The Rapid Inquiry Facility (RIF)
Version 4.0
How to use the RIF 4.0 client

Authors (2017):

Parkes, B., Morley, D.

Small Area Health Statistics Unit (SAHSU)
MRC-PHE Centre for Environment and Health
Department of Epidemiology and Biostatistics
School of Public Health
Imperial College London
Medical Faculty Building
St Mary's Campus, Norfolk Place
LONDON W2 1PG

Website www.sahsu.org

Contents

1. Introduction to the RIF

1.1 Purpose

The Rapid Inquiry Facility (RIF) is an automated tool to allow epidemiologists to rapidly address epidemiological and public health questions using routines collected health and population data.

The RIF can perform risk analysis around putative hazardous sources, and can be used for disease mapping. It generates indirectly 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…

This manual describes version **4.0** of the RIF (last update XXX). Supports browsers… versions…

1.2 Principal Features

- The system is designed using 3 tier architecture. The client runs entirely within the user's browser and does not require installation of any software on the client's machine. All data is stored in a secure database on the server with user access, security and data processing performed by the middleware, also running on the server. For more information see….
- 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.
- Within the risk analysis tool, the RIF performs test on the relative risks to assess for homogeneity and linear trend with exposure.
- Within the disease mapping tool, the RIF can perform empirical Bayes smoothing and full Bayes smoothing using the r-INLA library assessed from the middleware.
- The RIF can export data for further analysis in other (statistical) software packages such as SaTScan and WinBUGS.

1.3 Input Facilities

Provided with the RIF is a Data Loader tool that allows users to import their health and GIS data.

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 and SaTScan, as well as to Microsoft Excel for further processing.

1.5 Scope of this Manual

This manual explains how a user, typically an epidemiologist, would use the RIF client to set up a new study, run the study and examine the results. In addition there is an appendix covering the statistical methods used by the RIF when a study is run.

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

Beal 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 Perspectives*. Volume 116, number 8.

2.1 Disease mapping or risk analysis

The intention is to provide two types of study in the RIF, disease mapping and risk analysis. Initially the RIF v4.0 will only have disease mapping available.

The disease mapping approach can be used to visualise mortality/morbidity rates and risks across and 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.

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.

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 a 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 pot 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 name 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 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 EAs were grouped into zones of similar size to census tracts, but in a different spatial arrangement, it is likely to produce different statistical results 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

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 area 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 aid interpretation of the disease mapping output, the RIF can perform both empirical and full Bayesian 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 and 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 et al., 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 rate 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 and Best, 2003). Given that the it is intended for use at smaller geographical levels, the RIF uses indirect standardisation to calculate SMRs and does not currently perform direct standardisation.

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 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.

A full description of the statistical methods performed by the RIF is given in appendix A.

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. The map displays in the RIF can be configured with a wide range of base maps available. Additionally; it may be preferable for the user to export the data in order to import it into a GIS system with greater data symbolisation capabilities.

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 (Elliot and 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 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 the 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

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

Goldman DA & Brender JD. 2000. Are standardized mortaility 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.

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(10):40-46.

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

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

Richardson S, Thompson 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 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 and may not reflect patterns observed in reality.

The data consist of:

- Population data (by five year age group by gender[1]), for the period 1989-1996.
- Cancer incidence data for the period 1989-1996.
- Covariate data[2] on socio-economic status, ethnicity, and proximity to Toxic Release Inventory (TRI) 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 32,869 km[2]. The area of Sahsuland (Figure 1) uses for 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 LEVEL 2m 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 Table1).

Table 1. The census areas in Sahsuland

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)

All level IDs are stored as text values with LEVEL1 areas using two characters, LEVEL2 use 3 characters, which is joined with the LEVEL1 unique ID to make a LEVEL 2 unique ID of 2 and 3 characters separated by a dot. The Level 3 units use a 6 character value and the LEVEL4 is a single

---

[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+

[2] The RIF can handle ecological level covariate data

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 XXX.

Screen shots and examples in this manual are based on this Sahsuland data. Descriptive statistics of Sahsuland can be found in appendix B.

3.2 Logging in

Your RIF administrator should provide you with your user name, password and the correct URL to access the login page of your RIF installation. Type the URL in the address bar of your web browser and log in using your username and password.
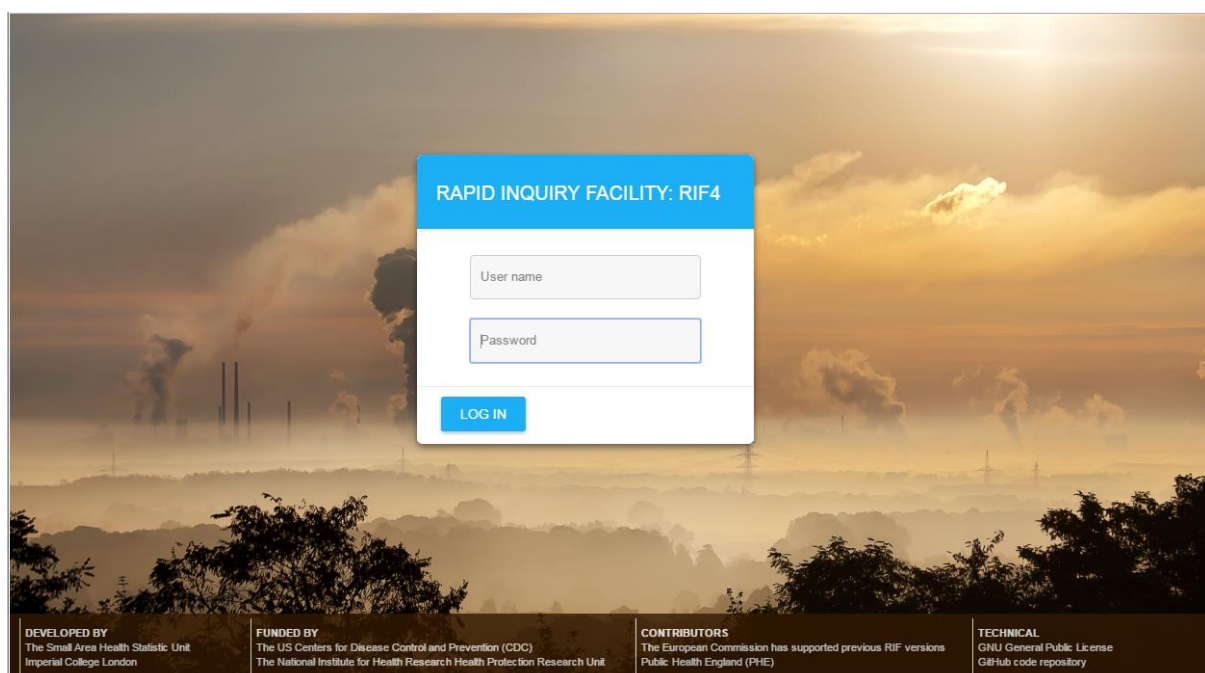

Figure XXX. The RIF login screen.

3.3 RIF mapping tools

The RIF uses an internet browser based map viewer to display and select study areas and to visualise results. These maps work in the same way as conventional map containers (e.g. Google Maps), the main difference being they use open source map data.

The following controls are common to all RIF maps. Other tools are specific to certain RIF functionality and these will be dealt with in the relevant sections.

+   **Zoom in:** Zoom map in.

−   **Zoom out:** Zoom map out.

Quick export map: Save the current map in view as a png file.

Zoom to selection: Zoom the map to the currently selected districts(s).

Zoom to study extent: Zoom the map to fit all districts used in this RIF study.

Zoom to full extent: Zoom the map to fit all districts in the current geography.

Clear selection: Deselect all currently selected districts.

Transparency: Change the transparency (opacity) of the district layer being mapped

Enter address: Zoom the map to a place name, geographical feature etc.

Full screen: Display the map in full screen mode (Esc to exit)

?   **Attribution:** Attribution (source, copyrights) information for the map layers

Base map: Opens the base map selection where the base map can be changed or removed.

Figure XXX. Base map settings.

4. Running a new RIF study

To create, then run a new RIF study, five steps will need to be completed:

- Enter study details such as 'geography' and study type
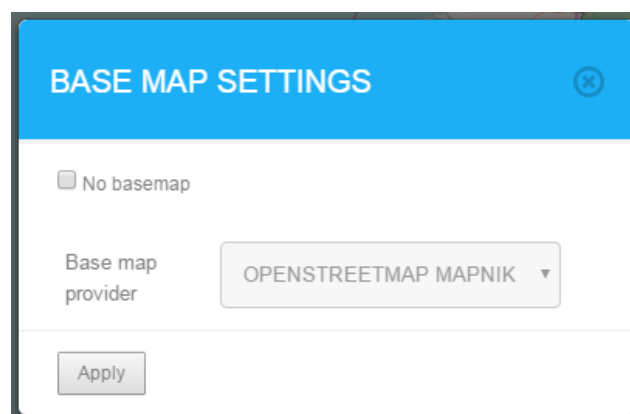- Defining the study area
- Defining the comparison area
- Set the investigation parameters
- Decide the statistical methods to be used

These steps can all be completed under the **Study Submission** tab. At any point all the details of the study can be cleared by clicking the **reset** link.

