Loading RIF data

Contents

[Overview 2](#_Toc468812432)

[Restrictions 2](#_Toc468812433)

[Pre Data Loader Build 2](#_Toc468812434)

[Loading Data 3](#_Toc468812435)

[Geospatial Data 3](#_Toc468812436)

[Health Study Themes 3](#_Toc468812437)

[Numerator data 4](#_Toc468812438)

[Numerator mandatory columns 4](#_Toc468812439)

[Denominator data 5](#_Toc468812440)

[Denominator mandatory columns 6](#_Toc468812441)

[Covariate Data 6](#_Toc468812442)

[Covariate mandatory columns 7](#_Toc468812443)

# Overview

RIF database data is of three types:

1. Study data. These tables always contain a *study\_id* column. To allow for multi user support the tables start t\_<view> with an DML capable view to the users to view and update only the data they can access.
2. Configuration data. This data is of two types – meta data for load data and other configuration data
3. Load data. This is data imported into the RIF by the data loader. Principally it is of four types:
   1. Numerator data
   2. Denominator data
   3. Covariate data
   4. Geospatial data

The geospatial data must be setup first. Additionally, numerators are allocated to health study themes to allow numbers of numerators to be supported and sorted for ease of access.

# Restrictions

* The data loader only supports one denominator per numerator. This is to allow the user to choose a numerator and the denominator can then be automatically pre-populated. These tables have the automatic flag set to Y in RIF40\_TABLES.
* The table T\_RIF40\_NUM\_DENOM is a per user table for extra numerator and denominator pairs not added automatically. Typically, these are the standard SAHSU tables with three geographies (ED91, OA2001, OA2011) that break the one denominator per numerator rule. This will not be supported by the data loader and will require further work in the front end and middleware.

# Pre Data Loader Build

Currently data is loaded by: *v4\_0\_postgres\_sahsuland\_imports.sql* and the empty data loader test database created by *v4\_0\_postgres\_sahsuland\_empty\_imports.sql*. Both scripts are in the directory *..rapidInquiryFacility\rifDatabase\Postgres\sahsuland*.

The script: *v4\_0\_postgres\_sahsuland\_imports.sql* will eventually call the data loader directly. It will then only load the standard RIF data (i.e. the contents of RIF40\_REFERENCE\_TABLES):

* Rif40\_PREDEFINED\_GROUPS
* RIF40\_AGE\_GROUP\_NAMES
* RIF40\_AGE\_GROUPS
* RIF40\_PROJECTS
* RIF40\_OUTCOMES
* RIF40\_OUTCOME\_GROUPS
* RIF40\_REFERENCE\_TABLES
* RIF40\_VERSION
* T\_RIF40\_PARAMETERS

This also includes schema metadata only used by the build process and not by the RIF itself:

* RIF40\_ERROR\_MESSAGES
* RIF40\_COLUMNS
* RIF40\_TAVBLES\_AND\_VIEWS
* RIF40\_TRIGGERS

The following are obsolete and will be removed:

* RIF40\_ICD10
* RIF40\_ICD9
* RIF40\_ICD\_O\_3
* RIF40\_OPCS4
* RIF40\_A\_AND\_E
* RIF40\_CHI2
* RIF40\_POIS\_DISTRIBUTION
* RIF40\_POPULATION\_EUROPE
* RIF40\_POPULATION\_US
* RIF40\_POPULATION\_WORLD
* T\_RIF40\_FDW\_TABLES
* RIF40\_DUAL

# Loading Data

## Geospatial Data

The following geospatial data is required by the data loader:

* Geography, in RIF40\_GEOGRAPHIES. The columns GEOGRAPHY and DESCRIPTION name and describe the geography respectively. The other columns are either related to the load process (e.g. MAX\_GEOJSON\_DIGITS) or set default behaviour in the runtime (e.g. DEFAULTSTUDYAREA) or provide the location for looking tables (e.g. TILETABLE). The SRID spatial reference identifier is used by load process and will in future be used by the runtime as it defines the map projection.
* Two or more geolevels in RIF40\_GEOLEVELS. The column GEOLEVEL\_NAME is the name of geolevel. This will be a column name in the numerator/denominator tables. GEOLEVEL\_ID is an identifier for ordering (1=lowest resolution). Up to 99 supported. DESCRIPTION is self evident. As above the other columns are either related to the load process (e.g. SHAPEFILE) or set default behaviour in the runtime (e.g. RESOLUTION) or provide the location for looking tables (e.g. LOOKUP\_TABLE). The spatial control fields are defined as follows:
  + RESTRICTED: Is geolevel access restricted by Information Governance restrictions (0/1). If 1 (Yes) then a) students cannot access this geolevel and b) if the system parameter ExtractControl=1 then the user must be granted permission by a RIF\_MANAGER to extract from the database the results, data extract and maps tables. This will be enforced by the RIF application.
  + RESOLUTION: Can use a map for selection at this resolution (0/1)
  + COMPAREA: Able to be used as a comparison area (0/1)
  + LISTING: Able to be used in a disease map listing (0/1)

The data loader is not concerned with the spatial control fields.

The following is the RIF 3.0 setup for SAHSU using the 2001 census administrative geography:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Geolevel Name | Geolevel ID | Description | Restricted | Resolution | Comparison area | Listing |
| SCNTRY2001 | 1 | 2001 super country (UK) | 0 | 1 | 1 | 1 |
| CNTRY2001 | 2 | 2001 country (e.g. England) | 0 | 1 | 1 | 1 |
| GOR2001 | 3 | 2001 Government office region (e.g. London) | 0 | 1 | 1 | 1 |
| LADUA2001 | 4 | 2001 local area district or unitary authority | 0 | 1 | 1 | 1 |
| WARD2001 | 5 | 2001 Census statistical ward | 1 | 1 | 1 | 0 |
| SOA2001 | 6 | 2001 Super output areas lower level | 1 | 0 | 1 | 0 |
| OA2001 | 7 | 2001 Census output areas | 1 | 0 | 1 | 0 |

Geospatial data must be loaded first.

## Health Study Themes

The data loader needs to either create a new theme or re-use a pre-exiting theme:

DELETE FROM rif40\_health\_study\_themes WHERE theme = 'SAHSULAND';

INSERT INTO rif40\_health\_study\_themes(theme, description) VALUES('SAHSULAND', 'SAHSU land cancer incidence example data');

Themes are intended to group numerator tables together for ease of use; e.g. England Hospital Episode Statistics, England Cancer Data

## Numerator data

The numerator table must be created, indexed, commented and access granted:

CREATE TABLE rif\_data.sahsuland\_cancer (

year smallint NOT NULL,

age\_sex\_group smallint NOT NULL,

level1 varchar(20) NOT NULL,

level2 varchar(20) NOT NULL,

level3 varchar(20) NOT NULL,

level4 varchar(20) NOT NULL,

icd varchar(4) NOT NULL,

total double precision NOT NULL

);

CREATE INDEX sahsuland\_cancer\_age\_sex\_group ON rif\_data.sahsuland\_cancer (age\_sex\_group);

CREATE INDEX sahsuland\_cancer\_icd ON rif\_data.sahsuland\_cancer (icd);

CREATE INDEX sahsuland\_cancer\_sex ON rif\_data.sahsuland\_cancer (trunc(age\_sex\_group/100));

CREATE UNIQUE INDEX sahsuland\_cancer\_pk ON rif\_data.sahsuland\_cancer (year,level4,age\_sex\_group,icd);

CREATE INDEX sahsuland\_cancer\_year ON rif\_data.sahsuland\_cancer (year);

CREATE INDEX sahsuland\_cancer\_age\_group ON rif\_data.sahsuland\_cancer (mod(age\_sex\_group,100));

CREATE INDEX sahsuland\_cancer\_level4 ON rif\_data.sahsuland\_cancer (level4);

CREATE INDEX sahsuland\_cancer\_level2 ON rif\_data.sahsuland\_cancer (level2);

CREATE INDEX sahsuland\_cancer\_level3 ON rif\_data.sahsuland\_cancer (level3);

