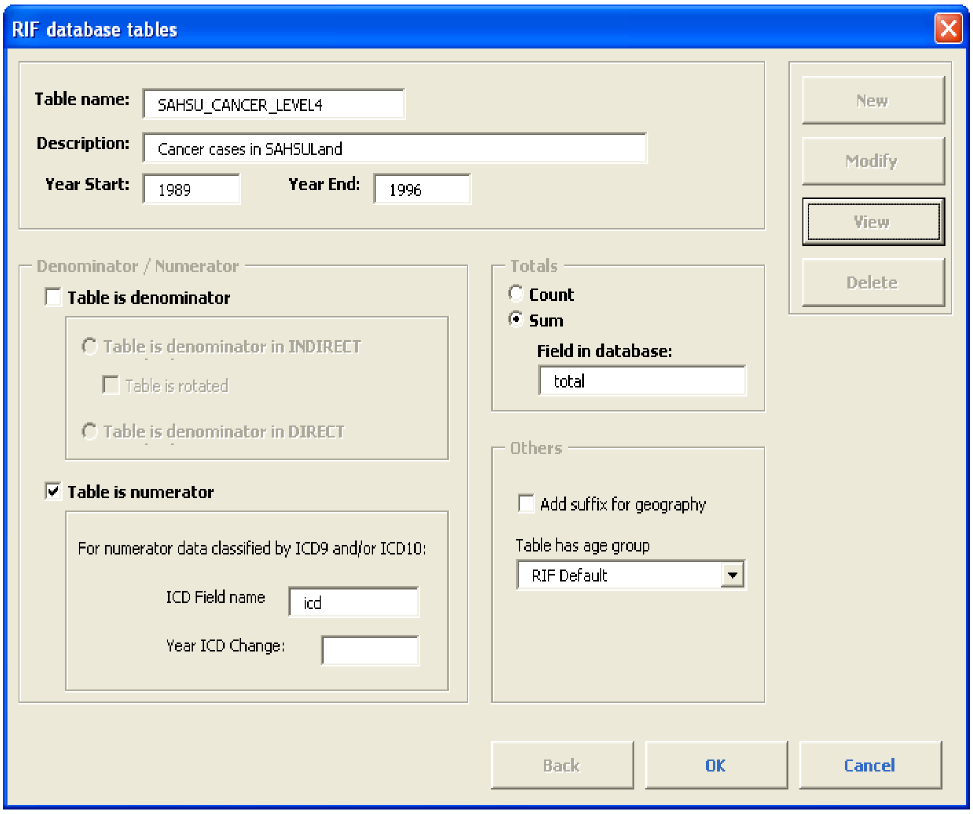


Table name: SAHSU_CANCER_LEVEL4

Description: Cancer cases in SAHSULand

Year Start: 1989 Year End: 1996

Denominator / Numerator

☐ Table is denominator

○ Table is denominator in INDIRECT

☐ Table is rotated

○ Table is denominator in DIRECT

☑ Table is numerator

For numerator data classified by ICD9 and/or ICD10:

ICD Field name: icd

Year ICD Change:

Totals

○ Count

● Sum

Field in database: total

Others

☐ Add suffix for geography

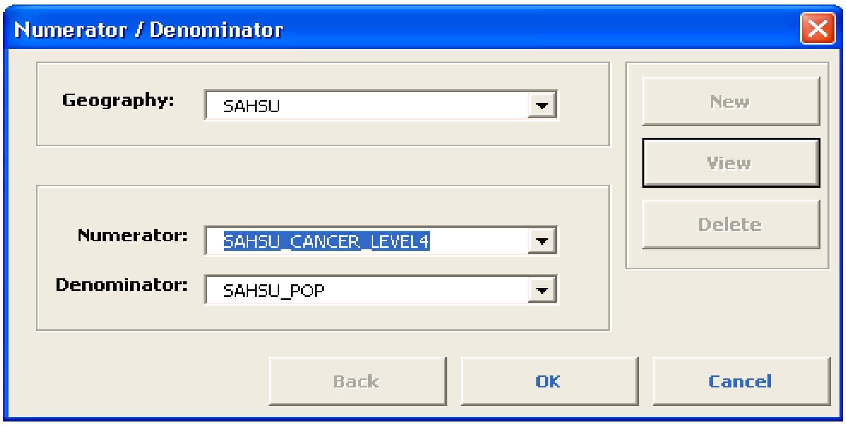
Table has age group: RIF Default

Figure 17. SAHSU_CANCER_LEVEL4

5.9 Numerator/Denominator configuration menu

This menu is used to define the relationship between the numerator and denominator data. The two tables **must** have already been setup as RIF database tables. These two tables must have the same age group classification.

Example: The Sahsuland example the pairing is between SAHSU_CANCER_LEVEL4 as the numerator and SAHSU_POP as the denominator data.



Geography: SAHSU

Numerator: SAHSU_CANCER_LEVEL4

Denominator: SAHSU_POP

Figure 18. Numerator/denominator tables

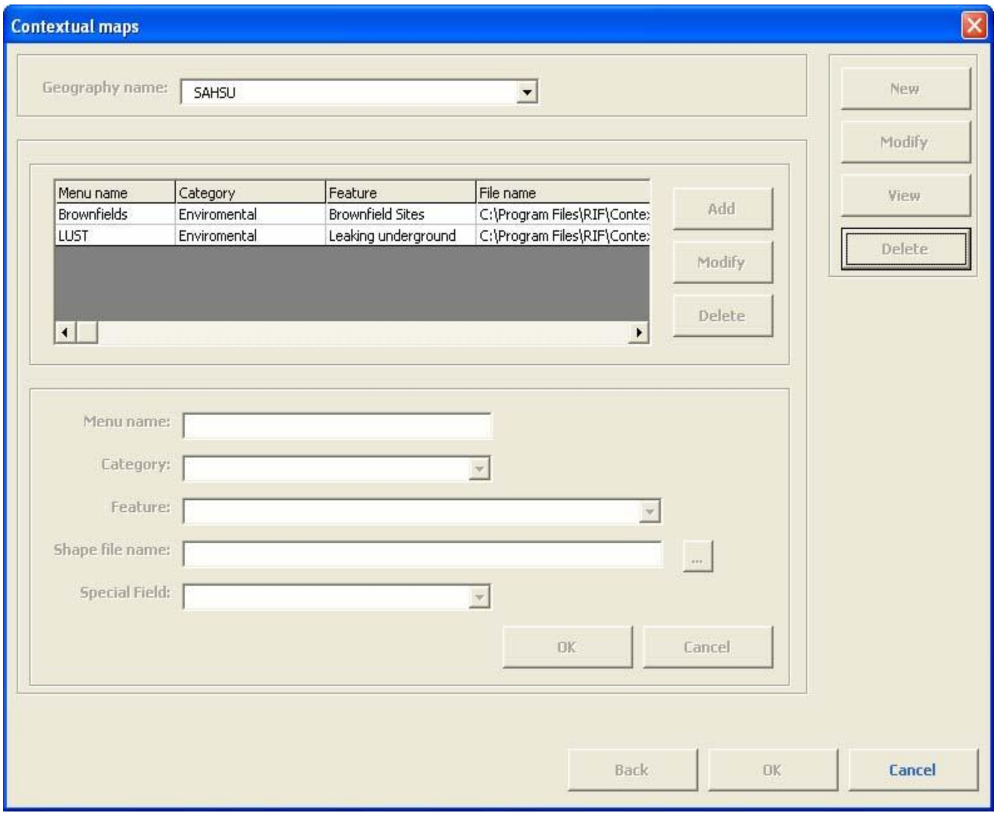
It is possible to define any number of pairs of numerator /denominator data for any chosen geography. These pairs will be available for use with any RIF study with that geography.

5.10 Contextual maps

Additional layers can be added after any RIF study, for example as a possible aid to interpretation of results.

The information that can be entered here are:

- *Geography name*: select, from the drop down menu, the geography to be associated with the contextual map(s)
- *Menu name*: select or enter a name that will appear in a drop-down menu
- *Category*: select a group for the contextual map. Different contextual map menus will use the same root menu name, e.g. environmental, utilities, etc.
- *Feature*: select or enter the feature type. A number of special symbols have been pre-programmed into the RIF and will use the internationally recognised symbols where they exist.
- *Shapefile name*: the path and name of the shape file containing the contextual information must be entered.
- *Special field*: area data can be displayed as categorical data, however the name of the field upon which that data should be categorised must be entered. This functionality has been included to allow the display of data such as population migration.



Geography name: SAHSU

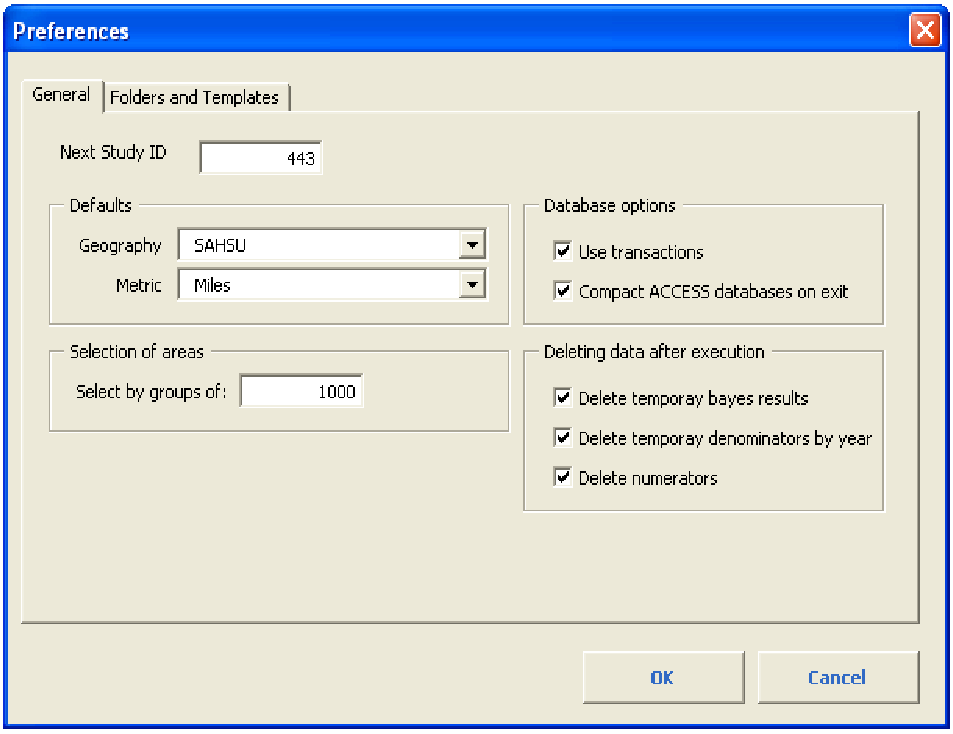
Menu name	Category	Feature	File name
Brownfields	Enviromental	Brownfield Sites	C:\Program Files\RIF\Conte
LUST	Enviromental	Leaking underground	C:\Program Files\RIF\Conte

Figure 19. Contextual maps

Example: In the Sahsuland example (Figure 19) we have two contextual maps: Brownfield sites and LUST (Leaking Underground Storage Tanks).

5.11 Preferences

This menu can be used to define some of the default settings and help to make the RIF more user friendly. They specify a number of features including the locations of files. There are two tabs, General and Folders and Templates.



Next Study ID: 443

Defaults
Geography: SAHSU
Metric: Miles

Database options
☑ Use transactions
☑ Compact ACCESS databases on exit

Selection of areas
Select by groups of: 1000

Deleting data after execution
☑ Delete temporay bayes results
☑ Delete temporay denominators by year
☑ Delete numerators

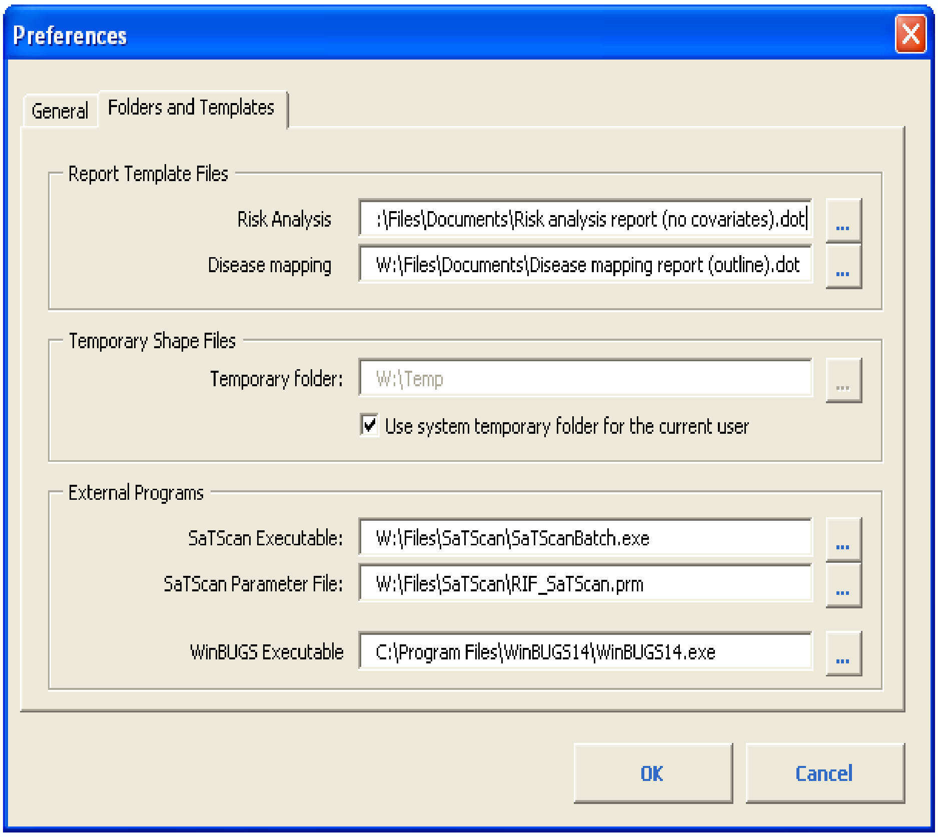
Figure 20. Preferences: General Tab

General:

- *Next study ID*: This shows the next number that will be used as the study ID but it can not be modified from this menu. It is only updatable directly in the table RIF_PARAMETERS in the database, for the entry 'nextstudyid'. This number auto increases for each new study.
- *Defaults:*
  - Geography: The default geography for RIF studies
  - Metric: the default measurement scale for RIF studies (Miles/Kms)
- *Selection of areas:*
  - *Select by groups of*: some of the GIS operations have to select areas from a shape file. Processing speed can be increased by selecting the areas in groups (bulks). Depending on the structure of the shape file (number of polygons, indexes, etc) and on the resources of the system (CPU, memory, disk space, etc) the performance of area selection can vary. Adjusting this value can help achieve better performance.
    Small values (less than 20) will make the selection of areas very slow, especially when the shape file contains a large number of areas. Large numbers may make the system run out of memory. This value should be estimated based on the most common sized selections.
    In our example the number selected is 1000 since studies done with Sahsuland LEVEL4 data means that thousands of LEVEL4 areas can be selected. Reducing this value may help, however, when running the RIF on a less powerful machine.
- *Database options*:
  - Use transactions: this allows users to activate or deactivate the use of database transactions when running a study. The default state is activated which ensures that if an error occurs all the operations carried out in the database will be rolled back to the starting state

    keeping the database clean of intermediate results. It could be activated for use when, for example, a system would not have sufficient memory.

    - *Compact ACCESS databases on exit*: Tick or untick the checkbox accordingly if you want the RIF to automatically compact the database or not. Access databases need to be compacted on a regular basis. The RIF may do this automatically as soon as the last user logs out from the database. Compacting is optional but highly recommended. This checkbox is greyed out when using Oracle databases.
- *Deleting data after execution*:
    - *Delete temporary Bayes results*: if this is activated data is deleted from a table used to perform Bayes smoothing. By default this button is selected, however, it can be deactivated when a check on intermediate results by database experts and users is required.
    - *Delete temporary denominators by years*: The RIF uses an intermediate table to retrieve data of denominators by years that is later aggregated. The default state is activated so that this temporary table is deleted. It can be activated to check intermediate results.
    - *Delete numerators*: The numerator data stored in the temporary numerator tables are not required after the execution of a study because the relevant data are in results tables, therefore, these tables are deleted by default. It can be activated to check intermediate results.



Report Template Files

Risk Analysis: :\Files\Documents\Risk analysis report (no covariates).dot

Disease mapping: W:\Files\Documents\Disease mapping report (outline).dot

Temporary Shape Files

Temporary Folder: W:\Temp

☑ Use system temporary folder for the current user

External Programs

SaTScan Executable: W:\Files\SaTScan\SaTScanBatch.exe

SaTScan Parameter File: W:\Files\SaTScan\RIF_SaTScan.prm

WinBUGS Executable: C:\Program Files\WinBUGS14\WinBUGS14.exe

Figure 21. Preferences: Folders and Templates

Folders and Templates:

- *Report Template Files:*
    - *Risk Analysis*: Users should specify the Microsoft Word template to be used to create any Risk Analysis study report. Two report templates are supplied and automatically installed with the RIF in the folder ‘Documents’ (see section 7.6.2 for more details).
    - *Disease Mapping*: Users should specify the Microsoft Word template to be used with any Disease Mapping study report. As with risk analysis, a file is automatically installed with the RIF in the folder ‘Documents’ (see section 6.6.2 for more details).
- *Temporary Shapefiles:*
    - *Temporary folder*: specifies the folder used for temporary files. **Note.** It is not required if the ‘system temporary folder for the current user’ is selected.
    - *Use system temporary folder for the current user*: Selects the temporary folder to be the folder that the operating system provides for the current user.
- *External Programs:*
    - *SaTScan Executable:* Full pathname for SaTScan executable batch program (SaTScanBatch.exe). A license must be obtained before using SaTScan from www.**satscan**.org/
    - *SaTScan Parameter File:* the name of the parameter file required by SaTScan to run the model.
    - *WinBUGS Executable:* full pathname for the WinBUGS executable file. A license must be obtained from http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml before WinBUGS can be used.

6. Running a new RIF study: Risk analysis

There are two types of study that the RIF can undertake, risk analysis and disease mapping (for details on the differences between these types of study, see section 2.1).

To undertake risk analysis, four steps will need to be completed:

1. Defining the study details such as ‘geography’ (in terms of the database to use) and study type.
2. Defining the study area.
3. Defining the comparison area.
4. Defining the health outcomes of interest.

This chapter will detail how to progress through each of these steps.

6.1 Study details

Start up the RIF (see section 3.3 Introduction to the RIF). Select the **Study** submenu, and then **New** to bring up the **Study Details** screen.

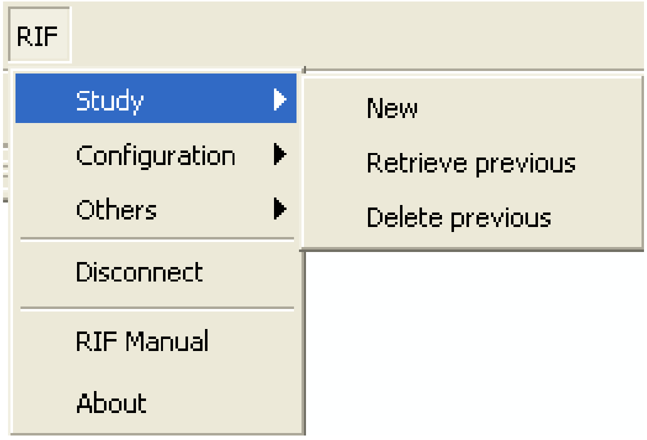
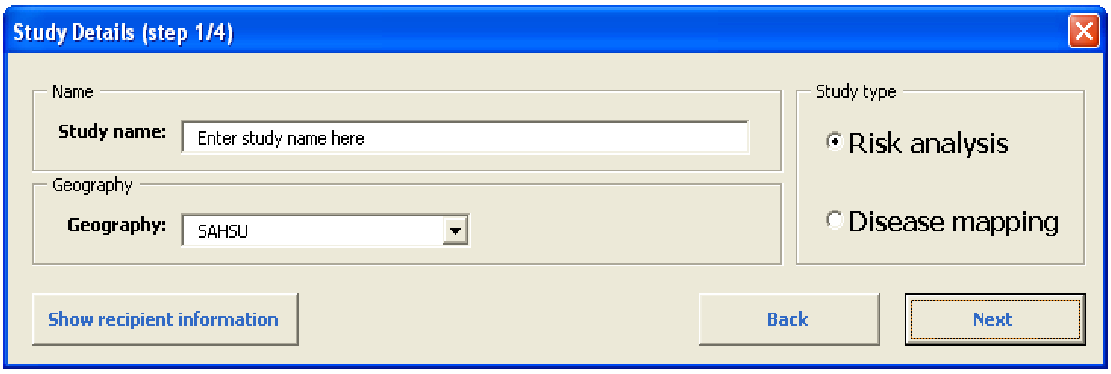


Figure 22. RIF study menu

The **Study Details** screen is used to enter a suitable study name, select the relevant geography, and add details of the recipient of the study report (these details are automatically entered into the study report created by RIF once the study has been run). Each study is automatically assigned a unique study ID (a numeric code), but a user defined study name giving a brief description of the investigation is useful if the study is likely to need to be retrieved after completion.



