CAS-SOP #5
Crude Incidence Rates

V1.1 2017-11-15

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, it is important to look at the **rate** of cancer - the number of cases of cancer divided by the at risk population.

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.

This SOP sets out basic methodology for calculating the crude rates of cancer.

It is accompanied by an Excel spreadsheet, *CASSOP 5 Crude Incidence Rates.xlsx* , which calculates rates and CIs for given numerators and denominators.

It is often useful to calculate Age Standardised Incidence rates as well as Crude Incidence rates. These are out of scope in the current version of this SOP, but should be born in mind when chosing the correct statistics to present. There is on-going work to generalise this SOP to cover Age Standardised Rates and other rates such as mortality rates.

2 Defining the Crude Rate

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. 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'.

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

For very 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.

3 Defining the population of interest

The population of interest must be defined in the same way for the numerator and the denominator of the crude rate. 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 from Cancerstats (where you will also find the age standardised rates for the same cohort). If it is not, you will need to extract the data from the CAS for your population of interest, by following the rest of this SOP.

The population of interest could also be defined by other charactoristics, 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'.

4 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. *(This is really useful when a requester seeks clarification several months after the information is first issued.)*

Please seek guidance from a member of the Analytical SLT if you think this snapshot will not meet your needs, so they can help you find a suitable alternative. [At the time of writing AV2015.AV_TUMOUR on CASREF01 was the most recent annual snapshot, and is based on the CAS1612 snapshot.]

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

Reference Code 1 gives example code for counting all liver cancers (C22) in 2013, for men aged 50 and over, for all England.

5 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 ons2015.populations_normalised on CASREF01.

Again, **always** document the population table and which snapshot was used.

Reference Code 2 gives example code for counting the at risk population in 2013, of men aged 50 and over, for all England.

6 Calculating the Crude Rate

The Crude Rate can be calculated by inserting the numbers produced in sections 4 and 5 into the formula defined in section 2:

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

An easy way to do this for a small number of rates is using the Excel spreadsheet, *CASSOP #5 - Crude incidence rates 1.1.xlsx.* There are also standard STATA scripts that are available to do these calculations (the next draft of this SOP should include more details on these, John Broggio can be contacted for more details).

7 Calculating the Confidence Intervals.

In some senses, the crude rate is a precise point measurement - the cancer registry collects population level data, so this is not an 'estimate' of the rate based

on sampling. However, as well as the crude rate being the measurement of what happened, the crude rate can be viewed as an estimate of the underlying risk of cancer in the population. When viewed in this second sense it is meaningful to put confidence intervals around the rate, to see if the risk in different populations is statistically significantly different.

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. The details of this are given in Appendix A.

These confidence intervals are implemented in the the Excel spreadsheet *CASSOP #5 - Crude incidence rates 1.1.xlsx.* It is also acceptable to use the standard STATA scripts to do these (a future draft of this SOP should include more details on these, John Broggio can be contacted for more details).

Appendix A

The 100(1–α)% confidence limits for the rate r are given by:

$$r_{lower} = \frac{O_{lower}}{n}$$

$$r_{upper} = \frac{O_{upper}}{n}$$

where:

O_{lower} *and* O_{upper} are the lower and upper confidence limits for the observed number of events.

Using Byar's method, the 100(1–α)% confidence limits for the observed number of events are given by:

$$O_{lower} = O \times \left(1 - \frac{1}{9O} - \frac{z}{3\sqrt{O}}\right)^3$$

$$O_{upper} = (O+1) \times \left(1 - \frac{1}{9(O+1)} + \frac{z}{3\sqrt{(O+1)}}\right)^3$$

where:

z is the 100(1–α/2)th percentile value from the Standard Normal distribution. For example, for a 95% confidence interval, α = 0.05 and z = 1.96 (i.e. the 97.5th percentile value from the Standard Normal distribution).

For small numerators, Byar's method can be less accurate and an exact method based on the Poisson distribution can be used. For 95% confidence intervals, Byar's method is within 0.1% of the exact value for numerators of 13 or more, but for 99.8% confidence intervals it is within 0.1% of the exact value for numerators of at least 44.

Using the link between the Poisson and χ^2 distributions, the equations for O_{lower} and O_{upper} above can be replaced by:

$$O_{lower} = \frac{\chi^2_{lower}}{2}$$

$$O_{upper} = \frac{\chi^2_{upper}}{2}$$

where:

χ^2_{lower} is the 100(1–α/2)th percentile value from the χ^2 distribution with 2O degrees of freedom;

χ^2_{upper} is the 100(α/2)th percentile value from the χ^2 distribution with 2O+2 degrees of freedom.

This spreadsheet uses Excel's built-in functions for exact probabilities for all cases based on numerators under 389, in order to give the most accurate results. For higher numerators, Excel's statistical functions fail (intermittently), and while macros are available to calculate exact Poisson probabilities, it is simpler to use Byar's method, and extremely accurate to do so.

Reference Code 1

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;
```

Reference Code 2

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
;
```