Figure XXX. Study Submission screen

4.1 Study details

The study details can be completed using the fields along the top of the **study submission** tab

Study Name. The study must be given a name which is types into the **study name** field. The name cannot exceed 20 characters in length.

Health Theme. This is defined during data loading as a means to group your relevant data sets together for ease of use. This will usually relate to a disease type, e.g. cancers.

Geography. Select the appropriate geography from the drop-down list of geographies (this will usually relate to the either the country in which your study is for or to a predefined representation of districts e.g. tracts, wards). The list consists of all the available geographies which you can access.

Numerator. Select the health outcome you are interested in mapping.

Denominator. The data to be used as a denominator. This cannot be changed and is auto-selected depending on which numerator table is being used. Relevant numerator-denominator pairs are decided in the data loading process.

4.2 Study area

Clicking the **study area** link will load the study area selection screen
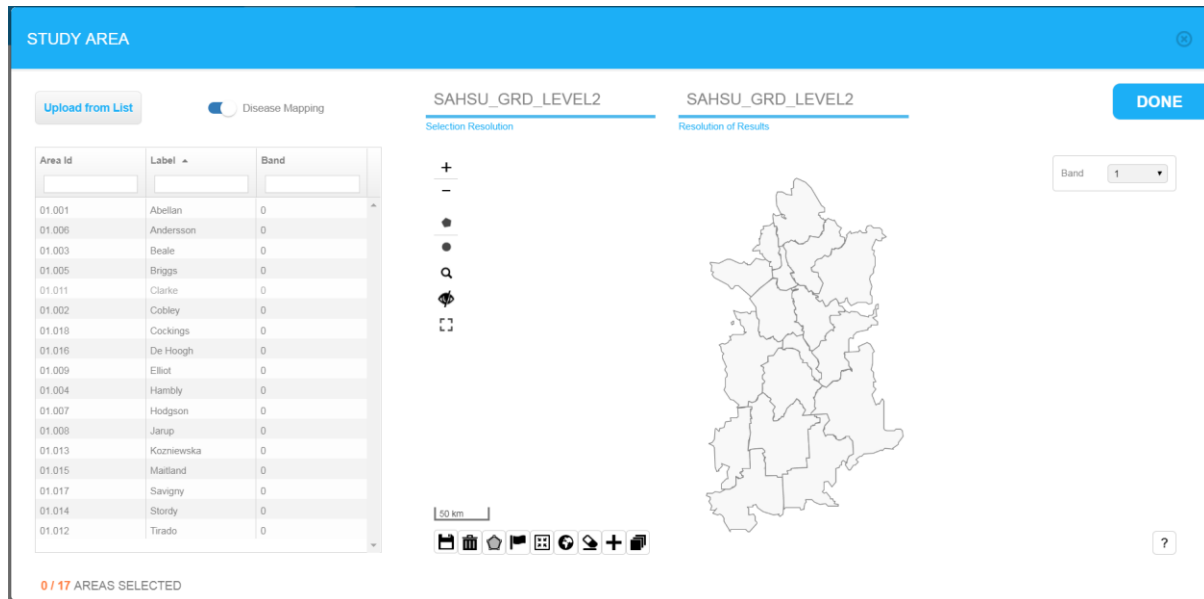
Figure XXX. Study area selection screen

The first thing to do is to set whether this is a **disease mapping** study or a **risk analysis** study using the switch at the top-left of the window. (See section XXX for more information on these study types).

Initially the whole geography is displayed in the map area at the default level of resolution. The purpose of this screen is to select which areas of the geography are to be investigated. For risk analysis, one to six bands (of multiple districts) may be specified. Disease mapping studies are not banded in this way, so only one selection band is available as default.

Area selection/deselection can be performed in a number of ways:

- Use the band drop-down to specify the band number
- Clicking directly onto the area in the map
- Selecting areas from the list displayed on the left side of the screen
- Selecting areas by defining a **freehand polygon** using the ⬠ button
- Select area within **concentric bands** using the ● button.

Note that selections using the boundary of another polygon based on intersection with a district's centroid. The ⚑ icon will toggle visibility of these centroid locations off and on.

The ✚ icon will select all the districts within a geography

By clicking 'Upload from list', a comma deliminated (csv) file can be used to select a predefined list of districts and bands. This must have just two columns ID and Band.

A zipped shapefile can also be used to define study areas. The icon brings up the open shapefile dialogue. The file can be points are polygon (see below), but must be in a zipped folder with extension .zip. Once displayed on the map, the uploaded layer can be removed with the icon.

- When loading point shapefile (e.g. incinerator locations), radii for exposure band(s) in metres need specified.
- When loading a polygon shapefile (e.g. output from an exposure model). Selections can be defined by the full extent of the areas with the shapefile, by a cut-off from an attribute within the file (e.g. a threshold value for a pollutant). This may be multiple (descending) cut-offs for risk analysis according to bands. In addition, in the case of a risk analysis study, if the polygons have a band attribute, this may be specified as well.

Clicking 'done' will store the selected study regions which define the study area and return the user to the study submission tab. The study area 'tree' should now be coloured to indicate that this part of the study submission is complete.

4.3 Comparison area

Clicking the **comparison area** tab on the study submission tab will load the comparison area selection screen for defining areas (populations) for the calculation of indirectly standardised risks. The comparison area screen is very similar to the study area selection screen with all the same methods of selection study regions that will for the comparison area. A comparison area is not banded in the same way as the study area, so by default only one selection band is possible. Note that the type of study can only be defined via the study area window.

Clicking 'done' will store the selected study regions which define the comparison area and return the user to the study submission tab. The comparison area tree should now be coloured to indicate that this part of the study submission is complete.

4.4 Investigation parameters

Clicking the **investigation parameters** link brings up the investigation parameters selection screen.

Figure XXX. Investigation parameters screen.

ICD10 codes of interest can be selected from the list and search box on the left hand side of the screen. Selected codes appear in the lower right side of the screen. An investigation name can be edited in the investigation name field. The **covariates** section allows the user to select male, female or both sexes and define one further covariate field. The range of ages and years the study is to cover are chosen in the **age range** and **year range** sections respectively.

Clicking 'done' will store the selected parameters and return the user to the study submission tab. The investigation parameter tree should now be coloured to indicate that this part of the study submission is complete.

4.5 Statistical methods

Clicking the statistical methods link brings up a screen allowing selection of the statistical methods the RIF will for the disease mapping process. By default the RIF always calculates the indirectly standardised rates and the relative risk ratios as well as performing Empirical Bayesian Smoothing. Either select one of the available procedures or the option not to apply smoothing if you do not to run an additional Bayesian method.

Details of the basic statistical methods and the full Bayesian smoothing options along with external references are included in the Technical Appendix.

Clicking 'done' will store the methods selected and return the user to the study submission tab. The statistical methods tree should now be coloured to indicate that this part of the study submission is complete.
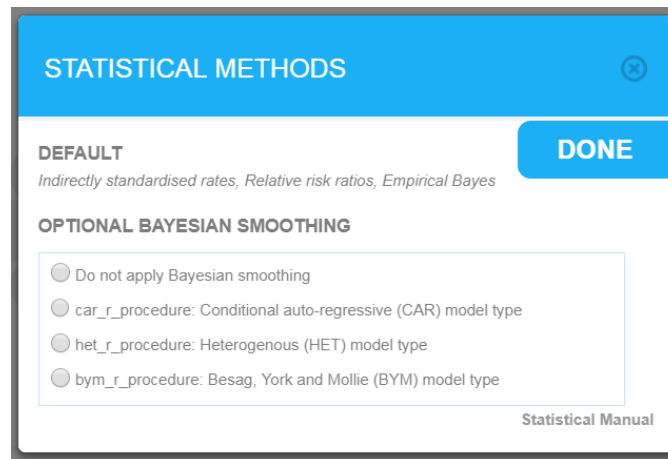
Figure XXX. Statistical methods selection screen.

