Public Health
England

Protecting and improving the nation’s health

National Cancer Registration and Analysis Service
Standard Operating Procedure
Crude Incidence, Age Specific and Age Standardised Rates

1. Summary

CASSOP #1 explains how to define a standard incidence cohort to count the number of cases of cancer diagnosed. When trying to compare the number of cases of cancer diagnosed across, for example, time or geographical regions, it is important to look at the **rate** of cancer - the number of cases of cancer divided by the at risk population.

This SOP sets out basic methodology for calculating crude rates and directly age standardised rates of cancer.

Crude rates are helpful in determining the cancer burden and service provision for a given population (compared with another population). However, sometimes it is useful to understand the cancer burden to only a subset of the population. If this subset is based on a specific age group, it is referred to as age specific rate.

It is often useful to calculate age standardised incidence rates as well as crude and age specific incidence rates. This is because disease and mortality rates may vary widely by age and this complicates comparisons made between two populations that have different age structures.

The most comprehensive way of comparing the disease experience of two populations is to present and compare their age specific rates. However, when the number of populations being compared increases, the volume of data that needs to be considered quickly becomes unmanageable. What is needed is a single, easily interpreted, summary figure for each population that is adjusted to take into account its age structure. Such summary figures are calculated using age standardisation methods. It is recommended to also standardise for sex. It is possible to standardise by other variables, such as level of deprivation,that may also potentially confound any comparisons

The two most common methods of age standardisation are:

- Indirect: The age specific rates of a chosen standard population (usually the relevant national population) are applied to the age structure of the subject population to give an expected number of events. The observed number of events is then compared to that expected and is usually expressed as a ratio (observed/expected). A common example is the standardised mortality ratio (SMR). Guidance for producing indirect standardised rates are detailed in this SOP.

- Direct: The age specific rates of the subject population are applied to the age structure of the standard population. This gives the overall rate that would have occurred in the subject population if it had the standard age-profile.

2. Method

Defining crude rates

The crude rate equals the total number of new cancer cases diagnosed in a specific year in the population category of interest, divided by the at-risk population for that category. This is expressed by the formula below. Sections in Part 3 detail how to gain the numerator and denominator for this.

$$Crude\ Rate = \frac{New\ Cancer\ Cases}{At\ Risk\ population} * 100{,}000$$

Cancer rates are usually given as 'rate per 100,000 people', and so are then multiplied by 100,000. This is to give a number that is a more 'natural' size, so we can talk about rates of, eg, '30', not '0.0003'.

For rare cancers it is possible to give a rate per 1,000,000 people, but this must be made very clear as it should not be accidently compared to rates per 100,000 people.

Defining Age specific rates

An age specific rate is calculated in the same manner as a crude rate. The number of cancer cases is divided by the at risk population, however the exception in age specific rates is that both the number of cancer cases and the at risk population are restricted to a certain age group. This is expressed in the formula below.

$$Age\ Specific\ Rate = \frac{Number\ of\ cancer\ cases\ in\ a\ specified\ age\ group}{Population\ of\ the\ specified\ age\ group} * 100{,}000$$

Defining Age standardised rates

The age standardised rate is a single summary measure for each population of interest that reflects the numbers of events that would have been expected if the populations being compared had the same age distribution. It therefore allows rates for differing geographies to be compared without bias resulting from the population age structures. It is also useful when comparing ASRs for one geography over time.

The European Standard Population (ESP 2013) is used to standardise the rates as below:

For each age band of each population being compared, age specific incidence rates are multiplied by the size of the standard population for that age band. This provides the number of cases expected in the standard population if it had the same incidence rates as the population/s of interest. Then, the total number of expected cases is calculated by summing all the values from the age specific calculations. Age-standardised rates are calculated by dividing the total number of expected cases by the total standard population size.

The ASR calculation, for information, can be found here.

Methods to calculate ASRs are presented in Part 4.

Calculating the Confidence Intervals for rates

As well as rates being the measurement of what happened, rates can be viewed as an estimate of the underlying risk of cancer in the population. Thus it is meaningful to put confidence intervals around the rate, to see if the risk in different populations is statistically significantly different.

NCRAS recommends using the Dobson method for calculating confidence intervals[1]. The details of this can be found here.

Confidence intervals using this method are produced via all methods detailed in this SOP.