Study Details (step 1/4)

Study name: Enter study name here

Geography: SAHSU

Study type: ● Risk analysis ○ Disease mapping

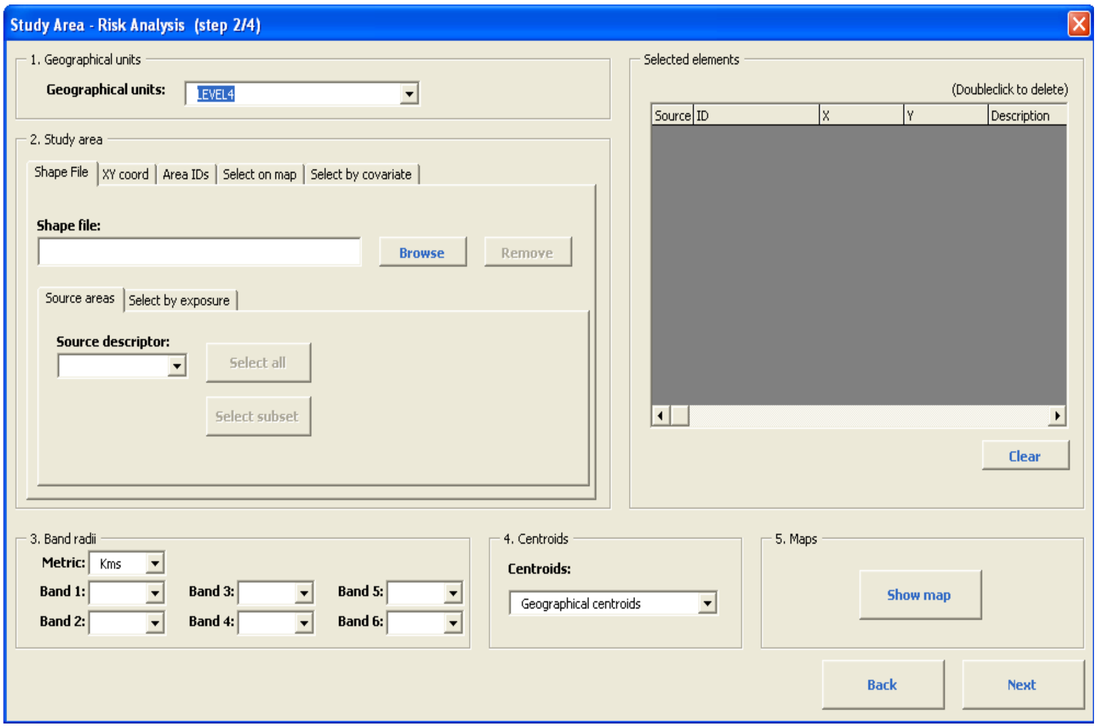
Figure 23. Study details screen

The **(step 1/4)** indicates that the **Study Details** screen is the first step out of four steps needed to define a RIF query.

- Add a suitable study name, select the ‘geography’ of interest, select the study type (**Risk Analysis** or **Disease Mapping**; see chapter 2) and then ‘Next’.
- To cancel the study, click **Back**.
- Additional recipient information can be added using the ‘*show recipient information*’, when the menu will automatically expand revealing text boxes where recipient details such as institution name, address etc can be added. These details are automatically added to the reports. To close this drop down section click **‘Hide recipient information’**.

6.2 Study Area - Risk Analysis

By ticking **Risk Analysis** at the **Study Details** screen and clicking **Next**, the **Study Area – Risk Analysis** screen is opened (step 2/4).



Study Area - Risk Analysis (step 2/4)

Geographical units: LEVEL4

Metric: Kms

Centroids: Geographical centroids

Figure 24. Study Area – Disease mapping screen

There are five sections to this screen:
1. Geographical units: These units refer to the geographical resolution used to select the study area. The drop down box indicates what levels of geography are available.

2. Study area: refers to the area to be investigated
There are several methods that can be used to select the **Study area**:

- *Shape file*: allows a shapefile to be imported to define the study area. The appropriate file can be browsed for by clicking the **Browse** button. All the centroids of the areas in the underlying geography (the chosen geographical units) that intersect (i.e. fall in the same geographical area as) the input shapefile are used as **Selected areas** in the Study area. Centroids can be geographic centroids or population-weighted centroids (selected in section 4 of the menu).
    - *Source areas*: If the shape file contains several fields you can optionally select a **Source descriptor** using the drop down menu (this information then appears in the selected elements box).

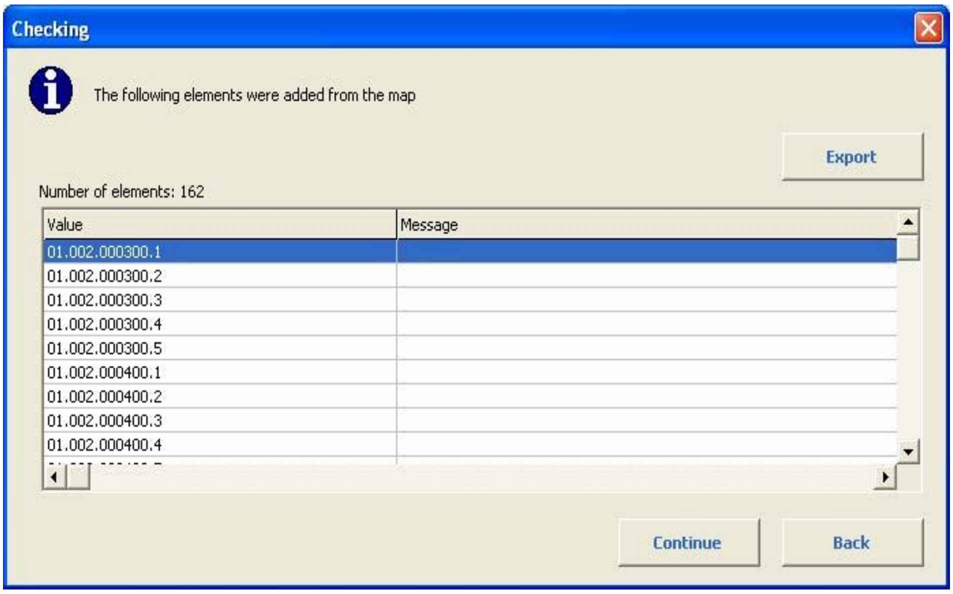
      The user must select **'Select all'** or **'Select subset'** to add areas from the chosen shapefile. Where the shape file contains several elements (e.g. several contours of exposure, or comprises more than one plume), click **Select all** to use all the elements from the shape file, or to choose specific elements from a map **Select subset** (from the map, to go back to the **Study Area - Risk Analysis** screen, go to the **RIF** menu, and select **Go back to last dialog**).
    - *Select by exposure*: a shapefile can be imported and the study area selected using attribute data for example, to define the various exposure bands. First, select the field that will be used to classify the area, then click **'Classify'**. Select the required classification scheme (manual, equal interval, defined interval, quantile, natural breaks (Jenks 1957), geometric interval (depending on ArcGIS version), standard deviation) and the number of classes, click **'OK'**. If any bands are not to be included, these can be removed by double clicking on the **'Selected elements'** list.
- *XY coord*: allows the x and y coordinate (and optionally a description) for one or several putative point sources to be added. This can be done by either typing in these details (using the sub-tab **Type in**); by loading a text file containing these details (using the sub-tab **Text File**); or by entering an SQL statement (using the sub-tab **SQL**).
- *Area IDs*: The first check box allows users to use the x and y coordinate that they may have in their database to define a point source in a known area. This option will only be available where x, y data has been set up on configuration (section 5.4). The first sub-tab (**Type in)**, allows the study area to be defined by entering area identifiers, or with the second sub-tab (**Text File**), load a text file containing these IDs, double click areas from a list (using the sub-tab **List**), or enter an SQL statement (using the sub-tab **SQL**).
- *Select on map*: To add a point to the map or multiple points, select the **Points** option, and click **Add**. Add a putative point source(s) to the map simply by clicking on the map in the appropriate place.

To zoom in, click the icon on the toolbar, and click on the map, or drag over the part of the map of interest. After using the zoom icon, re-activate the point source button by clicking the icon .

By clicking on the map, a small red square will be added to the map representing the location of the point source. To go back to the **Study Area Risk Analysis** screen, go to the **RIF** menu, and select '**Go back to last dialog'**.

To add an area to the map, select the **Area** option, and click **Add**. Click on the area of interest on the map, or hold down shift/drag the cursor to select several areas on the map. As with the points option, zoom in using the zoom icon; to re-activate the select features button click the icon . To go back to the **Study Area > Risk Analysis** screen, go to the **RIF** menu, and select '**Go back to last dialog'**.

Selected areas from the map appear in the pop-up window '**Checking'**, which provides information on the elements selected from the map.



The following elements were added from the map

Number of elements: 162

Value	Message
01.002.000300.1	
01.002.000300.2	
01.002.000300.3	
01.002.000300.4	
01.002.000300.5	
01.002.000400.1	
01.002.000400.2	
01.002.000400.3	
01.002.000400.4	

Figure 25. Checking screen with selected areas

These elements can be exported (so the same list can be imported into the RIF for subsequent studies, or used in other programs) by clicking the **Export** button. The exported elements are sent to the clipboard, and can be pasted into a notepad, excel or other file and then saved as appropriate (**Note**, the exported list has 'Tab' as the column delimiter).

Use the **Continue** or **Back** buttons to return to the **Study Area – Risk analysis** screen, and progress with the study.

- *Select by covariate*: allows the user to select the study populations on the basis of an 'exposure covariate'. Any covariate that has been setup in the covariate table in the database can be used to define a study area.

  Configured covariates will be listed in the **Select covariate** drop down menu with their accompanying min and max values. The exposure covariate and min/max values can be altered, and the covariate then selected by clicking **Add**. Numeric covariates can be n -tilised in the RIF (into 2 to 7 bands), or user specified cut-points can be chosen. For more details see section 4.2.2 and appendix B.3.

Which ever method is used to define the study area, the selected areas will be shown in the **Selected elements** part of the **Study Area - Risk Analysis** screen. If these elements are incorrect, individual elements can be deleted by double clicking; alternatively the whole list can be deleted by clicking the '**Clear'** button.

3. Band radii/Covariate cutpoints: Having selected the study area and the risk exposure sources, the area of potential exposure must be specified.

- When putative point source(s) have been used (using XY coordinates or by selecting a point (or points) from the map) or a shapefile has been imported, then distance bands must be used to specify the potential exposure area. **Band radii** are used for this. The appropriate distance **Metric** (Kms or Miles), and up to six bands of potential exposure or risk around the source can be selected using the drop down menus.

  Note: If a shapefile was imported that contains a large number of features (e.g. 10,000) an 'overflow error' may be encountered since the RIF must buffer all these features (although these buffers may subsequently be merged).

- Where the study area has been selected using a covariate, **Covariate cutpoints** will need to be defined. Similarly to the band radii, up to six covariate cutpoints can be selected from the drop down menus, or for continuous variables by typing in the desired value.

If the study area has been selected using area IDs or by selecting areas from the map, band radii are not defined since 'areas' of potential exposure have already been specified.

4. Centroids: the drop down menu allows the selection of either geographical or population-weighted centroids.

- The geographical centroids refer to the geographical mid-point of the area and are automatically calculated.
- The weighted centroid takes into account the population distribution within the small area, and so represents the location of the majority of the population (for more details see appendix B.2 Centroids). This is the recommended method where possible.

5. Maps: to view a map of the study area/s (i.e. the small areas with geographical or weighted centroids falling within the selected radii) click on the **Show map** button.

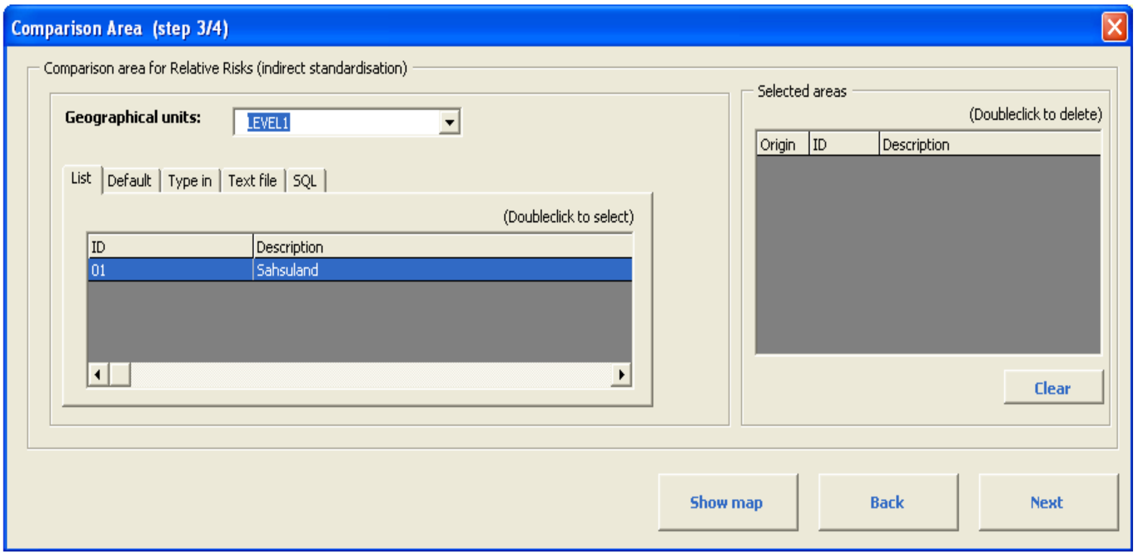
- If the study area has been selected using a shape file, XY coordinates or by selecting a point(s) from the map, the base map should show the putative point source(s)/output from the shape file as grey dots/areas. Around these sources will be grey lines indicating the selected radii. The shaded areas indicate the study area(s) that fall within the selected radii (i.e. those study areas with geographical or weighted centroids within each radii).

- If the study area has been selected using area IDs, by selecting areas from the map, or by selecting using a covariate, then the base map will show these selected areas using a greyscale colour ramp.

  To go back to the **Study Area - Risk Analysis** screen, go to the **RIF** menu, and select '**Go back to last dialog'**. To move to the next step, click **Next**.

6.3 Comparison Area

Step 3/4, like the previous step, requires the selection of a geographical area. The **Comparison Area** screen is used to define the reference area (population) used for the calculation of indirectly standardised risks (for more details of these calculations, see section B.1 Statistics).



Geographical units: LEVEL1

ID	Description
01	Sahsuland

Origin	ID	Description

Figure 26. Comparison Area screen

Comparison area for Relative Risks (indirect standardisation): this is the reference area from which the disease rates are calculated for application to the study area and calculation of expected counts and relative risks.

There are several methods that can be used to select the **Comparison area**, and these different methods can be used by selecting the appropriate tab:

- **List:** A list of all areas is provided in the tab **List**, and elements can be selected by double clicking from this list.
- **Default/Internal:** select the geographical areas of interest from the drop down menu, and then click the **Add** button. The default selection will add any areas at the selected resolution that contain all or part of the study area.
- **Type in:** The ID of the area or areas of interest can be typed in; click **Add** to create the study area.
- **Text file:** A text file created in another program or exported from a previous RIF study can be imported by browsing for the file containing the relevant comparison area IDs, by specifying the column delimiter (e.g. tab, comma), and by clicking **Load**.
- **SQL:** An SQL statement can be typed in, for example:

```
select zone_id from Sahsu_level2 where zone_id = '01.002'
```

SQL statements should only be used by those with extensive knowledge of their database. For more details on SQL statements see Appendix B4.

The selected areas are automatically shown in the **Selected areas** part of the **Comparison Area** screen. If these elements are incorrect, individual elements can be deleted by double clicking; alternatively the whole list can be deleted by clicking the **Clear** button.

To review the selected comparison area, click the '**Show map'** button.

The map showing the selected comparison area appears as a smaller pop-up box on top of the base map showing the study area.

To go back to the **Comparison Area** menu, go to the **RIF** menu, and select '**Go back to last dialog'**.

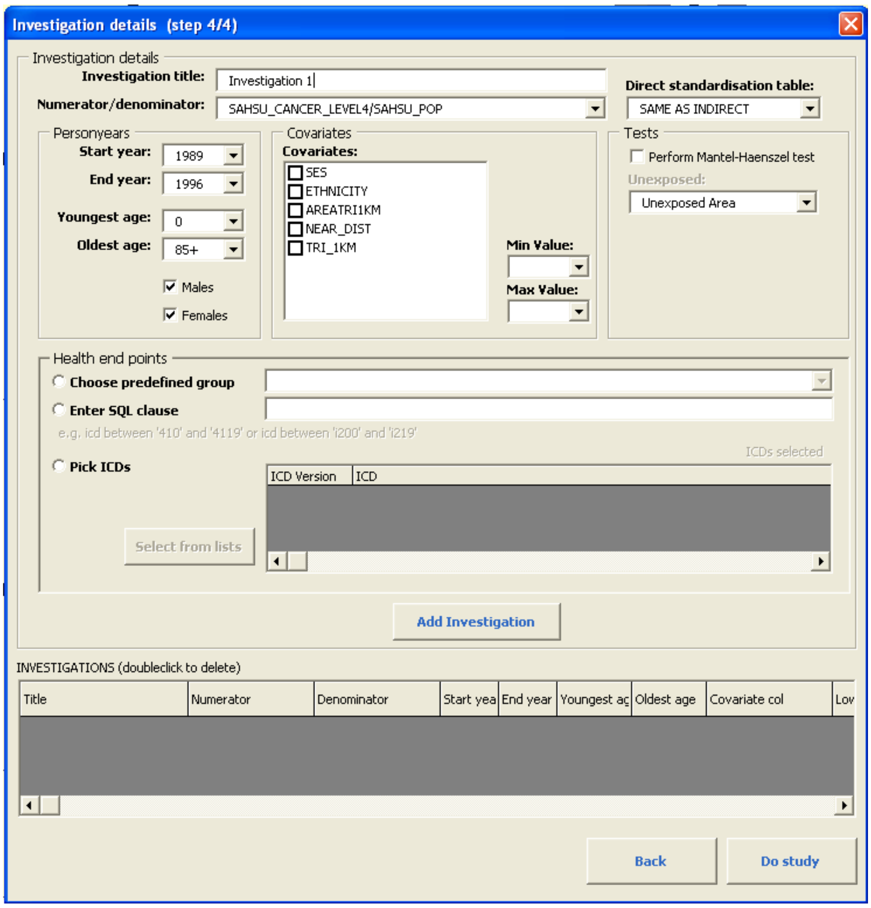
To move to the next step, click **Next**; to go back to the previous screen **Study Area – Risk Analysis**, click **Back**.

6.4 Investigation details

Step 4/4 allows the health outcomes of interest to be defined.

First, the investigation details must be specified:

- For each outcome to be studied, a unique **Investigation title** is required.
- The **Numerator/denominator** drop down box is where the health database (numerator) and population (denominator) *pairing* to be investigated is selected. This would normally take the form 'cancer'/'population' or 'congenital anomalies'/'births' etc, depending on what numerator/denominators pairings have been defined (see section 5.9).
- **Direct standardisation table**: select from the drop down list any predefined table that specifies the area to which disease rates from the study area are applied to give the directly standardised rates. Three tables are supplied with the RIF (World standard population, USA standard population, Europe standard population) but users can import and use their preferred population standard.
- The **Start year** and **End year** and **Youngest age group** and **Oldest age group** are numerator/denominator specific, and need to have been defined for each numerator/denominator pair when these were added to the RIF (see section 5.4). By changing the values of these parameters using the drop down menus, it is possible to customise the investigation to a specific range of years and age groups.


Investigation details (step 4/4)

Investigation details

Investigation title: Investigation 1

Numerator/denominator: SAHSU_CANCER_LEVEL4/SAHSU_POP

Direct standardisation table: SAME AS INDIRECT

Personyears

Start year: 1989

End year: 1996

Youngest age: 0

Oldest age: 85+

☑ Males

☑ Females

Covariates

Covariates:

□ SES
□ ETHNICITY
□ AREATRI1KM
□ NEAR_DIST
□ TRI_1KM

Min Value:

Max Value:

Tests

□ Perform Mantel-Haenszel test

Unexposed: Unexposed Area

Health end points

○ Choose predefined group

○ Enter SQL clause

e.g. icd between '410' and '4119' or icd between 'i200' and 'i219'

○ Pick ICDs

ICDs selected

ICD Version	ICD

INVESTIGATIONS (doubleclick to delete)

Title	Numerator	Denominator	Start yea	End year	Youngest ag	Oldest age	Covariate col	Lov

Figure 27. Investigation details screen

- The **Covariates** should already have been defined for the dataset (see section 5.5). By selecting a covariate the rates and risks will be generated with and without adjustment for these covariates. Selecting more than one covariate (e.g. SES and ethnicity) will generate a results table with both the age and gender only adjusted rates/risks, as well as the age, gender, SES and ethnicity adjusted rates/risks. The drop down menus can be used to customise the investigation to look at a subset of the population (e.g. only the most affluent) by selecting min or max values from the drop down menus.