CREATE INDEX sahsuland\_cancer\_level1 ON rif\_data.sahsuland\_cancer (level1);

COMMENT ON TABLE rif\_data.sahsuland\_cancer IS 'SAHSU land Population';

COMMENT ON COLUMN rif\_data.sahsuland\_cancer.age\_sex\_group IS 'Age sex group';

COMMENT ON COLUMN rif\_data.sahsuland\_cancer.icd IS 'ICD';

COMMENT ON COLUMN rif\_data.sahsuland\_cancer.level1 IS 'level1';

COMMENT ON COLUMN rif\_data.sahsuland\_cancer.level2 IS 'level2';

COMMENT ON COLUMN rif\_data.sahsuland\_cancer.level3 IS 'level3';

COMMENT ON COLUMN rif\_data.sahsuland\_cancer.level4 IS 'level4';

COMMENT ON COLUMN rif\_data.sahsuland\_cancer.total IS 'Total';

COMMENT ON COLUMN rif\_data.sahsuland\_cancer.year IS 'Year';

GRANT SELECT ON rif\_data.sahsuland\_cancer TO rif\_user, rif\_manager;

User may wish to have project and dataset specific grants. These will be supported by the information governance tool in a later release, access is current granted to the standard roles: *rif\_user* and *rif\_manager*.

Scripts exist to partition numerator data on Postgres. These will be supports by the data loader in a later release.

### Numerator mandatory columns

These columns must be present. They do not have to be populated in the data

* Year
* age\_sex\_group
* all <geolevel name> columns
* <outcome column> e.g. ICD
* <total field>, e.g. TOTAL if defined below. The RIF aggregates the data total using either COUNT(<primary key column>) or SUM(<total field>) if rif40\_tables.total\_field is defined.

If the primary key column is not defined then *COUNT(\*)* will be used. This is inefficient, especially on Postgres.

The table data must be loaded:

\COPY sahsuland\_cancer FROM '../sahsuland/data/sahsuland\_cancer.csv' WITH (FORMAT csv, QUOTE '"', ESCAPE '\');

Setup the numerator meta data:

DELETE FROM rif40\_tables WHERE table\_name = 'SAHSULAND\_CANCER';

INSERT INTO rif40\_tables(table\_name, description, theme, year\_start, year\_stop, total\_field, isindirectdenominator, isdirectdenominator, isnumerator, automatic, age\_group\_id)

VALUES (

'SAHSULAND\_CANCER','Cancer cases in SAHSU land', 'SAHSULAND', 1989,1996,'TOTAL',0,0,1,1,1);

Numerator RIF40\_TABLES column definitions

* THEME: Health study theme (RIF40\_HEALTH\_STUDY\_THEMES.THEME)
* TABLE\_NAME: RIF table name. Normally the schema owner will not be able to see the health data tables, so no error is raised if the table cannot be resolved to an accessible object. The schema owner must have access to automatic indirect standardisation denominators.
* DESCRIPTION: Description of table
* YEAR\_START: Year table data starts
* YEAR\_STOP: Year table data stops
* TOTAL\_FIELD: Total field (when used aggregated tables)
* ISNUMERATOR: Is table a numerator (0/1): Set to 1.
* ISINDIRECTDENOMINATOR: table a denominator to be used in indirect standardisation (0/1). This is always the opposite of ISDIRECTDENOMINATOR. 0 if the table is a denominator as only indirect standardisation is supported at present. Set to 0.
* ISDIRECTDENOMINATOR: Is table a denominator to be used in direct standardisation (0/1). E.g. POP\_WORLD, POP\_EUROPE. Always 0 (unless the denominator is a standard population). Direct standardisation using standard population is not currently supported. Set to 0.
* AUTOMATIC: Able to be used in automatic RIF40\_NUM\_DENOM (0/1, default 0). Cannot be applied to direct standardisation denominator. Restricted to 1 denominator per geography to prevent the automatic RIF40\_NUM\_DENOM having >1 pair per numerator. This restriction is enforced in RIF40\_NUM\_DENOM because of the “ORA-04091: table RIF40.RIF40\_TABLES is mutating, trigger/function may not see it”; error. A user specific T\_RIF40\_NUM\_DENOM is supplied for other combinations. The default is 0 because of the restrictions.
* SEX\_FIELD\_NAME: Name of SEX field. No default. AGE\_GROUP\_FIELD\_NAME must be set, AGE\_SEX\_GROUP\_FIELD\_NAME must not be set. Always NULL as only an age\_sex\_group\_field\_name of AGE\_SEX\_GROUP is currently supported.
* AGE\_GROUP\_FIELD\_NAME: Name of AGE\_GROUP field. No default. SEX\_FIELD\_NAME must be set, AGE\_SEX\_GROUP\_FIELD\_NAME must not be set. Always NULL as only an age\_sex\_group\_field\_name of AGE\_SEX\_GROUP is currently supported.
* AGE\_SEX\_GROUP\_FIELD\_NAME: Name of AGE\_SEX\_GROUP field. Default: AGE\_SEX\_GROUP; AGE\_GROUP\_FIELD\_NAME and SEX\_FIELD\_NAME must not be set. Always ‘AGE\_SEX\_GROUP’ as only an age\_sex\_group\_field\_name of AGE\_SEX\_GROUP is currently supported.
* AGE\_GROUP\_ID: Type of RIF age group in use. Link to RIF40\_AGE\_GROUP\_NAMES. No default.
* VALIDATION\_DATE: Date table contents were validated OK.

