

Appendix 02:

Statistical methodology

The statistical methods used in this report are those widely used by cancer registries throughout the world in describing the burden of cancer within their catchment area. Consequently a significant amount of literature is available on these techniques (see references 227 and 228 as introductory texts to general principles and survival analysis respectively). However while some of the methods, such as age-standardisation, have been used by epidemiologists for many years, some other techniques, such as relative survival period analysis, are relatively new. Additionally there are variations within the general methods such as the standard populations used or methodology behind deriving expected survival that can vary between studies. This appendix thus gives a very general overview of the techniques utilised in this report focusing on local variations and identifying which methods were selected. It should be noted however that this chapter is for reference only and to allow future reproduction of the results presented within the report and not meant to supplant the numerous (and better) texts on cancer registration techniques and medical statistics.

A2.1: Descriptive measures of incidence and mortality

The most common and useful measures of cancer levels in a population are the absolute number of cases (incidence) or deaths (mortality) in a given year. It is these very basic figures that allow planning by the health services of each country for each year and are the fundamental building blocks of any other analysis. However the number of diagnoses of cancer within a year compared to the size of the population of Ireland is relatively small. This can result in the number of events being studied fluctuating each year as a result of random factors, particularly for the less common cancers. This requires the population to be observed over a number of years in order to present a stable value for the number of cases diagnosed or number of deaths per year. Throughout this report a five-year annual average for the number of cases diagnosed or number of deaths from cancer has thus been used to represent the situation at a given point in time rather than using data for a particular year.

A2.1.1: Crude rates

While the absolute number of cases or deaths is useful for planning purposes these measures do not allow accurate comparison of populations of different size. A crude incidence/mortality rate compensates for this by presenting the number of cases/deaths per 100,000 members of the population and is based upon the ratio of events to members of the population. If we let R be the number of events in a given year and N be the population for that year then the crude incidence/mortality rate per 100,000 persons for that year, C, is given by:

$$C = \frac{R}{N} \times 100,000$$

In a situation where several years worth of data is required then R represents several years worth of events and the population used must reflect this by summing the populations of the years in question. In this event N is referred to as the number of person-years of observation.

A2.1.2: Age-specific rates

Crude rates are not always the best measure for comparative purposes as there is a very strong relationship between cancer and age, thus a younger population is more likely to have a lower number of cancers than an older population of the same size. The most useful and easiest to calculate measure that compensates for differences in the age-structures of two populations is a set of age-specific rates, which are calculated in a similar manner to crude rates.

If r_i is the number of events in age group i and n_i is the number of person-years of observation within which the events occur, then the age-specific rate for that age group, denoted by a_i is given by:

$$a_i = \frac{r_i}{n_i} \times 100,000$$

The draw back of the use of age-specific rates is of course the number of these that must be quoted in order to give a full picture of the cancer/population being studied, particularly since five-year age-groups are the most commonly used age breakdown. In addition the small numbers involved can cause very noticeable fluctuations over time, even when several years worth of data are used.

A2.1.3: Age-standardisation

A widely used technique, which provides a summary measure that allows for the changing or differing population age-structure, is age-standardisation. This does not completely overcome the difficulty in comparing rates between populations²²⁹ and is thus not a replacement for age-specific rates but does provide statistics that are more manageable and lend themselves to further analysis, particularly comparisons of many sets of incidence/mortality rates such as is required in trend and geographic analysis.

There are two methods of age-standardisation, direct and indirect, used in this report. The former is the most commonly used as it provides an absolute measure while the indirect method provides a value relative to some other measure and is thus very restricted in its range of applications.

Direct standardisation

The result of direct standardisation is known as an age-standardised rate (ASR), which refers to the number of events per 100,000 persons occurring in the population if the population possessed the same age structure as a standard population. There are two standard populations used in this report, the European standard population, which is used throughout the EU, and the World standard population, which is used for global comparisons of cancer rates. The former is the preferred measure used by NICR and NCRI, with the World standard only used for international comparisons of incidence. (Table A2.1)

The calculation of an age-standardised rate is based upon the age-specific rates introduced in section A2.1.2. These rates are multiplied by the standard population for that age class (also known as the weight), with the products summed and divided by the total standard population. In mathematical terms, if a_i is the age-specific rate for age class i and w_i is the standard population of age group i , with A the number of age intervals then the age-standardised rate, ASR, is given by:

$$ASR = \frac{\sum_{i=1}^A a_i w_i}{\sum_{i=1}^A w_i}$$

This value is the standard measure used for making comparisons between different populations, however while useful within this context it cannot be interpreted as a measure of the actual number of events within a population due to the removal of the age effect and corresponds to the crude rate in the standard population rather than that being studied.