4.6 Saving and reloading studies

At any point during the process of defining a new study submission, the details of the study can be saved locally on the user's machine by clicking the **save study** link. This brings up a **save as** dialog box allowing the user to select a local folder and file name, then click 'save' to write a local copy of the study.
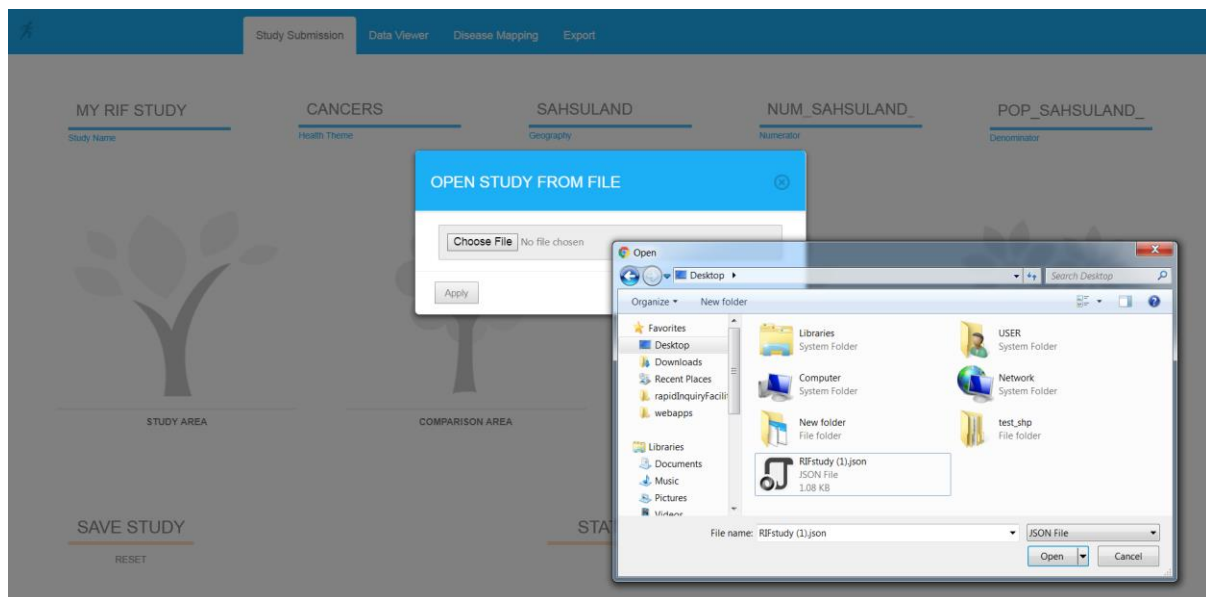
Figure XXX. Open from file dialog box for reloading a study during the submission process

At a later date the user can load the study by clicking the **open from file** link and navigating to the location of a locally stored file and clicking **open**. Depending on the browser being used, the name of the file can be changed. In future, this functionality will be improved to be more flexible.

4.7 Study status

The **status** link in the bottom, centre of the **study submission** screen brings up the **submission status** screen which lists all the user's studies that have been previously submitted and are still available to access on the system. It also shows any studies that have recently been submitted but are not yet

fully processed yet. The **study state** column provides the user with a short code to identify the status of the study:

- Study state: **R** means the study results have been computed and are ready to be used.
- Study state: **C** means the study has been created but not verified.
- Study state: **E** means the study has been extracted but results have not been generated.

Entries that are highlighted in pink are not available for mapping as they are currently being processed or have failed to be run.

Study Id	Study Name	Message	Date	Study State
19	USUAL	The study results have been computed and they are now ready to be used.	01-Sep-2017 00:00:00:...	R
18	Demoghadssghsadh	The study results have been computed and they are now ready to be used.	13-Jul-2017 00:00:00:000	R
17	DEMO	The study results have been computed and they are now ready to be used.	07-Jul-2017 00:00:00:000	R
16	monday2	The study results have been computed and they are now ready to be used.	03-Jul-2017 00:00:00:000	R
15	MONDAY	The study results have been computed and they are now ready to be used.	03-Jul-2017 00:00:00:000	R
14	TEST7	The study results have been computed and they are now ready to be used.	29-Jun-2017 00:00:00:...	R
13	TEST5	The study results have been computed and they are now ready to be used.	28-Jun-2017 00:00:00:...	R
12	test4	Study extracted imported or created but neither results nor maps have been created.	28-Jun-2017 00:00:00:...	E
11	TEST3	The study results have been computed and they are now ready to be used.	28-Jun-2017 00:00:00:...	R
10	TEST2	The study results have been computed and they are now ready to be used.	28-Jun-2017 00:00:00:...	R
8	HELLO	The study results have been computed and they are now ready to be used.	28-Jun-2017 00:00:00:...	R
7	Test	The study results have been computed and they are now ready to be used.	26-Jun-2017 00:00:00:...	R

Figure XXX. Submission status screen

4.8 Run study

Once the user is happy that all the study submission details are complete, the **run study** link can be clicked which brings up the run study screen. Here an optional description can be added and the details of the study can be checked by clicking the **view study submission summary** link. The study will be submitted when the user clicks 'run' and a message will be displayed: 'Success: study submitted'.

Figure XXX. Run study screen

Once the study has been submitted, the use will be returned to the study submission screen. The study may take some minutes to run depending on the size and complexity of the study. Clicking the **status** link allows the user to see the status of any recently submitted studies. When the status is listed as **R** the results of the study can be viewed under the **data viewer** tab.

5. Data viewer

The results of a study that has been submitted and run can be examined and analysed in detail under the **data viewer** tab. The header of the data viewer has two dropdown boxes allowing the user to select the study to view and to filter the results by sex. The main data viewer area is made up of four sections. Moving clockwise from top right, these are:

- configurable choropleth **map** of the study area
- data table of the regions included in the study area
- population pyramid of the whole study area
- frequency distribution of the outcome currently displayed in the choropleth map.
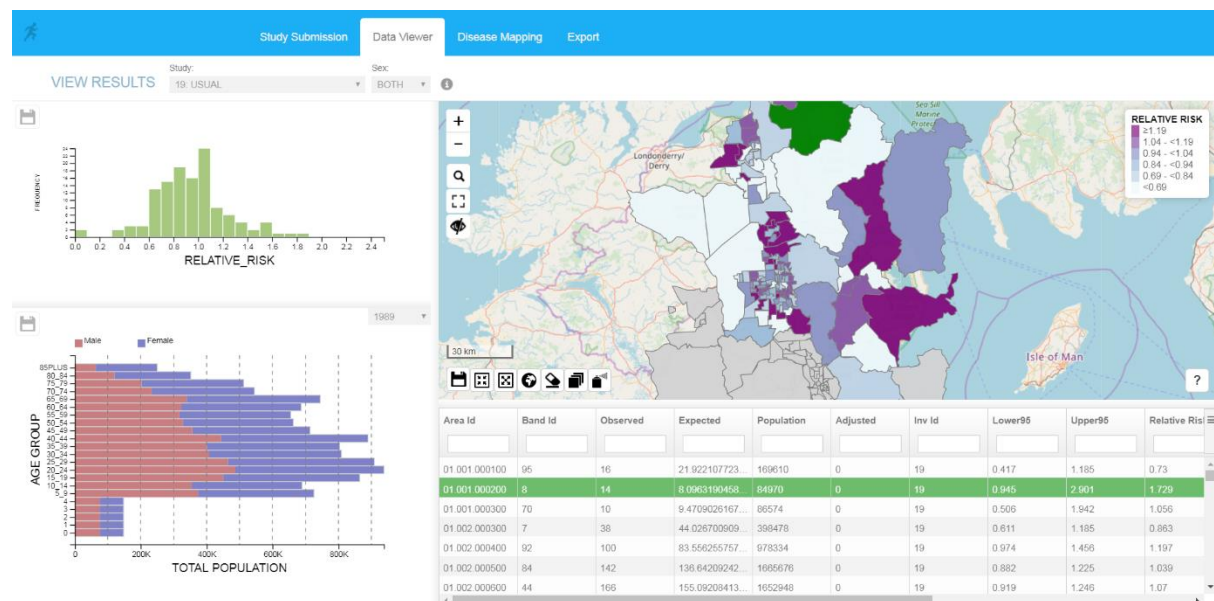
Figure XXX. An example study in the data viewer tab.

5.1 Choropleth map

The array of buttons above the map give the user several options to navigate the map and configure what information is displayed.