Integrate ICD data:

ICD columns must be added to the table RIF40\_TABLE\_OUTCOMES so that the data extract can create the correct outcome filter. There are two control tables: RIF40\_OUTCOMES for the differing types of ontology (e.g. ICD) and RIF40\_OUTCOME\_GROUPS for sets of one or more outcome filter column names.

Users either create a new outcome group, e.g.

<add>

Or use a pre-existing one, e.g. '*SINGLE\_ICD'.* In this case the expected column name is ICD and *current\_version\_start\_year* is the first year of ICD10 data.

DELETE FROM rif40\_table\_outcomes WHERE numer\_tab = 'SAHSULAND\_CANCER';

INSERT INTO rif40\_table\_outcomes(outcome\_group\_name, numer\_tab, current\_version\_start\_year)

VALUES ('SINGLE\_ICD', 'SAHSULAND\_CANCER', 1993);

In the case of multiple value …

## Denominator data

The denominator table must be created, indexed, commented and access granted:

CREATE TABLE pop.sahsuland\_pop (

year smallint NOT NULL,

age\_sex\_group smallint NOT NULL,

level1 varchar(20) NOT NULL,

level2 varchar(20) NOT NULL,

level3 varchar(20) NOT NULL,

level4 varchar(20) NOT NULL,

total double precision NOT NULL

);

CREATE INDEX sahsuland\_pop\_level4 ON pop.sahsuland\_pop (level4);

CREATE INDEX sahsuland\_pop\_age\_sex\_group ON pop.sahsuland\_pop (age\_sex\_group);

CREATE INDEX sahsuland\_pop\_level1 ON pop.sahsuland\_pop (level1);

CREATE INDEX sahsuland\_pop\_age\_group ON pop.sahsuland\_pop (mod(age\_sex\_group,100));

CREATE UNIQUE INDEX sahsuland\_pop\_pk ON pop.sahsuland\_pop (year,level4,age\_sex\_group);

CREATE INDEX sahsuland\_pop\_year ON pop.sahsuland\_pop (year);

CREATE INDEX sahsuland\_pop\_level2 ON pop.sahsuland\_pop (level2);

CREATE INDEX sahsuland\_pop\_sex ON pop.sahsuland\_pop (trunc(age\_sex\_group/100));

CREATE INDEX sahsuland\_pop\_level3 ON pop.sahsuland\_pop (level3);

COMMENT ON TABLE pop.sahsuland\_pop IS 'SAHSU land Population';

COMMENT ON COLUMN pop.sahsuland\_pop.age\_sex\_group IS 'Age sex group';

COMMENT ON COLUMN pop.sahsuland\_pop.level1 IS 'level1';