Standardised rate ratio

Given its purpose as a comparative measure, it is useful to introduce a derivative of the age-standardised rate known as the standardised rate ratio, which is the ratio of two age-standardised rates. It represents the relative risk of disease in one population compared to another and is beneficial for presentational purposes as it allows the presentation of a single percentage rather than two absolute values. Its calculation is straightforward; if ASR_1 is the age-standardised rate for population 1 and ASR_2 is the age-standardised rate for population 2 then the standardised rate ratio of population 1 compared to population 2, denoted SRR_{1-2} , is given by:

$$SRR_{1-2} = \frac{ASR_1}{ASR_2}$$

Table A2.1: Standard populations used in age-standardisation

Age class	European standard population	World standard population
0-4	8,000	12,000
5-9	7,000	10,000
10-14	7,000	9,000
15-19	7,000	9,000
20-24	7,000	8,000
25-29	7,000	8,000
30-34	7,000	6,000
35-39	7,000	6,000
40-44	7,000	6,000
45-49	7,000	6,000
50-54	7,000	5,000
55-59	6,000	4,000
60-64	5,000	4,000
65-69	4,000	3,000
70-74	3,000	2,000
75-79	2,000	1,000
80-84	1,000	500
85+	1,000	500
Total	100,000	100,000

This ratio can either be quoted as a ratio, be expressed as a percentage by multiplying by 100, with 100% referring to the events in the reference population, or be expressed as the percentage difference of one age-standardised rate compared to another by subtracting 1 and multiplying by 100. The later has been used in this report, as it is the easiest to interpret.

Indirect standardisation

The indirect method of age-standardisation is a comparison of the observed number of events within a population and the number of events expected in a reference population of the same size. When considering incidence of cancer the expected number is calculated by applying the age specific incidence rates of a reference population to the observed population (i.e. the population being studied). The formula for the standardised incidence ratio (SIR) is:

$$SIR = \frac{\sum_{i=1}^A r_i}{\sum_{i=1}^A \frac{a_i n_i}{100,000}}$$

where a_i is the age specific incidence rate in the reference population, n_i is the observed population in age group i and r_i is the observed number of cases in age group i . The formula is also valid for mortality data with a_i as the age specific mortality rate and the result known as the standardised mortality ratio (SMR).

The result is usually expressed as a percentage by multiplying by 100, with 100% referring to the events in the reference population. This measure is frequently used for geographic analysis with the reference population being that for an entire country (e.g. Ireland) and SIRs or SMRs calculated for smaller geographic units (e.g. district councils/counties) giving an indication of how cancer levels in these areas compare to that of the entire country.

A2.1.4: Cumulative risk

Another commonly used measure which is of particular interest to the general public, but is not as useful as age-standardised rates, is the cumulative risk, which gives the risk of an individual developing cancer during a particular age span (usually 0 to 74) assuming the absence of other causes of death. Like age-standardised rates it is based upon age-specific rates but is expressed as a percentage rather than a rate. It is derived using the formula:

$$CR_{0-74} = 100 \left[1 - \exp \left(- \frac{1}{100} \sum_{i=1}^A \frac{a_i t_i}{100000} \right) \right]$$

where a_i is the age-specific rate for age class i , t_i is the duration of age class i , A is the number of age intervals between 0 and 74 (or the upper age of the age span under consideration) and CR_{0-74} , is the cumulative risk of developing cancer before the age of 75.

A2.1.5: Unknown values

In the discussion on data quality in appendix A1 it was noted that both NICR and NCRI have a high level of completion in the data fields required for analysis. In particular both age and sex are 100% complete for all cancers (excluding non-melanoma skin cancer). Thus while corrections to age-standardised rates are possible in the event that age is missing in a small percentage of cases, these are not required for this report.