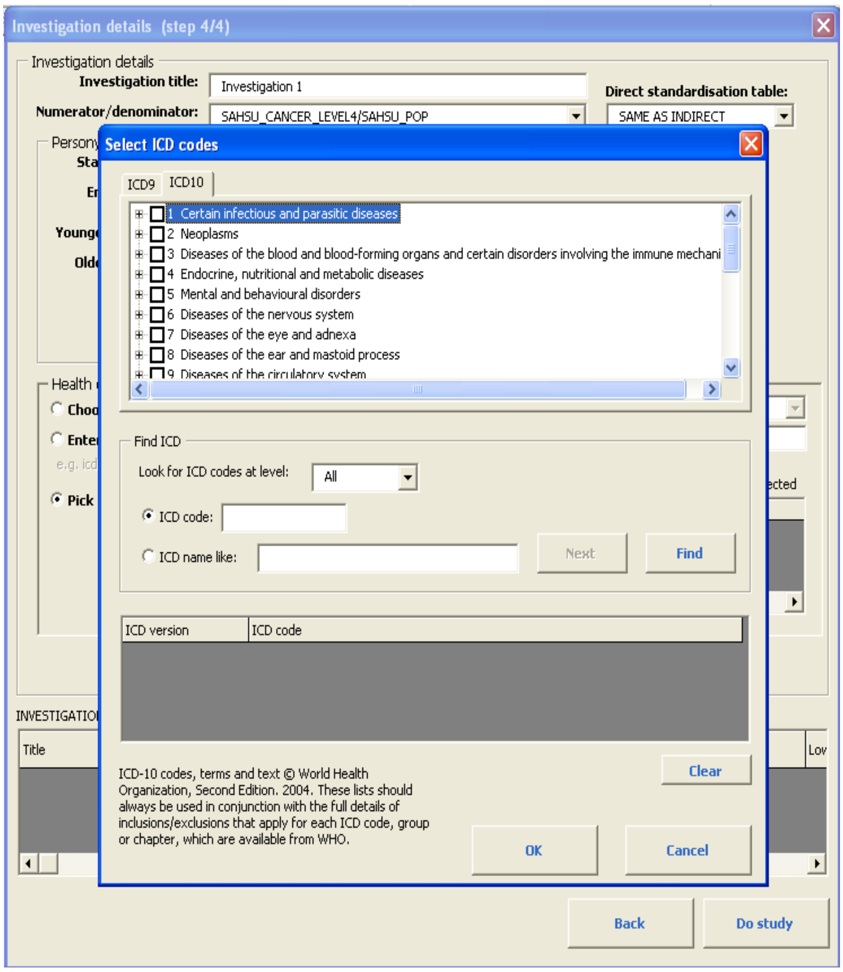
  Note. covariate data can only be used where it is at the same geographical resolution as the study area (for more details see technical appendix B.3). Where the study area has been selected using a covariate, the exposure covariate used will no longer appear as an adjustable covariate in the **Covariate** section of the **Investigation details** screen.

The sub-section **Health end points** is where the specific health end points of interest are selected. There are several ways of selecting the health outcomes of interest.

- *Choose predefined group:* Users can develop a list of frequently used end points by adding the appropriate SQL clause and description to the table RIF_PREDEFINED_GROUPS (supplied as part of the ‘rif_empty.mdb’).

Elements stored in this table can then be selected as the end point of interest from a drop-down list.

- *Enter SQL clause:* This option allows health end points to be selected using an SQL clause (for more details see appendix B.3). This option also allows very specific end points to be returned, such as renal disease in diabetics, or a specific cancer type by site (provided this level of detail is present in the database itself).
- *Pick ICDs:* Users can, if their health data are coded in revision 9 and/or 10 of the International Classification of Diseases (ICD), pick ICD codes from a pop-up box listing the ICD9 and ICD10 codes in expandable menus. Selecting the Pick ICDs option activates the **Select from lists** button. This button opens up a pop-up box listing the ICD9 and ICD10 codes in expandable menus, allowing the selection of any chapter, group, or specific 3 or 4 digit code simply by clicking in the box adjacent to the outcome of interest.



Investigation details (step 4/4)

Investigation title: Investigation 1

Numerator/denominator: SAHSU_CANCER_LEVEL4/SAHSU_POP

Direct standardisation table: SAME AS INDIRECT

Select ICD codes

ICD9 | ICD10

- ⊞ □ 1 Certain infectious and parasitic diseases
- ⊞ □ 2 Neoplasms
- ⊞ □ 3 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechani
- ⊞ □ 4 Endocrine, nutritional and metabolic diseases
- ⊞ □ 5 Mental and behavioural disorders
- ⊞ □ 6 Diseases of the nervous system
- ⊞ □ 7 Diseases of the eye and adnexa
- ⊞ □ 8 Diseases of the ear and mastoid process
- ⊞ □ 9 Diseases of the circulatory system

Find ICD

Look for ICD codes at level: All

● ICD code:

○ ICD name like:

ICD version	ICD code

ICD-10 codes, terms and text © World Health Organization, Second Edition. 2004. These lists should always be used in conjunction with the full details of inclusions/exclusions that apply for each ICD code, group or chapter, which are available from WHO.

Figure 28: Select ICD codes screen

The **Find ICD** facility allows the **ICD code** of interest to be searched for in the list, and the **ICD name like** option allows a key word search to be undertaken at the specified level (selected from the drop down menu). Selected ICD codes (and ICD version) appear in the list at the bottom of the page. To remove one of the items from the list, use the

Clear button. When happy with the list, click **OK** or **Cancel** to return to the **Investigation details** page.

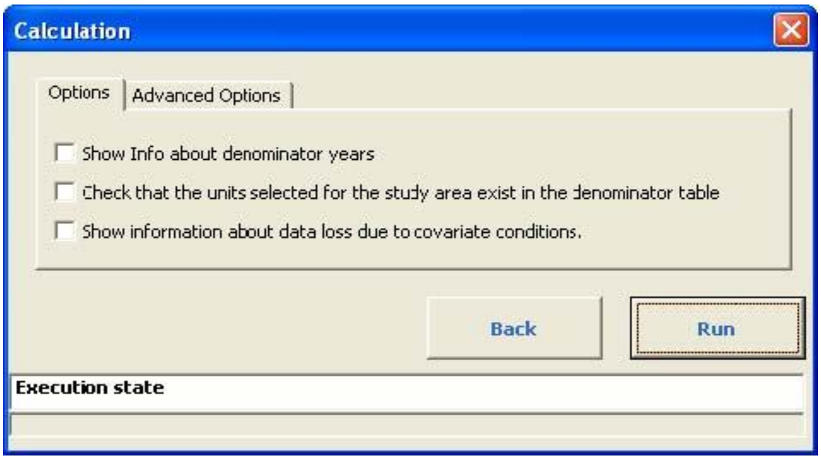
The option to select ICD codes from the drop down list has been added for the convenience of users with ICD coded data and who may not be confident using SQL clauses; however these lists should always be used in conjunction with the full details of inclusions/exclusions that apply for each ICD code, group or chapter, which are available from the WHO ICD books (http://www.who.int/classifications/icd/en/).

Once the health end point of interest has been selected, click the **Add investigation** button to add these details to the **INVESTIGATIONS** list at the bottom of the page. A number of investigations can be added in this way, from the same or different database pairings. Each investigation will use the same study and comparison area. Whilst there is no formal restriction to the number of investigations that can be run in any one study, the study will run more slowly if several investigations are run at the same time. To delete an investigation, double click it.

To run the study, click **'Next'**; to go back to the previous screen **Comparison area**, click **Back**.

6.5 Running the study

After clicking the **'Next'** button, the pop up box **Calculation** will appear.



□ Show Info about denominator years

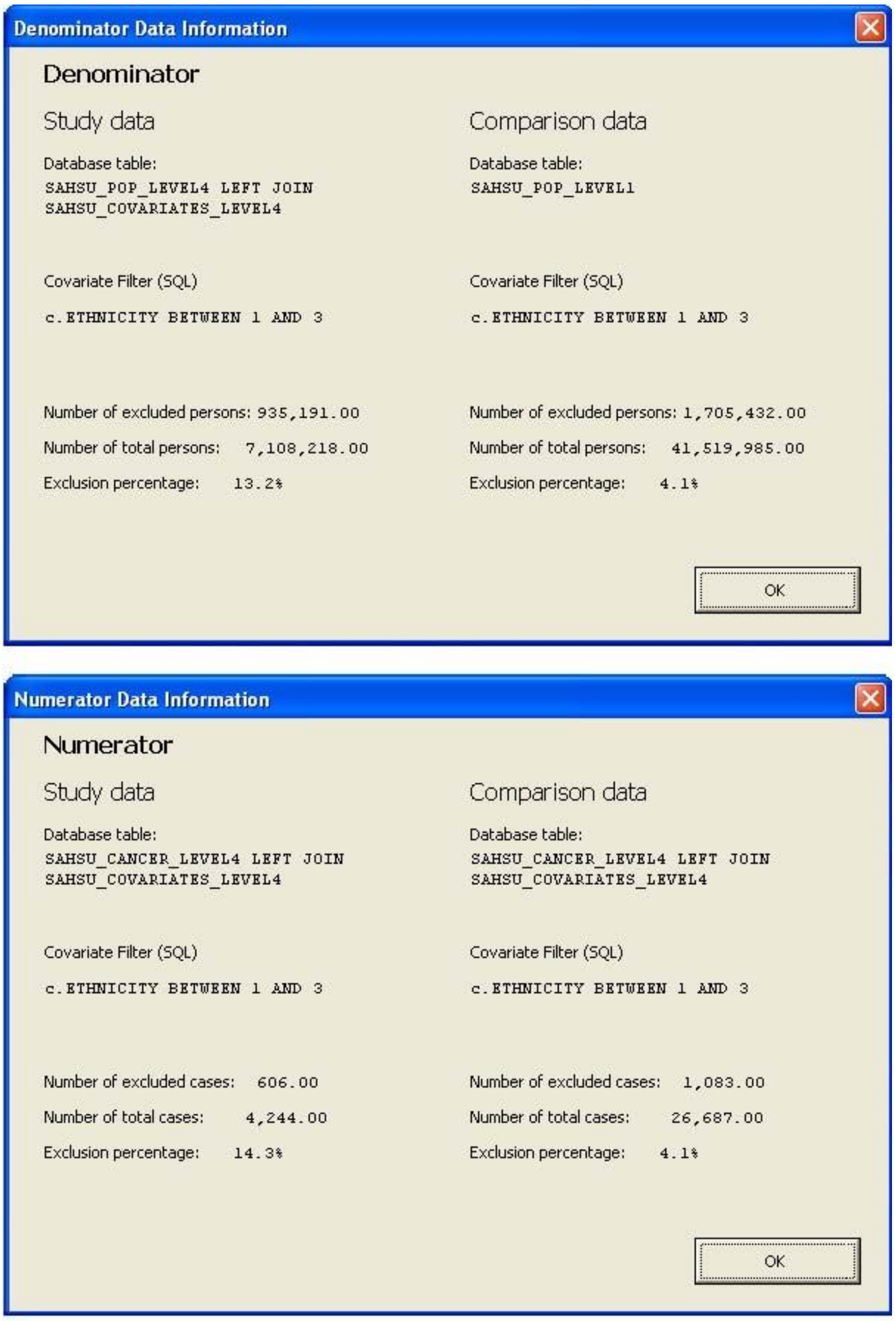
□ Check that the units selected for the study area exist in the denominator table

□ Show information about data loss due to covariate conditions.

Figure 29. Calculation pop up box

This box allows the selection of several options to provide feedback on potential problems encountered when running a RIF query:

- In the **Options** tab:
  - *Show information about data loss due to covariate conditions:* by selecting this option, two pop-up screens will appear to report on the health and population data excluded from investigation due to the covariate conditions. The first screen will report on the excluded denominator data for the study and comparison areas; the second on excluded numerator data.



Denominator

Study data

Database table:
`SAHSU_POP_LEVEL4 LEFT JOIN SAHSU_COVARIATES_LEVEL4`

Covariate Filter (SQL)

`c.ETHNICITY BETWEEN 1 AND 3`

Number of excluded persons: 935,191.00

Number of total persons: 7,108,218.00

Exclusion percentage: 13.2%

Comparison data

Database table:
`SAHSU_POP_LEVEL1`

Covariate Filter (SQL)

`c.ETHNICITY BETWEEN 1 AND 3`

Number of excluded persons: 1,705,432.00

Number of total persons: 41,519,985.00

Exclusion percentage: 4.1%

Numerator

Study data

Database table:
`SAHSU_CANCER_LEVEL4 LEFT JOIN SAHSU_COVARIATES_LEVEL4`

Covariate Filter (SQL)

`c.ETHNICITY BETWEEN 1 AND 3`

Number of excluded cases: 606.00

Number of total cases: 4,244.00

Exclusion percentage: 14.3%

Comparison data

Database table:
`SAHSU_CANCER_LEVEL4 LEFT JOIN SAHSU_COVARIATES_LEVEL4`

Covariate Filter (SQL)

`c.ETHNICITY BETWEEN 1 AND 3`

Number of excluded cases: 1,083.00

Number of total cases: 26,687.00

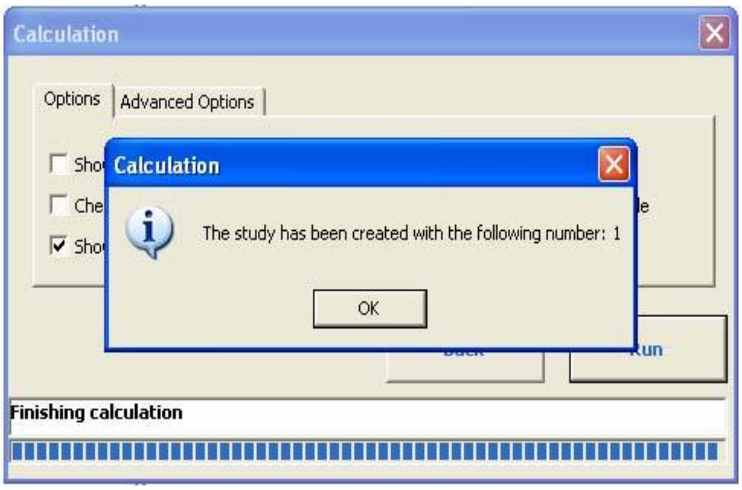
Exclusion percentage: 4.1%

Figure 30. Pop up screens showing details of data loss due to covariate conditions

A user unfamiliar with their data can use these screens to gain an understanding of the completeness of the covariate date, and decide whether the interpretability gained from adjusting for that covariate is worth the trade off in loss of power if a significant proportion of the data will be excluded. In addition, where the level of missing covariate data is known to be acceptable overall, these data screens can provide information on differential ascertainment between the study and comparison areas, which can also aid interpretability.

To run the study, click **Run**. To return to the previous menu **Investigation details**, click **Back**.

When the study calculations are complete, a pop up box will appear indicating the unique study ID. Click **OK** to continue to view the study output.



The study has been created with the following number: 1

Figure 31. Calculation pop up information box

6.6 Viewing the study output

Once the risk analysis study has successfully run, two additional menus are available above the tool bar – **Contextual Maps**, and **RIF Reporting**. These are described in turn below.

6.6.1 Contextual Maps

The **Contextual Maps** available will depend on what contextual data have been defined (see chapters 4 and 5).

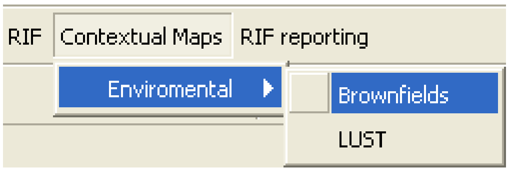


Figure 32. Example of the Contextual Maps menu

To view the contextual maps contained within the submenus, click on the element to be displayed. To remove the element from the map entirely, go back through the **Contextual Maps** menu and click on the element again.

When an element from the **Contextual Maps** submenu is added to the map, a **Layer** is added to the map. These layers are listed in the tab **Display**, in the box to the left of the map. The tab **Source** also lists the layers on the map, and provides information on where these data are stored.

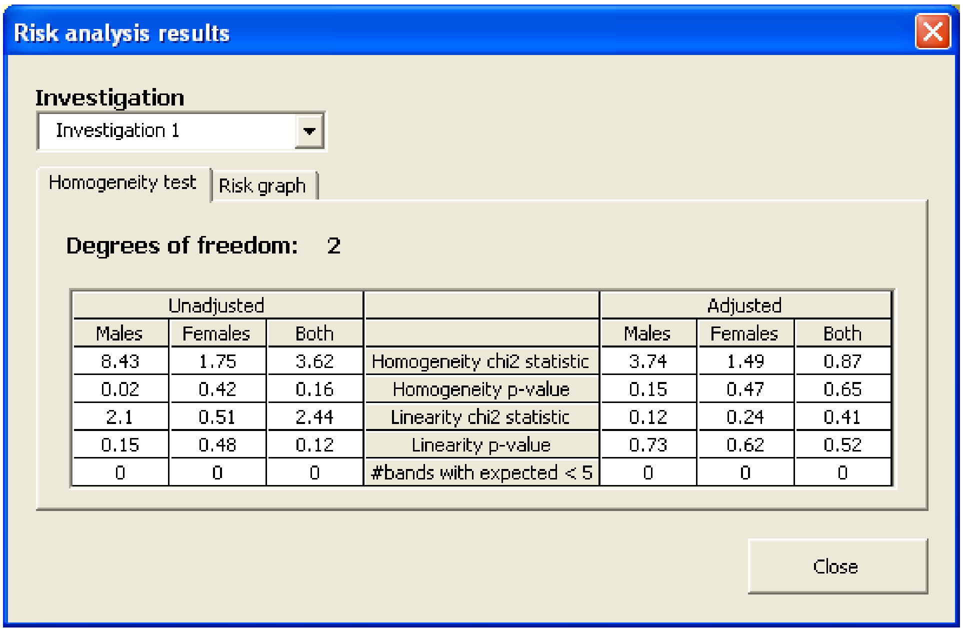
To keep the element available, but remove it from the current display, de-select this **layer** from the map. To do this, simply remove the tick from the box (by clicking in the box) next to the element. To view the element again, tick the box.

The layers on the map are easily manipulated to change the style, colour or size of the symbol. To alter these features use ArcGIS functions (see appendix B.5). Any other ArcGIS compatible data that has not been defined as contextual data can, of course, be added using standard GIS functions.

6.6.2 RIF results

The second additional menus added above the tool bar; RIF Reporting offers two further options: **Show RIF results** and **create report.**

- *Show RIF results:* To view the results of the test for homogeneity and Poisson trend statistic or to plot a risk graph, select the Show RIF results from the RIF reporting menu.
    - Homogeneity test: This will bring up the Risk analysis results window, which gives the results of the Chi square tests for homogeneity and linear trend (with accompanying p values), by gender, and for risks with and without adjustment for additional covariates. For more details on these tests, see appendix B.1).



Investigation

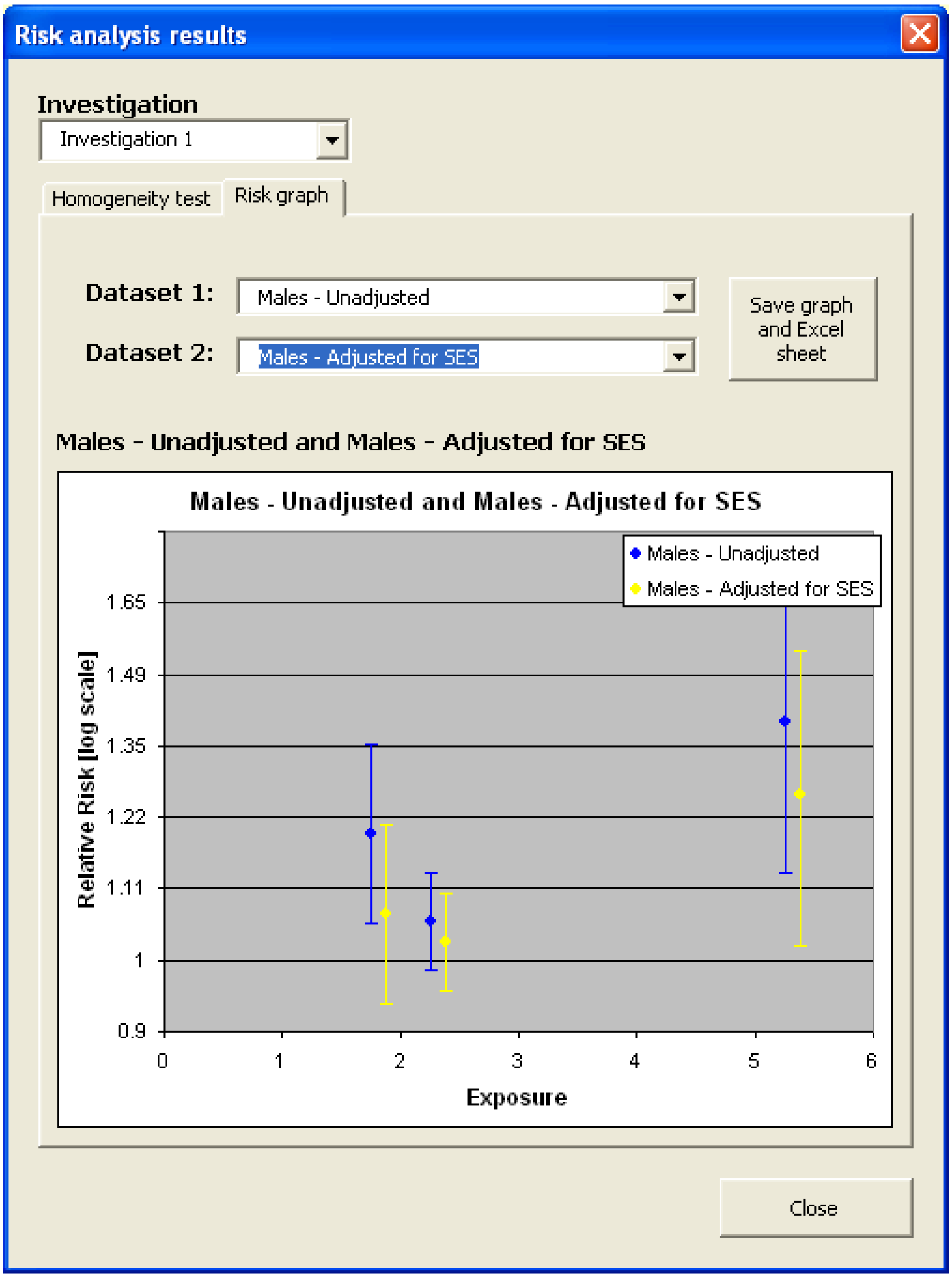
Investigation 1

Degrees of freedom: 2

Unadjusted				Adjusted		
Males	Females	Both		Males	Females	Both
8.43	1.75	3.62	Homogeneity chi2 statistic	3.74	1.49	0.87
0.02	0.42	0.16	Homogeneity p-value	0.15	0.47	0.65
2.1	0.51	2.44	Linearity chi2 statistic	0.12	0.24	0.41
0.15	0.48	0.12	Linearity p-value	0.73	0.62	0.52
0	0	0	#bands with expected < 5	0	0	0

Figure 33. Risk analysis results

-
    - By clicking the "Risk graph" tab, the user can create graphs of the risks as a function of exposure per band. The risks are plotted on a log-scale. The graph can plot one or two datasets and therefore risks for different genders and/or adjustments are easily compared.
      The user can also save a copy of the graph by clicking on the "Save graph and Excel sheet"-button. An Excel-file containing both the graph itself and the plotted data will then be saved along with a .gif-image of the graph.



