

---

# Standard Operating Procedure: Guidelines on Population Based Cancer Survival Analysis

---

This SOP has been developed by a subgroup of the UKIACR Analysis Group, with important contributions from Jason Poole<sup>1</sup> (co-lead), Finian Bannon<sup>2</sup>(co-lead), Sean McPhail (co-lead)<sup>3</sup>, Matthew Barclay<sup>4</sup>, Michel Coleman<sup>5</sup>, Marta Emmett<sup>1</sup>, Tim Evans<sup>1</sup>, David Greenberg<sup>6</sup>, Ula Nur<sup>5</sup>, Nick Ormiston-Smith<sup>7</sup>, Andy Pring<sup>1</sup>, Bernard Ratchet<sup>5</sup>, Rebecca Thomas<sup>8</sup>, Sarah Whitehead<sup>9</sup>. The Analysis Group will maintain and update this document.

<sup>1</sup>Knowledge & Intelligence, Public Health England (PHE); <sup>2</sup>Centre of Public Health, Queen's University Belfast; <sup>3</sup>National Cancer Intelligence Network, PHE; <sup>4</sup>Cambridge Centre for Health Services Research, University of Cambridge School of Clinical Medicine; <sup>5</sup>Cancer Research UK Cancer Survival Group, London School of Hygiene & Tropical Medicine; <sup>6</sup>National Cancer Registration Service, PHE; <sup>7</sup>Cancer Research UK; <sup>8</sup>Welsh Cancer Intelligence and Surveillance Unit; <sup>9</sup>National Statistician's Office (previously Health and Life Events Division, Office for National Statistics).

Further information on the content of this SOP is available from your regional or national cancer/public health intelligence lead. Their contact details are available from the UKIACR website (<http://www.ukiacr.org/>).

## Contents

Aim .....	3
Introduction .....	3
Methods of estimating net survival .....	5
Introduction .....	5
Observable net survival.....	5
Relative survival .....	5
Pohar-perme net survival estimator.....	6
Modelling approach to net survival estimation.....	7
Types of survival estimates.....	8
‘Cohort’ approach .....	9
‘Complete’ approach.....	9
Period approach.....	9
Hybrid approach.....	10
Data preparation for survival analysis .....	11
Data quality.....	11
Inclusion and exclusion criteria.....	11
Life-tables.....	12
Interval cut points .....	13
Age-standardisation.....	14
Software and worked examples .....	16
Secondary measures of survival .....	17
Avoidable deaths .....	17
Estimating “cure” from cancer.....	17
Training .....	18
References .....	19
Appendix 1: PHE cancer survival defaults – for survival analyses performed in 2015/16.....	22

## Aim

There are many different methods and techniques when approaching the analyses of population-based cancer survival data, and these can sometimes produce significantly different answers. The aim of this brief Standard Operating Procedure (SOP) is to make a recommendation on the best approach to analysing cancer survival data. It includes some background to cancer survival analyses, commonly used and recommended methods, and the latest training courses available for further study, along with relevant references.

This public access document is intended to be used by those cancer and public health analysts involved in the analysis of population based cancer survival primarily in the United Kingdom (UK) and Ireland, but also internationally. It is hoped that it will also be of wider interest to those tasked with compiling, understanding and interpreting cancer survival results and the different methods used to calculate these. For comparison, the PHE cancer survival analysis defaults are included (Appendix A) in this SOP.

If anything here is unclear or you feel that important information has not been included then we would like to hear from you. Please email: [f.j.bannon@qub.ac.uk](mailto:f.j.bannon@qub.ac.uk).

## Introduction

National health systems strive to prevent people dying from cancer. This is primarily carried out in two ways. Firstly, by reducing the risks of people getting cancer in the first place, mainly by avoiding life-style choices known to be associated with higher risk of cancer, e.g. smoking. And secondly, by providing the best evidence-based ways to detect cancer and cure patients, or at least extend their lives after diagnosis. Assessing how well the health system is achieving this is typically assessed by studying population-based incidence, mortality, and survival statistics; each statistic provides a different perspective on the cancer burden. Progress against cancer is reflected in reduced mortality – either by reducing incidence, increasing survival, or both. However, when comparing effectiveness of health systems in preventing cancer deaths between countries or time, it is desirable to have a measure that is consistently estimable and interpretable.

Incidence is generally considered a reasonable measure of the effects of cancer risk factors in the general population, while survival is generally considered a good measure of curing or prolonging life for cancer patients; the two measures, with their different formal objects (the general population and the cancer patient population), are generally considered independent of one another. On the other hand, mortality rates are difficult to interpret as they measure the cumulative and combined aspects of incidence and survival in the recent past. Furthermore, cancer mortality rate comparison rests upon the assumption that death-registration practice is consistent between countries—an assumption considered untenable in large international studies. However, at times mortality rates are indispensable for measuring cancer burden when either incidence or survival statistics are inflated by over-diagnosis following over-detection (see below).

The present SOP directs its attention on the survival of cancer patients following diagnosis, and hence the ability of the health system to cure cancer patients or prolong their life.

Cancer survival estimates are important for several reasons:

- To predict the survival for recently diagnosed patients.
- To assess the overall effectiveness of health systems; this includes public health programmes that raise the awareness of cancer symptoms and promote earlier diagnosis, screening, and efficient diagnosing and treating of cancer.
- To compare survival between sub-populations (ethnicity, socio-economic status) or time (trends).

Cancer survival estimation should be population-based, and reliant on complete and good quality data. The UK is widely acknowledged as having one of the most comprehensive cancer registration systems in the world. Regional cancer registries across the UK and Ireland (<http://www.ukiacr.org/>) have been collecting population-based cancer data for several decades. Survival estimates that are derived from a sample of the population are susceptible to biases. For instance, it is generally easier to collect information on good-prognosis patients. It is never certain that a sample of a population is truly representative of the entire population. For similar reasons, a population-based survival estimate should never be equated with survival estimates from randomised clinical trials in which highly-select patients, subject to inclusion and exclusion criteria, are treated within experimentally-controlled treatment regimes.