The **choropleth map** icon ▮ brings up the **choropleth map symbology** screen which allows the user to select which field is displayed on the map and how the values in the field are represented using colours on the map.
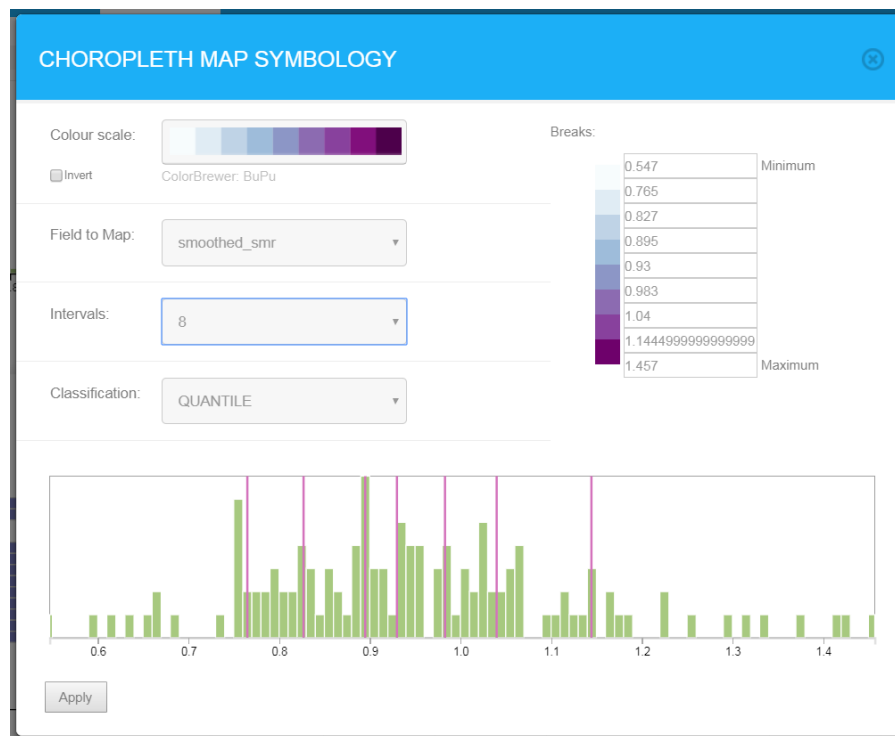
Figure XXX. Choropleth map symbology screen.

The **colour scale** dropdown selection lets the use choose from many different colour scales as defined by the Color Brewer project (http://colorbrewer2.org). The **field to map** dropdown lets the user select which data field is displayed on the map and in the frequency distribution graph. The **intervals** dropdown lets the user define the number of breaks in the data being displayed (the maximum is 11 but some colour scale have a lower limit). The **classification** dropdown lets the user define how the breaks in the data are defined. In each case the number of classifications is defined by the **intervals** dropdown.

- **Quantile**. Divided the data into quantiles with the same number of data points in each quantile.
- **Equal interval**. The difference between the lowest and highest values is divided equally into categories.
- **Jenks**. Uses the Jenks natural breaks classification method designed to determine the best arrangement of values into different classes (Jenks, 1967).
- **Standard deviation**. Calculated the mean and standard deviation of the underlying data, then divides the data into 5 categories (>2 standard deviations above/below mean, >1,<=2 standard deviations above/below mean, <=1 standard deviation away from mean). Always 5 categories.
- **Atlas relative risk**. The classification system used when displaying relative risks in the SAHSU Environment and Health Atlas (ref). Always 9 intervals.
- **Atlas probability**. The classification system used when displaying probabilities in the SAHSU Environment and Health Atlas (ref). Always 3 intervals.

Additionally, the breaks defined by the classification method can be manually edited in the **edit breaks** section.

5.2 Data table

The data table shows one row for each of the regions that make up the study area as selected during the study submission process. Clicking on rows in the data causes the corresponding regions to be highlighted in green on the choropleth map. The data table shows the population and health data (Area Id, Ban Id, Observer, Population, investigation id); basic statistics (Expected, Adjusted, Relative Risk, Lower95, Upper95); and the results of the Bayesian smoothing (Posterior probability, Smoothed Smr, Smoothed Smr Lower95, Smoothed Smr Upper95).

The data can be sorted ascending or descending by clicking on the column headings. There are filter boxes directly under the column names. Typing in a filter box will filter the results displayed in the data table. Note that the filters work using string filtering, i.e. typing 10 in the 'Band Id' filter will show all the rows that have the string '10' in the band id (e.g. 10, 100, 101, 110).

≡   The link to the right of the column heading gives the user an additional three menu options:

- **Clear all filters**. Removes an filter strings that have previously been entered.
- **Export all data as csv**. Allows the user to save a comma separated variable (csv) file of all the study data.
- **Export visible data as csv.** Allows the user to save a comma separated variable (csv) file of all the records currently in the data table.

Clicking on regions in the choropleth map or selecting the rows in the data table sets the value for that row in the 'selected' column to 1 (as well as highlighting the row). By filtering the data table so that it only shows the records where 'selected' is 1, then choosing '**export visible data as csv',** the user is able to effectively make a manual selection of regions in the map and export only the data associated with those selected regions.

5.3 Population pyramid

The population pyramid section displays a population pyramid showing the age distribution of the residents of the geography from which the study is taken. The dropdown box in the top right of the population pyramid shows the years for which there is population data and allows the user to view the population pyramids for the available years.
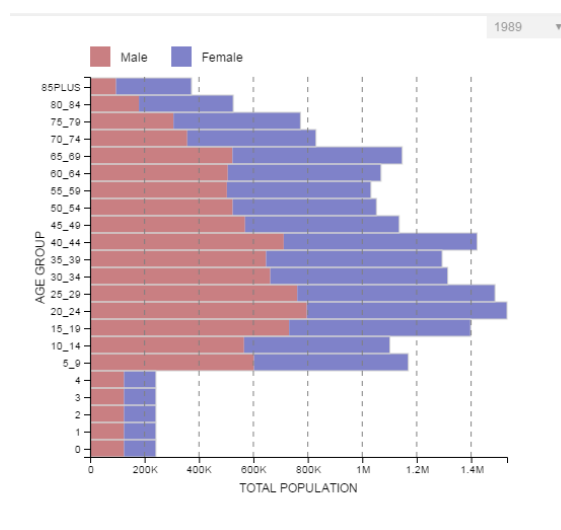
Figure XXX. Population pyramid.

5.4 Frequency distribution

The frequency distribution histogram shows the distribution of the data field currently being displayed in the choropleth map.
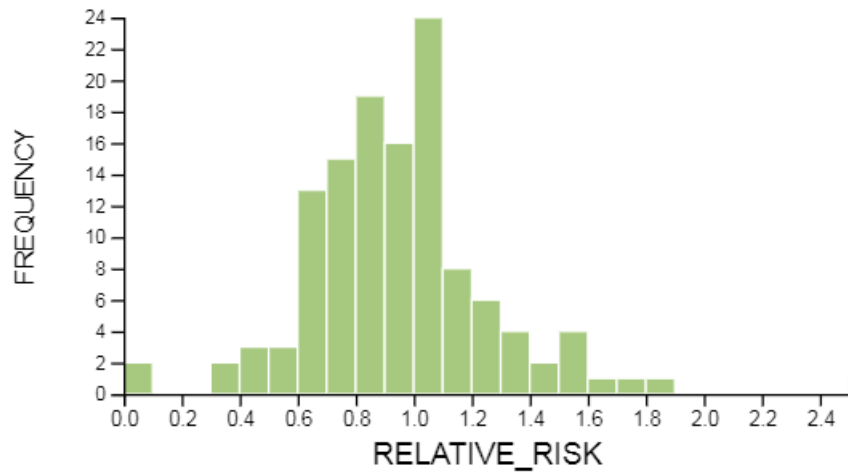
Figure XXX. Distribution histogram

6. Disease mapping

The **disease mapping** tab allows the user to compare two studies side-by-side or different data from the same study in two different maps. The disease mapping screen is divided vertically in two to give a **left display** and a **right display**. The header for each display allows the user to select the study and sex of the data displayed in the area below.
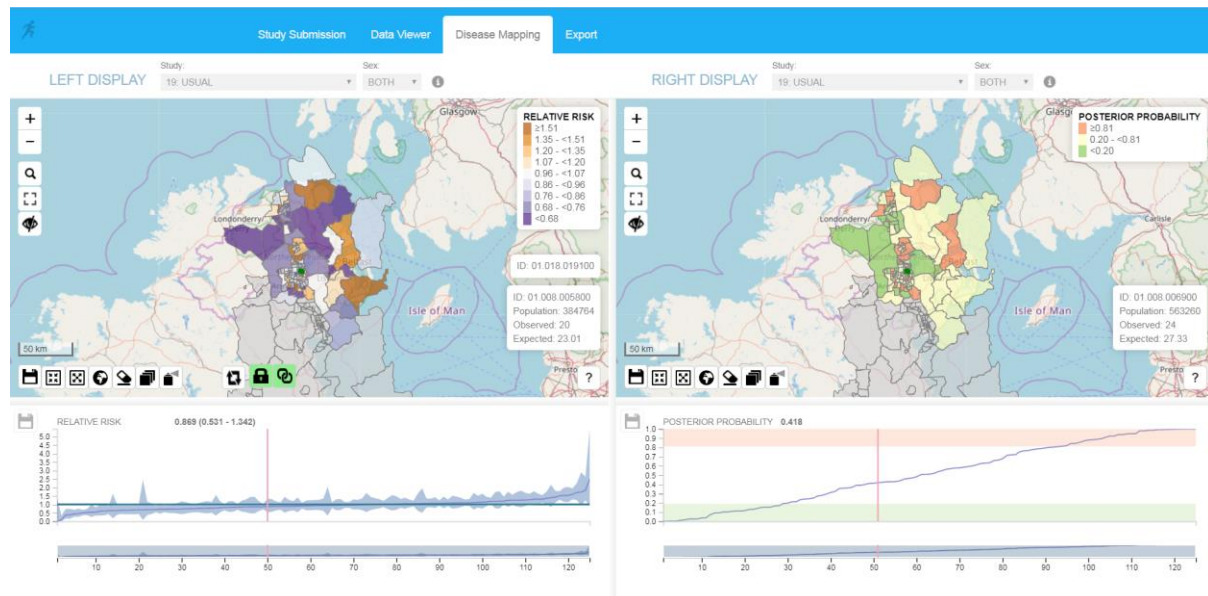
Figure XXX. Disease mapping tab.