Investigation: Investigation 1

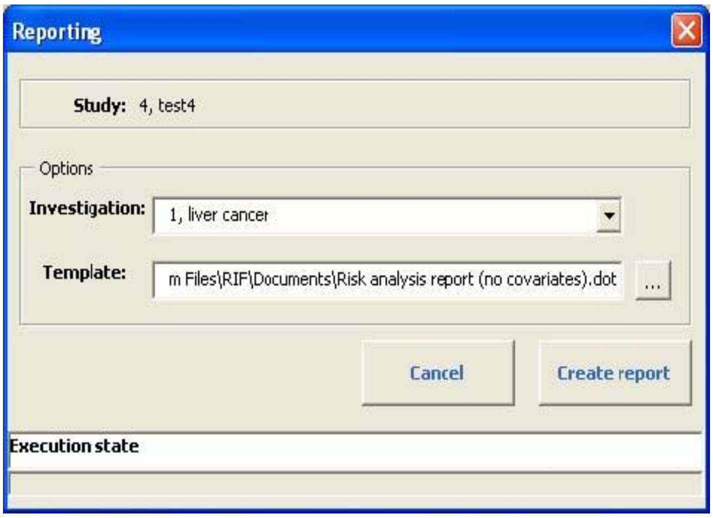
Dataset 1: Males - Unadjusted

Dataset 2: Males - Adjusted for SES

Males - Unadjusted and Males - Adjusted for SES

Figure 34. Risk graph

- *Create report*: To generate a study report detailing the study parameters, click on the **Study report** option. This will bring up the **Reporting** window, which allows the selection of the investigation for which the report will be generated, and allows the selection of the **Template** ... to be used to generate the report (to alter the default, click on the browse button and browse for the appropriate file). Once the appropriate template has been selected, click Create report to generate the report, or **Cancel** to close the **Reporting** window.



Study: 4, test4

Investigation: 1, liver cancer

Template: m Files\RIF\Documents\Risk analysis report (no covariates).dot

Figure 35. Reporting screen

Once the appropriate template has been selected, click **Create report** to generate the report, or **Cancel** to close the **Reporting** window.

The default template 'Risk analysis report (no covariates)' should be used when the investigation does not include adjustment for covariates. This report provides the following details:

- *Study information*: optional details entered in step 1/4.
- *Investigations summary:* investigation ID, name, numerator and denominator tables, health end points, years, age groups (defined in step 4/4).
- *Source Areas/Points*: a list of the areas/point sources used to define the study area (defined in step 2/3).
- *Comparison areas*: details of the comparison area (defined in step 3/4).
- *Relative risks (indirect standardisation):* observed, expected, relative risks and confidence intervals for each study area(s), with risks presented without adjustment for covariates, for males, females, and males and females combined.
- *Rates (direct standardisation):* observed, rate (per 100,000 person years) and confidence intervals for each study area(s). Risks are presented without adjustment for covariates, for males, females, and for males and females combined.
- *Map*: showing the study area(s).

The template 'Risk analysis report (covariates)' should be used when the investigation has included adjustment for covariates. This report provides the following details:

- *Study information/recipient information*: optional details entered in step 1/4.
- *Investigations summary*: investigation ID, name, numerator and denominator tables, health end points, years, age groups, covariate groups (defined in step 4/4).

- *Source Areas/Points*: a list of the areas/point sources used to define the study area (defined in step 2/3).
- *Comparison areas*: details of the comparison area (defined in step 3/4).
- *Relative risks (indirect standardisation)*: observed, expected, relative risks and confidence intervals for each study area(s), with risks presented with and without adjustment for covariates, for males, females, and males and females combined.
- *Rates (direct standardisation):* observed, rate (per 100,000 person years) and confidence intervals for each study area(s). Risks are presented with and without adjustment for covariates, for males, females, and for males and females combined.
- *Age bands:* a table showing the population by age band for the study area(s) and comparison area, as well as a graph showing how the age structure of the study and comparison area populations compare.
- *Sex*: a table showing the gender make-up of the study area(s) and comparison area, and a graph showing how the study and comparison areas compare in terms of gender make up.
- *Covariates*: a table showing the socio-economic make-up of the study area(s) and comparison area, as well as a graph showing how the covariates profile of the study and comparison areas compare.
- *Map*: showing the study area(s).

References:

Jenks, G F. 1967. "The Data Model Concept in Statistical Mapping", International Yearbook of Cartography 7: 186-190.

7. Running a new RIF study: Disease mapping

To undertake disease mapping, four steps will need to be completed:

- Defining the study details such as 'geography' and study type
- Defining the study area
- Defining the comparison area
- Defining the health outcomes of interest

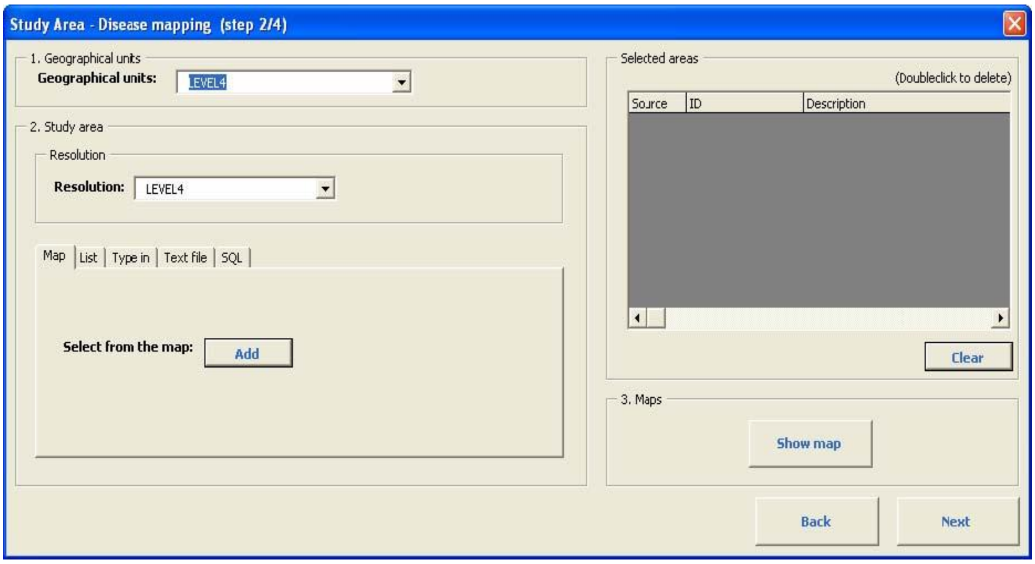
This chapter will detail how to progress through each of these steps.
Note. Where the set-up for a disease mapping study is the same as for a risk analysis study, detailed instructions will not be repeated; instead reference to the previous relevant sections will be provided.

7.1 Study details

This step is the same as described in section 6.1, but tick **Disease Mapping** instead of **Risk Analysis**.

7.2 Study Area: Disease mapping

By ticking **Disease Mapping** on the **Study Details** screen and clicking **'Next'** the **Study Area - Disease mapping** screen is opened.



Study Area - Disease mapping (step 2/4)

1. Geographical units

Geographical units: LEVEL4

2. Study area

Resolution

Resolution: LEVEL4

Selected areas

(Doubleclick to delete)

Source	ID	Description

3. Maps

Figure 36. Study Area – Disease Mapping screen

There are three sections to this screen:
1. **Geographical units**: these refer to the geographical resolution of the disease map.

- The drop down box indicates what levels of geography are available for disease mapping.

2. Study area: the study area refers to the area of interest for investigation.

- *The Resolution* can be altered to help select the study area however this will not affect the resolution at which the mapping is carried out (defined by the **Geographical units** above).

  For example, if the **Geographical units** are selected to be LEVEL4, and a **Resolution** of LEVEL2 is selected for the **Study area**; the resulting disease map will be of disease risk, by LEVEL4 area, giving all LEVEL4 areas within the selected LEVEL2 area.

  If the **Resolution** is set to LEVEL2, then the list is by LEVEL2 area. If the **Geographical units** for the mapping are set to LEVEL4, then by double clicking on a LEVEL2 area from the list, all LEVEL4 areas that fall within that LEVEL2 area will be selected.

  Note. for high resolution areas, where this list may contain thousands of elements, memory problems can occur when the RIF tries to populate the list. If a geographical level has more than 2000 elements, the listing capability should be deactivated (see section 5.8 for more details).

There are several methods that can be used to select the **Study area**, and these different methods can be used by selecting the appropriate tab:

- *Map:* this tab allows the study area to be selected from a map. To show the map, click the **Add** button on the **Map** tab.

  By clicking on the map, the study area can be selected. If the **Resolution** is set to LEVEL2, then the mapped units selected are by LEVEL2 area. If the **Geographical units** for the mapping are LEVEL4, then by selecting a LEVEL2 area from the map, all LEVEL4 areas within that LEVEL2 area will be selected.

  To choose more than one area, hold down the shift key, and click on each area to be selected, or hold down the left hand mouse button and drag the pointer over the area of interest. Use the zoom icons (appendix B5) to zoom in or out on the map. To re-activate the selection tool, click on selection tool icon.

  To go back to the **Study Area - Disease mapping** screen, go to the **RIF** menu, and select **Go back to last dialog**. Any selected areas will be listed in the pop-up window **Checking**, which provides information on the elements selected from the map.

- Additionally, as with the comparison area in a risk analysis, the study area can by selected by picking units from a **List**, by **Type in** the area ID, by importing a **Text file**, or by writing an **SQL** statement (for details of these methods of selection, see section 6.2 above).

  All these methods will reflect the chosen **Resolution** but remember the study resolution is set by the selected **Geographical units**.

  So for example, when typing in a unit the ID must be at the **Resolution** scale although the study will be carried out at the scale of the **Geographical units**.

3. Maps: Click *Show map* to view a map of the study area/s. the map will automatically zoom to the extent of any selected areas (rather than any input point source(s)/area(s)).

To go back to the **Study Area > Risk Analysis** screen, go to the **RIF** menu, and select '**Go back to last dialog'**. To move to the next step, click **Next**.

7.3 Comparison Area

The next step, defining the areas (populations) for the calculation of indirectly standardised risks is exactly the same as in the risk analysis approach (see section 6.3).

7.4 Investigation details

The next step, where the health outcomes of interest are selected, is exactly the same as in with risk analysis (see section 6.4 Investigation details).

7.5 Running the study

The options to provide feedback on potential problems running the RIF query are the same as in risk analysis (see section 6.5).

7.6 Viewing the study output

Once the risk analysis study has successfully run, three additional menus are added above the tool bar – **Contextual Maps**, **RIF reporting and Disease Mapping**.

7.6.1 Contextual Maps

For details on viewing the **Contextual Maps** see section 6.6

7.6.2. RIF results: reporting, exporting, and running SaTScan and WinBUGS.

The **RIF results** menu has the submenus: **Create report , Export and Run external program.** These options are discussed below.

- *Create report*
  To generate a study report, click on the Create report option. This will bring up the **Reporting** window, which allows the selection of the investigation for which the report will be generated as well as the selection of the **Template** ... to be used to generate that report. To alter the default template click on the button and browse for the appropriate file.
  The report generated using the default template (Disease mapping analysis report (outline)) provides the following details:
  - Study information/recipient information (if details were entered in step 1/4).
  - Investigations summary (investigation ID, name, numerator and denominator tables, health end points, years, age groups and covariates (the details entered in step 4/4)).
  - Study areas: details of the mapped areas (defined in step 2/3).
  - Comparison areas: details of comparison area (defined in step 3/4).
  - Map: showing the mapped area.

  Note. when undertaking a disease mapping study, it can become very unwieldy and time consuming to generate a result table for each mapped unit; furthermore, the generated data can be difficult to interpret in this format. The mapped data is usually much easier to interpret with the rates and risks for each mapped unit being easily viewed from a pop-up box from the map. Data tables can easily be exported to excel for the generation of additional tables and charts. As such, the default template for the disease mapping study now provides only details of the study parameters.

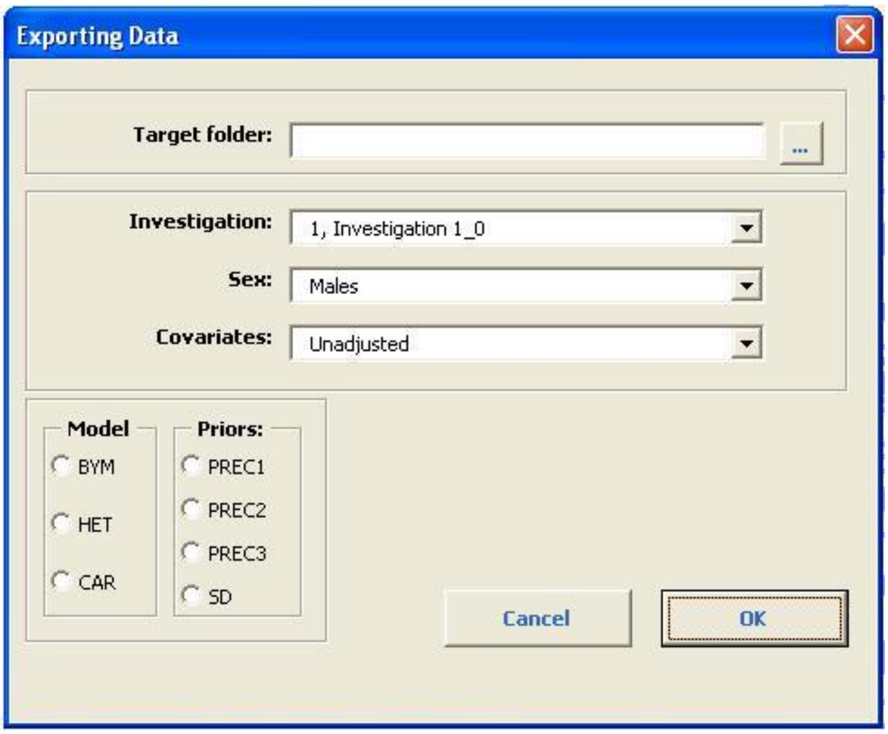
- *Export*

  To export the data generated from the RIF investigation for further analysis in WinBUGS, SaTScan, or Excel, click **Export ► To WinBUGS, Export ► To SaTScan** or **Export ► To Excel** as appropriate.

  When exporting to **WinBUGS** or **SaTScan**, the following data screen is used: Select the Target folder using the browse button ... . Use the drop down menus to choose the investigation, sex and covariate adjustment of the data set to be exported, then click **OK**. A pop up window will appear indicating that the files have been exported successfully.

  When exporting to **WinBUGS**, five files are exported to the target folder:

  - *data.txt*: file containing the observed, expected and adjacency matrix for each geographical unit.
  - *model.txt*: file containing the WinBUGS code of the BYM model.
  - *inits1.txt, inits2.txt, inits3.txt*: initial values for the parameters in the model; aimed to run three chains.
  - *disease mapping scripts.txt*: WinBUGS code to run the models in batch mode.
  - *id_file.txt*: area identifiers in the same order as the data.txt file.



Exporting Data

Target folder:

Investigation: 1, Investigation 1_0

Sex: Males

Covariates: Unadjusted

Figure 37. WinBUGS or SaTScan exporting data screen

Note. More details can be found in the WinBUGS user manual (WinBUGS: D. Spiegelhalter, A. Thomas, N. G. Best, and D.J. Lunn. WinBUGS Version 1.4.3 User's Manual, 2007. MRC Biostatistics Unit, Institute of Public Health, Cambridge; Rolf Nevanlinna Institute, University of Helsinki; and Department of Epidemiology and Public Health, Imperial College London, available from http://www.mrc-bsu.cam.ac.uk/bugs). Details on the BYM model can be found in J. Besag, J. York, and A. Molié. A bayesian image restoration with two applications in spatial statistics. Annals of the Institute of Statistics and Mathematics, 43:1–59, 1991.

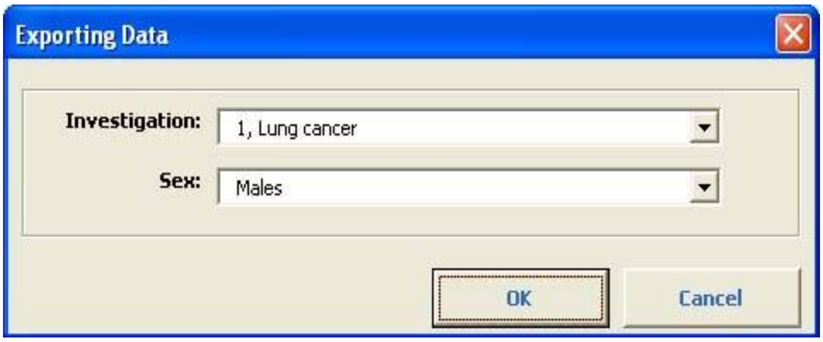
When exporting to **SaTScan**, the relevant information (population and case data plus x,y coordinates of area centroids) is exported to three files:

- study<study number >.geo - a tab delimited file with area ID (e.g. LEVEL4 identifier), and the xy-coordinates of the area centroids (dynamically calculated geographical centroids).
- study<study number>.cas – a tab delimited file which holds the area ID and the number of observed cases for each area.
- study<study number>.pop – a tab delimited file with area ID, a dummy year (required by SaTScan), and expected number of cases for each area.

  The population file is aimed to provide information on the background population in order to calculate expected counts, which can also be adjusted for further covariates within SaTScan (adjustment covariates are passed to SaTScan in both the cases and the population files). Therefore, providing SaTScan with just the expected counts (already adjusted within the RIF) is equivalent to providing it with person-years and the adjustment covariates. For this reason we have chosen the former, because although SaTScan will recalculate the expected counts (so that the sum is equal to the total number of observed cases), less information need to be exported.

More details can be found at www.satscan.org.

When exporting to **Excel**, the following data screen is used:



Investigation: 1, Lung cancer

Sex: Males

Figure 38. Excel exporting data screen

Use the drop down menus to choose the investigation and gender of the data set to be exported, then click **OK**. An Excel window will appear containing the selected results, which can be saved to an appropriate folder as required.

The excel file contains the following information (exported from the spatial data):

FID	- unique area identifier
PERIMETER	- length of each polygon outline
<level4>	- Area level identifier
<level2>	- Area level identifier
<level1>	- Area level identifier <level3> - Area level identifier
Observed -	Observed cases
EXP_UNADJ	Expected cases (indirect standardisation), unadjusted
RR_UNADJ	Indirectly standardised relative risk (unadjusted)
RRL95UNADJ	Lower 95% CI (indirect standardisation), unadjusted
RRU95UNADJ	Upper 95% CI (indirect standardisation), unadjusted
SMRR_UNADJ	- Empirical Bayesian smoothed indirectly standardised relative risk (unadjusted)
RATE_UNADJ	Directly standardised rate (unadjusted)
RATEL95UNADJ	Lower 95% CI (indirect standardisation), adjusted
RATEU95UNADJ	Upper 95% CI (indirect standardisation), unadjusted
EXP_ADJ	Expected cases (indirect standardisation), adjusted
RR_ADJ	Indirectly standardised relative risk (adjusted)
RRL95ADJ	Lower 95% CI (direct standardisation), adjusted
RRU95ADJ	Upper 95% CI (direct standardisation), adjusted
SMRR_ADJ	Smoothed relative (Empirical Bayes) adjusted for additional covariates
RATE_ADJ	Directly standardised rate (adjusted)
RATEL95ADJ	Lower 95% CI (indirect standardisation), adjusted
RATEU95ADJ	Upper 95% CI (indirect standardisation), adjusted

- *Run external program*

It is possible to run further complementary analyses using external programs such as SaTSCan and WinBUGS from the RIF. To do so, click **Run external program ► Run SaTScan** or **► Run WinBUGS**. When running **WinBUGS** or **SaTScan** directly from the RIF, a similar data screen as that used when exporting data will appear.