Survival is not a straightforward indicator. The cancer patient's survival time, defined as the time between diagnosis and death, is sensitive to any factor that may affect either of these events. Considering the diagnosis event, screening and sensitive diagnostic techniques may lead to a cancer being diagnosed much earlier and asymptotically, and therefore increase survival time even though the natural course of the disease remains unchanged – so called *lead time bias*. Another bias, *length bias*, occurs in screening programmes, where slow-growing, less aggressive tumours are more likely to be detected (success in detecting aggressive tumours is sensitive to the length of time between screenings); these cancers, which may never be life-threatening, will inflate cancer survival estimates. Considering the death event, if death information is not being matched correctly, this will extend patient survival time, and inflate survival estimates. As mentioned above, if these biases are known to be large, survival estimates can be biased; in this case, mortality rates are considered a more sound way of appraising cancer burden.

Population-based observed or crude survival is a valuable statistic when advising patients about their prognosis; all causes of mortality are implied and this is appropriate as cancer patients can die from any cause. However, in order to assess health systems, it is desirable to remove the effect of competing causes of death which can differ markedly from country to country. Competing causes of death are approximately equal to population mortality rates (found in a national life table), and their removal in the estimation of survival leads to a quantity known as *net survival*. Net survival is a quantity better suited for international comparison, or sub-group analysis within a population.

Further information on useful recent publications of cancer survival data are available in the National Cancer Intelligence Network (NCIN) report 'What cancer statistics are available, and where can I find them?' (<http://www.ncin.org.uk/publications/reports/>). This includes references to results within and for the UK as a whole, and for international comparisons. Other examples, cited at the end of this document (1–8), include:

- Estimation of differences in survival by type of cancer, between the sexes, or between regions of a country

- Time trends in survival
- The number of avoidable premature deaths by ethnicity, region or socio-economic status, in comparison with another population or country where survival is higher
- For certain cancers, the proportion of patients who may be considered “cured”

## Methods of estimating net survival

### Introduction

Implicit in a survival estimate is a mortality rate. The living cohort of patients is continually being depleted by a mortality rate, according to the following formula (when the rate is considered as a continuous function of time):

$$S(t) = e^{-\int \lambda(t)dt}$$

where  $t$ =time,  $S(t)$  is proportion of patients alive, or survival at  $t$ ,  $\int \lambda(t)dt$  is the cumulative mortality rate at time  $t$ . Cancer patients’ mortality rate,  $\lambda(t)$ , is the sum of their cancer-related death or excess mortality,  $\lambda_E(t)$ , and their competing causes of death [approximated by],  $\lambda_P(t)$ <sup>1</sup>, the background population mortality rate. Net survival (9) can be defined as the survival of cancer patients in the hypothetical situation in which cancer is the only possible cause of death, i.e. the effects of competing causes of disease,  $\lambda_P(t)$ , are removed.

### Observable net survival

If the underlying cause of death is accurately known, that is properly registered on the death certificate, for all cancer patients, *observed net survival* can be estimated by the cause-specific approach using the Kaplan-Meier method, in which deaths attributed to (“caused by”) the cancer are counted as *events*, while deaths attributed to other causes are *censored*. However, this approach can lead to a biased estimate of net survival because the censoring mechanism is driven partly by  $\lambda_P(t)$ , which is often associated with  $\lambda_E(t)$ , the quantity driving the net survival estimate. In practice, older patients who have high  $\lambda_E(t)$  often have high  $\lambda_P(t)$ , and therefore more likely to be censored and therefore not contribute as they should to the net survival curve as follow-up time progresses. In this setting, the censoring process becomes “informative”. Moreover, it should be borne in mind, the cause of death as registered in death certificates may be inaccurate.

**Recommendation: avoid estimating observable net survival**

### Relative survival

Relative survival derives its name from its approach to estimating net survival as a ratio of observed (or crude) survival to ‘competing causes of death’ survival in cancer patients. If the observed mortality rate is the sum of excess mortality and ‘competing causes of death’ mortality rate,  $\lambda_O(t)=$

<sup>1</sup> In this SOP, the subscript ‘P’ derived from population mortality rate, will be considered equivalent to competing causes of death (see section on ‘Lifetables’ for a more comprehensive explanation). The ‘P’ is retained as a reminder that this information is derived from a life table of population mortality rates. Therefore  $\lambda_P(t)$  will mean ‘competing causes of death’ mortality rates,  $S_P(t)$  will mean survival from ‘competing causes of death’.

$\lambda_E(t) + \lambda_P(t)$ , then the observed survival is the product<sup>2</sup> of net survival and ‘competing causes of death’ survival, so that:

$$S_E(t) = \frac{S_O(t)}{S_P(t)}$$

While this relationship is true for an individual cancer patient, it is not true on a cohort level unless every patient shared the same characteristics: sex, age, year of diagnosis. The most common relative survival estimator, *Ederer II*, proceeds by taking the patients alive at the start of an interval and estimating a) their observed survival over that interval, b) the mean of their individual probabilities of surviving that interval based on the ‘competing causes of death’ mortality rate. The two estimated quantities then form a ratio called [conditional] relative survival; the product of these ratios over all intervals gives the final relative survival estimate. There are two potential biases with this approach. Firstly, the population net survival should be the mean of a sum of individual patient ratios, not the ratio of two population ‘mean’ values (10).

$$\frac{1}{N} \sum_{i=1}^N S_{E_i}(t) = \frac{1}{N} \sum_{i=1}^N \frac{S_{O_i}(t)}{S_{P_i}(t)} \neq \frac{\frac{1}{N} \sum_{i=1}^N S_{O_i}(t)}{\frac{1}{N} \sum_{i=1}^N S_{P_i}(t)}$$

Secondly, like the observed net survival estimator (see above), informative censoring is occurring in the Ederer II estimator also because the censoring mechanism is driven partly by  $\lambda_P(t)$ , which is often associated with  $\lambda_E(t)$ , the quantity driving the net survival estimate. When patients in survival estimation are homogeneous in their demographics, i.e. have similar age, same sex, year of diagnosis, the relative survival estimator becomes an adequate estimator of net survival. Typically, there is very little difference in age-standardised (see below) estimates of relative survival and net survival, demonstrating that age is the chief source of informative censoring. By age-standardising, conditional independence can be assumed<sup>3</sup> meaning that there are no factors associated with both cancer mortality and ‘competing causes of death’ mortality other than those factors that have been controlled for in the estimation (e.g., via stratification, regression modelling or appropriate weighting). In the present SOP, we will continue to consider age-standardised relative survival as a useful estimator of net survival in circumstances where the version of software or computing capacity does not support other options.