Interpreting crude rates, age specific and age standardised rates

Confidence intervals are used to interpret whether a crude rate, age specific rate or ASR is statistically higher or lower than another. If the confidence intervals of the crude rate, age specific rate or ASR have no overlap with comparison rate confidence intervals then it is statistically significantly higher/lower than the comparison. If there is overlap then there is no statistically significant difference between them.

Crude and age specific rates based on a numerator of less than three cases should be considered for suppression or flagged as potentially unreliable. It is recommended that ASRs calculated with fewer than ten cases across all age bands are suppressed. In both cases, when the rates are based on such low numbers, the rates are susceptible to inaccurate interpretation.

[1] The NCRAS method is closely aligned to the PHE method; the latter using Byar's approximation.

3. Extracting the data from CAS

Defining the population of interest

The population of interest must be defined in the same way for the numerator and the denominator for calculating rates. You will need to specify the population by defining at least

- Time period (usually a single year, or a range of years)
- Geography (such as all England, a region, or a CCG)
- Sex (persons, males or females)
- Age groups (all ages, or specific age ranges. Note our population file only allows five year age bands for age groups)

If your population of interest is one of the standard populations that is available in CancerStats, this SOP recommends you extract the crude rates and age standardised rates from Cancerstats. If it is not, you will need to extract the data from CAS for your population of interest by following the rest of this SOP.

The population of interest could also be defined by other characteristics, such as deprivation or ethnicity. The important thing is that the population as defined for the numerator ('new cancer cases') must be equivalent to the population defined for the denominator ('at risk population').

Total number of new cancer cases

The total number of new cancer cases should be counted following CASSOP #1, defining a standard incidence cohort for your cancer of interest, with filters for your population of interest.

You should choose the snapshot to run your query on carefully. The data should be signed off as finished for the time period you wish to calculate incidence rates for. Generally, the most recent annual snapshot in CASREF01 is the recommended choice. This snapshot is the basis for figures in CancerStats/CancerData, the National Statistics and will normally be available for longer than a monthly snapshot.

Please seek guidance from a member of the Analytical SMT if you think this snapshot will not meet your needs, so they can help you find a suitable alternative.

Always document the snapshot and table that was used in your code and metadata.

See Reference code 1 and 2 for examples of extracting numbers of cases.

At Risk Population

The total number of people in the at risk population should be summed from the population tables on CASREF01, using the same filters for the same population of interest as were used to count the cancer cases.

The most recent population tables should be used in almost all circumstances. At the time of writing this was `ons2019.populations_normalised` on CASREF01.

See Reference code 1 and 2 for examples of extracting population data.

4. Calculating crude and age standardised rates

There are multiple ways to calculate age standardised rates. All the methods detailed produce crude rates too.

We recommend that R or Stata is used to calculate ASRs. Reference code 2 provides sql code which can be used to extract incidence and population data from the CAS system in a format to be inputted into either of these programmes. We have then included basic ASR calculations in R code and Stata which can be used as a base for your work.

Additionally Ruby code is provided as another means of calculating ASRs. Ruby code is used to produce the rates for the CancerStats website.

Also included is an Excel spreadsheet for calculating ASRs, whereby incidence and population data can be inputted and rates are automatically calculated. This is not one of the preferred methods as the method is less transparent in terms of QA'ing and there is a risk of copy and paste errors. However, this method is recommended for: colleagues who are new and haven't coded before and would like to understand how to calculate ASRs; double-checking values calculated in R or Stata; checking a value or two where it will save considerable time to use this method.

RStudio

To calculate ASRs in R, the following steps can be used. Example R code can be found in the R notebook file embedded below:

R code for
calculating ASRs_fin

1) Set up your R script
    a. Set your working directory
    b. Install required packages (only needs to be done once)
    c. Load required packages

2) Set up data for your population of interest
    a. Import the observed cases as a data frame
    b. You need to ensure there is a row for every combination of your variables (e.g. age, sex, year). If there isn't you will need to create new rows with zero counts.

3) Import the mid-year population estimates from your population
    a. Import the population as a dataframe.
    b. Ensure there is a row for every combination of your variables (e.g. age, sex, year). If there isn't you will need to create new rows with zero counts.