Run SaTScan

It is possible to run further complementary analyses using external programs such as SaTSCan and WinBUGS from the RIF. To do so, click **Run external program ► Run SaTScan** or **► Run WinBUGS**. When running **WinBUGS** or **SaTScan** or **RunINLA** directly from the RIF, a similar data screen as that used when exporting data will appear.

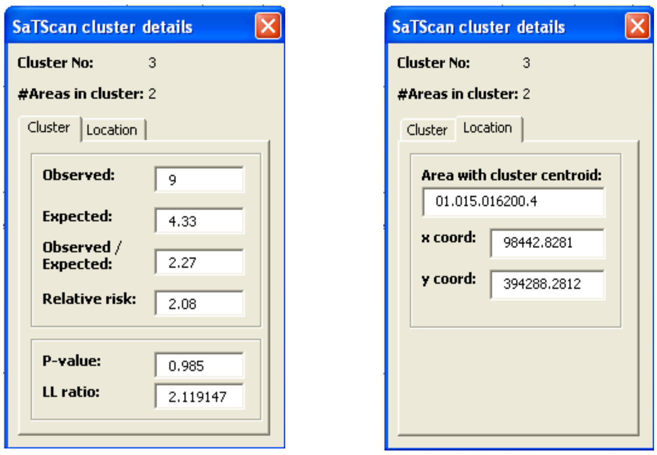
Run SaTScan

SaTScan is free software (distributed under license agreement) to scan a region to detect and locate clusters of disease (Kulldorff, 1997). It can be downloaded from www.satscan.org. Once installed, the RIF can run it in batch mode. The SaTScan windows interface is bypassed and the RIF uses data selected by the user input for the RIF study to construct the SaTScan file format input files and a SaTScan parameter file. The SaTScan calculation engine is executed when launched. When the analysis is done, the RIF reads in both the SaTScan cluster information and the SaTScan location information output files, and presents these results as two shapefiles; one showing any identified clusters and one of the study area.

Select the Target folder for the output results from SaTScan, using the browse button .... Use the drop down menus to choose the investigation, sex and covariate adjustment of the data set to be exported, and finally the spatial window shape to be used in SaTScan (Elliptic or Circular), then click **OK**. The data is then exported to SaTScan. SaTScan then carries out the calculations; RIF users will have to sign the SaTScan license agreement before they can use it.. .

A pop up window will appear indicating that SaTScan has run and the results have been imported back into the RIF successfully.

The SaTScan results are displayed in two new layers named *SaTScanClusters* and *SaTScanAreas*. To view the SaTScan results, contained in each shapefile, simply highlight the layer name in the Table of Contents (e.g. SaTScanClusters) by clicking on it. Then select the SaTScan reporting tool and click on, for this example, a cluster. The corresponding details that were calculated in SaTScan will be shown.



SaTScan cluster details

Cluster No: 3

#Areas in cluster: 2

Cluster | Location

Observed: 9

Expected: 4.33

Observed / Expected: 2.27

Relative risk: 2.08

P-value: 0.985

LL ratio: 2.119147

SaTScan cluster details

Cluster No: 3

#Areas in cluster: 2

Cluster | Location

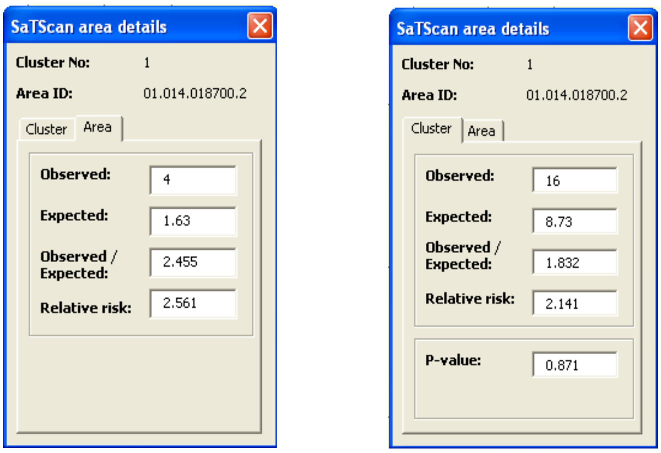
Area with cluster centroid: 01.015.016200.4

x coord: 98442.8281

y coord: 394288.2812

Figure 39. SaTScan cluster details

Details for individual areas can also be viewed by simply clicking on the SaTScanAreas layer name in the table of contents and then using the reporting tool .



Cluster No: 1
Area ID: 01.014.018700.2

Observed: 4
Expected: 1.63
Observed / Expected: 2.455
Relative risk: 2.561

Cluster No: 1
Area ID: 01.014.018700.2

Observed: 16
Expected: 8.73
Observed / Expected: 1.832
Relative risk: 2.141

P-value: 0.871

Figure 40. SaTScan cluster details

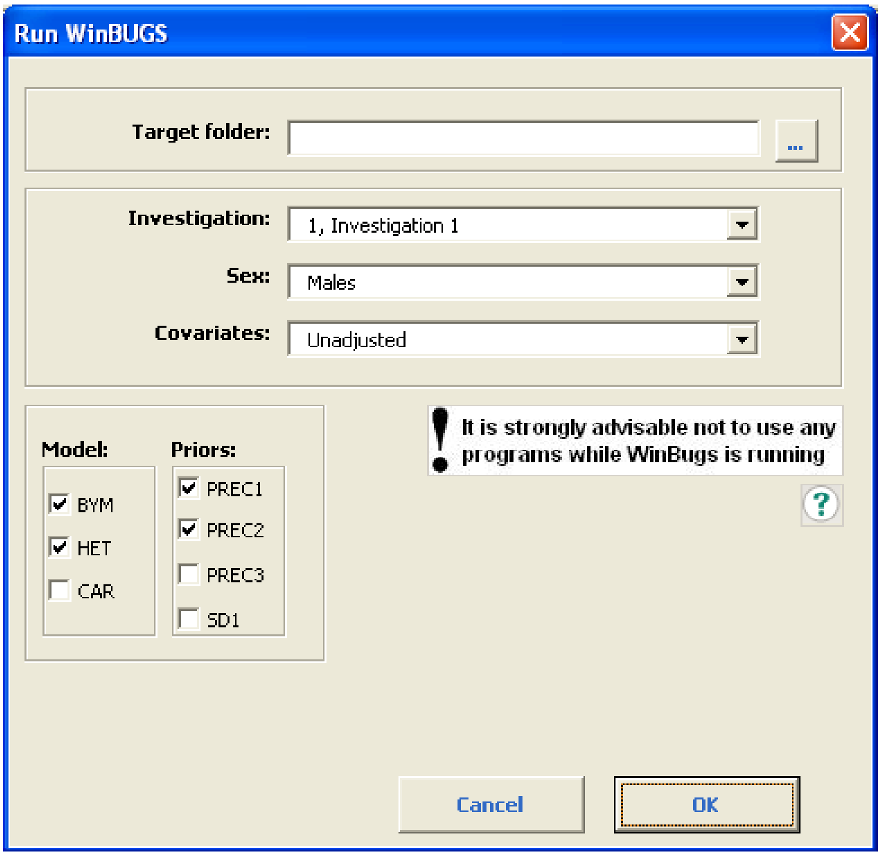
The disease mapping results (see section 7.6.3) can be viewed together with the SaTScan output.

References:

Kulldorff, M. A spatial scan statistic. Communications in Statistics: Theory and Methods, 1997, 26:1481–1496.

Run WinBUGS

WinBUGS is a Bayesian statistical package. It is freely distributed under license agreement. It can be downloaded from http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml. Provided that it is installed, the RIF can use it to run a fully Bayesian disease mapping model in batch mode. In the export WinBUGS form, the RIF fits three different models: the *BYM* (Besag, York and Mollié, 1991), *HET* (Heterogeneity) and *CAR* (conditional autoregressive) model, each with 4 different Priors.



Target folder:

Investigation: 1, Investigation 1

Sex: Males

Covariates: Unadjusted

Model: ☑ BYM ☑ HET □ CAR

Priors: ☑ PREC1 ☑ PREC2 □ PREC3 □ SD1

It is strongly advisable not to use any programs while WinBugs is running

The values of the included priors, by model, are:

Besag, York and Mollié Model (BYM)

	a1	b1	a2	b2	c1	c2
BYM_PREC1	0.5	0.005	0.5	0.005		
BYM_PREC2	0.1	0.01	0.01	0.01		
BYM_PREC3	0.1	0.1	0.1	0.1		
BYM_SD1					0.1	0.1

Heterogeneity Model (HET)

	a1	b1	c1
HET_PREC1	0.5	0.005	
HET_PREC2	0.01	0.01	
HET_PREC3	0.1	0.1	
HET_SD1			0.1

Conditional autoregressive Model (CAR)

	a2	b2	c1
CAR_PREC1	0.5	0.005	
CAR_PREC2	0.01	0.01	
CAR_PREC3	0.1	0.1	
CAR_SD1			0.1

To start the analysis a minimum of one model and one prior must be selected. In addition an **empty** output folder must be selected (a new folder can be created by the user at model run time). Within this folder model runs will create a new folder, containing results, for every model/prior combination selected.

Users are strongly advised not to use any other applications while running the WinBugs export tool from the RIF (this warning appears on the menu), as the program requires keystrokes and errors could occur. Furthermore depending on the speed of your computer WinBUGS may take a few seconds to re-start between model and model please be patient, and do not start using another application.

The RIF-WinBUGS integration calculates and maps the posterior median of the (smoothed) relative risks (RR) and the exceedence posterior probabilities Pr(RR>1 | data). (Note: this model is fitted using Markov Chain Monte Carlo (MCMC) simulation techniques. This approach requires assessment of convergence of the chains of simulations.

The process is composed of three steps:

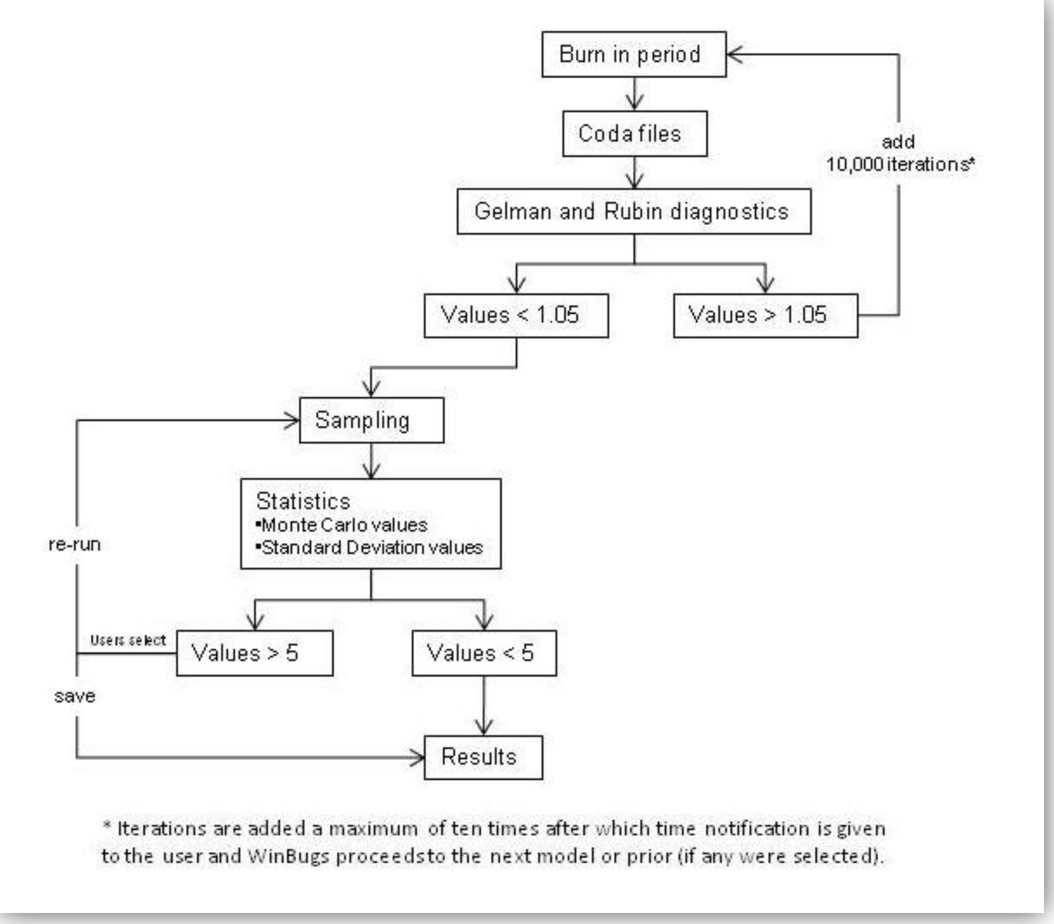
- Burn in period
- Sampling
- Results output

The Burning period produces a set of coda files and the values are checked using the Gelman and Rubin (GR) diagnosis.

If the GM test is successful, the Sampling script is run. This creates a file called log_autocor.txt which stores all the statistics for each node including the Monte Carlo error and SD value. The ratios between these two set of values are then calculated. If all the calculated ratios are less than 5, the results are automatically saved. If any of the ratios are greater than 5 the user must decide whether to re-run the sampling script or to save the analysis results.

In the case of one or more values being greater than 1.05 the GR test fails and the model will automatically update and run a further 10000 iterations. When the new coda files have been created and the model has finished updating itself the GM test is repeated, once again if this fails a further 10000 iterations are run. This cycle is repeated a maximum of ten times after which time notification is given to the user and WinBugs proceeds to the next model or prior (if any were selected).

NOTE:The whole process can be stopped by pressing the ‘Esc’ button while WinBUGS is updating a model.


Burn in period
Coda files
add
10,000 iterations*
Gelman and Rubin diagnostics
Values < 1.05
Values > 1.05
Sampling
re-run
Statistics
•Monte Carlo values
•Standard Deviation values
Users select
Values > 5
Values < 5
save
Results
* Iterations are added a maximum of ten times after which time notification is given to the user and WinBugs proceedsto the next model or prior (if any were selected).

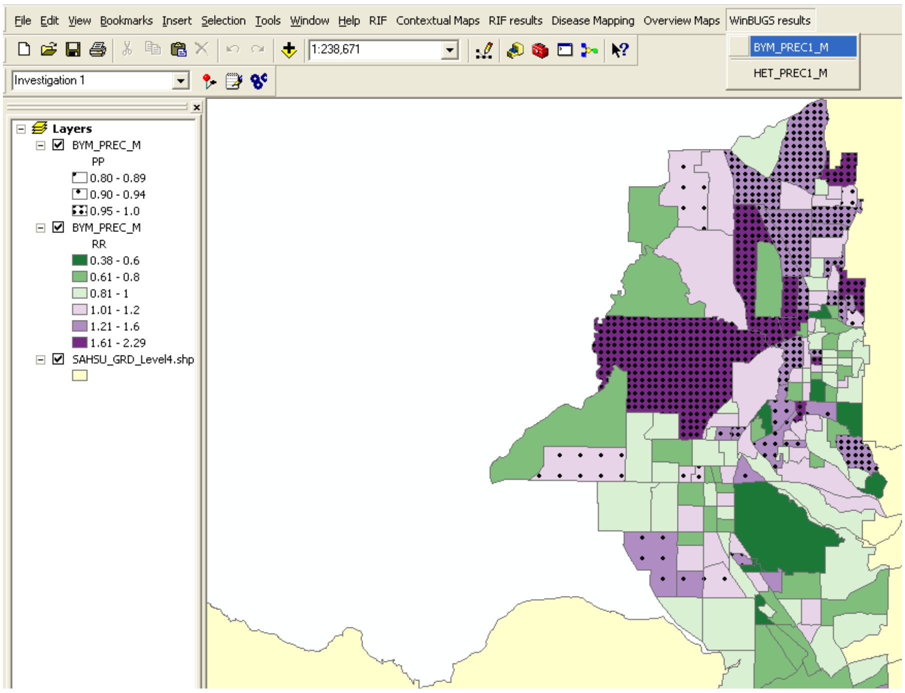
When the analysis is complete, within your selected directory you will find as many folders as models were successfully run. In each folder, the following files will be present:

- data.txt containing the observed, expected and adjacency matrix for each geographical unit
- model.txt file containing the WinBUGS code of model
- inits1.txt,inits2.txt initial values for the parameters in the model ;
- Ids_File.txt: area identifiers in the same order as the data.txt file
- Coda files:output from burn in process
- Log_autocor.txt: output from Sampling process used to determine ratio between MC error and SD
- Results.txt: final result files storing AreaID, posterior probabilities, residual relative risk and relative risk
- Log.txt:burning script
- Samples.txt:sampling script
- saveScript.txt: script which save results to selected directory
- Shapefile of the analysed area with attribute table containing the analysis results

Where a model has failed to pass the GM test, the above list of files will not be created; instead an info_*model*.txt file notifying the problem occurred will be generated.

After WinBUGS has been run a new menu called ‘WinBUGS results’ will appear on the right hand side of the main tool bar.
By default results from the BYM model for smoothed RR results and, the corresponding exceedence posterior probabilities (EPP) will be added automatically to the map. If a BYM model has not been run, no results are automatically displayed. All other layers e.g. those corresponding to other models are available for selection under the WinBUGS results menu.

Note. Only one model can be displayed at a time. To see multiple maps at the same time, user the overview maps.



These are displayed as a texture map at the same time as the RR results as a measure of uncertainty. Only the areas where the likelihood the true RR exceeds 1.0 is greater than 0.8 (see Richardson et al, 2004) will be displayed and these are subdivided into three bands:

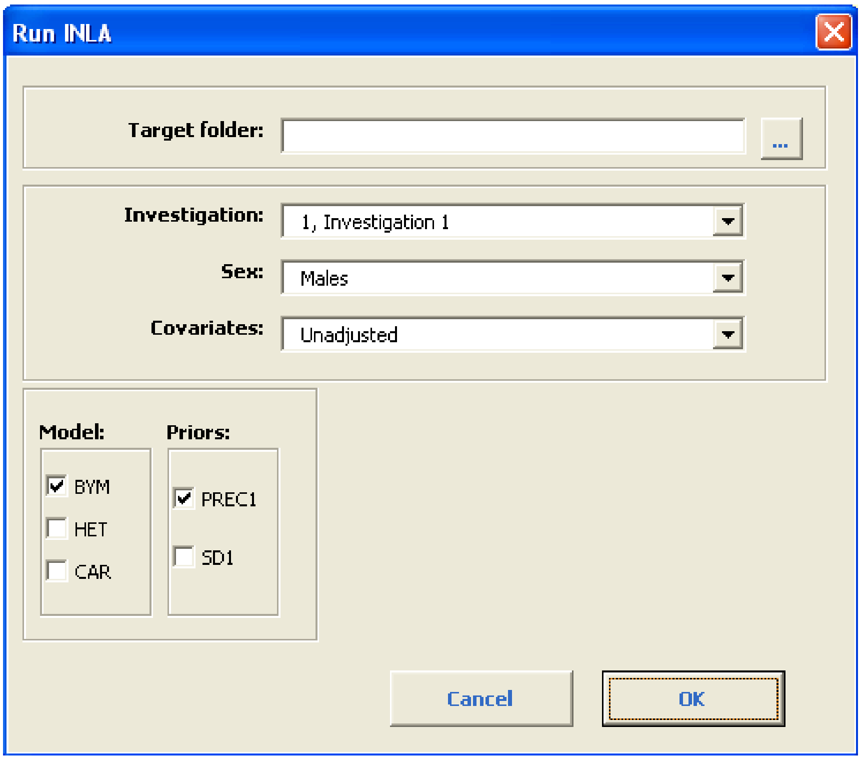
0.80 – 0.89
0.90 – 0.94
0.95 – 1.00

However, these bands can be changed by the user by changing the shapefile display properties (see Appendix B.5.1). For details on how to view the exact smoothed standardised mortality/morbidity ratio for each area see section 7.6.3 below.

References:

Best NG, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Statistical Methods in Medical Research 2005; 14: 35-59.
Lawson, A. B. Biggeri, A. Boehning, D. et al.(2000) Disease Mapping Models: an empirical evaluation Statistics in Medicine, 19,17/18,2217-2242.
Richardson S, Thomson A, Best N, Elliott P. 2004. Interpreting posterior relative risk estimates in disease-mapping studies. Environmental Health Perspectives 112(9):1016-1025.

Run Inla



Target folder:

Investigation: 1, Investigation 1

Sex: Males

Covariates: Unadjusted

Model:
☑ BYM
□ HET
□ CAR

Priors:
☑ PREC1
□ SD1

As with WinBUGS, INLA needs a minimum of one model selected and one prior, in addition to an empty directory as target folder. Again, within it as many folders as the combination between models and priors will be created.