Geographic and socio-economic information is less complete, however incomplete records must be catered for otherwise any ASRs or SIR/SMRs will be an underestimate of the true value. In this event records with an unknown district council, county or deprivation quintile are redistributed within the relevant country according to the distribution of the records with a known district council, county or deprivation quintile. This assumption is not completely justified as missing geographic information is more likely to occur for the elderly and in rural areas but the adjustment will bring the estimated rate closer to the true value.

A2.1.6: Confidence intervals and statistical significance

This highlights an important factor of age-standardised rates in that they are only estimates of the true value, as uncertainty exists due to random fluctuations in the number of events between different populations. In order to quantify this uncertainty any rates in this report are accompanied by 95% confidence intervals to indicate the range within which there is a 95% probability that the true value is likely to fall. The size of the confidence intervals depends upon the number of events and the size of the population within

which they occur, with rates made up of a small number of observations within a large population being less stable and having large confidence intervals. The formulae used to calculate confidence intervals for the incidence and mortality measures used in this report are given in table A2.2.

Rates for two different time periods or population groups are considered to differ only if the 95% confidence intervals for the two age-standardised rates do not overlap. Alternatively, in the case of ratios, the rates for a population differ from those of the reference population only if the confidence interval does not include 100%. This is known as statistical significance and for significant differences the level of certainty about any difference can be quantified by calculating the p-value. This measure provides the probability that any difference observed between two rates is due to chance. Thus a p-value of 0.001 indicates a 99.9% probability that differences are genuine and not a result of random factors.

Table A2.2: Formulae for confidence intervals of incidence and mortality measures

Measure	95% confidence interval
Age-standardised rate (ASR)	$ASR \pm 1.96\sigma$ where $\sigma^2 = \frac{\sum_{i=1}^A [a_i w_i^2 (100000 - a_i) / n_i]}{\left(\sum_{i=1}^A w_i\right)^2}$
Standardised rate ratio (SRR ₁₋₂)	$\exp \left[\text{Ln}(SRR_{1-2}) \pm 1.96 \sqrt{\frac{\sigma_1^2}{ASR_1^2} + \frac{\sigma_2^2}{ASR_2^2}} \right]$
Standardised incidence ratio (SIR)	$SIR \pm 100 \left[1.96 \times \frac{\sqrt{\sum_{i=1}^A r_i}}{\sum_{i=1}^A \frac{a_i n_i}{100000}} \right]$

A2.1.7: Trend analysis

Trends in ASRs are assessed by calculation of the annual percentage change (APC), which is the percentage increase, or decrease per year in the age-standardised rate. From our earlier discussion regarding fluctuations in rates over time it is not appropriate to select the rates corresponding to the beginning and end of the trend and calculate the percentage difference. Using an average over several years provides a better estimate; however a much better approach is through the use of curve fitting or regression.

A full discussion of regression is well outside the scope of this appendix but in summary it is the mathematical technique that allows a series of points in a trend to be estimated by a simple formula. In this case we are assuming that the age-standardised rate, ASR, depends upon the calendar year according to the equation

$$\text{Ln}(ASR) = mx + b$$

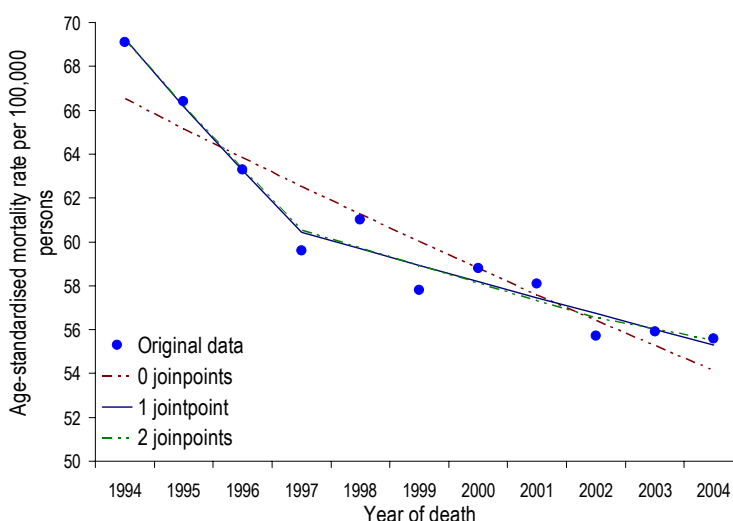
where x is the calendar year, b is a constant and the annual percentage change (APC) is given by

$$APC = 100 \times (e^m - 1)$$

The formula, or model, is known as a log-linear model and is not the only type we could have chosen, however the curve it creates is a good fit for the data available and is a commonly used model in cancer incidence and mortality studies.