COMMENT ON COLUMN pop.sahsuland\_pop.level2 IS 'level2';

COMMENT ON COLUMN pop.sahsuland\_pop.level3 IS 'level3';

COMMENT ON COLUMN pop.sahsuland\_pop.level4 IS 'level4';

COMMENT ON COLUMN pop.sahsuland\_pop.total IS 'Total';

COMMENT ON COLUMN pop.sahsuland\_pop.year IS 'Year';

GRANT SELECT ON pop.sahsuland\_pop TO rif\_user, rif\_manager;

Scripts exist to partition denominator data on Postgres. These will be supports by the data loader in a later release.

### Denominator mandatory columns

These columns must be present. They do have to be populated in the data, i.e. all years, age\_sex\_group and geolevel column present. If this is not so, cases may be excluded from the study even if the population is zero. The years present should equal or exceed the numerator data.

* Year
* age\_sex\_group
* all <geolevel name> columns
* <total field>, e.g. TOTAL if defined below. The RIF aggregates the data total using either COUNT(<primary key column>) or SUM(<total field>) if rif40\_tables.total\_field is defined.

If the primary key column is not defined then *COUNT(\*)* will be used. This is inefficient, especially on Postgres.

The table data must be loaded:

\COPY sahsuland\_pop FROM '../sahsuland/data/sahsuland\_pop.csv' WITH (FORMAT csv, QUOTE '"', ESCAPE '\');

Setup the denominator meta data:

DELETE FROM rif40\_tables WHERE table\_name = 'SAHSULAND\_POP';

INSERT INTO rif40\_tables(table\_name, description, theme, year\_start, year\_stop, total\_field, isindirectdenominator, isdirectdenominator, isnumerator, automatic, age\_group\_id)

VALUES (

'SAHSULAND\_POP','SAHSU land population', 'SAHSULAND', 1989,1996,NULL,1,0,0,1,1);

For other RIF40\_TABLES column definitions see numerator data:

* ISNUMERATOR: Is table a numerator (0/1): Set to 0.
* ISINDIRECTDENOMINATOR: table a denominator to be used in indirect standardisation (0/1). This is always the opposite of ISDIRECTDENOMINATOR. 0 if the table is a denominator as only indirect standardisation is supported at present.
* ISDIRECTDENOMINATOR: Is table a denominator to be used in direct standardisation (0/1). E.g. POP\_WORLD, POP\_EUROPE. Always 0 (unless the denominator is a standard population). Direct standardisation using standard population is not currently supported.

## Covariate Data

Covariate data is required if *covariate\_table\_name* is defined for a geolevel in rif40\_covariates. Covariates are normally only defined for the higher resolution geoelvels and must not be defined for the first level (low resolution, e.g. country).

The covariate table(s) must be created, indexed, commented and access granted:

CREATE TABLE rif\_data.sahsuland\_covariates\_level3 (

year smallint NOT NULL,

level3 varchar(20) NOT NULL,

ses smallint NOT NULL,

ethnicity smallint NOT NULL

);

CREATE UNIQUE INDEX sahsuland\_covariates\_level3\_pk ON rif\_data.sahsuland\_covariates\_level3 (year,level3);

CREATE TABLE rif\_data.sahsuland\_covariates\_level4 (

year smallint NOT NULL,

level4 varchar(20) NOT NULL,

ses smallint NOT NULL,

areatri1km smallint NOT NULL,

near\_dist double precision NOT NULL,

tri\_1km smallint NOT NULL

);

CREATE UNIQUE INDEX sahsuland\_covariates\_level4\_pk ON rif\_data.sahsuland\_covariates\_level4 (year,level4);

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level3.ethnicity IS 'Ethnicity % non white - 1: <5%, 2: 5 to 10%, 3: >= 10%';

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level3.level3 IS 'Level3';

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level3.ses IS 'Social Economic Status (quintiles)';

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level3.year IS 'Year';

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level4.areatri1km IS 'Toxic Release Inventory within 1km of area (0=no/1=yes)';

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level4.level4 IS 'Level4';