The output folder for each model will contain the following:

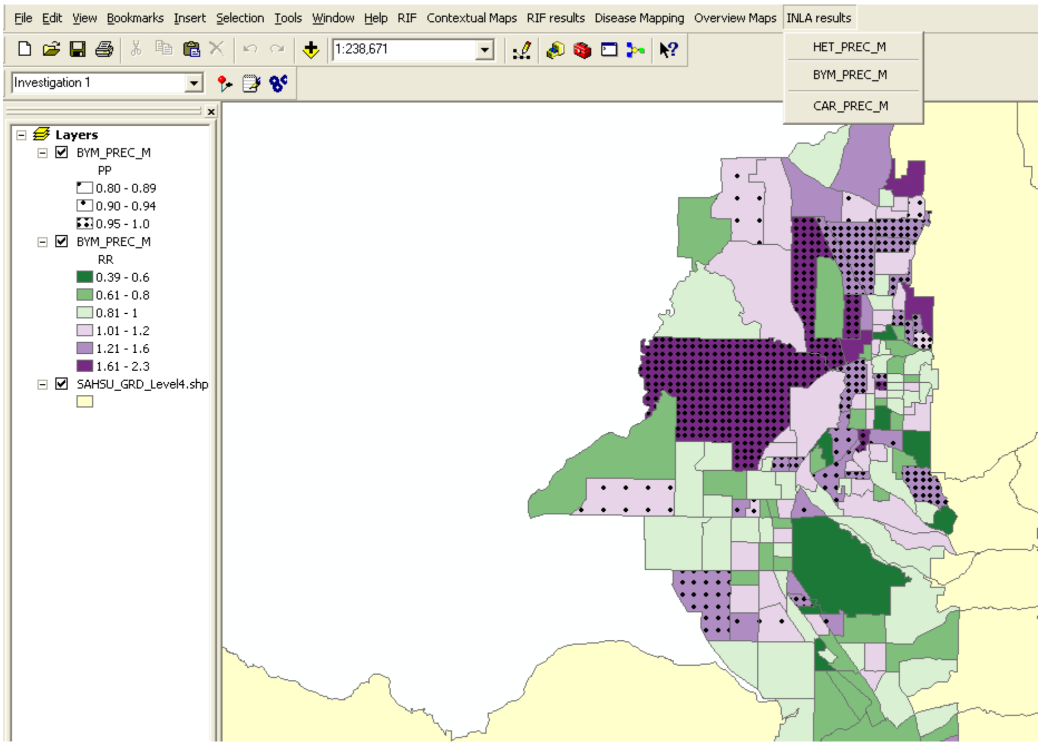
- Data_INLA directory
- Res_INLA directory
- Log.txt: Raw result from INLA. It stores the mean, SD and median values for RR and ResidRR of each node
- INLAresults.txt: stores the AreaID,InlaID,RR,SmSMRl95,SmSMRu95,ResidRR,ResidRRl95 and ResidRRu95Results values derived from the following files:
  - \data_INLA\ID.txt
  - \intercept\summary.dat

  - \predictor-user-scale\summary.dat
  - \predictor-user-scale\quantiles.dat
- INLA_IDs_file.txt:List of unique identifiers of the study area
- Results.txt: Final attribute table which is joined to the output shapefile and used to map the results. It stores the same list of IDs as the INLA_IDs file as well as the values for the residual relative risk (residRR) and relative risks (RR) extracted from the log.txt file.
  In addition, INLA calculates the cumulative distribution function (cdf) (in res_INLA\predictor-user-scale\cdf.DAT) from which the posterior probability can be calculated using:

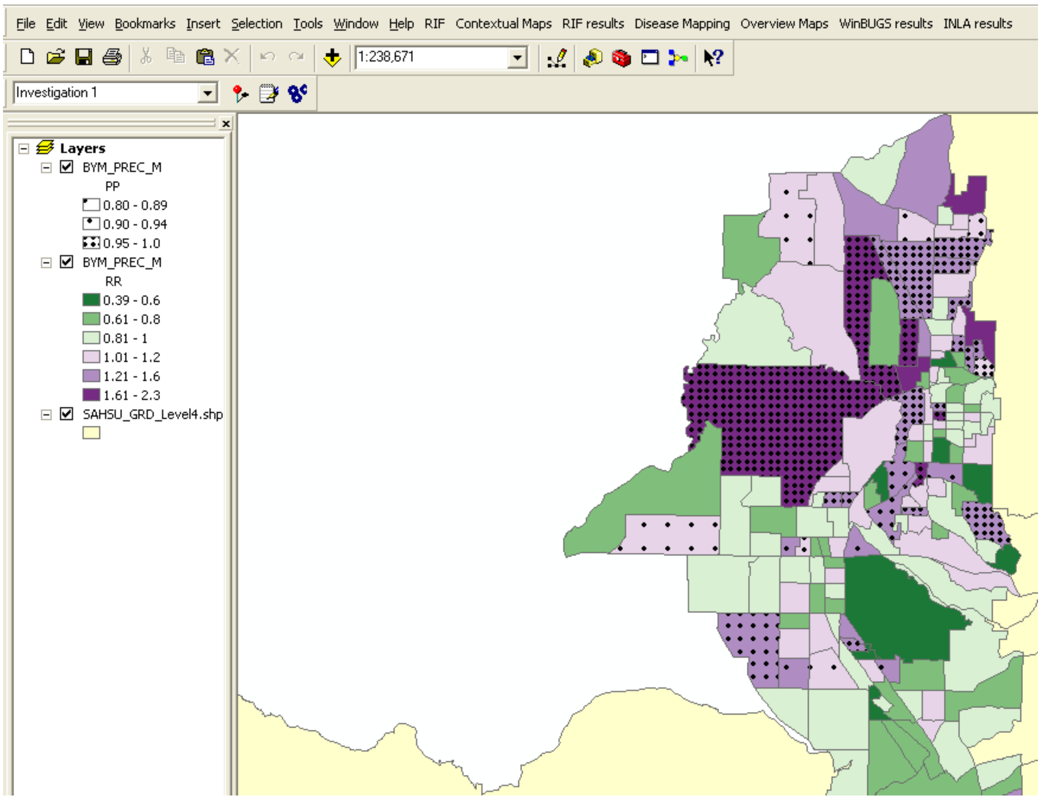
$$PP(X>x) = 1\text{-cdf}.$$

- Shapefile of the analysed area with Results.txt joined as attribute table.

On the Main toolbar a new menu appears: ‘INLA results’. In the same way, as with WinBUGS by default the layers corresponding to the BYM model are displayed to the map. All others can be added by simply clicking on one of the entry under the INLA results menu.

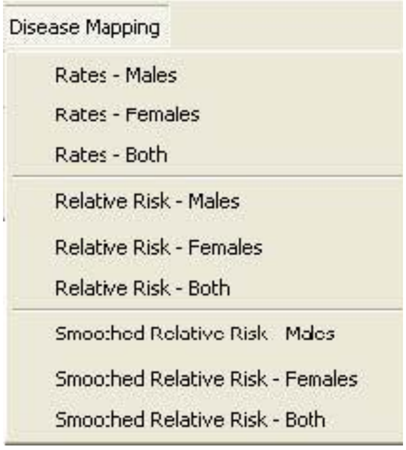


INLA and WinBUGS can both be run for the same investigation. Both menus (WinBUGS results and INLA results) will be added to the main tool bar (shown below).



7.6.3. Disease Mapping

The **Disease Mapping** menu contains a list of the disease rates and unsmoothed and smoothed risks (see Appendix B.1 for details) calculated in the RIF. If no covariates have been selected, then this menu will be as follows:



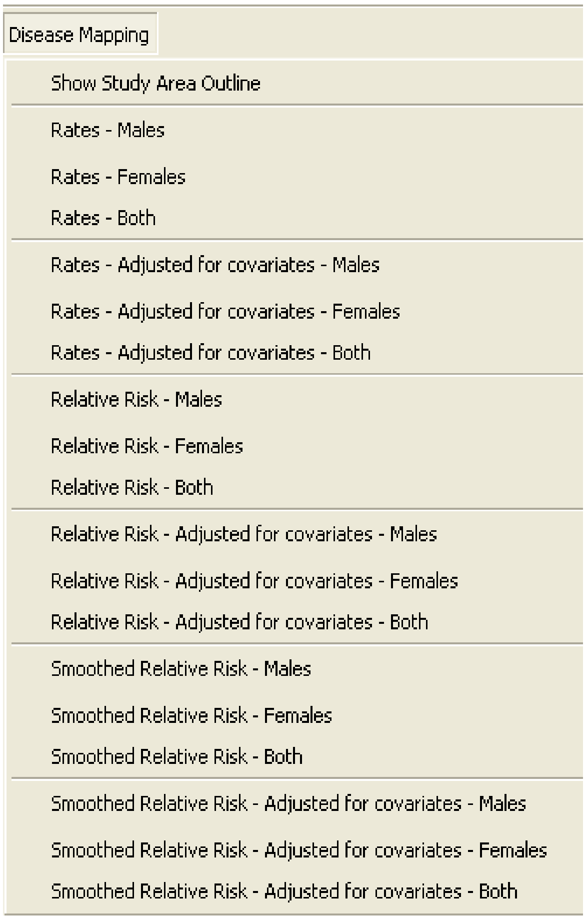
Description

The directly age standardised rate.

The indirectly standardised age adjusted relative risk.

The smoothed indirectly standardised age adjusted relative risk.

If covariates have been selected then this menu will be as follows:


Disease Mapping
Show Study Area Outline
Rates - Males
Rates - Females
Rates - Both
Rates - Adjusted for covariates - Males
Rates - Adjusted for covariates - Females
Rates - Adjusted for covariates - Both
Relative Risk - Males
Relative Risk - Females
Relative Risk - Both
Relative Risk - Adjusted for covariates - Males
Relative Risk - Adjusted for covariates - Females
Relative Risk - Adjusted for covariates - Both
Smoothed Relative Risk - Males
Smoothed Relative Risk - Females
Smoothed Relative Risk - Both
Smoothed Relative Risk - Adjusted for covariates - Males
Smoothed Relative Risk - Adjusted for covariates - Females
Smoothed Relative Risk - Adjusted for covariates - Both

Description

The outline of the selected study area

The directly age standardised rate.

The directly age and socio-economic status standardised rate.

The indirectly standardised age adjusted relative risk.

The indirectly standardised age and socio-economic status adjusted relative risk.

The smoothed indirectly standardised age adjusted relative risk.

The smoothed indirectly standardised age and socio-economic status adjusted relative risk.

If several investigations have been carried out as part of one RIF study, then use the drop down box just to the right of the **Disease Mapping** menu in the toolbar to choose between the different investigations to be viewed.

To map the study results, click on the rates or risks of interest to add these as a layer to the map. As with the contextual maps, these layers can be de-selected to determine what will be viewed on the map at any time; similarly the colours of each category can be changed by double clicking on the colour key in the **Display** tab to the left of the map (see appendix B.5).

The rates calculated in the RIF are automatically categorised into quintiles, but the categories can easily be altered by double clicking on the relevant layer (Rate, or RateCov) in the **Display** tab to the left of the map to bring up a window **Layer Properties** (see appendix B.5).

All the relative risks calculated in the RIF are automatically displayed in default categories (see below) to allow for direct comparison between males and females, unadjusted and adjusted risks and so on. These default categories will not be appropriate for every outcome or investigation, but these categories can easily be altered by double clicking on the relevant layer (RR, RRCov, SmRR, SmRRCov) in the **Display** tab to bring up the window **Layer Properties** (see appendix B.5).
Default relative risk categories:

Min – 0.6
0.61 – 0.8
0.81 – 1
1.01 – 1.2

1.21 – 1.6
1.61 – Max

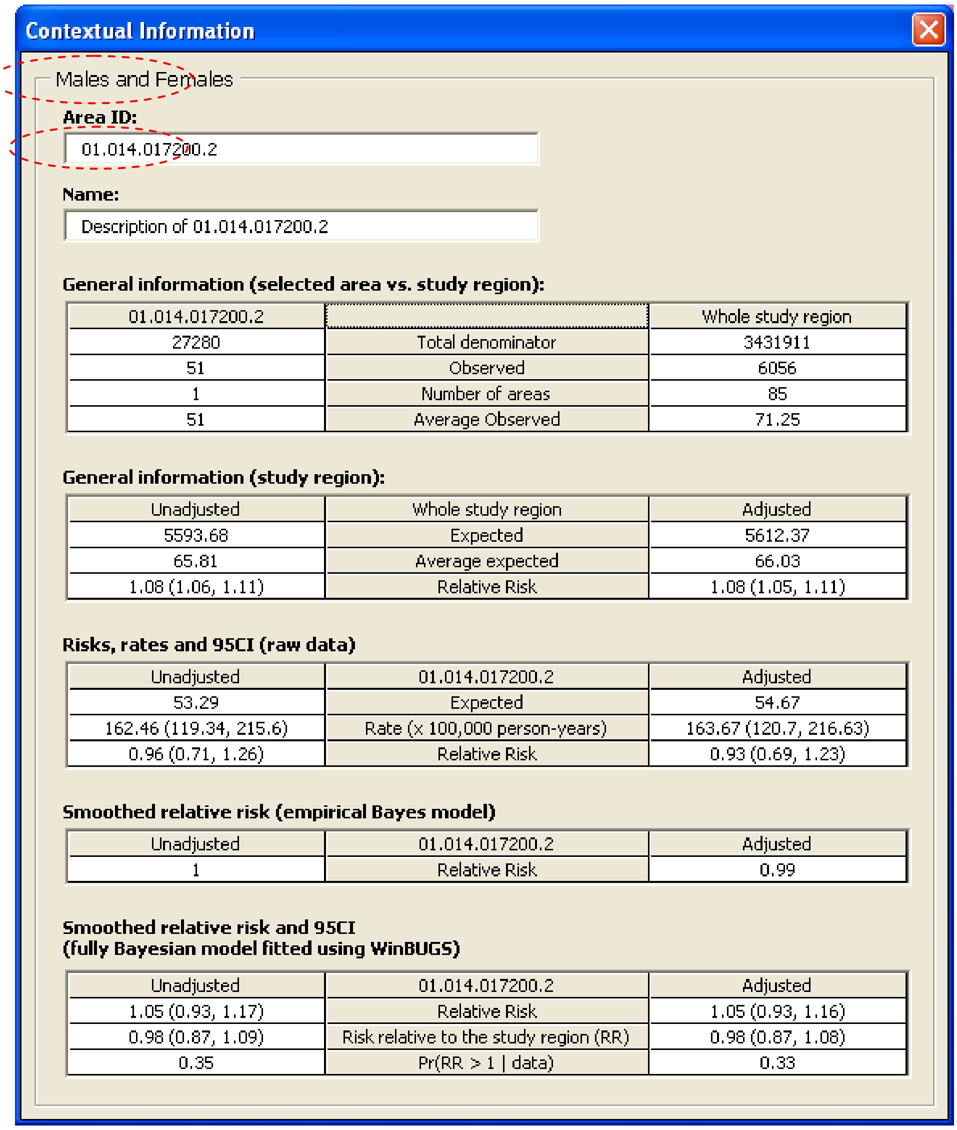
While the colouring can show which range of risks or rates each mapped area can be classified into and can reveal interesting patterns of risk, it is often useful to be able to view the exact risk/rate and associated 95% confidence intervals for each area.

To view this information, click on the notebook icon to the right of the drop down box showing the name of the investigation being displayed, then click on one of the mapped areas.

A window **Contextual Information** will appear (see Figure 41). The **Contextual Information** window provides details related to the investigation being viewed. The reported results refer to the **gender** being viewed (shown at the top) and the **area** that has been ‘clicked on’.

There are four tables that will appear, although where no results exist e.g. adjustment for covariates was not chosen in the study, then no results will be shown in the relevant table. Furthermore, a fifth table is included when WinBUGS has been run and the results from using the fully Bayesian BYM model are reported. This fifth table reports the results from the database that relate to the gender of the **RIF disease map** that is currently being displayed. So, in cases where the study has been retrieved and the WinBUGS model has previously been run, these results can be reported on this menu. Since, the RIF reports the results from the database please note that the results on the menu may not match a WinBUGS output shapefile present in the table of contents, where the gender differs from that of the disease map.

The first table of the contextual information shows general information about the selected area compared with the study area as a whole. A second table outlines details from the study region. The risks, rates and 95% confidence intervals for both adjusted and unadjusted results for the selected area are shown in the third table and the results from the smoothed relative risk (using empirical Bayes) are in the fourth table.



Males and Females

Area ID:

01.014.017200.2

Name:

Description of 01.014.017200.2

General information (selected area vs. study region):

01.014.017200.2		Whole study region
27280	Total denominator	3431911
51	Observed	6056
1	Number of areas	85
51	Average Observed	71.25

General information (study region):

Unadjusted	Whole study region	Adjusted
5593.68	Expected	5612.37
65.81	Average expected	66.03
1.08 (1.06, 1.11)	Relative Risk	1.08 (1.05, 1.11)

Risks, rates and 95CI (raw data)

Unadjusted	01.014.017200.2	Adjusted
53.29	Expected	54.67
162.46 (119.34, 215.6)	Rate (x 100,000 person-years)	163.67 (120.7, 216.63)
0.96 (0.71, 1.26)	Relative Risk	0.93 (0.69, 1.23)

Smoothed relative risk (empirical Bayes model)

Unadjusted	01.014.017200.2	Adjusted
1	Relative Risk	0.99

Smoothed relative risk and 95CI
(fully Bayesian model fitted using WinBUGS)

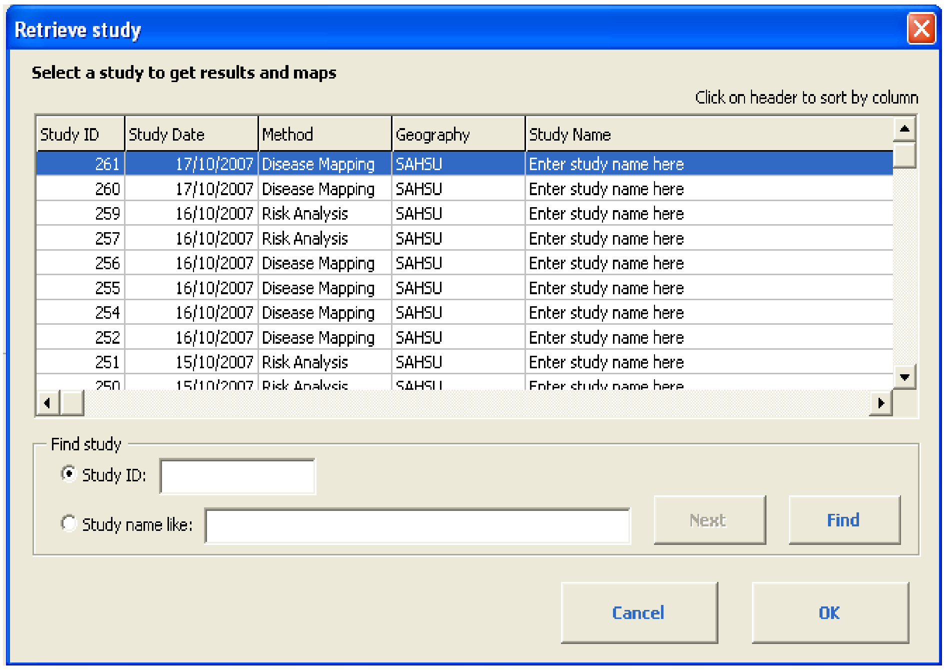
Unadjusted	01.014.017200.2	Adjusted
1.05 (0.93, 1.17)	Relative Risk	1.05 (0.93, 1.16)
0.98 (0.87, 1.09)	Risk relative to the study region (RR)	0.98 (0.87, 1.08)
0.35	Pr(RR > 1 | data)	0.33

Figure 41. Contextual Information pop-up box

8. Retrieving or deleting a study

8.1 Retrieving a study

To retrieve a previously run study open the **RIF ►** menu, and select **Retrieve previous**. This will bring up a pop-up screen **Retrieve study**, which lists all previously run studies, and provides details of the study ID, study date, method (i.e. risk analysis or disease mapping), the geography used and the user defined study name. Studies are sorted by study ID by default, but the list can be re-sorted by any of these parameters by clicking on the appropriate column. Additionally there is a **Find study** facility, whereby the study ID or part of the study name can be used to locate any particular study from the list.