6.1 Choropleth maps

Choropleth maps are displayed for both the left and right display with the same navigation and selection functions as the choropleth map in the **data viewer** tab (see section 5.1). In addition there additional functions to facilitate the simultaneous use of two maps.

Lock and **unlock selection**. When locked, clicking on the map region in the left display will also highlight the same region in the right display and vice-versa. If the studies selected for the left and right displays have been defined at different geographical levels, **lock selection** will not work and a warning will be displayed.

Lock and **unlock map extents**. When locked, both maps will always display the same extents. Zooming or scrolling on one map will cause the other map to move such that both maps display the same area.

Copy symbology. Copies the symbology settings from the **left display** to the **right display**.

6.2 Disease map charts

The charts displayed below the maps summarise the risk field data across the whole study area. The charts display all the values of the risk field show in the map above as well as the upper and lower confidence intervals. The data in the chart is ordered from lowest to highest risk (moving left to right). Clicking on a point in the chart moves the red line to that data point, displays the risk and

confidence intervals above the chart and selects the same region in the map above. Similarly, clicking on the region in the map moves the red line to the equivalent data point in the chart.
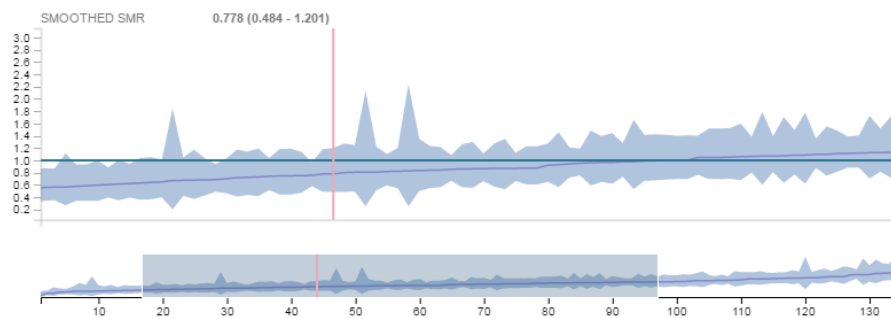
Figure XXX. Disease map chart.

The smaller chart displayed below the main chart acts as a navigation panel for the main chart above. By moving the mouse so it is above the left or right edge of the shaded area in the navigation panel, the user can click and drag to make the shaded area narrower or wider. This will increase or decrease the zoom in the main chart above.
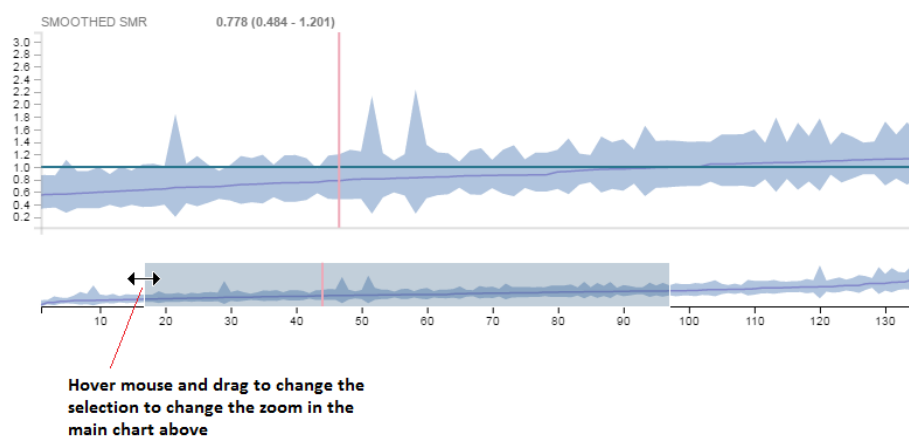
Figure XXX. Disease map chart showing how to alter the zoom.

7 Export

The export tab allows a user to export the results of a completed study as a zip file. Select the study to display using the dropdown at the top of the page. Initially a preview of the extract (input data) and map (result) tables are shown. Enter a range of rows and click the refresh button to preview further rows. The map container shows either the study or comparison area.

On clicking 'Export study tables', the full map and extract tables are downloaded as csv files and the study and comparison areas are downloaded in GIS format (geoJSON) at the specified detail level. All files are saved in a zipped folder prefixed with the study name and date in your specified output folder (see the RIF set up instructions for how to change this).
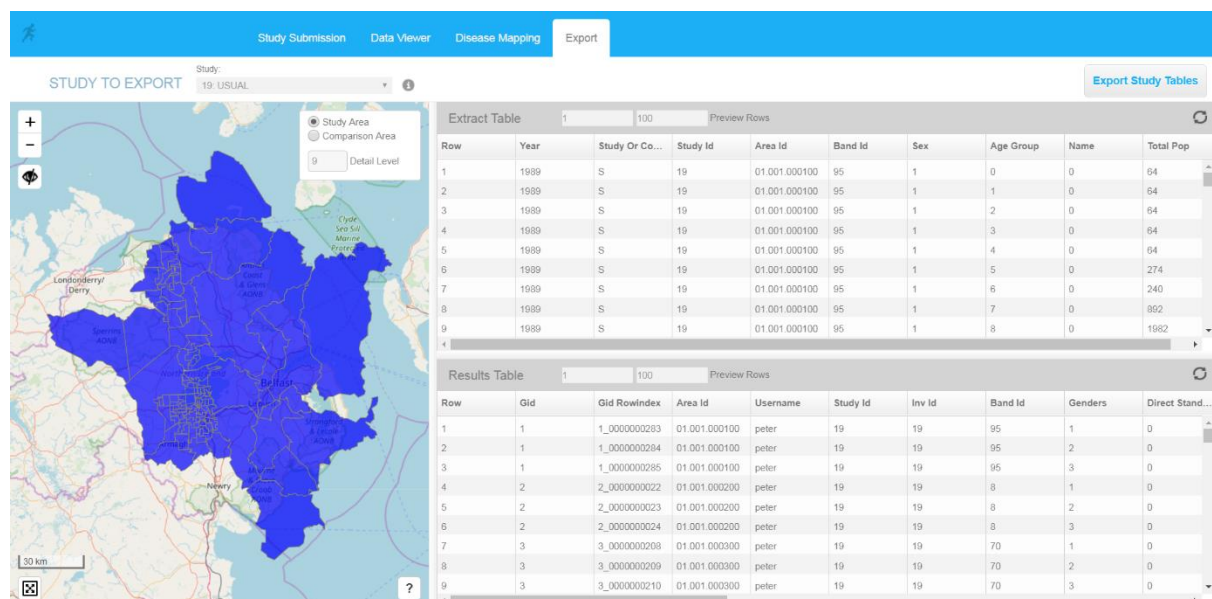
Figure XXX. Export tab.

Appendices

Appendix A. Statistical methods

Disease mapping

Disease maps aim at representing the geographical distribution of the incidence of disease. In the frame of the RIF software, only maps produced with counts of data are considered. Counts of disease cases are reported for a list of regions, denoted here areas, and delimited by geographical boundaries. The easiest way to map the geographical variations of the disease would be to directly map the counts. However, these counts depend strongly of the age-sex composition of the population at risk within each area, and cannot be directly compared. Consequently it is necessary to apply the use of **standardisation** and **standardised rates** to exclude the effect of populations. Standardisation requires the definition of a 'comparison' population associated to each area. The comparison population may be the total population of all study areas, or subsets of study areas.

Standardised disease rates in populations can be calculated using **direct** and **indirect standardisation**. Direct standardisation involves applying the disease rates found in the study areas to a standard population. This is not always available so, currently the RIF version 4.0 does not attempt direct standardisation of disease rates. When applying indirect standardisation, the standard disease rates from a comparison population are applied to the study population to give the **expected** disease counts such as the **standardised mortality rates (SMR)** or the **standardised incidence rates (SIR)**.

