Public Health England

Protecting and improving the nation's health

National Cancer Registration and Analysis Service's Cancer Analysis System (CAS)-SOP #1

Counting cancer cases

About Public Health England

Public Health England exists to protect and improve the nation's health and wellbeing, and reduce health inequalities. We do this through world-leading science, knowledge and intelligence, advocacy, partnerships and the delivery of specialist public health services. We are an executive agency of the Department of Health, and are a distinct delivery organisation with operational autonomy to advise and support government, local authorities and the NHS in a professionally independent manner.

Public Health England, Wellington House, 133-155 Waterloo Road, London
SE1 8UG Tel: 020 7654 8000 www.gov.uk/phe
Twitter: @PHE_uk Facebook: www.facebook.com/PublicHealthEngland

Prepared by: National Cancer Registration and Analysis Service (NCRAS).

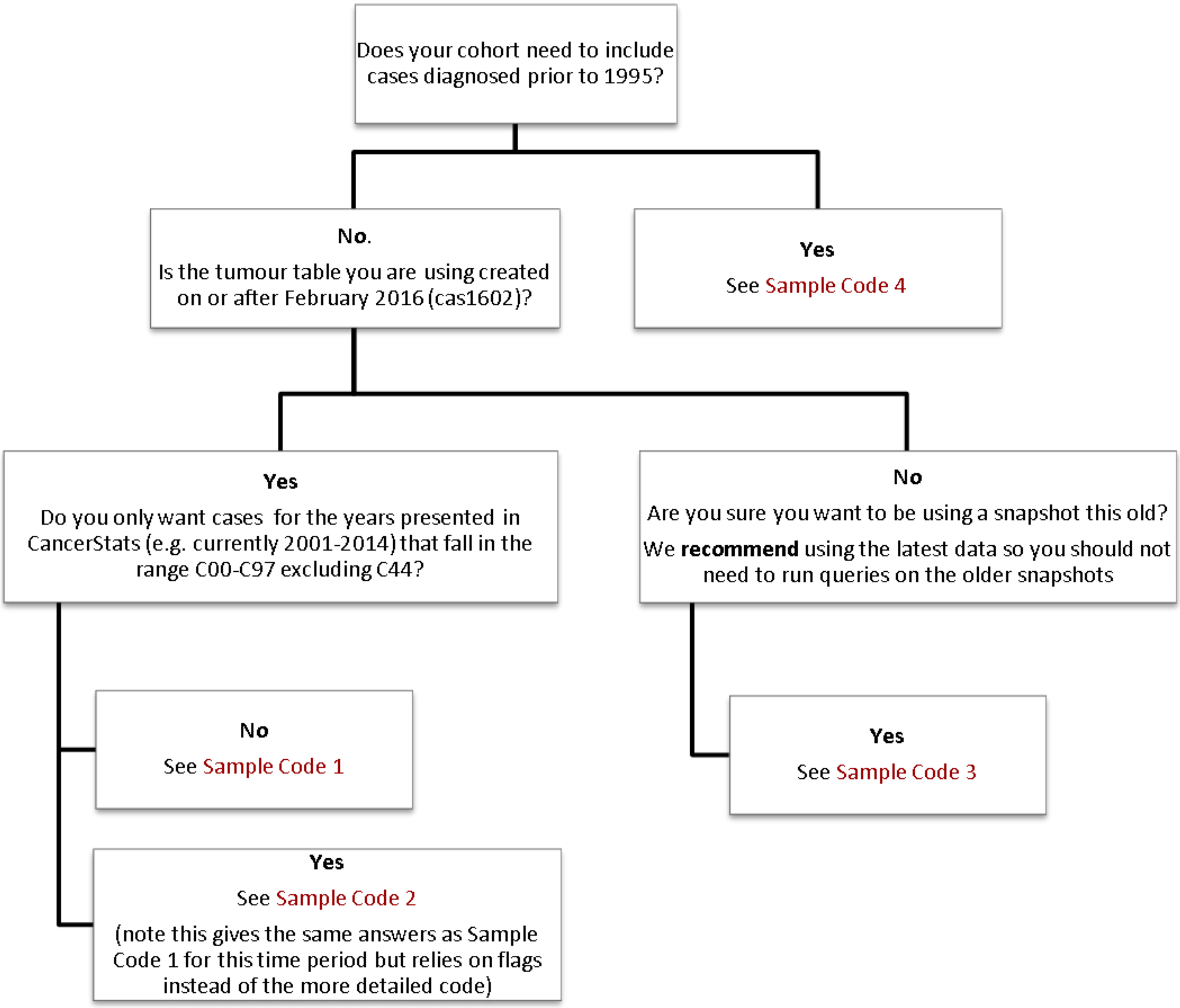
Published: November 2017
PHE publications
gateway number: 2017617