Retrieve study

Select a study to get results and maps

Click on header to sort by column

Study ID	Study Date	Method	Geography	Study Name
261	17/10/2007	Disease Mapping	SAHSU	Enter study name here
260	17/10/2007	Disease Mapping	SAHSU	Enter study name here
259	16/10/2007	Risk Analysis	SAHSU	Enter study name here
257	16/10/2007	Risk Analysis	SAHSU	Enter study name here
256	16/10/2007	Disease Mapping	SAHSU	Enter study name here
255	16/10/2007	Disease Mapping	SAHSU	Enter study name here
254	16/10/2007	Disease Mapping	SAHSU	Enter study name here
252	16/10/2007	Disease Mapping	SAHSU	Enter study name here
251	15/10/2007	Risk Analysis	SAHSU	Enter study name here
250	15/10/2007	Risk Analysis	SAHSU	Enter study name here

Find study

● Study ID:

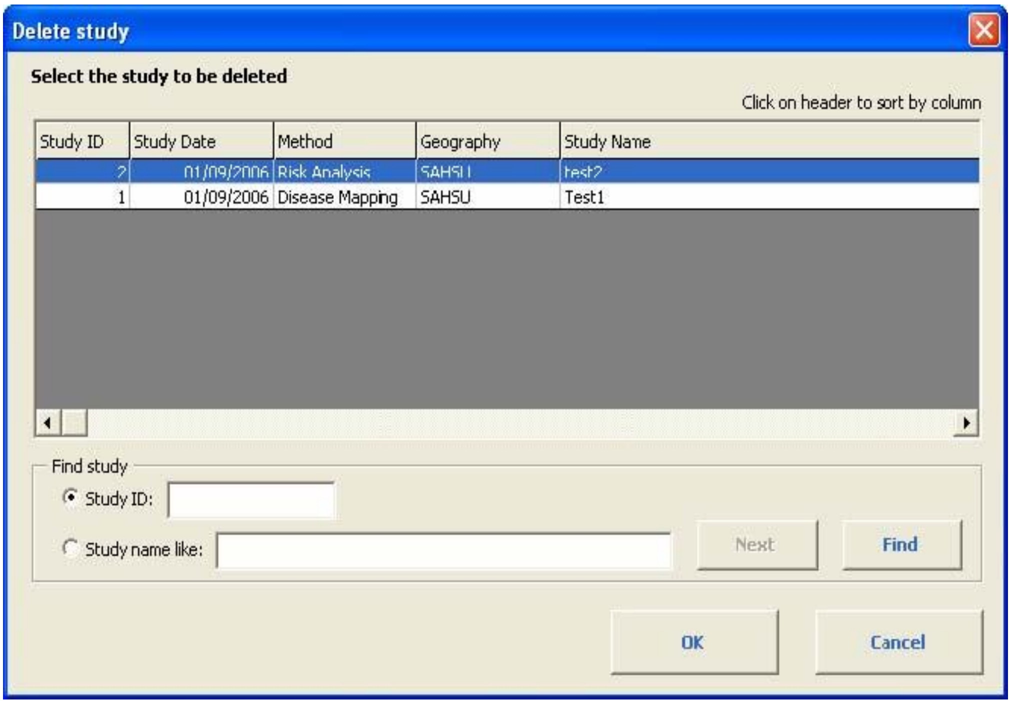
○ Study name like:

Figure 42. Retrieve screen

To retrieve the study once it has been located, click on the study to be retrieved and then click **OK**. To close down the **Retrieve** pop-up box, click **Cancel**.

8.2 Deleting a study

To delete a previously run study open the **RIF ►** menu, and select **Delete previous**. This will bring up the pop-up box **Delete**, which again lists all previously run studies with details of the study ID, study date, method, geography and user defined study name. Studies are sorted by study ID by default, but the list can be re-sorted by any of these parameters by clicking on the appropriate column. There is a **Find study** facility, whereby the study ID or part of the study name can be used to locate any particular study from the list.



Select the study to be deleted

Click on header to sort by column

Study ID	Study Date	Method	Geography	Study Name
2	01/09/2006	Risk Analysis	SAHSU	test2
1	01/09/2006	Disease Mapping	SAHSU	Test1

Find study

● Study ID:

○ Study name like:

Figure 43. Delete screen

To delete a study, click on the study to be deleted and click **OK**. A pop-up box **Deleting study** requires confirmation 'Do you really want to delete this study?' Click **Yes** to delete the study, or **No** to exit the **Delete previous** pop-up box.

Appendices

A Trouble shooting

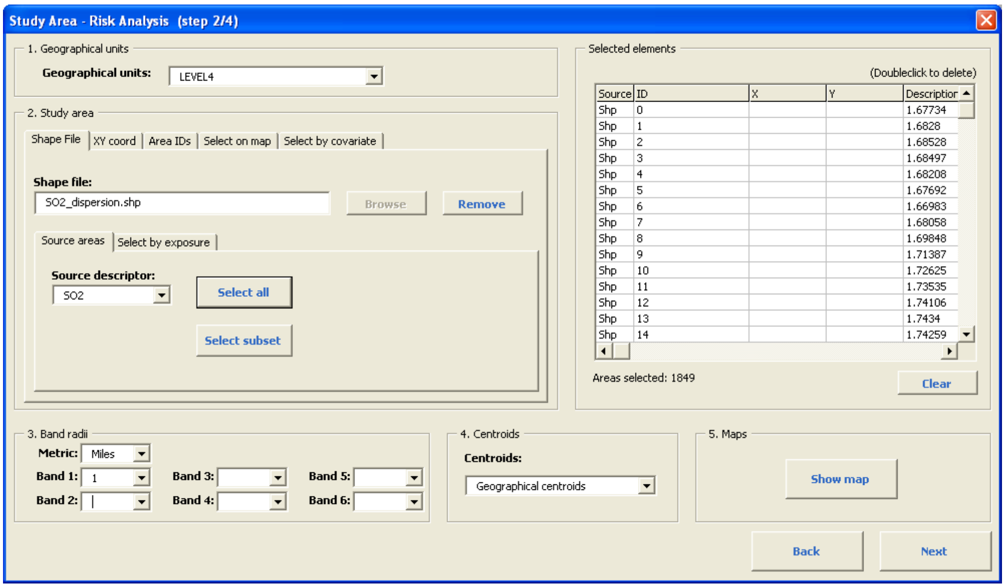
There are some known problems that can occur when running the RIF. These are listed below with a brief indication as to what the problem is, and what should be done to rectify it.

For help with other error messages, contact (rif@imperial.ac.uk).

A.1 Known problems

A.1.1 Overflow when selecting a study area by shapefile

This error occurs when selecting a study area by using distance bands around a source shapefile (See Figure 42).


Study Area - Risk Analysis (step 2/4)

1. Geographical units
Geographical units: LEVEL4

2. Study area
Shape file: SO2_dispersion.shp
Source descriptor: SO2

Selected elements
(Doubleclick to delete)

Source	ID	X	Y	Description
Shp	0			1.67734
Shp	1			1.6828
Shp	2			1.68528
Shp	3			1.68497
Shp	4			1.68208
Shp	5			1.67692
Shp	6			1.66983
Shp	7			1.68058
Shp	8			1.69848
Shp	9			1.71387
Shp	10			1.72625
Shp	11			1.73535
Shp	12			1.74106
Shp	13			1.7434
Shp	14			1.74259

Areas selected: 1849

3. Band radii
Metric: Miles
Band 1: 1
Band 2:
Band 3:
Band 4:
Band 5:
Band 6:

4. Centroids
Centroids: Geographical centroids

5. Maps

Figure 44. Selecting a study area using distance bands around a source shapefile

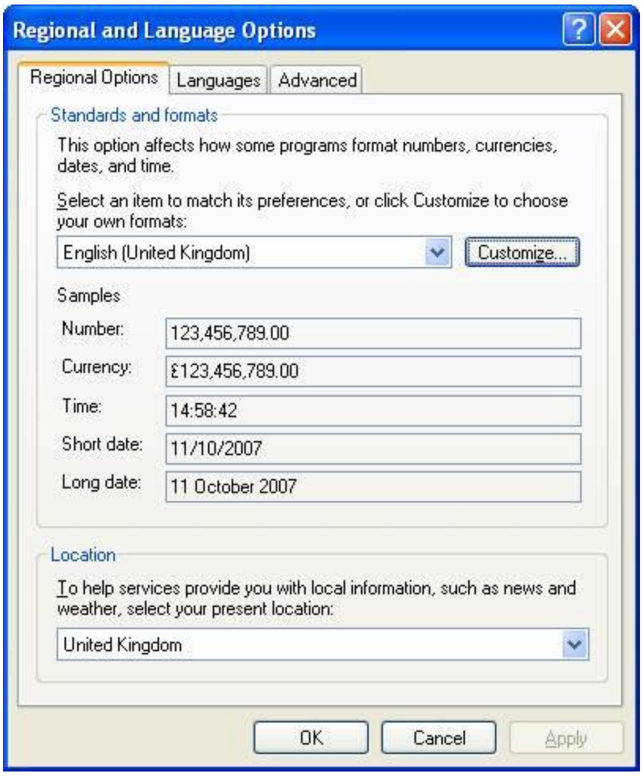
The RIF throws an Overflow-error when trying to buffer around it if the number of features in the shapefile is too large. A suggested workaround is to, if possible, 'merge' some features in the shapefile and use that as the new source shapefile.

A.1.2 The coordinates or measures are out of bounds

When buffering around shapefiles or points, the RIF sometimes throws an error message saying "The coordinates or measures are out of bounds". This error does *not* occur in ArcGIS 9.0.
One workaround is to change the distance band cutpoints very slightly (2 – 10 metres) and try buffering again.

A.1.3 [ODBC Microsoft Access Driver] Too few parameters

This error occurs just after the user clicks "Run" to run the study. This is caused by the fact that the RIF assumes floating point numbers to use a point as a decimal separator and the operating system is set to use commas. To change this in Windows XP, Open the Control Panel->Regional and Language Options->Regional Options dialog in XP (shown in Figure 43).



Regional and Language Options

Regional Options | Languages | Advanced

Standards and formats

This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

English (United Kingdom)

Samples

Number:	123,456,789.00
Currency:	£123,456,789.00
Time:	14:58:42
Short date:	11/10/2007
Long date:	11 October 2007

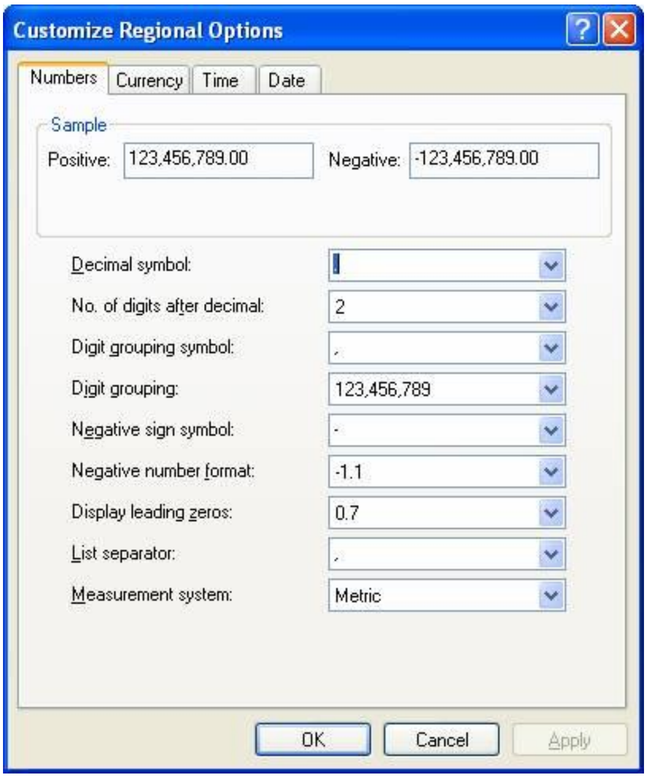
Location

To help services provide you with local information, such as news and weather, select your present location:

United Kingdom

Figure 45. The Regional and Language Options screen in XP (English version)

Click “Customize” and change the “Decimal symbol” to a point and run the RIF again.



Customize Regional Options

Numbers | Currency | Time | Date

Sample

Positive: 123,456,789.00 Negative: -123,456,789.00

Decimal symbol:	.
No. of digits after decimal:	2
Digit grouping symbol:	,
Digit grouping:	123,456,789
Negative sign symbol:	-
Negative number format:	-1.1
Display leading zeros:	0.7
List separator:	,
Measurement system:	Metric

Figure 46. The Customize Regional Options screen.

A.2 Optimising the performance of the RIF in MS Access

Within Microsoft Access there are limitations on the database file size which will limit the use of Microsoft Access as a platform for running very large datasets in RIF[1]. In practice, running a disease mapping study will yield the highest database workload.

To optimise Microsoft Access for use with the RIF, make the following configuration changes:

A.2.1 MaxLocksPerFile

1. Close all connections to the database.
2. Open AdvancedArcMapSettings (typically located in C:\Program Files\ArcGIS\Utilities\AdvancedArcMapSettings.exe
3. Click on the “editor” tab.
4. Change the “JET engine max # of records to calculate:” to the appropriate value (see calculation below).
5. Click Apply and close AdvancedArcMapSettings.

A.2.2 Calculating JET engine max # of records

The MAX_NUMBER_OF_ROWS it is roughly equal to:

In a disease mapping study:
#STUDY_AREAS * #YEARS * #AGE_SEX_GROUPS

In a risk analysis study:
# BANDS * #YEARS * #AGE_SEX_GROUPS * #COVARIATE_COMBINATIONS

Divide these values by 20 to give the ‘JET engine max # of records to calculate’ to enter at step 4 above.

A.2.3 MaxBufferSize

1. Close all connections to the database
2. Open the ODBC configuration dialogue (typically found in Start->Settings->Control Panel->Administrative Tools->Data Sources (ODBC)
3. Select the database you want to configure.
4. Click Configure.
5. Click Options>>
6. Change the Buffer size to 0 and click OK.
7. Close the ODBC configuration tool.
8. The database is now configured and ready to use.

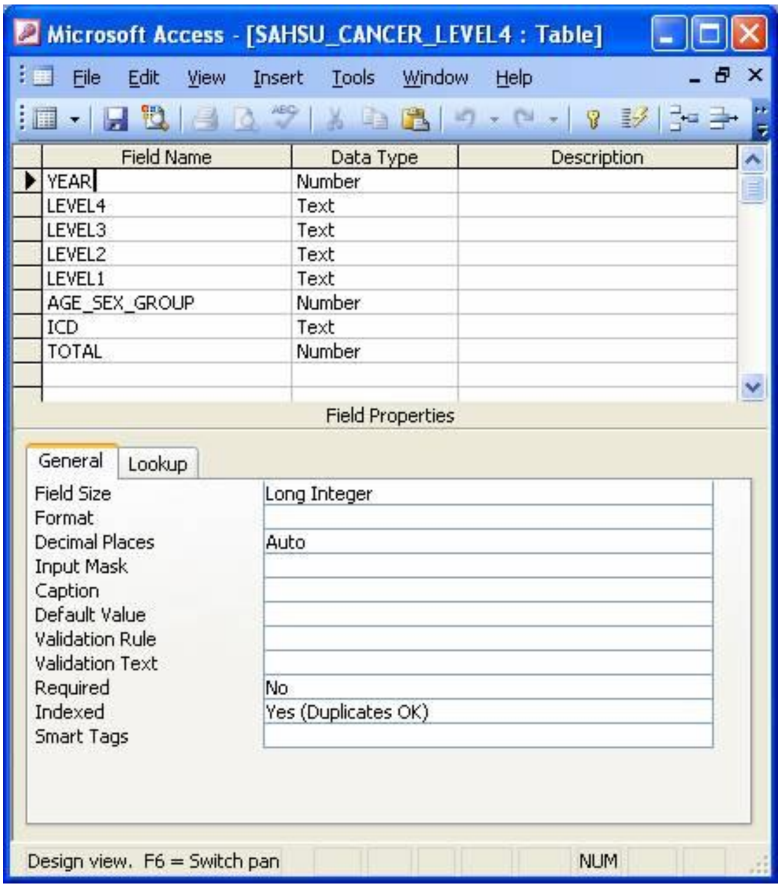
Some common error messages and advised actions regarding the ODBC connection are listed below:

[1] Oracle V10.0 is advised for very large datasets

Error message	Action
[Microsoft][ODBC Microsoft Access Driver] File sharing lock count exceeded. Increase MaxLocksPerFile registry entry.	The MaxLocksPerFile parameter is set too low. Increase its value (a good rule of thumb is to set it to MAX_NUMBER_OF_ROWS_PER_TRANSACTION / 20.
[Microsoft][ODBC Microsoft Access Driver] Invalid argument.	The database file has grown to its limit (~2GB). Compact the database and delete old studies if possible.
[Microsoft][ODBC Microsoft Access Driver] Not enough space on temporary disc.	The MaxBufferSize has been set to a finite value that is too low. Set it to 0 (zero) (see instructions above) to let the operating system optimise the buffer size automatically.
[Microsoft][ODBC Microsoft Access Driver] System resource exceeded.	The MaxBufferSize has been set to a finite value that is too high. Set it to 0 (zero) (see instructions above) to let the operating system optimise the buffer size automatically.

A.3 Creating Indexes

It is strongly recommended to create indexes for your data tables. This will often speed up calculations the RIF does significantly. To add an index on a table, first open the table in Access and switch to the Design view by clicking on the icon. Then click on the icon to enter the Indexes screen. An example of how this screen may look like is shown in Figure 45.


Microsoft Access - [SAHSU_CANCER_LEVEL4 : Table]
File
Edit
View
Insert
Tools
Window
Help
Field Properties
General
Lookup
Design view. F6 = Switch pan
NUM

Field Name	Data Type	Description
YEAR	Number	
LEVEL4	Text	
LEVEL3	Text	
LEVEL2	Text	
LEVEL1	Text	
AGE_SEX_GROUP	Number	
ICD	Text	
TOTAL	Number	

Field Size	Long Integer
Format	
Decimal Places	Auto
Input Mask	
Caption	
Default Value	
Validation Rule	
Validation Text	
Required	No
Indexed	Yes (Duplicates OK)
Smart Tags	

Figure 47. The Indexes screen in Access for the table SAHSU_CANCER_LEVEL4.

Add indexes for all columns that are frequently queried. As a general rule, always create and index for the year column and for all columns containing geographical data (LEVEL4, LEVEL3, LEVEL2). For covariate tables, create an index on all covariate columns. For numerator tables, create an index on the AGE_SEX_GROUP-column and the ICD-column.

B Technical Appendix

B.1 Statistics

The disease mapping function of the RIF allows a user to produce maps of directly standardised rates and of indirectly standardised relative risks, two of the most widely used disease risk indicators in Epidemiology. It also allows smoothing of indirectly standardised relative risks via a Poisson-Gamma models using an empirical Bayes inferential approach (Clayton & Kaldor 1987), to account for some of the instability inherent to this indicator when using small areas. The risk analysis function of the RIF allows the user to calculate directly standardised rates and indirectly standardised relative risks in user defined study area/s, and to test for homogeneity and linear trend in these risks by exposure.

B.1.1.Directly standardised rates

Relevant health outcomes, ‘cases’, occurring over the study period are among residents in each study population to give the observed count. Rates of disease are calculated by linking these events to the underlying populations at risk.

Directly standardised rates are calculated by applying the age, gender and covariate specific disease rates of the study population to the age, gender and covariate make up of the comparison population. The directly standardised rates of two or more study populations (e.g. different ‘bands’ of exposure around a putative source of pollution), can be directly compared, because they are calculated for the same standard population (the comparison population).

$$DSR_i = \frac{\sum_j N_j^* r_{ij}}{\sum_j N_j^*} \times 100{,}000$$ to obtain rate per 100,000 person years

where DSR_i is the directly standardised rate in the study area i

and N_j^* is the population-years at risk in age/sex/SDI strata j in the comparison population

and r_{ij} is the rate of disease D in age/sex/SDI strata j in the study area i

and × 100,000 to obtain rate per 100,000 person years

95% confidence intervals (95% CI) are calculated as a measure of the degree of uncertainty in the estimation of the DSR_i. For areas with more than 100 cases (O_i > 100), the limits of the intervals are calculated as follows:

lower 95% CI = $\frac{DSR_i}{e^{1.96\,SE(\log(DSR_i))}}$

upper 95% CI = $DSR_i e^{1.96 SE(\log(DSR_i))}$

where SE(log(DSR_i)) = $\dfrac{\sqrt{\sum_j (N_j^*)^2 r_{ij}(1-r_{ij})/N_{ij}}}{\sum_j N_j^* r_{ij}}$

[NB These results assume that $\log(DSR_i)$ is approximately normally distributed.]

For areas with less than or equal to 100 cases ($O_i \leq 100$), confidence intervals are obtained by using statistical tables (see below) for the Poisson Distribution (Dobson et al 1991):

lower 95% CI = DSR_i + ($\dfrac{SE(DSR_i)}{\sqrt{O_i}}$ × (LCIP(O_i) - O_i))

upper 95% CI = DSR_i + ($\dfrac{SE(DSR_i)}{\sqrt{O_i}}$ × (UCIP(O_i) - O_i))

where SE(DSR_i) = $\dfrac{\sqrt{\sum_j (N_j^*)^2 r_{ij}(1-r_{ij})/N_{ij}}}{\sum_j N_j^*}$ × 100,000

and LCIP is the lower 95% CI of the Poisson distribution around O_i e.g. 1.088 where O_i = 4

and UCIP is the upper 95% CI of the Poisson distribution around O_i e.g. 10.24 where O_i = 4

B.1.2. Indirectly standardised risks

Relevant health outcomes, ‘cases’, occurring over the study period are among residents in each study population to give the observed count.

Expected numbers of cases in the study area are derived from the user defined comparison population. Cases occurring in the comparison population are located for males and for females in each five year age band and per covariate. Rates of disease in the comparison population are calculated for each gender, age and covariate by dividing number of cases in each group by the total population in each corresponding gender, age and covariate. These disease rates are then applied to the gender, age and covariate make up of the study population/s to get the expected number of cases. For example the rate of mortality from kidney disease in 60-64 year old males in one covariate value derived from the comparison population is multiplied by the number of 60-64 year old males with the same covariate value in the study population, and so on for each age, gender, covariate value and study population.