The calculation makes the assumption that the age-standardised rates increase or decrease at a constant rate over the period examined. While this is a reasonable assumption for incidence and mortality rates, it is not reasonable to assume that there is no change in the trend during the time period for which data exists. To investigate whether changes in trends occur during the years for which data exists the JoinPoint regression program developed by the US National Cancer Institute²³⁰⁻²³¹ was used. This software reads trend data and divides the trend into an increasing number of separate sections, which are connected by points known as joinpoints. For each number of joinpoints (starting at 0) it fits the simplest model that the data allow up to a maximum number dictated by the user (with 3 selected for this report)

Figure A2.1: Example of trend analysis using joinpoint regression (lung cancer mortality in Ireland). The addition of one joinpoint is statistically significant; however the addition of a second is not. Thus the results indicate a single change of trend between 1994 and 2004 that occurred in 1997.



giving a set of possible fits to the data (4 in the analysis in this report) ranging from the best possible fit using a straight line to the best possible fit using the maximum number of joinpoints. Statistical tests are conducted to assess whether the addition of joinpoints from one model to another is statistically significant thereby allowing the user to test whether an apparent change in trend is statistically significant. See figure A2.1 for an example.

A2.2: Treatment analysis

The majority of analysis of treatment is through the derivation of the number of patients receiving different types of treatment and their characteristics, with these numbers frequently presented as a percentage of the total number. While this is fairly straightforward, random fluctuations in values (see section A2.1.6) mean that caution needs to be exercised when comparing either two proportions or the overall distribution of treatment (or other factors) between two sets of patients. Statistical tests exist for both scenarios and are utilised in this report to identify those differences that are statistically significant. Statistical decisions with regard to differences in proportions are based upon the assumption that any differences are normally distributed about zero, while the chi-square test is used to test for differences between the distribution of patient or tumour characteristics of two different cohorts. In both cases a 95% confidence level is applied.

Details on both the normal distribution with its use in testing for differences in proportions and on the chi-square test for differences between distributions can be found in numerous statistical texts²³².

A2.2.1: Relative risk

In analysis of cancer incidence, age-standardisation is used when making comparisons between incidence rates by different population groups, as age is a very strong factor in the development of cancer. Likewise there are many factors that can influence whether a patient receives treatment. For any thorough understanding of differences between patient groups these factors must therefore be identified and controlled for. This is done through logistic regression, which in this report is used to identify and quantify the degree to which various patient characteristics (e.g. age, sex) and tumour details (e.g. stage, basis of diagnosis) can influence treatment receipt while controlling for the interaction between these factors.

The methodology behind logistic regression is similar to other forms of regression. However the outcome and variables used in the model are dichotomous (i.e. have only two possibilities) rather than continuous which was the case for the regression models used for trend analysis of incidence (see section A2.1.7). This means that potential factors are split into more than one variable in the event that more than two divisions are insufficient to fully represent the group of cancer patients being studied. Thus while sex can be represented by a single variable, age is split into several distinct groups. Additionally the models developed in this report to identify factors influencing treatment are multivariate with age and sex initially assumed to be factors and further variables added to improve the models predictive power. The set of variables used in the multivariate models are sex, age, stage at diagnosis, basis of diagnosis, deprivation quintile, cell type, year of diagnosis and country. The later is omitted in separate models for Northern Ireland and Republic of Ireland. It is recognized however that these models could be further improved by the addition of further variables.