PHE supports the UN
Sustainable Development Goals

Contents

Introduction

Every year, data on over 300,000 cases of cancer are systematically registered by the National Cancer Registration and Analysis Service (NCRAS). These registrations include details on the patient, their type of cancer, how advanced it is and the treatment they receive. This Standard Operating Procedure (SOP) covers the process for counting cancer cases and extracting data on cancer incidence from the NCRAS Cancer Analysis System (CAS). It is not mandatory and depending on the nature of the request, a different approach may be adopted. It exists to outline a set of rules that can be followed to produce consistent and replicable results. The method used will depend on the diagnosis years you are interested in and the iteration of the cancer registration dataset you intend to use. The flow diagram will guide you to the relevant sample code. Note you may need to use combinations of the code depending on your project.


Does your cohort need to include cases diagnosed prior to 1995?
No.
Is the tumour table you are using created on or after February 2016 (cas1602)?
Yes
See Sample Code 4
Yes
Do you only want cases for the years presented in CancerStats (e.g. currently 2001-2014) that fall in the range C00-C97 excluding C44?
No
Are you sure you want to be using a snapshot this old?
We **recommend** using the latest data so you should not need to run queries on the older snapshots
No
See Sample Code 1
Yes
See Sample Code 3
Yes
See Sample Code 2
(note this gives the same answers as Sample Code 1 for this time period but relies on flags instead of the more detailed code)

For cases diagnosed in or after 1995

This SOP exists to outline the suggested exclusions that should be applied to the tumour table in CAS in order to define cancer cases for use in cancer incidence statistics.

1) **People who are resident outside England**
   The SOP recommends only including residents in England, so selecting records with a country code of E. In earlier datasets where the country code does not exist LSOA codes beginning with 'E' can be used instead.

2) **Cancer cases that the registration officers have not finalised**
   Provisional cases are registered but not confirmed to be cancer until they are finalised so important details about the cancer case may be subsequently added. Therefore, this SOP recommends only using finalised cases. This may not be the best course of action in all projects and there may be good reasons to include the provisional cases in some analyses.

3) **Cases that are considered to be duplicate records**
   The English cancer registration system holds data from the 8 former regional databases. A lot of work has been done to deduplicate these datasets but there are still duplicates in the data before 2012. The dedup_flag was developed to flag up records identified as duplicate records. Separate documentation is available for this field but briefly the flag takes account the following issues. For tumours diagnosed between 1995 and 2011, only those that can be traced in the 2013 ONS data will be counted. Cases sent late to ONS with a valid ONS ID are also included as cases.

   A small quality issue with the dedup flag occurs when no ONS ID is available and some cases are potentially identified as duplicates in error. This issue applies to cancer registrations diagnosed in areas covered by West Lancashire CCG and, to a lesser extent, Eastern Cheshire, South Cheshire, and Vale Royal. This will also affect local authority data and regional data that cover this area. Cancer cases may be missing from the cohort of patients diagnosed prior to 2008. This only relates to a relatively small number of cases and so will not impact greatly on national figures, but may mean that cancer incidence is significantly underestimated in these areas.

4) **Cases with suspected incorrect age at diagnosis**
   This SOP recommends including records of patients aged between 0 and 200.

5) **Cases with unknown sex**
Cancer cases with an unknown sex are excluded.

6) **Cases where the sex is incompatible to the tumour site**
For example, male patients with female reproductive cancers or female patients with testicular or prostate cancer. Sometimes this is a data quality issue but it is also possible that the registration was based on sex at birth instead of their current sex. Due to the very small numbers and sensitivity around these cases, they are excluded from most analysis. Therefore, you should not include females with site codes in the range of C60-C63 or males with a site code in the range C51-C58.

7) **Non-invasive tumours or non-melanoma skin cancers (C44)**
Perfomance indicators and incidence trends of cancer generally focus on invasive cancers (C codes excluding C44). Non-invasive tumours (D-codes) tend to have trends over time that are affected by data quality so any analysis about these groups should be done with great care. For the purpose of this SOP, only tumours with a site code beginning with C (excluding C44 non-melanoma skin cancer) should be counted.