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level4.near\_dist IS 'Distance (m) from area centroid to nearest TRI site';

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level4.ses IS 'Social Economic Status (quintiles)';

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level4.tri\_1km IS 'Toxic Release Inventory within 1km of areai centroid (0=no/1=yes)';

COMMENT ON COLUMN rif\_data.sahsuland\_covariates\_level4.year IS 'Year';

GRANT SELECT ON rif\_data.sahsuland\_covariates\_level3 TO PUBLIC;

GRANT SELECT ON rif\_data.sahsuland\_covariates\_level4 TO PUBLIC;

### Covariate mandatory columns

These columns must be present. They do have to be populated in the data, i.e. all years, age\_sex\_group and geolevel columns present and the values must be within the specified ranges. If this is not so, cases may be excluded from the study if the covariate is NULL, zero out outside of the defined maximum and minimum values.

* Year
* <geolevel name>
* <covariate fields>, e.g. SES

Annualised data (e.g. incomes) and non-annualised data is supported. For non-annualised data an entry must be present for all years covered by the denominator data.

The table data must be loaded:

\COPY sahsuland\_covariates\_level3 FROM '../sahsuland/data/sahsuland\_covariates\_level3.csv' WITH (FORMAT csv, QUOTE '"', ESCAPE '\');

\COPY sahsuland\_covariates\_level4 FROM '../sahsuland/data/sahsuland\_covariates\_level4.csv' WITH (FORMAT csv, QUOTE '"', ESCAPE '\');

Covariate data is already partly setup in rif40\_geolevels. To complete the setup the individual covariate columns must be added to RIF40\_COVARIATES

--

-- RIF40\_COVARIATES

--

DELETE FROM rif40\_covariates WHERE geography = 'SAHSU';

INSERT INTO rif40\_covariates(geography, geolevel\_name, covariate\_name, min, max, type)

SELECT 'SAHSU', 'LEVEL3', 'SES', MIN(ses), MAX(ses), 1 FROM sahsuland\_covariates\_level3;

INSERT INTO rif40\_covariates(geography, geolevel\_name, covariate\_name, min, max, type)

SELECT 'SAHSU', 'LEVEL3', 'ETHNICITY', MIN(ethnicity), MAX(ethnicity), 1 FROM sahsuland\_covariates\_level3;

INSERT INTO rif40\_covariates(geography, geolevel\_name, covariate\_name, min, max, type)

SELECT 'SAHSU', 'LEVEL4', 'SES', MIN(ses), MAX(ses), 1 FROM sahsuland\_covariates\_level4;

INSERT INTO rif40\_covariates(geography, geolevel\_name, covariate\_name, min, max, type)

SELECT 'SAHSU', 'LEVEL4', 'AREATRI1KM', MIN(areatri1km), MAX(areatri1km), 1 FROM sahsuland\_covariates\_level4;

INSERT INTO rif40\_covariates(geography, geolevel\_name, covariate\_name, min, max, type)

SELECT 'SAHSU', 'LEVEL4', 'TRI\_1KM', MIN(tri\_1km), MAX(tri\_1km), 1 FROM sahsuland\_covariates\_level4;

INSERT INTO rif40\_covariates(geography, geolevel\_name, covariate\_name, min, max, type)

SELECT 'SAHSU', 'LEVEL4', 'NEAR\_DIST', MIN(near\_dist), MAX(near\_dist), 2 FROM sahsuland\_covariates\_level4;

RIF40\_COVARIATES column definitions:

* GEOGRAPHY: Geography (e.g. EW2001)
* GEOLEVEL\_NAME: Name of geolevel. This will be a column name in the numerator/denominator tables
* COVARIATE\_NAME: Covariate name. This will be a column name in RIF40\_GEOLEVELS.COVARIATE\_TABLE
* MIN: Minimum value
* MAX: Maximum value
* TYPE: TYPE of covariate (1=integer score/2=continuous numeric variable). Min must be less than max. Max and min precision is appropriate to type. Continuous variables are not currently supported. Integer scores can be a binary variable 0/1 or an NTILE e.g. 1 to 5 for a quintile.