With the factors influencing treatment identified, odds ratios (the ratio of the odds of an event occurring in one patient group to the odds of it occurring in a baseline patient group) are derived from the coefficients to the variables of the logistic regression model by taking the exponential function of these coefficients. However while analysis of odds ratios can lead to useful statistical conclusions in their own right, a more useful and easy to interpret measure of the likelihood of a group of patients receiving treatment relative to a baseline group is the relative risk, which can be derived from the odds ratio using the approach suggested by Zhang and Yu²³³. If OR is the odds ratio of a group of patients receiving treatment compared to a baseline group and the proportion of cases treated in the baseline group is given by P_0 , then the relative risk RR of the patient group to the baseline is given by:

$$RR = \frac{OR}{(1 - P_0)(OR \times P_0)}$$

Analysis of relative risk will lead to the same conclusions as examination of odds ratios; however the relative risk can be interpreted as the percentage of patients receiving treatment compared to the baseline value, or as a percentage difference between the two groups.

When comparing relative risks from two different logistic models (e.g. for Northern Ireland and Republic of Ireland) the confidence intervals accompanying the relative risk give some indication of whether or not there is a significant difference between the two, with confidence intervals that do not overlap indicating a significant difference. A better comparative measure however is known as the test for interaction²³⁴ which gives the p-value for one relative risk RR_1 compared to another RR_2 . This test is based upon the normal distribution with a z-value given by:

$$z = \frac{\ln(RR_1) - \ln(RR_2)}{\sqrt{\sigma(\ln(RR_1))^2 + \sigma(\ln(RR_2))^2}}$$

A significant difference indicated by confidence intervals around relative risks RR_1 and RR_2 will also be identified via this test; however the test may also identify significant differences in cases where the confidence intervals overlap by small amounts.

A2.3: Survival analysis

Survival refers to the proportion of patients who survive a given amount of time after a diagnosis of cancer. It is one of the best indicators as to the efficiency of diagnostic and treatment methods in a geographic area and is widely used by cancer registries as a broad indicator as to the effectiveness of health services in the treatment of cancer. Unfortunately it is also one of the most difficult cancer measures to calculate, with many different techniques and types of measure in existence.

A2.3.1: Observed survival

The most fundamental, and perhaps of most relevance to patients, is observed survival, which is the probability that a patient with cancer will be alive at the end of a particular length of time as measured from the date of diagnosis. It is independent of the cause of death and can be calculated using several different techniques. In this report the Kaplein-Meier method has been used to calculate the observed survival, S_i , for a time i after the date of diagnosis. Using this method S_i is calculated using the formula

$$S_i = \prod_{k=1}^i \left(1 - \frac{d_k}{n_k - \frac{1}{2}w_k} \right)$$

where k is a predefined time interval between the date of diagnosis and i , d_k is the number of deaths from any cause occurring during interval k , n_k is the number of patients alive entering interval k and w_k is the number of patients withdrawn alive during the k^{th} interval.

It is worth noting how the observed survival is calculated by breaking the overall time period being measured, i , into intervals of length k . The choice of these intervals can have an impact upon the final result, albeit a small one. In this report we have used intervals of three months for the first year after diagnosis, six months for the next two years and one-year after that point.

The number of patients, n_k , who are still alive entering interval k is dependent upon the survival experience of each individual patient. This is determined by assessing whether each patient is alive or dead at the date that the start of interval k refers to by using the date of death for that patient. Those alive at the start of interval k but who have died by the end of the interval contribute to the value of d_k for that interval while those alive at the end of interval k contribute to the value of n_{k+1} for the next interval. However these values can only be determined for intervals during which follow up data (i.e. alive or dead status) for each patient is known. The date beyond which this information is not available for a patient is known as the censor date, with the alive or dead status of the patient on this date known as the vital status. Due to follow up data on patients coming from death registrations in Ireland, the censor date is the same for all patients, although allowance is made in the Kaplein-Meier method for some patients, w_k , to be withdrawn alive from an interval for reasons such as emigration (known as lost to follow up). For this report the censor date is 31st December 2004 while no patients have been withdrawn due to being lost to follow up.

Having to apply a censor date however does mean that there are restrictions as to the length of survival time that can be calculated. For example due to the lengthy follow up time required to derive five-year observed survival, it is only possible to report on the survival experience of patients diagnosed 7-8 years in the past. Thus for this report the most up to date five-year observed survival data is for patients diagnosed in 1999.