**Recommendation: use age-standardised relative survival when Pohar-Perme estimator equivalent is not available**

### Pohar-perme net survival estimator

A non-parametric approach, the Pohar-Perme estimator (PPE), addresses the biases mentioned above in the relative survival estimator in order to achieve a non-biased estimator of net survival (11, 12). At each observed event time [death or censoring] marking the end of an interval since the

$$^2 S_O(t) = e^{-\int \lambda_O(t) dt} = e^{-\int \lambda_E(t) + \lambda_P(t) dt} = e^{-\int \lambda_E(t) dt} \times e^{-\int \lambda_P(t) dt} = S_E(t) \times S_P(t)$$

<sup>3</sup> In age-standardised relative survival, there will still be some residual informative censoring occurring within the defined age-groups, but in practice any bias is so small that it can be ignored.

previous event, three quantities, namely, cumulative observed deaths and [expected] deaths from ‘competing causes of death’, and the at risk population are inflated by inverse-weighting the individuals [in each quantity] with their individual probability of their surviving from deaths from ‘competing causes of death’ since diagnosis,  $S_p(t)$ . Intuitively, the effect of the weights is to inflate the observed person-time and number of deaths in order to account for person-time and deaths not observed as a result of mortality due to competing causes (10). The three inflated quantities are combined to estimate cumulative excess mortality, and hence net survival. The individual inverse-weighting addresses simultaneously the two biases mentioned in the relative survival estimate. The non-parametric PPE is data- and life table-driven, requiring no data modelling assumptions (see modelling approach below). This estimator is suitable for official statistics.

It has been observed with the PPE method that in estimating long-term survival, the estimate can become unstable in the older patient cohorts (13). However, adherents of PPE claim that this simply reflects the inherent difficulty in estimating long-term (10-20 year) net survival in this age group. The number of patients in the risk group becomes small due to high competing causes of death at that age. In addition, the  $S_p(t)$  weightings of these patients can vary widely because the ‘competing causes of death’ mortality rates vary much more with age in this age group. Based on these two realities, the particular deaths or the survival of some very old patients in a small risk group can have a large influence. The solution is to obviate such a situation by assessing whether the expected ‘competing causes of death’ survival, i.e. survival constructed from life table mortality rates, of a cohort of cancer patients indicates that there are enough patients, independent of the excess mortality rates, to estimate net survival. While long-term (for example, 10 year estimates of patients >85, e.g. prostate cancer) age-standardised Ederer II survival estimates appear to be more stable, the level of bias present from the two biases aforementioned is unknown.

**Recommendation: use Pohar-Perme estimator as the preferred method of net survival estimation**

### Modelling approach to net survival estimation

In the modelling approach of net survival devised by Lambert and Royston (14), a fully-parametric model describes the relationship between net survival and follow-up time. The approach uses restricted cubic splines to capture the non-linear relationship between the continuously changing mortality rate and follow-up time<sup>4</sup>; this relationship can be allowed to vary for different types of patients (time-dependent effects). Each patient’s time-to-event in the analysis is offset by its ‘competing causes of death’ mortality rate from the life table (at the time of the event) in order to give an unbiased estimate of the excess cancer rate.

An adequately fitted model, can then predict the net survival of each patient at a fixed follow-up time, the mean of these predictions yields the population net survival at that fixed time. It is obviously important, that the fitted model accurately captures all the systematic (i.e. non-random)

<sup>4</sup>In fact the relationship modelled is presented below where the function  $s \{ \ln(t) \mid \gamma, k_0 \}$  represents restricted cubic spline function of  $\ln(t)$  and  $x$  represent covariates, e.g. sex, age etc:

$$\log \left( \int \lambda(t) dt \right) = s \{ \ln(t) \mid \gamma, k_0 \} + x_i \beta$$

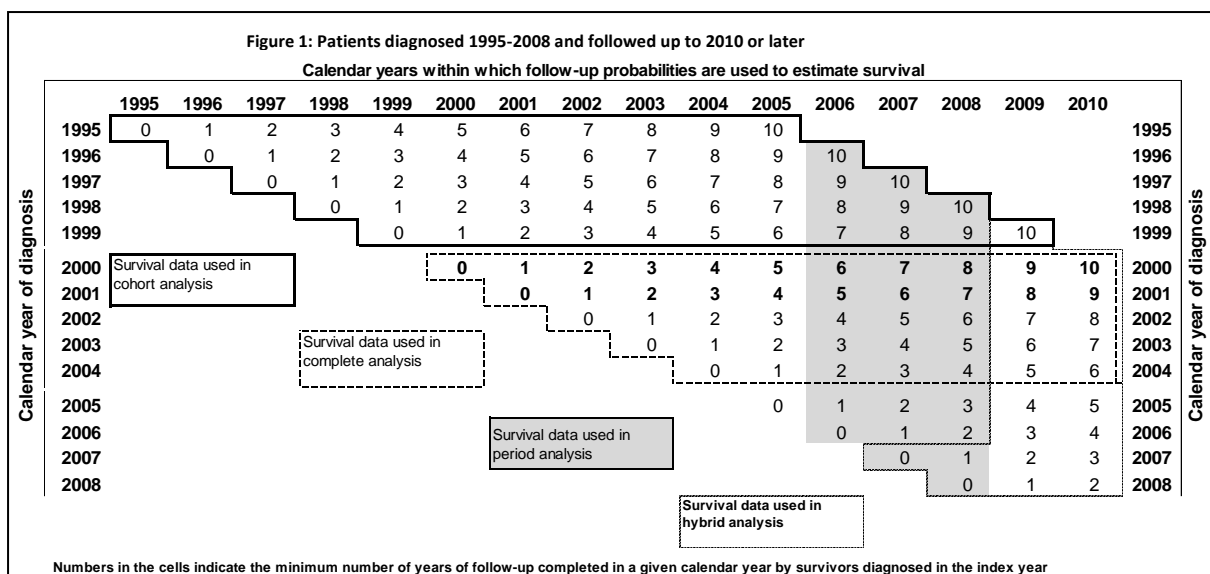
variation that arises from the demographic effects (year of diagnosis, sex, year, and follow-up time), in order to give an unbiased estimate of population net survival. Restricted cubic splines can also be used to describe any non-linearity in the effects and their interactions.

A high degree of experience and expertise is required in such modelling. For example, decisions have to be made on (a) what covariates to include, (b) how to model age (grouped or continuous), (c) if continuous, what functional form to use, (d) similar decision for other continuous variables (e.g. year of diagnosis), (e) whether to incorporate time-dependent effects and how to model these if so, (f) are interactions necessary, e.g. is it sensible to assume that the effect of calendar time of diagnosis is the same at each age of diagnosis. The approach is time-consuming, and each cancer site requires individual attention. However, it is an excellent research tool in the study of net survival.