Sample code 1: This code will count cancer cases diagnosed from 1995 onwards using recent snapshots (CAS1602 onwards).

Sample code 1

```
select  SITE_ICD10_O2_3CHAR, diagnosisyear, count(tumourid)
from av2014.av_tumour@CASREF01
where
ctry_code ='E' -- England residents using country code
--and SUBSTR(LSOA11_CODE, 1, 1) ='E'  --England residents using LSOA
and STATUSOFREGISTRATION ='F' -- Finalised cases
and dedup_flag=1 -- Excluding duplicates, note quality issue in text above
and age between 0 and 200 -- Sensible age
and sex in (1,2) -- Known sex
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'))) -- Sex doesn’t agree with tumour site
and (diagnosisyear>1994 and diagnosisyear<2016) -- Years of interest
and substr(site_ICD10_O2,1,1)= 'C' and substr(site_ICD10_O2,1,3)<> 'C44' -- all
malignant neoplasms (excl non-melanoma skin cancer)
group by SITE_ICD10_O2_3CHAR, diagnosisyear
order by SITE_ICD10_O2_3CHAR, diagnosisyear;
```

Sample code 2: If you want to make sure your figures align with those in CancerStats ie currently for diagnosis years between 2001 and 2014, you can use the criteria where CASCADE_INCI_FLAG equals 1. This applies to many of the filters in Sample Code 1 so is a shorter code to do the same thing. It is also designed to give identical numbers with those in CancerStats (which used to be called Cascade) and CancerData. The flag will include both C and D site codes, so be aware of this when creating site groups.

Sample code 2

```
select  SITE_ICD10_O2_3CHAR, diagnosisyear, count(tumourid)
from av2014.av_tumour@CASREF01
where
cascade_inci_flag=1 -- England, finalised cases, non-duplicates, sensible age, known
sex, correct sex specific cancers, diagnoses after and including 2001. Please note the
cascade_inci_flag will not restrict to C codes only, some D codes will be included
and substr(site_ICD10_O2,1,1)= 'C' and SITE_ICD10_O2_3CHAR <> 'C44' -- all
malignant neoplasms (excl non-melanoma skin cancer)
and (diagnosisyear>2000 and diagnosisyear<2015) -- Years of interest
group by SITE_ICD10_O2_3CHAR, diagnosisyear
order by SITE_ICD10_O2_3CHAR, diagnosisyear;
```

Sample code 3: Before the dedup_flag was included in the tumour table, we needed more complex code to identify and remove duplicates. This process used the ONS dataset to help with the deduplication. We **do not** recommend using this code unless absolutely necessary but it is included for completeness. It can be used on the CAS snapshots including 1502 or AV2013. Note some specific IDs in the code are only available internally.

Sample code 3

```
with
tidycanregcodes as(
select decode(av.centre, '0402', '0401', '0403', '0401', '0404', '0401', av.centre)  as
canreg , substr(av.onsid, 5, 11) as canregno , tumourid
from av2013.av_tumour av where av.centre is not null and substr(av.onsid, 5, 11) is not
null
)
, findpairs as(select canreg, canregno, count(*) as paircount from tidycanregcodes
group by canreg, canregno
)
, dupflags as(select canreg, canregno, case when paircount =1 then 0 else 1 end as
dupflag from findpairs
)
```

```
, table1 as (select * from (
(
select tumourid, diagnosisyear, substr(site_ICD10_O2,1,3) as site3
from av2013.av_tumour T INNER JOIN ONS2012.ONSINCIDENCE@CASREF01 N ON
DECODE(T.CENTRE, '0402', '0401', '0403', '0401','0404', '0401', T.CENTRE) =
N.CANREG AND SUBSTR (T.ONSID, 5, 11) = N.CANREGNO
left outer join dupflags d on d.canreg = N.canreg and d.canregno = N.canregno
where diagnosisyear>1994 and diagnosisyear<2012 --1995-2011 cases
and SUBSTR(T.LSOA11_CODE, 1, 1) ='E'
and substr(site_ICD10_O2,1,1)= 'C' and substr(site_ICD10_O2,1,3)<> 'C44' and
not(T.tumourid between [specific ID 1] and [specific ID 2] and substr(T.onsid, 1, 4) =
'[Specific ONS ID]' and dupflag = 1) and STATUSOFREGISTRATION='F')
union
(select tumourid, diagnosisyear, substr(site_ICD10_O2,1,3) as site3
from av2013.av_tumour T
where diagnosisyear>2011 and diagnosisyear<2014  --2012-2013 cases
and SUBSTR(T.LSOA11_CODE, 1, 1) ='E'
and substr(site_ICD10_O2,1,1)= 'C' and substr(site_ICD10_O2,1,3)<> 'C44' and
STATUSOFREGISTRATION ='F')))
,
table2 as (select diagnosisyear, site3, count(tumourid) from table1
group by diagnosisyear, site3 order by diagnosisyear, site3)

select * from table2;
```