The observed number of cases divided by the expected number of cases gives an indirectly standardised risk for the study population/s. The risks obtained for two or more study populations (e.g. different ‘bands’ of exposure around a putative source of pollution), should not be directly compared as they are not based on the same standard population (i.e. the age, gender and covariate make up between the populations being compared are not exactly the same).

$$ISRR_i = \frac{O_i}{E_i}$$

where $ISRR_i$ is the indirectly standardised rate ratio in area i
O_i is the number of observed cases in area i
E_i is the number of expected cases in area i

and $E_i = \sum_j N_{ij} r_j^*$

where N_{ij} is the population-years at risk in age/sex/SDI strata j in area i

and r_j^* is the rate of disease D in strata j in the **reference region**

Again 95% confidence intervals are calculated. For areas with more than 100 cases (O_i > 100)

lower 95% CI = $\dfrac{O_i/E_i}{e^{1.96(\sqrt{1/O_i})}}$

upper 95% CI = $O_i/E_i \times e^{1.96(\sqrt{1/O_i})}$

[NB These results assume that log $ISRR_i$ is approximately normal.]

For areas with less than or equal to 100 cases (O_i <= 100), confidence intervals are obtained by using statistical tables for the Poisson Distribution (calculated by Chi-squared method):

Observed	Lower 95% CI	Upper 95% CI
0	0	3
1	0.025	5.57
2	0.242	7.22
3	0.618	8.76
4	1.088	10.24
5	1.620	11.65
…	…	…

So where O_i is 4,
the lower CI of the $ISRR_i$ is given by 1.088/E_i
the upper CI of the $ISRR_i$ given by 10.24/E_i

B.1.3 Empirical Bayes Analysis

Empirical Bayes analysis using the so-called Poisson-Gamma model has been integrated into the system via a PL/SQL script, which applies the empirical Bayes Poisson-Gamma smoothing model by Clayton and Kaldor (1987) to the relative risks calculated in the disease mapping analysis. This hierarchical approach provides more precise estimates of relative risk and more accurate assessments of significant changes than standard methods, by accounting for differential variability in the data.

Specifically, if λ_i represents the relative risk in area i, the Poisson-Gamma model assumes

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$
$$\lambda_i \sim \text{Gamma}\ (\alpha, \beta)$$

Using the empirical Bayes approach, estimates of the relative risks are obtained through the following iterative procedure:

1. Start with initial values for the relative risks λ_i. For instance $\lambda_i = O_i / E_i$.
2. Obtain estimators $\hat{\alpha}$ and $\hat{\beta}$ using equations

$$\frac{\hat{\alpha}}{\hat{\beta}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}_i \quad \text{and} \quad \frac{\hat{\alpha}}{\hat{\beta}^2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(1 + \frac{\hat{\beta}}{E_i}\right)\left(\hat{\lambda}_i - \alpha\big/\beta\right)^2$$

3. Obtain new estimated values for the relative risks

$$\hat{\lambda}_{\mathrm{i}} = \frac{O_i + \hat{\alpha}}{E_i + \hat{\beta}}$$

4. Repeat steps 2 and 3 until estimated values for $\hat{\alpha}$ and $\hat{\beta}$ do not change significantly.

B.1.4. Running a disease mapping model in WinBUGS

After running a disease mapping study in the RIF, it is possible to fit a different disease mapping models using WinBUGS. When running a disease mapping study using WinBUGS, the model by Besag et al (1991) is fitted. This model is a Bayesian hierarchical model. Using the same notation as above, its specification is as follows:

- in the first layer, the number of observed cases in each area is assumed to follow a Poisson distribution with mean the expected counts times the relative risks. These in turn are decomposed as the sum of an overall intercept plust two random effects, one spatially structured and one unstructured.

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$
$$\log(\lambda_i) = \mu + u_i + v_i$$

- in the second layer, priors are assigned to the intercept and the spatially unstructured and structured random effects: a flat prior for the intercept, a normal distribution for the unstructured random effects, and an intrinsic conditional autoregressive model (CAR) for the spatially structured random effects – the latter uses an neighbourhood structure based on adjacency.

$$\alpha \propto 1$$

$$u_i \sim \text{Normal}(0,\ \sigma^2_{\text{u}})$$

$$v_i \mid \mathbf{v}_{-i} \sim \text{Normal}\left(\sum_{j\sim i} v_j / n_i, \sigma^2_v/n_i\right)$$

- Finally, prior distributions are in turn assigned to the hyperparameters

$$\sigma^2_u, \sigma^2_v \sim \text{IGamma}\ \ (0.5, 0.0005\ )$$

The RIF runs three chains of 20,000 simulations starting from as many different sets of initial values for the parameters. The first 10,000 are discarded and rest are kept. Inference for each parameter is therefore based on a sample of 30,000 simulations. The RIF does assess convergence of the simulations. Several convergence checks (history, autocorrelation and the Brooks-Gelman-Rubin diagnostic) for the smoothed relative risks are invoked in WinBUGS, and the corresponding output is saved in the file *log.odc* if the user wants to look at them.

B.1.5.Testing for homogeneity in the relative risks

In the risk analysis module, areas are grouped into distance/exposure bands, and rates and relative risks as well as 95% confidence intervals are calculated for each band, as described above. The aim of this type of analysis is to assess whether there exists some association between distance/exposure and disease risk. This can not be addressed by comparing the risk and confidence interval in each band with those of the other bands, as this would imply multiple testing, that is, we risk identifying truly non-significant distance/exposure as significant with a probability higher than the nominal 5% of each pairwise band comparison test. One way to overcome this problem is to carry out a global test of homogeneity of bands to ensure that the overall probability of type I error equals the nominal significance level of the test. Chi-square tests for homogeneity and linear trend (Breslow and Day, 1987; Bland, 2000) have been implemented in the RIF to test global association between a distance/exposure and relative risks:

- *Homogeneity test.* This tool is intended to test whether the relative risks are homogeneous across all bands or not.
  In general, if the user selects *$K \geq 2$* bands, after calculating observed O_k and expected counts E_k in each band *$i = 1, . . . , K$* for each adjustment scenario, the chi-square statistic and the sampling distribution under the null hypothesis are:

$$\chi^2_{\text{hom}} = \sum_{k=1}^{K} \frac{(O_k - E_k^*)^2}{E_k^*} \sim \chi^2_{K-1}$$

where $E_k^* = E_k \cdot O_+ \big/ E_+$, and $O_+ = \sum_{k=1}^{K} O_k$, $E_+ = \sum_{k=1}^{K} E_k$.

- *Linear trend test.* The aim of this method is to test the null hypothesis of all risks being homogeneous across all bands versus the alternative of there being a linear trend in them.

  To carry out this test, we again need to know the number of observed *O_k* and expected cases *E_k*, and also the average exposure value *Z_k* in each band *i = 1, ... ,K (K≥3).* Then the expression for the chi-square statistic and the sampling distribution under the null are

$$\chi_{lt}^2 = \frac{\left(\sum_{k=1}^{K} Z_k \left(O_k - E_k^* \right) \right)^2}{\sum_{k=1}^{K} Z_k^2 E_k^* - \left(\sum_{k=1}^{K} Z_k E_k^* \right)^2 / O_+} \sim \chi_1^2$$

Notes:
For distances, the RIF uses the mean distance from all selected area centroids by exposure band for Zk.
For exposure variables the RIF uses the mean value from all selected exposure areas, by exposure band for Zk.
For other variables (e.g. for categories that are ordered but not based on numerical values (e.g., low, middle, high)) the RIF assigns Zk=k
Users should be aware that the results of the test are sensitive to the choice of values for Zk, and thus results should be carefully interpreted.

References

Bland, M. (2000) *An introduction to medical statistics,* (Third Edition). Oxford University Press, Oxford.

Breslow, N.E. & Day, N.E. (1987) *Statistical Methods in Medical Research Vol. II.* International Agency for Research on Cancer. Lyon.

Clayton, D.G. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**:671–681.

Dobson AJ, Kuulasmaa K, Eberle E, Scherer J. (1991) Confidence intervals for weighted sums of Poisson parameters. *Statistics in Medicine* **10**: 457-462.

B.2 Centroids

In a risk analysis, the ideal study population would include all people living within the selected study area, and no one out side this area. Unfortunately, population data are rarely available at the resolution of the individual (where this ideal could be achieved), and so the RIF uses buffers to select areas (and associated populations) falling inside the user defined area using ‘centroids’.

There are different ways of defining the centroid of a small area, the geographical centroid, and the population-weighted centroid. The geographical centroid refers to the geographical mid point of the area. The weighted centroid takes into account the population distribution within the small area, and so represents the location of the majority of the population.

Figure B.2.1 shows the shapefile of the TCE contaminated ground water plume at Hill Air Force Base, Utah and an example selection of a potentially exposed population using a 1km buffer around this potential pollution source. This example demonstrates how different areas and therefore populations are selected when using the geographical versus population-weighted centroids. A total of 30 census block areas with an area of 63.63km^2 are selected for the study when the geographical centroids are used to select the exposed census block groups. Using population weighted centroids this is increased to 31 census block groups but the total area is reduced to 42.54km^2.

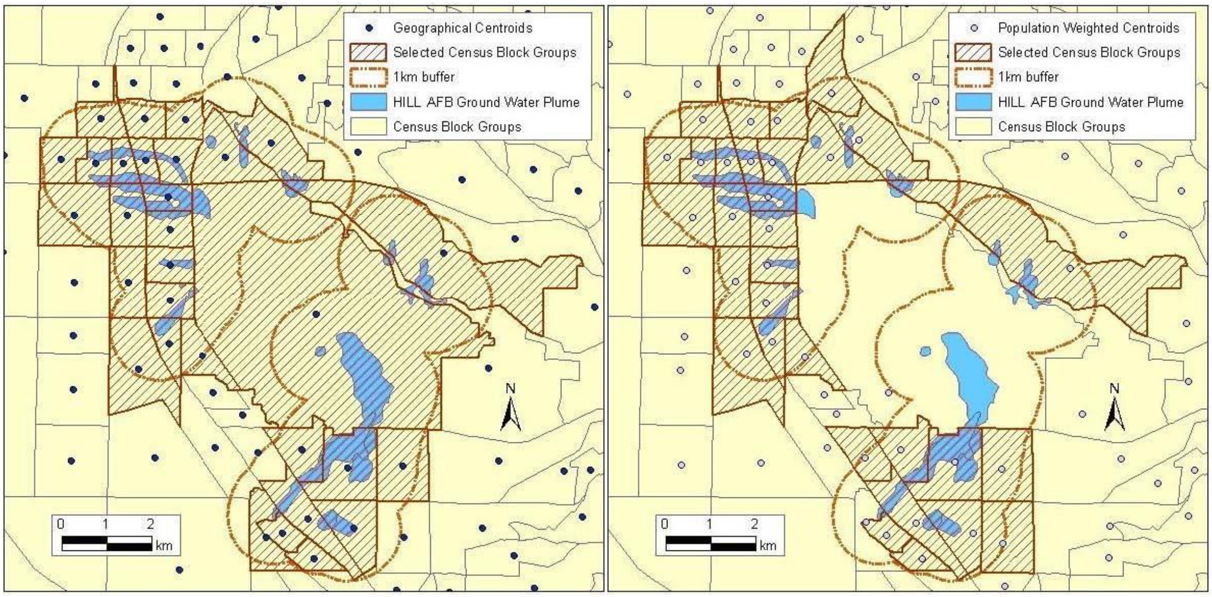

Geographical Centroids
Selected Census Block Groups
1km buffer
HILL AFB Ground Water Plume
Census Block Groups
N
0 1 2
km
Population Weighted Centroids
Selected Census Block Groups
1km buffer
HILL AFB Ground Water Plume
Census Block Groups
N
0 1 2
km

Figure 48. Geographical versus population-weighted centroids – This figure shows the different ‘exposed’ populations selected when using geographical or population weighted centroids falling within a 1km buffer around TCE contaminated ground water plumes at Hill Air Force Base, Utah.

B.3 Covariates

B.3.1. Covariate resolution

The RIF can read covariate information at the geographical area level (ecological level). The covariate resolution has to be the same as the study area resolution.

B.3.2. Adjustable and exposure covariates

Covariate tables can contain adjustable and exposure covariates, and in fact there is no distinction between adjustable covariates and exposure covariates in the RIF other than that numeric covariates cannot be adjusted for at the area level.

With respect to adjustable covariates, there is no theoretical limit to the number of covariates the RIF can adjust for. However, the user should note that for any reasonable dataset, data sparseness may well become an issue for even with only a small number of covariates.

With respect to exposure covariates, these can be numeric or categorical. Categorical exposure covariates may be either ordinal or nominal. Exposure covariates, regardless of type can not be time dependent. Numeric exposure covariates can be n-tilised into 2-7 bands in the RIF.

When testing for association between an exposure covariate and disease risk the RIF does distinguish between ordinal and categorical nominal covariates. For categorical nominal exposure covariates, the linearity test is not run, neither is it run for risk analyses with less than three exposure bands.

The statistic of the chi-square test for linear trend is based on points (Zk, RRk), where RRk is the relative risk in distance/exposure band k=1,..,K, and Zk is a representative value of the distance/exposure for areas in band k. For distances, the RIF uses the mean distance from all selected area centroids by exposure band; for exposure variables the RIF uses the median value; otherwise the RIF assigns Z_k=k.

B.3.3. Covariate tables

The covariate tables contain links between areas and covariate values. There are no numerator or denominator data in the tables and there can be only one value for each covariate per area and year.
For these tables, there can be one table for each geographical level for which there are covariate data (e.g. SAHSU_COVARIATES_LEVEL4 and SAHSU_COVARIATES_LEVEL3 hold covariate information at their geographical levels respectively). Please note that these tables are not geographically aggregated; SAHSU_COVARIATES_LEVEL3 still has one and only one covariate value per year per area. Covariate data at different geographical levels (e.g. LEVEL4 SES and LEVEL3 INCOME) cannot be used within the same study.

B.3.4. Data resolution

The RIF is able to read numerator and denominator data aggregated to any geographical level, but covariate information has to be stored internally in numerator/denominator tables at a lower resolution than the covariate data due to the lack of information about the combination of multiple covariates.

Example 1: comparison area and covariate data at highest resolution
Here is a typical example of what a denominator/covariate table looks like in the RIF database:

(1) DENOMINATOR table at LEVEL4

YEAR	**LEVEL4**	**LEVEL3**	**POP**
1995	A1	A	700
1995	A2	A	600
1995	A3	A	500
1995	B1	B	900
1995	B2	B	800
1995	B3	B	800

(2) COVARIATE table at Level4

YEAR	**LEVEL4**	**LEVEL3**	**COV1**	**COV2**
1995	A1	A	1	6
1995	A2	A	2	5
1995	A3	A	2	5
1995	B1	B	2	5
1995	B2	B	1	5
1995	B3	B	1	6

Using this denominator table (1) along with covariate table (2) it is possible to calculate how many cases or people there are for a certain value of the covariate. The result of joining these tables together is:

(3) RESULT-table

YEAR	**LEVEL4**	**LEVEL3**	**COV1**	**COV2**	**POP**
1995	A1	A	1	6	700
1995	A2	A	2	5	600
1995	A3	A	2	5	500
1995	B1	B	2	5	900
1995	B2	B	1	5	800
1995	B3	B	1	6	800

Example 2: comparison area at lower resolution than covariate resolution
Instead, if we want to use a lower resolution geography than that at which the covariates exist, our DENOMINATOR table will typically look like this:

(4) DENOMINATOR table at Level3

YEAR	**LEVEL3**	**POP**
1995	A	1800
1995	B	2500

There is now no way this table can be used together with the covariate table (2) to produce a table similar to the RESULT table (3) which is what is required.

To overcome this problem, the highest resolution geography (level4) table (1) can be used, however this is an inefficient in terms of data processing. Alternatively, covariate information can be stored internally in the lower resolution numerator/denominator tables (5).

(5) DENOMINATOR, grouped to LEVEL3 with internally stored covariates

YEAR	**LEVEL3**	**COV1**	**COV2**	**POP**
1995	A	1	6	700
1995	A	2	5	1100
1995	B	1	5	800
1995	B	1	6	800
1995	B	2	5	900

When using numerator/denominator data of a lower resolution than the covariate data, the RIF will read the covariate information internally from the appropriate grouped tables.

B.4 SQL statements and clauses

SQL statements (a complete command) and SQL clauses (a sub-section of a statement) can be used at various stages to help define a RIF study.

SQL *statements* can be used to select the study and comparison areas using the **SQL** tab in the **Study Area** and **Comparison Area** screens. SQL statements take the form:

```
select <column1> from <table> where <column2> = <’x’>
```

For example, to select the level 2 area ‘Cobley’ from Sahsuland, this statement would be:

```
select zone_id from Sahsu_level2 where name = ‘Cobley’
```

SQL *clauses* can be used when selecting the specific health end point of interest from the **Investigation Details** screen when using either the **Choose a predefined group** or **Enter an SQL clause** option. The keyword "WHERE" should NOT be included, as it will be added automatically by the RIF when it creates the full SQL statement.

An SQL query is built up using logical operators AND, OR and filter conditions (LIKE, =, IN and BETWEEN). As such, SQL queries can be used to extend the scope of the study, for instance to utilise secondary end point code fields or cancer histology codes (if available).

The RIF will check the syntactical accuracy of the SQL statement entered, and will return a warning if the clause is invalid.

The examples given below relate to International Classification of Diseases (ICD) codes, as the example numerate dataset Sahsuland is coded in ICD9 and 10; however the same principals apply no matter how the numerator data are coded.

The LIKE condition is a simple pattern match: e.g. `icd LIKE 'C34%'`. The '%' wildcard matches any number of characters. It can be used as the first or last character in the character string, e.g. 'wh%' finds what, white, and why, but not awhile or watch. '_' matches any single alphabetic character, e.g. 'b_ll' finds ball, bell, and bill.

The = (EQUALS) operator is a simple equality. Wildcards will have no effect.

The IN predicate signifies a list of valid (ICD) codes, e.g. `icd IN ('7871', '7873')`

The BETWEEN predicate signifies a range of values e.g. `icd BETWEEN ('7871' AND '7874')`

Note:

- The data type of the LIKE pattern can only be used for textual data. Textual data MUST be enclosed in single quotes.
- The match is CASE sensitive.
- Beware of operator precedence - always bracket logical sections together, e.g. `(icd_code1 LIKE 'C34%' OR icd_code1 LIKE 'C35%') OR (icd_code2 LIKE 'C34%' OR icd_code2 LIKE 'C35%')`. Be very careful when mixing AND and OR in the same clause - it is very easy to create a syntactically correct but logically impossible filter: `icd_code1 LIKE 'C34%' AND icd_code1 LIKE 'C35%'`.
- The "~" (tilde) is not a valid wildcard. The effect of tilde is purely alphabetic, i.e. `icd BETWEEN '7870' and '7889'` will return 18 distinct 4 character ICD codes, whereas `icd BETWEEN '7870' and '788~'` will only return icd codes in the 788% category as "~" follows the numbers alphabetically.
- Unexpected (and different) results might occur if your (ICD) codes are five character or contain non numerics in position 4. It is logically safer to write:

  ```
  icd IN ('787', '788') OR (icd BETWEEN '7870' AND '7889')
  ```

The full list of Oracle SQL conditions may be found at: http://download-east.oracle.com/docs/cd/B19306_01/server.102/b14200/conditions.htm#i1066777

Access supports both SQL-89 and SQL-92 standards depending on what type of file is opened, and on what compatibility is set. The wildcards supported by Access include all those detailed above. For more information see: http://office.microsoft.com/en-us/assistance/HP051881851033.aspx#AboutANSI

B.5 Basic ArcGIS functions

These icons from the RIF and ArcGIS tool bars are likely to be useful when viewing and manipulating the RIF output, and are very easy to use. More details can be accessed from the help files in ArcGIS, from the ArcGIS user guides, or from www.esri.com.

RIF icons:

Add point to the map (e.g. add a point source for risk analysis).

View contextual information (e.g. to view rates and risks by small area from the disease map).

ArcGIS icons:

Add data (e.g. shapefile)

Zoom in

Zoom out

Fixed zoom in

Fixed zoom out

Pan

Full extent

Go back to previous extent

Go to next extent

Select features

Select elements

Identify results

Find

Measure

Layout view

Data view

Refresh

B.5.1. Changing relative risk categories:

The categories for relative risk have been pre-defined; however, these can be easily altered using the following steps:

- Double click on the risk or rate layer in the Display tab to the left of the map to bring up the **Layer Properties** window
- Click the tab “Symbology”
- In the classification box click “Classify”
- The both classification method and number of classes can be changed
- If manual selection is required the breakpoints can be typed in

B.5.2. Changing symbols/colours:

In the Display tab, double click on the layer to be altered to bring up the **Laye**r **Properties** window. Use the symbology tab to view the options for changing the symbol size, shape, colour, colour ramp as appropriate (e.g. see ColorBrewer, a web tool for selecting color schemes for thematic maps http://www.personal.psu.edu/cab38/ColorBrewer/ColorBrewer.html).

B.5.3. Open attribute table:

While the button ℹ can be used to identify any element of the map, it is sometimes desirable to open up the attribute table behind the mapped layer. This can be achieved by clicking the right hand mouse button, and selecting Open Attribute Table from the menu.