A2.3.2: Relative survival

Observed survival for cancer patients includes death from causes other than cancer, some of which may be related to cancer or its cause (e.g. other smoking related illnesses) or may even be completely unconnected to the disease (e.g. accidental death). It is thus not the best survival measure for monitoring the effectiveness of treatment of the disease or its impact on society. Instead measures that remove other causes of death from survival figures are preferred. The most commonly used of these types of measures, but not the only one, is relative survival, which is used in this report. (Fig. A2.2)

Relative survival is the ratio of the observed survival of a given group of patients to the expected survival for a group of persons in the general population with the same characteristics (usually sex and age, but also country in this report). The expected survival can be calculated using several different techniques. The method used in this report is the Ederer II method²³⁵ which is calculated in a similar way to observed survival by using the formula:

$$E_i = \prod_{k=1}^i \left(1 - \sum_{h=1}^{n_k} \frac{P_k(h)}{n_k} \right)$$

where E_i is the expected survival for a time i after the date of diagnosis, k is the same predefined time interval between the date of diagnosis and i that is used in the calculation of observed survival, n_k is the number of patients alive entering interval k and $P_k(h)$ is the probability of a similar person, h , in the general population surviving to the end of period i . This latter value is taken from life tables derived by the Office of National Statistics (ONS)²³⁶ (for Northern Ireland) and Central Statistics Office (CSO)²³⁷ (for Republic of Ireland) which use information on deaths from all causes in the general population along with mid-year population estimates to develop estimates of life expectancy by age and sex.

Confidence intervals

As with the other statistical measures used in this report observed and relative survival values are accompanied by 95% confidence intervals. These are derived directly from the standard error using the following formula:

$$RS \times \exp \left[\pm 1.96 \frac{\sigma}{RS} \right]$$

where RS refers to the relative survival rate with σ as the standard error.

Age-standardisation

Survival from cancer is dependent upon age at diagnosis. Thus when comparing survival from different populations the same difficulties that occur when comparing incidence and mortality rates are also apparent, with populations having high percentages of younger people having better survival than those with high percentages of older people. To compensate we thus apply the direct age-standardisation approach that was used with incidence and mortality rates to all relative survival rates. As before this involves the application of a standard population to age-specific relative survival rates. However it is necessary to bear in mind that the population at risk when investigating survival differs from that for incidence in that the latter refers to the entire population while survival relates only to patients with cancer. Thus different standard populations are required, not only from that used for the age-standardisation of incidence data but also for those cancer sites that have significantly different age distributions for patients than usual, such as testicular cancer which is more predominant in young men and prostate cancer which is more common in the very elderly. In this report we use the same standard populations as those used in the EUROCARE-IV²³⁸ study, which were those suggested by Corazziari et al²³⁹. (Tab. A2.3)

Figure A2.2: Typical survival curves using observed and relative survival

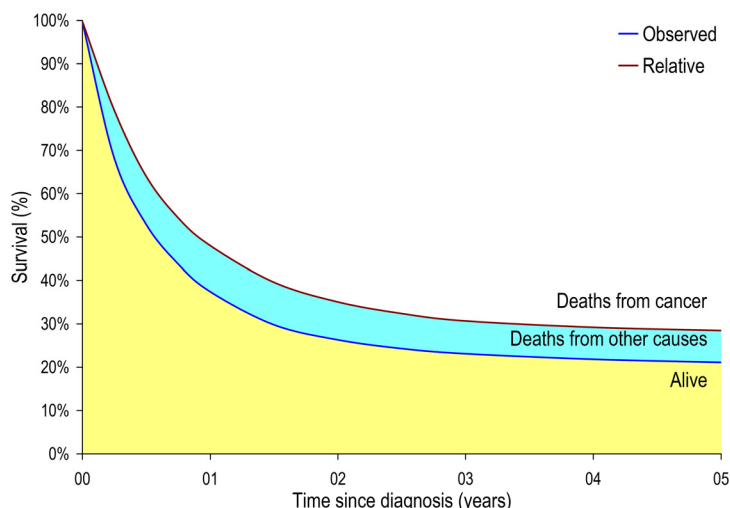


Table A2.3: Standard cancer populations used in age-standardisation of relative survival