4) Combine the two dataframes using the strata of interest (sex, age)

5) Create the standardised population
    a. Use the European Standard Population (ESP) for five year age bands repeated twice, once for males, once for females. This will allow the ASR to be sex as well as age standardised.

   b. And/or, if you need sex specific cancer ASRs, use the ESP without replicating it

6) Calculate ASRs

   a. Create a function to produce one ASR

   b. Create an ASR function, which uses the function above, with the method for using the dataframe, grouping variables, standardisation variables and standard population.
      This function calculates confidence intervals using the recommended Byar’s method with Dobson method adjustment method

   c. Create ASRs using the ASR function

Stata

There are various ways to calculate ASRs in Stata. Like R it is easy to produce lots of combinations of ASRs. To calculate ASRs in Stata, the following steps and code should be used. Note you may have to adapt the code slightly dependent on what you want to standardise on and what geographic breakdown you are interested in.

Example Stata code can be found here:

Stata code for ASR calculation.do

1) Set up your Stata script
   a. Set your working directory

2) Set up data for your population of interest
   a. Import the observed cases
   b. You need to ensure there is a row for every combination of your variables (e.g. age, sex, year). If there isn’t you will need to create new rows with zero counts.
   c. Save the data as a Stata data file.

d. Import the mid-year population estimates from your population as a dataframe.

e. You need to ensure there is a row for every combination of your variables (e.g. age, sex, year). If there isn’t you will need to create new rows with zero counts.

f. Save the data as a Stata data file.

3) Merge the datasets together

a. Merge the two datasets using the strata of interest (sex, age)

b. Save the data as a Stata data file.

4) Calculate ASRs

a. You need to save the ado file asr.ado to your ado folder on your computer – this is located here “C:\ado\Personal”

b. Use the asr command to calculate ASRs and confidence intervals. Use split to group the data as needed.

Ruby

Ruby is another programming language which is used to create the ASRs for Cancerstats. If you wish to use this please find the ASR code at the link below:

asr_functions.rb

Excel

The embedded excel spreadsheet allows users to input incidence/mortality and population data. The spreadsheet will then automatically calculate the ASRs for you.

ASR calculation spreadsheet.xlsx

Reference Code 1: extracting data from CAS for crude rates

This code will give a count of the incidence of cancers for a specified cohort. This is example code for counting all liver cancers (C22) in 2013, for men aged 50 and over, for all England.

```
select  site_ICD10_O2_3char, diagnosisyear, ctry_code as geography
, sex
, '50+' as age
, count(tumourid)
from av2015.av_tumour@CASREF01
where
-- ***CASSOP 01 RULES***
-- English
ctry_code ='E'
-- Final
and STATUSOFREGISTRATION ='F'
-- Not duplicates
and dedup_flag=1
-- Exclude patients with age over 200
and age between 0 and 200
-- Known sex
and sex in (1,2)
-- Sex agrees with cancer type
and ((sex = '2' and site_ICD10_O2_3char not in ('C60','C61','C62','C63'))
or (sex = '1' and  site_ICD10_O2_3char not in ('C51','C52','C53','C54','C55','C56','C57','C58')))

-- ***COHORT OF INTEREST***
-- Cancer site of interest (Liver in this example)
and site_ICD10_O2_3char = 'C22'
-- Geography (all England in this example)
and ctry_code ='E'
-- Sex (male in this example)
and sex = 1
-- Year of interest (2013 in this example)
and diagnosisyear= 2013
-- Age (50+ in this example)
and age > = 50
group by site_ICD10_O2_3char, diagnosisyear, ctry_code, sex
order by site_ICD10_O2_3char, diagnosisyear, ctry_code, sex;
```

This code will give a count of at risk population for a specified cohort. This is example code for counting all men aged 50 and over, for all England.

```
select sum (popcount) from  ons2015.populations_normalised
-- Geography
-- All England in this example
-- (can join on LSOA look ups for other geographies)
where substr(lsoa11,1,1) = 'E'
-- Time period of interest  (2013 in this example)
and year = '2013'
-- Sex (male in this example)
and sex = '1'
-- Age (50+ in this example)
-- Note that these are coded, '1' is 0-4s, '2' is 5-9s etc
and quinaryagegroupint >= 11
;
```

Reference Code 2: extracting data from CAS for ASRs (and crude rates and age specific rates)