**Recommendation: use modelling approach only with sufficient expertise**

### Types of survival estimates<sup>5</sup>

Aside from the method of estimating survival (see above), different types of survival estimates are distinguished by their timely use or recency of cancer registry information. The following example (Figure 1) shows the structure of a particular data set in which patients diagnosed during the period 1995-2008 have been followed up for their vital status to the end of 2010. Numbers in the cells indicate the *minimum number of complete years of follow-up* data that are available for patients who were diagnosed in a given year between 1995 and 2008 (rows) and who survived to the end of a given year (column) up to the end of 2010. In Figure 1, four sets of survival information are identified corresponding to the four survival types explained below. Further information on the comparison of these approaches is published (15). *(Please print out this figure to view properly).*



<sup>5</sup> The nomenclature of the types of survival presented here is not universally agreed and hence we have placed terms in inverted commas to alert the reader. However, we have adopted the common meaning in cancer registries, all the while defining exactly what is meant.



### **'Cohort' approach**

The cohort approach identifies a cohort of patients, and follows them each up for the same length of time. All patients in the diagnosis period 1995-1999 have 10 years of follow-up by the end of 2009 (stepped horizontal wedge bounded by a solid line in Figure 1). Each patient, irrespective of their actual year of diagnosis, will contribute survival information at each point in follow-up time that, taken cumulatively, makes up the survival estimate at 10 years. This approach, grounded on the full follow-up of a clearly-defined group of patients, is attractive when comparing patient survival experience between different populations of patients.

However, the necessity to collect 10-years follow-up data on every patient in the diagnosis period means that cohort survival estimates summarises more historic information on patients than what the registry currently holds. The registry will have more up-to-date information on patient survival for follow-up time of less than 10 years of patients diagnosed more recently than the diagnosis period; the next approaches seek to incorporate this information.

**Recommendation: use 'cohort' approach when publishing standard and routine data tables, and when making international comparisons**

### **'Complete' approach**

The complete approach will take more recently diagnosed patients than the cohort approach to estimate survival (partially stepped horizontal wedge in Figure 1); hence using more completely the available information. For patients diagnosed 2000-2004, the 'cohort' approach can only estimate five-year survival, but with the 'complete' approach ten-year survival can be estimated. However, with the 'complete' approach, more information is available to estimate survival in the early follow-up than late, leading to variation in statistical power along the survival curve. In addition, the estimation of survival in later follow-up uses proportionately more historic information than earlier follow-up, which can make interpretation of estimates difficult when survival over calendar time is rapidly changing for some cancers; this imbalance is not a feature of the cohort estimate.

**Recommendation: Use cohort or period (see below) in preference to complete estimates as the former are more easily interpretable**

### **Period approach**

The period approach (16) estimates the excess mortality rates from deaths and person-time<sup>6</sup> of follow-up in a defined calendar period (see gray-shaded area in Figure 1), whereas the cohort approach estimates excess mortality from deaths and person-time of follow-up in a defined cohort of patients. The person-time can be divided into intervals of follow-up time. Even if the 'period' is only one calendar year, person-time over the full range of follow-up (0-10 years) will be available. For instance, patients diagnosed [at least] 9 years before the period will be contributing person-time from [at least] 9 years follow-up time onwards for a year. By classifying the deaths and the person-

---

<sup>6</sup> Person-time is a measurement combining the number of persons and their time contribution in a study. It is the sum of individual units of time that the persons in the study population have contributed to the denominator of a mortality rate, for instance.

time by follow-up time it is possible to estimate a mortality rate over follow-up time, and hence a period survival estimate.

The period estimate combines survival information up to 10 years after diagnosis that were observed within the period 2006-2008, but for patients who were diagnosed during the period 1996 to 2008 (Figure 1, shaded area). This estimate can be interpreted as a prediction of ten-year cohort survival estimate for patients diagnosed during 2006-2008, on the assumption that the excess mortality rates remain stable from 2006 to 2018, when all patients will have at least 10 years of follow-up.

Period survival uses more up-to-date data to estimate the long-term survival outcomes; this makes this estimate useful in advising patients of their prognosis. However, it still may not capture recent changes to survival that occur later after diagnosis. It is a measure analogous to the life-time risk of getting cancer (17), which uses the most recent age-specific incidence rates to predict the lifetime risk of getting cancer; however, when the cohort is followed-up, after many years, the true lifetime risk will have been determined also by changes in the incidence rates since the prediction was made. The period survival estimate is conceptually more difficult to explain than 'cohort' survival.

**Recommendation: use the 'period' approach for expert audiences or where most up-to-date prognostic information is needed**

### Hybrid approach

The hybrid approach (18) is used where the follow-up information is more recent (2010) than the incidence data (2008). A period estimate for 2008-2010 would have three years of available data to estimate excess mortality at two years or more of follow-up, but only one or two years of available data for earlier follow-up periods; the imbalance could supply potential for bias. By contrast, a hybrid estimate provides three years of available data for all follow-up since diagnosis, by combining the cohort approach up to two years after diagnosis (patients diagnosed 2007-2008) and the period approach for two or more years of follow-up (patients diagnosed 1999-2008; dotted outline).

Hybrid survival uses the most up-to-date data to estimate the long-term survival outcomes, the most recent changes to survival outcomes that occur late after diagnosis is not captured, and this approach is the most conceptually difficult to explain.

**Recommendation: use only in the specialised circumstances requiring it, i.e. where death follow-up information is more recent than incidence**

## Data preparation for survival analysis

### Data quality

A standard quality control procedure should ensure that tumour records meet basic criteria of data quality (e.g. not a duplicate, consistent tumour site/morphology/behaviour/patient sex). The International Association of Cancer Registries has 'Check and Conversion Programs for Cancer Registries' software that runs internal validation and consistency checks on tumour records (see: <http://www.iacr.com.fr> under 'Support for registries').

Net survival estimates will be inflated if there is not complete and accurate follow-up of cancer patients' living status. To ensure the best possible follow-up for all patients, it is necessary to check that all notifications of death have been received and recorded; this is more relevant in registries that do not perform a complete linkage of their incidence register to a national death register on a regular basis (e.g. annually).

### Inclusion and exclusion criteria

Even when a tumour record is correct, there are still criteria to determine if the record should be included in the survival analyses. Ineligible records should be tabulated in order to document quality, and then excluded from survival analysis (19). Here is a list of the commonly applied inclusion and exclusion criteria.