Indirectly standardised risks

The standard disease rates are taken from the regions defined by the comparison area when the study was submitted. Cases occurring in the comparison population are located for male and females in each five year age band and per covariate. Standardised rates of disease in the comparison population are calculated for each gender, age and covariate by dividing the number of

cases in each group by the total population in each corresponding gender, age and covariate stratum. The comparison area populations need to be large enough such that the age-sex-covariate specific disease rates are reliable. These standardised rates (r_j*) are applied to the study area population strata (j) to calculate the **expected** counts (E_i),

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

Where N_{ij} is the study population in strata j of area i. The SMR for the study population(s) is then simply given by the ratio of the observed (O_i) to expected counts (E_i):

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

Values of the relative risk larger than one indicate an excess of risk relatively to the underlying 'comparison population', whereas values smaller than 1 indicate a deficit of risk. Since each observation, is divided to the expected counts given the structure of the population, this variable has no unit, and comparisons between areas can be done. 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 population being compared are not exactly the same).

The uncertainty associated with the SMR estimate is quantified by calculating the 95% confidence. Here we note that relative risk RR_i is the parameter of a Poisson distribution,

$$O_i \sim P(E_i RR_i).$$

- If $O_i < 100$, confidence intervals are found by using the Chi-squared method.

    For instance, if RR_i^U is the upper bound, then, by definition, for any $X_i \sim P(E_i RR_i^U)$, $P(X_i > O_i) = \frac{\alpha}{2}$.

    We know that $P(X_i \leq O_i) = 1 - P(Y \leq 2\,E_i RR_i^U)$, with $Y \sim X^2_{2O_i+2}$. Consequently,

    - upper 95% CI $= \frac{1}{2E_i} q^{X^2_{2O_I+2}}_{\alpha/2}$
    - lower 95% CI $= \frac{1}{2E_i} q^{X^2_{2O_I+2}}_{1-\alpha/2}$.

    For confidence level 100(1-α)%.

- If $O_i \geq 100$, a Gaussian approximation of the log relative risk is done, $\log(O_i/E_i)$ is assumed to follow a Gaussian distribution with mean $\log(RR_i)$, and variance $1/O_i$. Then,
    - lower 95% CI $= \dfrac{O_I/E_I}{\exp(1.96\sqrt{1/O_i})}$
    - upper 95% CI $= \dfrac{O_I}{E_I} \times \exp\left(1.96\sqrt{\dfrac{1}{O_i}}\right)$

Empirical Bayes Analysis

The maps of the standardised mortality or incidence ratio may lead to misinterpretations, since the extreme values are more often the consequence of small counts than a true extreme relative risk. Consequently, a non-significantly positive standardized mortality risk may be higher than a significant one for which the population at risk is higher. To reduce the influence of the small counts, Clayton and Kaldor (1987) proposed empirical Bayes estimates of the relative risk. They are based on a Poisson-Gamma hierarchical model. By accounting for differential variability in the data, this hierarchical approach provides more precise estimates of relative risk and more accurate assessments of significant changes that the standard methods. The estimates are smoothed toward a global value by assuming that all the relative risks are sampled in the same gamma distribution. Moreover, the smaller the count, the stronger the shrinkage effect. In detail, relative risks RR_i are assumed to come from a single Gamma distribution of scale α and shape β,

$$O_i \sim Poisson(E_i RR_i)$$

$RR_i \sim Gamma(\alpha, \beta)$.

Approximation of the posterior mean of relative risk RR_i are given by the empirical Bayes estimates,

$$E\big(RR_i \big| O_i, E_i, \hat{\alpha}, \hat{\beta}\big) = \frac{O_i + \hat{\beta}}{E_i + \hat{\alpha}}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood estimates of α and β.

In practice, these estimates of the relative risk are obtained through the following iterative procedure:

1. Start with initial values for the relative risk RR_i. For instance

$$\widehat{RR}_\iota = \frac{O_i}{E_i}.$$

2. Obtain estimators $\hat{\alpha}$ and $\hat{\beta}$ using equations

$$\frac{\hat{\alpha}}{\hat{\beta}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{RR}_\iota$$

   and

$$\frac{\hat{\alpha}}{\hat{\beta}^2} = \frac{1}{n-1} \sum_{i=1}^{n} \left(1 + \frac{\hat{\beta}}{E_i}\right) \left(\widehat{RR}_\iota - \frac{\hat{\alpha}}{\hat{\beta}^2}\right)$$

3. Obtain new estimated values for the relative risks

$$\widehat{RR}_\iota = \frac{O_i + \hat{\beta}}{E_i + \hat{\alpha}}$$

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

Full Bayesian smoothing

If some degree of spatial dependence of the risk is assumed, meaning that the risks in close areas are similar, we can estimate the risk in one area with information borrowed from its neighbours.

Since these estimates are based on more information than SMR estimates, they are more robust. The BYM model (Besag, York, and Mollié 1991) was developed to address this issue of spatial dependence of risk. First a neighbourhood structure of the study area must be defined which specifies the neighbour relationships between areas. Here, we denotes ∂_i the set neighbours for area i. The log relative risk is then split into two terms, u_i a spatial term which accounts for the spatial variations within the risk, and v_i a noise which accounts for independent local variations,

$$\log(RR_i) \ = \ u_i + v_i.$$

The spatial term is modelled with an intrinsic conditional auto-regressive (CAR) model, meaning that for each area i, u_i is given conditionally to its neighbours. Specifically, it is assumed to follow a normal distribution with mean, the neighbour mean and variance, a variance parameter, σ_u^2, divided by the number of neighbours, n_i,

$$u_i | u_{-i} \sim N \left(\frac{\sum_{k \in \partial_i} u_k}{n_i}, \frac{\sigma_u^2}{n_i} \right).$$

The larger the number of neighbours, the smaller the variance. The Gaussian noises, denoted v_i, are supposed to be independent and identically distributed with variance σ_v^2.

One can consider the map of the uniquely spatially structured term given by $(\exp(u_i))$. They display uniquely the part of the risk which has a smooth spatial distribution. The independent term is then seen as residual. But one can also consider the estimate of the entire relative risk which leads to robust estimate thanks to the spatial CAR term, but also allows individual variability.

Other models are also proposed in the RIF, the 'CAR' model, in which the log relative risk is only modelled by the conditional autoregressive u_i term. In this model, the relative risk is assumed to be spatially smooth without any local independent variations. In the CAR and the BYM model, estimates are smoothed towards a local mean.

Finally the 'HET' (heterogeneous) model is also proposed. In this model, the log relative risk is only composed of the independent term v_i. As for empirical Bayes estimates, with this model the log relative risk estimates are smoothed toward a global mean. These two models are thus similar, they only differ by their prior (Gaussian vs gamma), and their inference method (fully Bayesian vs empirical Bayes).

Prior specification

Minimally informative independent inverse gamma $IGamma(0.5, 0.0005)$ priors are assigned to the model parameters σ_u^2 and σ_v^2.

R and R-INLA

The statistical calculations described above are performed by the RIF server calling an instance of a R procedure (R Core Team, 2015). The full Bayesian smoothing is performed using the integrated nested Laplace approximations (INLA), proposed by Rue et al. (2009) . Whereas Bayesian inference often makes use of Markov Chain Monte Carlo (MCMC) simulation methods (Casella and George, 1992), the increasing size and related high spatial resolution of the datasets supported by the RIF mean that even state of the art, high powered servers would take several days to perform Bayesian

inference via MCMC. Since INLA uses a deterministic algorithm it produces accurate results much faster than MCMC methods (Blangiardo and Cameletti, 2015). The INLA functionality is delivered through and R package called R-INLA. The website www.r-inla.org is a useful source of further information; it provides many papers, tutorials and examples that assist in the understanding and implementation of INLA.