This code will produce population counts and incidence counts for use in ASR calculations

```
--- Select population counts for use in ASRs
SELECT
SUM (popcount) AS POPULATION
,Year
--add whichever geographies you would like through look up
,ccg17cdh as ccg_code
,CCG17NM as ccg_name
,Sex
--this case statement is necessary as excel converts some of these range to dates on output

,case when replace (quinaryagegroup,' ','') = '0-4' then 1
 when replace (quinaryagegroup,' ','') = '5-9' then 2
 when replace (quinaryagegroup,' ','') = '10-14' then 3
 when replace (quinaryagegroup,' ','') = '15-19' then 4
 when replace (quinaryagegroup,' ','') = '20-24' then 5
 when replace (quinaryagegroup,' ','') = '25-29' then 6
 when replace (quinaryagegroup,' ','') = '30-34' then 7
 when replace (quinaryagegroup,' ','') = '35-39' then 8
 when replace (quinaryagegroup,' ','') = '40-44' then 9
 when replace (quinaryagegroup,' ','') = '45-49' then 10
 when replace (quinaryagegroup,' ','') = '50-54' then 11
 when replace (quinaryagegroup,' ','') = '55-59' then 12
 when replace (quinaryagegroup,' ','') = '60-64' then 13
 when replace (quinaryagegroup,' ','') = '65-69' then 14
 when replace (quinaryagegroup,' ','') = '70-74' then 15
 when replace (quinaryagegroup,' ','') = '75-79' then 16
 when replace (quinaryagegroup,' ','') = '80-84' then 17
 when replace (quinaryagegroup,' ','') = '85-89' then 18
 when (replace (quinaryagegroup,' ','')) = '90+' then 19
 else null end as fiveyearageband


FROM

--joining from LSOA to LSOA need to use distinct command, as there will be several rows with the
same LSOA code. Create a table within a query,
(select distinct lsoa11,sex,quinaryagegroup,year,popcount from
ons2016.POPULATIONS_NORMALISED@casref01.encore.nhs.uk) a

inner join ANALYSISNCR.LSOA_CA_201706 b on a.LSOA11 = b.lsoa11cd

--filter for what years you would like to study, you can also restrict geographies here
where
year between '2010' and '2016'
and a.lsoa11 like 'E%'

GROUP BY
Year
----add whichever geograpies you would like through look up
----,a.lsoa11
,ccg17cdh
,ccg17nm
,sex
,quinaryagegroup
;

--Generic incidence code for ASR from av_tumour, edit as neccessary, gives counts instead of
individual records

select
```

```
        count (distinct tumourid)
        ,diagnosisyear as year --to match name in populations table
        ,site_icd10_o2_3char
--        ,(replace (fiveyearageband,' ','')) as fiveyearageband -- so its has the same format as
in populations table
        , case when replace (fiveyearageband,' ','') = '0-4' then 1
 when replace (fiveyearageband,' ','') = '5-9' then 2
 when replace (fiveyearageband,' ','') = '10-14' then 3
 when replace (fiveyearageband,' ','') = '15-19' then 4
 when replace (fiveyearageband,' ','') = '20-24' then 5
 when replace (fiveyearageband,' ','') = '25-29' then 6
 when replace (fiveyearageband,' ','') = '30-34' then 7
 when replace (fiveyearageband,' ','') = '35-39' then 8
 when replace (fiveyearageband,' ','') = '40-44' then 9
 when replace (fiveyearageband,' ','') = '45-49' then 10
 when replace (fiveyearageband,' ','') = '50-54' then 11
 when replace (fiveyearageband,' ','') = '55-59' then 12
 when replace (fiveyearageband,' ','') = '60-64' then 13
 when replace (fiveyearageband,' ','') = '65-69' then 14
 when replace (fiveyearageband,' ','') = '70-74' then 15
 when replace (fiveyearageband,' ','') = '75-79' then 16
 when replace (fiveyearageband,' ','') = '80-84' then 17
 when replace (fiveyearageband,' ','') = '85-89' then 18
 when replace (fiveyearageband,' ','') = '90+' then 19
 else null end as fiveyearageband
        ,sex
        ,ccg_code
        ,ccg_name
        ,ora_database_name

from  --analysisncr.at_tumour_england
AV2016.av_tumour@CASREF01

where
      diagnosisyear between 2010 and 2016
      and site_icd10_o2_3char like 'C34'
      and statusofregistration = 'F'
      and lsoa11_code like 'E%'
      and dedup_flag = '1'

group by
diagnosisyear
        ,site_icd10_o2_3char
        ,fiveyearageband
        ,sex
        ,ccg_code
        ,ccg_name
        ,ora_database_name
;
```