#### *Inclusion criteria*

1. Age 15-99 years, i.e. includes adults (not children).
2. Tumour topography code (as defined by either International Classification of Diseases (ICD) or International Classification of Diseases for Oncology (ICDO)) belongs to the appropriate definition of cancer site. Variations in coding definitions can affect survival estimates depending on the patient-prognoses arising from tumours included in the definitions. Only patients with primary tumours are included, i.e. tumours that have originated in the organ of the cancer site defined, and are not spread from another organ in the body (secondary). The patient's survival time begins on the diagnosis date of their first primary tumour [of the cancer of interest] that occurred in the period [of diagnosis] of interest, e.g. 2000-2005. Patients are not excluded if they had
  - further primary tumours of the cancer of interest diagnosed later in the period of interest.
  - any primary tumour of another cancer site diagnosed in the period of interest.
  - any type of primary tumour diagnosed before or after the period of diagnosis.

The rationale for these criteria is twofold. Firstly, with survival improving all the time, an increasing proportion of cancer patients will have had another cancer and not including them will bias-upwards the site-specific survival estimate. Secondly, in international studies, older registries are more likely than younger registries to register previous cancers, thus by not excluding patients who had a previous tumour a potential bias in country comparisons (20, 21) is avoided.

3. More complete guidance on dealing with multiple primaries, including definitions of synchronous tumours, is available at the Surveillance, Epidemiology, and End Results Program (SEER) website: <http://seer.cancer.gov/tools/mphrules/>. IARC has provided guidance also (22).
4. Only patients with tumours of an invasive, primary, and malignant behavioural code (=3) in ICD O (<http://www.who.int/classifications/icd/adaptations/oncology/en/>) are included. It is worth noting that sometimes in revisions of ICD O, the behavioural codes have been changed. In ovarian cancer some behavioural codes 3 were reclassified as 1 (uncertain behaviour) when moving from the 2<sup>nd</sup> to 3<sup>rd</sup> edition of ICD O; this has relevance when comparing historic estimates to recent estimates.
5. Patients whose survival time is zero (date of diagnosis is the same as the date of death), but which are not Death Certificate Only (DCO) registrations (see 6 below), should be included in survival analyses. Stata's *stset* command does not accept zero-day survivor patients, it is necessary to add one day to the recorded date of death to include these patients.

#### *Exclusion criterion*

6. The accuracy in estimating survival is highly dependent on correctly identifying and recording the first diagnostic episode. This is particularly important in cancers where the death certificate is the first notification of a cancer that a cancer registry may receive; thorough investigation of such cancers with hospitals and general practitioner (GP) practices may identify previous diagnostic episodes. If such earlier diagnostic episodes are not traced, cancer cases are called death certificate only (DCO) registrations and excluded from survival analysis. The PHE National Cancer Registration Service is currently implementing processes to follow up all death certificate notifications of cancer within three months.
7. Exclude patient if the following information is missing or imputed: sex, date of diagnosis, date of birth or age.

#### **Recommended data processing for site-specific analysis:**

##### **Include**

1. Patients aged between 15-99 years (inclusive)
2. The earliest primary tumour in the period of diagnosis with eligible topography code
3. Invasive, primary, and malignant behavioural code (=3) tumours

##### **Exclude**

4. Death certificate only (DCO) registrations
5. Missing or imputed sex, date of diagnosis, date of birth or age information

## **Life-tables**

When estimating net survival for a particular cancer, say kidney cancer, we need to remove the contribution of the 'competing causes of death' mortality rate (which includes death from other types of cancer). The general population mortality rates are a good approximation of the competing causes of death in kidney cancer patients, since the kidney cancer deaths will make up a negligible proportion of the general population rate. A life table tabulates the general population mortality

rate by various demographics usually age, sex, and calendar period, but also sometimes by sub-region, ethnicity, socio-economic deprivation.

It is important that the life table reflects the competing causes of death of the cancer patients otherwise there will be an under- or over- estimation of net survival. Therefore, when estimating survival by sub-groups in the population, such as socio-economic status, life tables that are specific for each socio-economic population sub-group should be used. Another example, in a comparison of cancer survival between the South West and the North East of England, it would be preferable to use the corresponding regional life tables, rather than a single national life table, since overall mortality in the North East is higher than in the South West. A similar argument applies to analysis of trends in survival – life tables for each calendar year are preferable to a single life table applied over a long period of time. Extra demographic variable classification in the life-table (for instance, life table by age, sex, calendar year, and socio-economic deprivation), correspondingly matched to the patients, will assign a more individualised ‘competing causes of death’ mortality rate to the patient. This will give rise to an even less biased net survival estimate, even if there is no intention to report survival by the extra demographic variables.

There are various sources of published life tables, including the Office for National Statistics (23) and the World Health Organisation (24). Many life tables for the UK are also available on the Cancer Research UK Cancer Survival Group web-site (25), broken down by country and regional geographies, deprivation (usually by quintile, using Carstairs (26) or Townsend indices, or the Indices of Multiple Deprivation) and ethnicity (White, Black, South Asian), for all calendar years since 1971. The Cancer Survival Group (London School of Hygiene and Tropical Medicine) also has a suite of life tables catering for particular populations and calendar times (<http://www.lshtm.ac.uk/eph/ncde/cancersurvival/tools/index.html>).

The calendar years during which follow-up occurs, i.e. up to the last matching of death information or the general censoring date, require a corresponding life-table. In the case of a national life-table not having the most recent year, the previous year is used again.

**Recommendation: Use the life-table that most closely matches the ‘competing causes of death’ in the cancer patients; where possible match patients to a lifetable including region and socio-economic deprivation.**

## Interval cut points

The net survival estimate can, broadly speaking<sup>7</sup>, be put-together by 1) estimating the [excess] mortality at each event time and summing to give the cumulative mortality (and then survival, see PPE), or 2) by estimating cumulative mortality (or conditional [relative] survival) in piece-wise intervals and combining to give a survival estimate. In the case of the latter, interval cut points define the interval lengths, in which the excess mortality rate is considered constant over the interval; thus the relationship of excess mortality rate to follow-up time is described by a step-function. Therefore, if the excess mortality rate is changing rapidly, the interval widths need to be narrower if the step-function is to remain an adequate description of the underlying continuous

---

<sup>7</sup> A third approach is modelling the underlying excess mortality rate with a continuous parametric model.

function,  $\lambda_E(t)$ . For lethal cancers such as those of lung or pancreas, the excess mortality will change quite rapidly early in the follow-up.