Standard 1		Standard 2		Standard 3		Standard 4	
Age class	Population	Age class	Population	Age class	Population	Age class	Population
15-44	7,000	15-44	28,000	15-44	60,000	15-54	19,000
45-54	12,000	45-54	17,000	45-54	10,000	55-64	23,000
55-64	23,000	55-64	21,000	55-64	10,000	65-74	29,000
65-75	29,000	65-75	20,000	65-75	10,000	75-84	23,000
75+	29,000	75+	14,000	75+	10,000	85+	6,000
Total	100,000	Total	100,000	Total	100,000	Total	100,000
Cancer sites							
All except for those requiring standards 2-4		Nasopharynx, soft tissues, melanoma, cervix uteri, brain, thyroid, bone		Testes, Hodgkin's disease		Prostate	

Source: Corazziari et al²³⁹

A2.3.3: Conditional survival

Observed and relative survival is measured from the point that patients are diagnosed with cancer; however the start date in survival analysis can be any date that has relevance to the patient provided that the length of time being analysed is clearly defined. Using this flexibility the long-term survival of patients who have already survived a certain amount of time can be determined by using start dates for survival analyses that are a certain amount of time after diagnosis. In other words it is possible to derive the probability that a patient will survive a certain amount of time if that patient has already survived to a certain point. This measure is known as conditional survival and in this report we present conditional survival data for patients who are alive at one-year increments from date of diagnosis up to a maximum of five-years.

Patients who survive a minimum amount of time obviously have higher overall relative survival than all patients measured from diagnosis. However the benefit of examining conditional survival is in the possible identification of a point where 100% of those who survive to that point are from then on cancer free. For example suppose five-year relative survival from diagnosis for all patients was 50% but six-year relative survival from diagnosis was 70% for patients who survived a minimum of one-year. This would show that the group of patients who were alive one-year from diagnosis was much closer to being cancer-free than the group of patients alive at diagnosis. Extending further if seven-year relative survival was 100% for those were still alive two years after diagnosis it could be safely concluded this group of patients was cancer-free. Given that this is an investigation of long-term survival the analysis in this report is thus restricted to those diagnosed in 1994-96.

A2.3.4: Period analysis

The method of deriving survival results thus far described is known as cohort analysis and is the method widely used by cancer registries in survival analysis. One recognised disadvantage of using this method is that due to the lengthy follow up time required to derive five-year relative and observed survival, it is only possible to report on the survival experience of patients diagnosed 7-8 years in the past.

Period analysis was introduced in 1997 by Brenner & Gefeller²⁴⁰ as a method for obtaining more up-to-date estimates of survival, which can complement those obtained by traditional methods. This approach involves using the year that patients survive to instead of the year that they are diagnosed with cancer. Thus survival data for patients diagnosed in 2000-2004 can be estimated using the period approach by examining the survival experience of patients who have survived to 2000-2004. (Fig. A2.3)

Figure A2.3: Method of deriving most up to date survival estimates using cohort analysis (1997-1999, solid box) and period analysis (2000-2004, dashed box). Cells represent the minimum and maximum years of follow up data available for each year of diagnosis.

Year of diagnosis	Year of follow up											
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	
1994	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10	
1995		0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	
1996			0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	
1997				0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	
1998					0-1	1-2	2-3	3-4	4-5	5-6	6-7	
1999						0-1	1-2	2-3	3-4	4-5	5-6	
2000							0-1	1-2	2-3	3-4	4-5	
2001								0-1	1-2	2-3	3-4	
2002									0-1	1-2	2-3	
2003										0-1	1-2	
2004											0-1	

A2.3.5: Modelling of excess mortality

Survival is dependent upon many factors, age having one of the biggest impacts. However there are other factors that potentially have an even larger impact (e.g. stage at diagnosis). While these are investigated in this report primarily by calculating age-standardised relative survival for these factors (when data is available) a full understanding on how these factors interact can only be obtained by attempting to model survival using these factors. Modelling of any description can be particularly tricky, especially in the case of survival, which is a continuous variable rather than a binary one as in the case of treatment. Traditionally Cox's proportional hazards model is used to model observed survival, however in this report we have focused on relative survival and thus use techniques appropriate to this measure. As with most other survival techniques thus far encountered there are several possible approaches to this task. In this report we have selected the method utilised by Dickman et al²⁴¹, which relies upon Poisson regression to model the number of deaths and patients entering each survival interval and calculate excess mortality for each factor incorporated into the model. The excess mortality is inversely related to survival with low survival compared to the baseline inducing relatively high excess mortality. The ratio of excess mortality to that of the baseline is known as the excess hazard ratio.