Appendix B. Descriptive analysis of Sahsuland

Sahsuland is a fictitious island nation of approximately 32860 km^2 comprising 4 different hierarchical levels of geography.
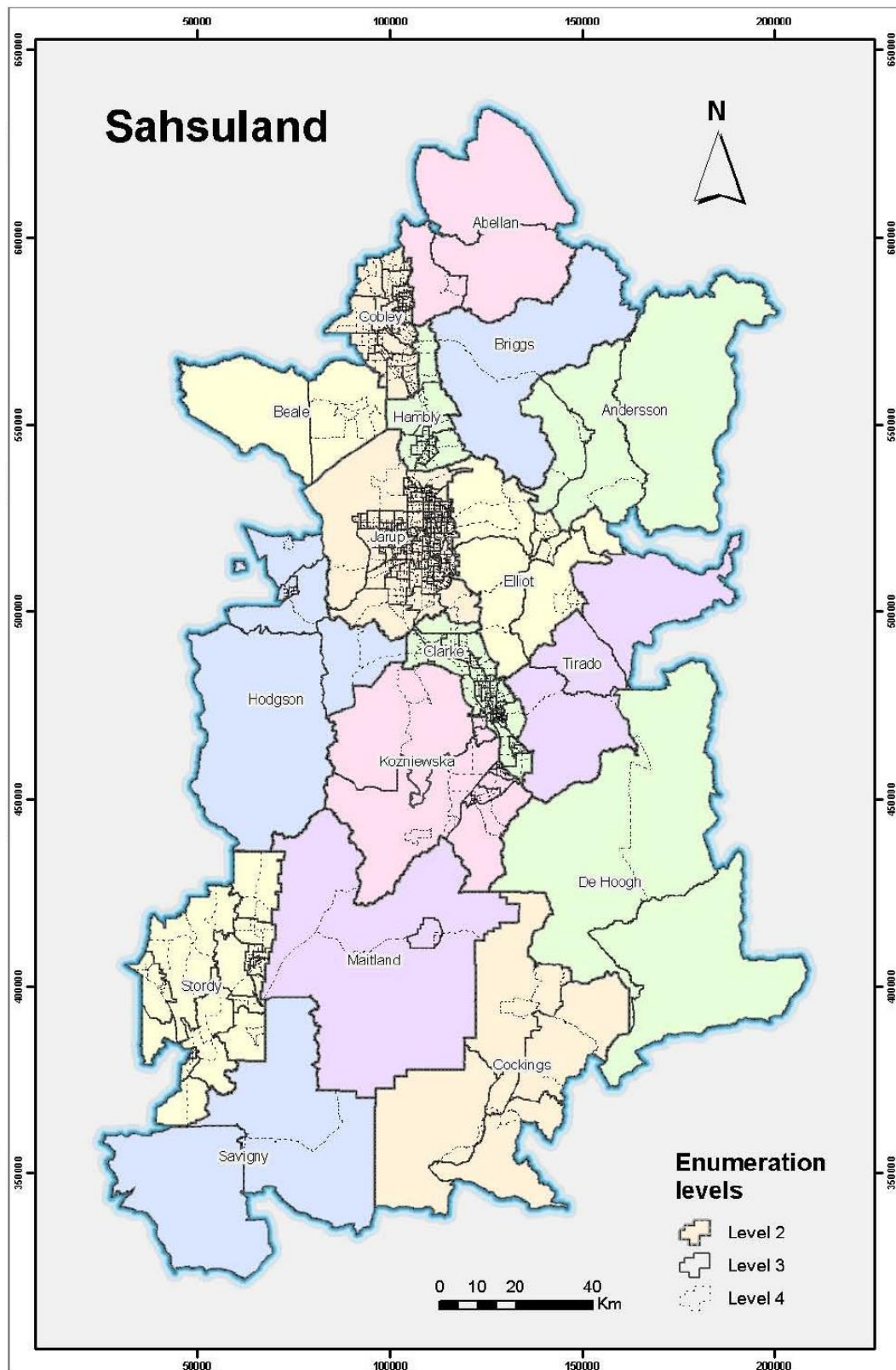
Figure XXX. Sahsuland

Sahsuland population

Population data is available for the period 1989-1996 in five year age groups from ages 5 to 85 and one year age groups for the ages 0 to 4.

Table XXX. Sahsuland population by age.

Year	0-4	5-9	10-19	20-39	40-59	60-79	80+	Total
1989	622,270	600,646	1,283,290	2,893,970	2,380,576	1,961,654	460,856	10,203,262
1990	631,650	602,524	1,261,454	2,899,864	2,416,484	1,960,170	472,292	10,244,438
1991	642,260	607,206	1,239,456	2,914,374	2,450,368	1,961,088	484,834	10,299,586
1992	644,900	616,862	1,223,550	2,941,456	2,477,394	1,954,206	497,464	10,355,832
1993	639,190	631,042	1,216,676	2,949,324	2,508,448	1,942,956	511,690	10,399,326
1994	637,090	645,432	1,212,252	2,966,698	2,548,060	1,931,200	523,196	10,463,928
1995	628,860	655,334	1,220,428	2,979,138	2,579,936	1,925,502	533,890	10,523,088
1996	617,540	661,308	1,235,714	2,964,174	2,610,400	1,922,848	538,526	10,550,510

Table XXX. Suhsuland population by level 2 geography (17 divisions, equivalent to US Counties)

Year	Population					Divisions		
	Min	Max	Quartiles			Number with pop. less than:		
			1st	Median	3rd	100,000	500,000	1 million
1989	7,848	4,531,618	59,719	145,340	620,080	6	13	14
1990	7,942	4,546,944	60,344	146,260	621,506	6	13	14
1991	8,060	4,561,132	60,948	146,930	625,866	6	13	14
1992	8,076	4,584,918	61,080	147,852	628,008	6	13	14
1993	8,096	4,604,640	61,194	148,660	629,334	6	13	14
1994	8,096	4,629,846	61,260	149,888	630,394	6	13	14
1995	8,100	4,659,062	61,357	150,888	631,606	6	13	14
1996	8,138	4,668,972	61,315	152,100	633,123	6	13	14

Table XXX. Suhsuland population by level 3 geography (202 divisions, equivalent to US Tract)

Year	Population					Divisions		
	Min	Max	Quartiles			Number with pop. less than:		
			1st	Median	3rd	10,000	50,000	100,000
1989	2,760	203,848	26,000	42,746	62,519	9	122	183
1990	2,820	206,622	26,271	42,917	62,459	9	123	183
1991	2,856	209,214	26,484	43,119	63,308	8	124	183
1992	2,880	208,694	26,621	43,297	63,684	8	123	183
1993	2,900	208,924	26,612	43,458	64,134	8	123	183
1994	2,950	209,526	26,843	43,746	64,505	8	121	182
1995	2,980	210,038	27,207	44,281	64,837	8	121	182
1996	3,016	209,140	27,397	44,457	65,221	8	121	182

Table XXX. Suhsuland population by level 4 geography (1230 divisions, equivalent to US Census Block Group)

Year	Population	Divisions

	Min	Max	Quartiles			Number with pop. less than:		
			1st	Median	3rd	2,500	5,000	20,000
1989	128	33,588	3,567	6,120	10,759	58	517	1122
1990	144	33,662	3,578	6,091	10,656	58	515	1121
1991	146	34,150	3,610	6,148	10,760	58	512	1120
1992	146	34,226	3,636	6,182	10,774	55	509	1119
1993	154	34,516	3,648	6,217	10,724	52	506	1119
1994	160	34,622	3,702	6,300	10,804	47	501	1117
1995	156	34,806	3,744	6,323	10,842	46	499	1116
1996	146	35,092	3,750	6,347	10,894	44	495	1116

Sahsuland numerator data

Numerator data consists of cancer incidences. For the early years (1989 – 1994) the data is in 45 different ICD-9 codes. Until a taxonomy service is written for ICD-9 codes, this data is cannot be used. The years 1995 and 1996 have cancer data for 41 different ICD-10 codes.

Table XXX. Total cases in Sahsuland, ICD-10 codes and health conditions covered by the Environment and Health Atlas*.

ICD-10	Description	Cases	
		1995	1996
C220	Liver cell carcinoma	162	156
C221	Intrahepatic bile duct carcinoma	186	182
C223	Angiosarcoma of liver	2	4
C229	Malignant neoplasm of liver, not specified as primary or secondary	74	96
C33	Malignant neoplasm of trachea	22	6
C340	Malignant neoplasm of main bronchus	422	496
C341	Malignant neoplasm of upper lobe, bronchus or lung	1,164	1,166
C342	Malignant neoplasm of middle lobe, bronchus or lung	120	178
C343	Malignant neoplasm of lower lobe, bronchus or lung	572	648
C348	Malignant neoplasm of overlapping sites of bronchus and lung	74	38
C349	Malignant neoplasm of unspecified part of bronchus or lung	4,258	3,728
C500	Malignant neoplasm of nipple and areola	220	206
C501	Malignant neoplasm of central portion of breast	506	452
C502	Malignant neoplasm of upper-inner quadrant of breast	430	526
C503	Malignant neoplasm of lower-inner quadrant of breast	190	238
C504	Malignant neoplasm of upper-outer quadrant of breast	1,596	1,596
C505	Malignant neoplasm of lower-outer quadrant of breast	352	358
C506	Malignant neoplasm of axillary tail of breast	50	78
C508	Malignant neoplasm of overlapping sites of breast	118	176
C509	Malignant neoplasm of breast of unspecified site	3,914	3,954
C64	Malignant neoplasm of kidney, except renal pelvis	1,014	862
C65	Malignant neoplasm of renal pelvis	78	48
C670	Malignant neoplasm of trigone of bladder	26	42
C671	Malignant neoplasm of dome of bladder	14	22
C672	Malignant neoplasm of lateral wall of bladder	186	110
C673	Malignant neoplasm of anterior wall of bladder	44	30
C674	Malignant neoplasm of posterior wall of bladder	88	92
C675	Malignant neoplasm of bladder neck	42	40