The cut-points (breaks) for the intervals, within each of which survival will be separately estimated, should therefore be chosen to generate short (e.g. monthly) intervals at the beginning of follow-up, and longer intervals (e.g. quarterly, six-monthly or annual) thereafter. For example, in 2008 the Cancer Research UK Cancer Survival Group and colleagues (4) reported their methods for calculating 10-year survival for 20 of the most common cancers. Their interval structure strategy was chosen as monthly up to 6 months after diagnosis, then at 3-monthly intervals up to 2 years after diagnosis, 6-monthly during 2 to 5 years, then yearly up to 10 years.

Strel (27, see *Software and Worked Examples* below) estimates relative survival using intervals. The estimates of excess mortality may be unreliable if there are very few patients or very few deaths (for example less than 10) within a given time interval, because the sparsity of data prevents convergence of the maximum likelihood algorithm. Strel has the capability to progressively group the time intervals, provided that estimates are maintained at 1, 5 and 10 years after diagnosis, so that convergence can be achieved.

Interval cut points are also required if available survival time is not continuous (in days) but in discrete intervals (months or years). Dickman and Coviello (10) have implemented a PPE estimator when only such discrete survival time information is available.

**Recommendation for estimating survival using discrete interval approach: monthly up to 6 months after diagnosis, then at 3-monthly intervals up to 2 years, 6-monthly during 2 to 5 years, then yearly up to 10 years**

## Age-standardisation

Net survival of cancer patients is related to age at diagnosis in many cancer sites. Therefore, if two cohorts of cancer patients have a different age distribution or structure, their net survival will likely be different. Age-standardised net survival estimates are the estimates that would occur if that population of cancer patients had a standard population age structure. They are estimated by firstly estimating age-group specific survival estimates (e.g. 15-44, 45-54, 55-64, 65-74, 75+) which are then weighted by standard weights, which reflect the proportion of cancer patients in that age-group in the standard population, and summed to give an age-standardised survival estimate. Age-standardised survival estimates allow us to compare territories, populations, or time periods such that any observed differences are not attributed to different age structure, but something else, e.g. health system, ethnicity.

The International Cancer Survival Standard (ICSS) weights (28) are recommended for international comparisons of cancer survival (1, 2). The weights correspond to five age-groups that classify patients diagnosed from 15-99 years of age. Four sets of standard age weights were derived from discriminant analysis which sought to find the smallest number of sets that would give age-standardised estimates of survival across a range of different cancers that broadly reflect the unstandardised estimates. The same age weights can be used for men and women, and for different ethnicities.

The ICSS weights are an optimal set of weights derived from cancer patients in Europe. If looking at trends in a particular country or region over time, it might be more attractive to design a set of weights that reflect better the population structure in that region. In short, the adoption of weights in an analysis depends on the aim of the analysis being undertaken.

The weights described above are sometimes referred to as external weights meaning that in some way they will remain fixed and imposed across regions or time periods, and thus allowing comparisons to be made. Survival estimates can be internally standardised meaning that weights reflect exactly the age-group proportions of cancer patients in the analysis. The reason for doing this is to produce an estimate of the unstandardised relative survival of all the patients, but removing the informative censoring arising from age mentioned above, and so be an adequate estimate of unstandardised net survival.

The question arises when estimating an age group specific survival, usually the youngest age-group, what information threshold should be adopted below which survival is considered inestimable. An immediate solution is to collapse neighbouring age-groups, but with small regions, even with doing this, the question may still remain. A number of diagnostics have been suggested, e.g. the number of deaths, or number of patients surviving. However, both these measures depend on the excess mortality rate which can vary between cancer site, e.g. lung cancer versus prostate cancer. A diagnostic that is independent of excess mortality rate is the probability of their surviving from 'competing causes of death' since diagnosis, i.e.  $S_p(t)$ , which is derived using life table information. Multiplying this probability by the number in the initial cohort will indicate the expected number of patients alive at time  $t$  surviving from 'competing causes of death'. If this number is too small then it is likely impossible that excess mortality (and net survival) at time  $t$  is estimable, i.e. there are no patients in existence to die, irrespective of if they die from the cancer or 'competing causes of death'. It is good practice also to graph the survival curve, and assess its stability. By doing this for a number of cancer sites, varying in prognosis, it should be possible to choose a general cut-off point defined by the expected number of patients alive at time  $t$  surviving 'competing causes of death' (<10, for instance). When an age-standardised survival estimate is un-estimable, the unstandardised estimate can be reported as the next best thing.

Note that for estimating age-standardised incidence or mortality rates, the age structure of the general population is standardised, whereas with cancer survival it is the age structure of the cancer patient population. The weights used for age standardisation of cancer survival estimates are thus completely different from those required for standardising incidence or mortality rates.

**Recommendation: use the ICSS weights to age-standardise net survival estimates to maximise their comparative potential**

**Recommendation: calculate the expected number of patients to survive the competing causes of death, and observe the survival curve, to check if survival is estimable**

## Software and worked examples

We present several commands that are currently available in Stata for the estimation of net survival; in addition, the website for SEER\*Stat net survival software is given. The different commands embody different approaches to estimating net survival, each with its own strengths and weaknesses. To assist analysts who might be new to 'net survival' estimation, Web Appendix 1 (Note: under construction) describes worked examples of each command using the patient survival data from four different cancer sites (female breast, lung, malignant melanoma, and pancreas) registered in Northern Ireland. The do-files used to analyse the data can be downloaded from the appendix; the data files are available from the NICR. In addition, the net survival estimates obtained from each command are compared in order to gain insights into any differences.

