

Get Data Out technical documentation

Incidence

GDO 0021, 2021-04-30

Background

The incidence counts and crude incidence rates are calculated for each Get Data Out grouping. Crude rates are helpful in determining the cancer burden and specific needs for services for a given population, compared with another population, regardless of size.

The [NCRAS SOP - Crude incidence rates, age specific rates and ASRs](#) was followed. This technical documentation should be read together with the SOP as it provides additional information about the snapshots that were used and the code that was run.

1. Populations of interest

The populations of interest were defined as in the documents below which are available on the Get Data Out website:

[Bladder_Grouping.docx](#),
[Bone_Grouping.docx](#),
[Brain_Grouping.docx](#),
[Head_and_neck_Grouping.docx](#),
[Kaposi_Grouping.docx](#),
[Kidney_Grouping.docx](#),
[Oesophageal_and_Stomach_Grouping.docx](#),
[Ovarian_Grouping.docx](#),
[Pancreatic_Grouping.docx](#),
[Prostate_Grouping.docx](#),
[Sarcoma_Grouping.docx](#),
[Soft_Tissue_Grouping.docx](#),
[Testicular_Grouping.docx](#).

Data for 2013-2018 are included. The documents explain the partition of cases into groups. Incidence counts, crude rates and confidence intervals were produced for all these groups.

2. Counting the number of cancer cases

The number of cancer cases were taken from AV2018.AV_TUMOUR on CASREF01. The *CAS SOP #1 – Counting Cancer Cases* was followed, with the additional exclusion of testicular tumours in female patients and ovarian tumours in male patients for tumours whose ICD code starts with D. (The SOP handles C-coded tumours but does not yet perform any exclusion on D-coded tumours: sex-exclusions are performed on all tumours for our dataset).

The code used to extract the cohort (and organise by cancer type) are in the following documents:

GDO_0021_inc_Bladder_13-18_sql.txt,
GDO_0021_inc_Bone_13-18_sql.txt,
GDO_0021_inc_Brain_13-18_sql.txt,
GDO_0021_inc_Head_neck_13-18_sql.txt,
GDO_0021_inc_Kaposi_13-18_sql.txt,
GDO_0021_inc_Kidney_13-18_sql.txt,
GDO_0021_inc_Oesophagus_Stomach_13-18_sql.txt,
GDO_0021_inc_Ovary_13-18_sql.txt,
GDO_0021_inc_Pancreas_13-18_sql.txt,
GDO_0021_inc_Prostate_13-18_sql.txt,
GDO_0021_inc_Sarcoma_13-18_sql.txt,
GDO_0021_inc_Soft_Tissue_13-18_sql.txt,
GDO_0021_inc_Testes_13-18_sql.txt

3. Counting At Risk populations

The number of people in the At Risk population was taken from ONS2018.POPULATIONS_NORMALISED on CASREF01. The codes above also extract the populations.

4. Calculating the Crude Rates and Confidence Intervals

The crude rates and their confidence intervals were calculated in R Studio via PHE rate from the **PHEindicatormethod** package. For numerators that were equal

to or greater than 10 Byar's method is applied. For small numerators, Byar's method is less accurate and so an exact method based on the Poisson distribution is used. As discussed in the **APHO Technical Briefing 3 - Commonly Used Public Health Statistics and their Confidence Intervals**, the rates can be approximately described by a Poisson distribution, and confidence intervals around them produced according to this. For small observed counts a precise confidence interval can be calculated using Poisson functions, and for larger numbers Byar's approximation can be used. This methodology agrees with the **NCRAS SOP - Crude incidence rates, age specific rates and ASRs**.

5. QA

The QA was done internally by the Get Data Out team.

6. Alterations to technical specification

Column names for the incidence data have been changed since our previous publication. We know this may make difficulties for people leading the data into standard templates, and we aim to minimise this. The rationale for these changes are:

- Removal of '(95%)' from column names – this is standardising with the Routes file. The characters '(', ')' and '%' in column names cause difficulties when loading data into some software such as R

Changes to the column names from the previous release are documented below:

Previous column name	New column name
Incidence Rate LCI (95%)	Incidence Rate LCI
Incidence Rate UCI (95%)	Incidence Rate UCI

Any questions, please contact sally.vernon@phe.gov.uk, charlotte.eversfield@phe.gov.uk, anna.cooper@phe.gov.uk,