C676	Malignant neoplasm of ureteric orifice	64	38
C678	Malignant neoplasm of overlapping sites of bladder	142	84
C679	Malignant neoplasm of bladder, unspecified	2,500	2,390
C710	Malignant neoplasm of cerebrum, except lobes and ventricles	92	52
C711	Malignant neoplasm of frontal lobe	188	136
C712	Malignant neoplasm of temporal lobe	98	78
C713	Malignant neoplasm of parietal lobe	146	72
C714	Malignant neoplasm of occipital lobe	32	30
C715	Malignant neoplasm of cerebral ventricle	10	14
C716	Malignant neoplasm of cerebellum	50	40
C717	Malignant neoplasm of brain stem	26	14
C718	Malignant neoplasm of overlapping sites of brain	42	12
C719	Malignant neoplasm of brain, unspecified	134	306
	Total	19,448	18,794
	EHA health conditions		
C34	Lung cancer	6,610	6,254
C50	Breast cancer	7,376	7,584
C67	Bladder cancer	3,106	2,848
C71	Brain cancer	818	754
C22	Liver cancer	424	428

* There are no cases of prostate cancer, skin cancer, Leukaemia or mesothelioma in the Sahsuland data.

Table XXX. Summary of cancer cases for **1995**, by **level 2** geography. All cases, top 6 individual ICD codes and health conditions covered by the Environment and Health Atlas

ICD-10 code	Cases					Divisions				
	Min	Max	Quartiles			Number with cases less than:				
			1st	Median	3rd	10	50	150	500	1500
All	8	9,000	119	298	1,255	1	2	6	10	13
C349	2	2,112	18	60	279	2	7	11	15	16
C509	4	2,218	21	56	219	2	8	12	16	16
C679	0	1,216	12	22	172	3	10	13	16	17
C504	0	592	7	26	125	6	10	13	16	17
C341	0	488	6	20	67	5	12	15	17	17
C64	0	466	5	10	67	7	12	16	17	17
EHA										
C34	2	3,060	37	88	419	1	6	10	13	16
C50	4	3,528	46	120	444	1	6	9	13	16
C67	2	1,350	20	44	235	3	9	13	16	17
C71	0	366	2	10	59	7	13	16	17	17
C22	0	180	2	8	28	9	15	16	17	17

Table XXX. Summary of cancer cases for **1996**, by **level 2** geography. All cases, top 6 individual ICD codes and health conditions covered by the Environment and Health Atlas

ICD-10 code	Cases					Divisions				
	Min	Max	Quartiles			Number with cases less than:				
			1st	Median	3rd	10	50	150	500	1500
All	10	8,962	93	228	1,172	0	2	6	11	13
C349	2	1,794	18	68	242	2	8	12	15	16
C509	4	2,174	16	46	248	3	9	12	16	16
C679	0	1,250	10	28	154	2	11	13	16	17
C504	0	690	3	14	122	7	11	13	16	17
C341	0	542	6	16	53	6	11	14	16	17
C64	0	380	5	10	52	8	13	16	17	17
EHA										
C34	2	2,926	28	96	360	1	6	11	14	16
C50	4	3,708	39	72	513	1	5	10	13	16
C67	0	1,354	14	44	177	1	10	13	16	17
C71	0	388	3	8	41	9	13	16	17	17
C22	0	178	3	6	31	11	15	16	17	17

Table XXX. Summary of cancer cases for **1995**, by **level 3** geography. All cases, top 6 individual ICD codes and health conditions covered by the Environment and Health Atlas

ICD-10 code	Cases							Divisions			
	Min	Max	Centiles					Number with cases less than:			
			5%	25%	50%	75%	95%	1	5	25	100
All	0	412	20	52	81	116	262	1	2	14	127
C349	0	106	2	10	16	28	65	3	24	145	201
C509	0	90	2	8	14	24	56	4	26	154	202
C679	0	50	0	4	10	16	38	14	54	182	202
C504	0	38	0	2	6	12	25	39	93	192	202
C341	0	36	0	2	4	8	18	39	113	198	202
C64	0	24	0	2	4	8	14	40	122	202	202
EHA											
C34	0	172	4	16	26	41	100	2	12	99	192
C50	0	132	6	20	30	46	91	1	6	68	195
C67	0	54	2	8	12	20	44	8	38	171	202
C71	0	22	0	2	4	6	12	49	141	202	202
C22	0	12	0	0	2	4	8	87	176	202	202

Table XXX. Summary of cancer cases for **1996**, by **level 3** geography. All cases, top 6 individual ICD codes and health conditions covered by the Environment and Health Atlas

ICD-10 code	Cases							Divisions			
	Min	Max	Centiles					Number with cases less than:			
			5%	25%	50%	75%	95%	1	5	25	100
All	4	360	22	44	79	115	267	0	2	15	131
C349	0	72	2	8	14	24	52	6	24	154	202

ICD-10 code	Min	Max	5%	25%	50%	75%	95%	1	3	5	15	50
C509	0	130	2	10	14	24	54	6	23	153	200	
C679	0	70	2	6	10	16	36	9	50	185	202	
C504	0	62	0	2	4	12	28	48	105	189	202	
C341	0	32	0	2	4	8	20	43	117	197	202	
C64	0	24	0	0	4	6	12	52	131	202	202	
EHA												
C34	0	132	6	14	24	41	84	1	9	107	195	
C50	0	160	6	18	32	48	100	1	9	79	191	
C67	0	74	2	6	12	20	41	7	35	177	202	
C71	0	26	0	0	2	6	12	58	141	201	202	
C22	0	14	0	0	2	3	8	90	174	202	202	

Table XXX. Summary of cancer cases for **1995**, by **level 4** geography. All cases, top 6 individual ICD codes and health conditions covered by the Environment and Health Atlas

ICD-10 code	Cases							Divisions				
	Min	Max	Centiles					Number with cases less than:				
			5%	25%	50%	75%	95%	1	3	5	15	50
All	0	70	2	6	12	22	42	38	115	237	719	1204
C349	0	22	0	0	2	6	12	374	729	918	1208	1230
C509	0	26	0	0	2	4	10	402	736	952	1214	1230
C679	0	18	0	0	2	4	8	573	897	1085	1228	1230
C504	0	16	0	0	0	2	6	747	1041	1149	1229	1230
C341	0	14	0	0	0	2	4	820	1115	1190	1230	1230
C64	0	8	0	0	0	2	4	854	1126	1204	1230	1230
EHA												
C34	0	26	0	2	4	8	17	228	536	733	1151	1230
C50	0	32	0	2	4	8	16	181	413	637	1150	1230
C67	0	24	0	0	2	4	8	476	804	1025	1226	1230
C71	0	8	0	0	0	2	4	918	1154	1215	1230	1230
C22	0	6	0	0	0	0	2	1,050	1204	1224	1230	1230

Table XXX. Summary of cancer cases for **1996**, by **level 4** geography. All cases, top 6 individual ICD codes and health conditions covered by the Environment and Health Atlas

ICD-10 code	Cases							Divisions				
	Min	Max	Centiles					Number with cases less than:				
			5%	25%	50%	75%	95%	1	3	5	15	50
All	0	70	2	6	12	22	42	38	115	237	719	1204
C349	0	22	0	0	2	6	12	374	729	918	1208	1230
C509	0	26	0	0	2	4	10	402	736	952	1214	1230
C679	0	18	0	0	2	4	8	573	897	1085	1228	1230
C504	0	16	0	0	0	2	6	747	1041	1149	1229	1230
C341	0	14	0	0	0	2	4	820	1115	1190	1230	1230
C64	0	8	0	0	0	2	4	854	1126	1204	1230	1230
EHA												
C34	0	26	0	2	4	8	17	228	536	733	1151	1230
C50	0	32	0	2	4	8	16	181	413	637	1150	1230

C67	0	24	0	0	2	4	8	476	804	1025	1226	1230
C71	0	8	0	0	0	2	4	918	1154	1215	1230	1230
C22	0	6	0	0	0	0	2	1,050	1204	1224	1230	1230

References

Besag J, York J, Mollie A. (1991). Bayesian image restoration, with applications in spatial statistics. Annals of the Institute of Statistical Mathematics, **43**, 1-59.

Blangiardo M. and Cameletti M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley and Sons Ltd.

Casella G. and George E. (1992). Explaining the Gibbs sampler. American Statistician, **46**, 167-174.

Clayton DG and Kaldor J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics **43**, 671-681

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rue H, Martino S, and Chopin N. (2009). Approximate Bayesian inference for latent Gaussian model by using integrated Laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B, **71**, 319-392.