- **strel** (28) is a Stata command developed by Bernard Rachet and Milena Falcaro of the Cancer Research UK Cancer Survival Group, London School of Hygiene and Tropical Medicine (LSHTM), for the estimation of the excess mortality hazard and relative survival. It implements the maximum-likelihood estimation approach for individual records (9) and assumes the excess hazard to be a step-wise constant function. Version II of **strel** replaces the original command, and incorporates a multivariable functionality, enabling estimation of covariable-specific excess hazards of death. Categorical covariates can be incorporated into the model, allowing the user to obtain interval- and covariate-specific estimates of the quantities of interest; however, the model assumes proportional excess hazards between categories. It is particularly convenient for those who may not have strong statistical skills and who want to analyse very large data sets.
- **stpm2** (14) is a user-written Stata program written by Paul Lambert and Patrick Royston which employs restricted cubic splines to fit flexible parametric survival models. It is a flexible command that can be used to fit complex multi-variable models, time-varying effects and multiple time scales, and has a powerful post-estimation command for predictions. Internal age-standardisation can be performed using the `postestimation –meansurv-` option and external age-standardisation with the addition of the `–meansurvwt()` option.
- **strs** (10) is a user-written Stata program written by Paul Dickman and Enzo Coviello supports both cohort and period estimation, crude probabilities of death, relative/net survival using a number of approaches, and prepares data for tabular/graphical presentation and modelling. Details of all options can be found on the **strs** help file.

Age-standardisation requires a set of survival estimates for each age group. It is not always possible to obtain an estimate for each combination of cancer, age group, sex and for example calendar year of diagnosis with small populations, because of the limited number of cases. Analysts should take extra care in these situations as the age standardised estimates using the **standstrata** option could be incorrect.

- **stns** (29) is a user-written Stata program written by Michel Grzebyk and Isabelle Urmès that implements the method of net survival estimation proposed by Maja Pohar Perme, Janez Stare, and Jacques Esteve (11). Stata version 13 required for **stns** to carry out period survival.



strs (10) implements a modified version of the Pohar Perme estimator that is optimised for the situation where continuous survival times are not available.

- The SEER\*Stat website (<http://seer.cancer.gov/seerstat/>) contains software for implementing net survival.

**Recommended Stata software:**

1. `stns` for net survival using continuous survival time (days)
2. `strs` for net survival using discrete survival time (months, years)
3. `strs` for age-standardised or internally-standardised relative survival estimates (when `stns` is not an option)

## Secondary measures of survival

### Avoidable deaths

Net survival reflects the excess mortality among cancer patients, over and above the background mortality. "Avoidable" cancer-related deaths are the number of deaths that would have been avoided by a time after diagnosis if net survival, at that time, were as high as that achieved by a comparable population (30). Trends in avoidable mortality can be seen as an overall comparative measure of progress in cancer control strategy of a country. Similarly, the public health impact of the differences in survival between deprivation groups can be addressed by considering the number (or proportion) of deaths attributable to cancer that might be avoidable if patients in all groups of society were to have the same survival as that actually observed for patients in the most affluent category (31).

### Estimating "cure" from cancer

The proportion of patients who may be considered "cured" of the disease (the "cure fraction"), is a useful measure to monitor trends in survival. This proportion is estimated using a special type of statistical model: these 'cure models' (32) enable estimation of the level of the asymptote in a curve of relative or net survival as it approaches a plateau, indicating that the cancer patients surviving up to the point of "cure" no longer have significant excess mortality over that of the general population.

## Training

**Cancer survival: principles, methods and applications (Cancer Survival Group, LSHTM <http://www.lshtm.ac.uk/study/cpd/scspma.html>)**

A one-week course on the principles, methods and applications of cancer survival using population-based data, with lectures, computer-based analytic exercises with real data, review sessions and a session for participants to present their own work or ideas. Net survival is the main approach to analysis, with discussion of recent methodological developments (e.g. net survival). This course has been running since 2006, most recently in collaboration with IARC, Lyon. Further details, including dates and location, are available via the website.

**Statistical methods for population-based cancer survival analysis (<http://cansurv.net/>)**

An intensive 5-day course on the principles, methods and application of statistical methods in population-based cancer survival analysis. Central concepts will be covered, such as how to estimate and model relative survival, as well as recent methodological developments. The course consists primarily of lectures and hands-on computing sessions with a focus on individual instruction and discussion. It runs as part of the Summer School on Modern Methods in Biostatistics and Epidemiology. Further details, including dates and location, are available via the website.

## References

1. EUROCARE. Survival of Cancer patients in Europe [Internet]. 2013 [cited 2013 Oct 25]. Available from: <http://www.eurocare.it/>
2. Allemani C, Weir HK, Carreira H, Harewood R, Spika D, Wang XS, Bannon F, Ahn JV, Johnson CJ, Bonaventure A, Marcos-Gragera R, Stiller C, Azevedo E Silva G, Chen WQ, Ogunbiyi OJ, Rachet B, Soeberg MJ, You H, Matsuda T, Bielska-Lasota M, Storm H, Tucker TC, Coleman MP; the CONCORD Working Group. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet*. 2014 doi: 10.1016/S0140-6736(14)62038-9. [Epub ahead of print].
3. Lambert PC, Dickman PW, Osterlund P, Andersson T, Sankila R, Glimelius B. Temporal trends in the proportion cured for cancer of the colon and rectum: a population-based study using data from the Finnish Cancer Registry. *Int J Cancer*. 2007;121(9):2052–9.
4. Rachet B, Woods LM, Mitry E, Riga M, Cooper N, Quinn MJ, et al. Cancer survival in England and Wales at the end of the 20th century. *Br J Cancer*. 2008;99 Suppl 1:S2–10.
5. Verdecchia A, Baili P, Quaglia A, Kunkler I, Ciampichini R, Berrino F, et al. Patient survival for all cancers combined as indicator of cancer control in Europe. *Eur J Public Health*. 2008;18(5):527–32.
6. Rachet B, Maringe C, Nur U, Quaresma M, Shah A, Woods LM, et al. Population-based cancer survival trends in England and Wales up to 2007: an assessment of the NHS cancer plan for England. *Lancet Oncol*. 2009;10(4):351–69.
7. Rachet B, Ellis L, Maringe C, Chu T, Nur U, Quaresma M, et al. Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer*. Nature Publishing Group; 2010 Aug 10;103(4):446–53.
8. Maringe C, Walters S, Butler J, Coleman MP, Hacker N, Hanna L, et al. Stage at diagnosis and ovarian cancer survival: Evidence from the International Cancer Benchmarking Partnership. *Gynecol Oncol*. 2012;127(1):75–82.
9. Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: Elements for further discussion. *Stat Med*. 1990;9(5):529–38.
10. Dickman P, Coviello, 2014. Estimating and modelling relative survival. *The Stata Journal* 20:1-29.
11. Perme MP, Stare J, Estève J. On estimation in relative survival. *Biometrics*. 2012;68(1):113–20. Dickman P, Coviello, 2014. Estimating and modelling relative survival. *The Stata Journal* 20:1-29.
12. Roche L, Danieli C, Belot A, Grosclaude P, Bouvier A-M, Velten M, et al. Cancer net survival on registry data: use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. *Int J cancer*. 2013;132(10):2359–69.
13. Dickman PW, Lambert PC, Coviello E, Rutherford MJ. Estimating net survival in population-based cancer studies. *Int J Cancer*. 2013 Jul 15;133(2):519–21.

14. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stata J.* 2009;(2):265–90.
15. Brenner H, Gefeller O, Hakulinen T. Period analysis for “up-to-date” cancer survival data. *Eur J Cancer.* 2004;40(3):326–35.
16. Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. *Cancer* 1996; 78: 2004-11.
17. Day, 1987. Cumulative rates and cumulative risk. In *Cancer Incidence in Five Continents*, Muir C, Waterhouse J, Mack T, Powell J, Whelan S [eds] vol. V, pp 787-789. International Agency for Research on Cancer. IARC Scientific Publications No. 88: Lyon
18. Brenner H, Rachet B. Hybrid analysis for up-to-date long-term survival rates in cancer registries with delayed recording of incident cases. 2004; 40: 2494-2501.
19. Li R, Abela L, Moore J, Woods LM, Nur U, Rachet B, Allemani C, Coleman MP. Control of data quality for population-based cancer survival analysis. *Cancer Epidemiology* 2014; 38: 314-20.
20. Brenner H, Hakulinen T. Patients with previous cancer should not be excluded in international comparative cancer survival studies. *International Journal of Cancer* 2007 15; 121: 2274-8.
21. Rosso S, De Angelis R, Ciccolallo L, Carrani E, Soerjomataram I, Grande E, Zigon G, Brenner H; EUROCARE Working Group. Multiple tumours in survival estimates. *Eur J Cancer* 2009; 45:1080-94.
22. International Agency for Research on Cancer, World Health Organisation. International rules for multiple primary cancer (ICD-0 Third Edition). Internal Report No. 2004/02. IARC, Lyon, 2004.
23. ONS. Office of National Statistics (ONS) National Life Tables, 2010-2012 [Internet]. 2014 [cited 2014 Apr 3]. Available from: <http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-352834>
24. WHO. World Health Organisation (WHO) mortality data and statistics [Internet]. 2013 [cited 2013 Nov 18]. Available from: <http://www.who.int/healthinfo/statistics/mortality/en/index3.html>
25. CRUKCSG. Life tables for England and Wales by sex, calendar period, region and deprivation. [Internet]. 2011 [cited 2012 Nov 27]. Available from: <http://www.lshtm.ac.uk/eph/ncde/cancersurvival/tools/index.html>
26. Carstairs V. Deprivation indices: their interpretation and use in relation to health. *J Epidemiol Community Health.* 1995;49(Suppl 2):3–8.
27. Hills M, Rachet B, Falcaro M. strel 2: a command for estimating excess hazard and relative survival in large population-based studies. *The Stata Journal* 2013; 14:176-190.
28. Corazziari I, Quinn M, Capocaccia R. Standard cancer patient population for age standardising survival ratios. *Eur J Cancer.* 2004;40(15):2307–16.

29. Clerc-Urmès I, Grzebyk M, Hédelin G. Net survival estimation with stns. *The Stata Journal* 2014; 14: 87-102.
30. Abdel-Rahman M, Stockton D, Rachet B, Hakulinen T, Coleman MP. What if cancer survival in Britain were the same as in Europe: how many deaths are avoidable? *Br J Cancer*. Nature Publishing Group; 2009 Dec 3;101 Suppl (S2):S115–24.
31. Ellis L, Coleman MP, Rachet B. How many deaths would be avoidable if socioeconomic inequalities in cancer survival in England were eliminated? A national population-based study, 1996–2006. *Eur J Cancer*. 2012;48(2):270–8.
32. Lambert PC. Modeling of the cure fraction in survival studies. *Stata J*. 2007;(3):1–25.

## Appendix 1: PHE cancer survival defaults – for survival analyses performed in 2015/16.

Version 1, 9/4/15

This document details a set of processes that can be (but don't have to be) followed to yield consistent cancer survival results. This document applies to analyses performed in 2015/16, for tumours diagnosed from 1995, and is expected to change over time. Three stages of survival analysis are identified below:

### Stage 1: pulling cohort from AV\_TUMOUR using SQL

1. The ICD10\_02\_3CHAR field should be used to filter the cohort of interest by tumour type.
2. Filter for residents of England by looking for a first character of "E" in the LSOA11 code.
3. Filter for a CREG\_CODE which is one of the English registries
4. Tumours diagnosed 1995-2011 should be joined to the ONS data via an inner-join to only include cases present in *both* CAS and in the ONS data. Outside this range of years just use the CAS records.
5. Include only final registrations with STATUSOFREGISTRATION='F'
6. Exclude known duplicates from historic ECRIC/THAMES boundary change.
7. Limit ALIASFLAG in AV\_TUMOUR to be equal to '0' only to avoid including referenced-off duplicates still present in the patient table.
8. (Don't exclude tumours due to DCO status or tumour order at this point – wait until stage 2 to allow the exclusion to be logged in the Stata output.)
9. Check that all vital status update dates are after the censor date. If not then arrange for tracing of all cases with vital status date < censor date and merge traced deaths into cohort.

### Stage 2: processing cohort for analysis using Stata

**Note – the bulk of the exclusionary criteria are applied at this stage so that the number of exclusions and reason are captured by the Stata output.**

1. Drop cases that could not be traced.
2. Drop DCOs
3. Drop tumours if sex is unknown
4. Drop tumours if diagnosis date is unknown (retain cases with imputed dates)
5. Drop tumours if date of birth is unknown (retain cases with imputed dates)
6. Drop tumours for persons aged <15 or >99 at diagnosis
7. Drop second and subsequent tumours where the tumour has the same topography code as an earlier tumour in the period of the study. Do this after exclusions above.
8. Set zero-survivors survival time to 1 day.

### Stage 3: running survival analyses using Stata (currently assumes relative cohort survival)

1. Use breakpoints of br(0(0.08333)0.5, 0.75(.25)1.75, 2(0.5)5, 6(1)11)
2. Use LSHTM 2011 lifetables.
3. For 2012 and 2013 cases just use the 2011 lifetables as they are.
4. Merge lifetables using mergeby(year sex age GOR quintile)

5. If performing age standardisation use the Corazziari method and weightings.
6. Do not report survival estimates based on less than 10 deaths in survival period.