For cases diagnosed between 1971-1994

Due to historical duplicates on CAS and the dedup_flag only being available for cases back to 1995, it is necessary to use the Office for National Statistics incidence data to count cases between 1971 and 1994. The ONS dataset is stored in the ONS1971_1994 schema in CASREF01. The numbers of cases produced by the code in this section should be the same as the ONS publication covering the same period.

Things to include/exclude:

1) The dataset includes 710 cases registered in Wales, identified by a specific canreg number, but these **should be retained** to make sure the numbers agree to previous ONS publications using this data.

2) There are several filters applied to the more recent data that do not need to be applied to the ONS data. For example, ONS data only includes records with a known sex and duplicates have already been removed. Cases where the sex does not agree with the tumour site have also been accounted for. For information ICD8/9 codes 185-187 are specific to men only and 179-184 are specific to women only (ICD8 174 was only split between men and women in the 1979 version of ICD9).

3) Cases are coded in ICD8 for 1971-1978 and in ICD9 for 1979-1994. For ICD8 invasive cancers include 140-207 (excluding 173) and ICD9 includes codes 140-208 (excluding 173). To compare the site distribution with more recent data mapping the ICD 8/9 codes to ICD10 / ICD10-02 will be necessary. Similarly for morphology (type5) as this is coded in MOTNAC in 1971-1989 and ICD00 in 1990-1994. If you are interested in both invasive and in-situ cancer registrations you will need to include the ICD range of 140 to 239 e.g. substr(site4,1,3) > '139' and substr(site4,1,3) < '240'.

4) Age at diagnosis needs to be derived from date of birth (DOB) and date of diagnosis (DIAGDATE). There are 2 dates of birth (DOB1 & DOB2) in the dataset. This is because it might be known that a patient died in a particular month ie April but not exactly when. So DOB1 will be 1 April and DOB2 will be the 30 April. Therefore, the mid-point between these 2 DOB should be used to derive age at diagnosis. Only patients with ages between 0 and 200 at diagnosis should be included.

Sample code 4: This code will allow you to count cancer cases diagnosed from 1971 to 1994 using the ONS incidence data.

Sample code 4

```
select extract(year from DIAGDATE) as diagyear, substr(site4,1,3) as site, count(*)
from ONS1971_1994.ONSINCIDENCE
where Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12)>0 and
Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12)<=200
and substr(site4,1,3) != '173'
and (((substr(site4,1,3) < '208' and (substr(site4,1,3)>'139'
and to_char(DIAGDATE, 'yyyy') < '1979'))
   or (substr(site4,1,3) < '209' and (substr(site4,1,3)>'139'
and to_char(DIAGDATE, 'yyyy') > '1978'))))
group by extract(year from DIAGDATE), substr(site4,1,3), sex
```

For information:

1) Deriving age at diagnosis in the ONS dataset:

```
Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) as diagage
```

2) Deriving age group:

```
case when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 0 and 39 then '<39'
     when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 40 and 44 then '4044'
     when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 45 and 49 then '4549'
     when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 50 and 54 then '5054'
     when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 55 and 59 then '5559'
     when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 60 and 64 then '6064'
     when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 65 and 69 then '6569'
     when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 70 and 74 then '7074'
     when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 75 and 79 then '7579'
     when Trunc(trunc(months_between(diagdate, (dob1+(dob2-dob1)/2)))/12) between 80 and 84 then '8084'
     else '85+' end ageg,
```

Note: You may need to extend these age groups to 90+ depending on your project.

3) Deriving date of death:

```
(dod1+(dod2-dod1)/2) as dod
```