The variables used in these investigations are those that are readily available in the compiled dataset and are known to contribute to cancer survival. For the majority of cancer sites these are the same variables as used in modelling of treatment, with receipt of treatment also included as a potential factor. However the models presented in this report are still of a very basic nature, particularly for those cancers where treatment and stage data is incomplete and it is therefore acknowledged that there is potential for considerable improvement in this area.

A2.4: Prevalence

Prevalence refers to the number of people living in a population with a diagnosis of cancer. Most cancer registries have difficulty in providing an exact figure for this value for a variety of reasons. In the context of Ireland the problems are threefold:

- There is no point at which cancer is considered cured. While some people diagnosed with cancer may be cancer free within a few years, others may need treatment for a considerable length of time. Thus in order to develop prevalence figures, either an assumption must be made as to an average "cure" point (sometimes arbitrarily taken as being five-years) or all people who have been diagnosed with cancer and are still alive at a certain point must be included.
- The cancer registries in Ireland have information on people diagnosed with cancer from 1994 onwards (Northern Ireland has data from 1993 but is excluded from this report for the purpose of creating data for all of Ireland). Unfortunately with regard to measuring prevalence, this means that there is no information on members of the population who had a diagnosis of cancer prior to 1994. Thus any prevalence figures produced would be an undercount of the true value.
- Neither NICR nor NCRI have information on those cancer patients who have emigrated from Ireland since diagnosis, which might result in a slight inflation of the prevalence figures.

Figures for overall prevalence are thus not provided in this report, however prevalence figures for people diagnosed within the most recent eleven-years (1994-2004) and five-years (2000-2004) are provided. These would be equivalent to prevalence figures that assume that a patient can be considered cancer free within eleven-years and five-years respectively. More detailed analysis is provided using the later definition, as IARC occasionally uses this definition to estimate prevalence²⁴².

A2.5: International comparisons

Cancer statistics on incidence and survival are available from the cancer registries in most countries in the European Union as well as North America and Australia; however caution needs to be exercised when making comparisons between statistical measures in Ireland and these countries for a variety of reasons:

- Incidence and survival rates in different countries use different diagnostic periods to those in this report. Given that cancer rates change over time any differences observed between countries could be the result of differences in the time period being examined as opposed to regional variations.
- In some cases incidence and survival rates from other countries only represent a fraction of the population, as the cancer registries do not always cover the entire country.

- Confidence intervals for rates from other countries are not always provided. In such an event it cannot be ascertained whether or not differences between countries are statistically significant.
- Different age structures exist in different countries. Given that cancer is strongly dependent upon age any differences in rates may be a factor of differences in the age distribution of the population. While most cancer registries regularly produce age-standardised rates to compensate for this, different standard populations are used in different parts of the world.
- While most cancer registries use the ICD10 classification for recording cancer, as illustrated in Appendix 1 coding techniques can differ between countries.

These problems are widely recognised, thus various international collaborations are regularly undertaken to address these issues (with the exception of complete coverage of a country which little can be done about). While still flawed, the results provide the best possible comparisons of incidence and survival between countries. The results from these collaborations are used in this report and are briefly described below.

A2.5.1: Incidence

International data for incidence of cancer in various countries comes from the IARC publication "Cancer incidence in five continents: Volume IX"²⁴³ which collated and published information on 60 countries world wide using data from 225 cancer registries. Both NICR and NCRI supplied data to this volume, which was published in 2007 and reported on cancers diagnosed in 1998-2002.

The primary measure produced by IARC for this publication was age-standardised incidence rates, which were standardised using the world standard population. The ICD10 classification was used to identify each cancer site, although for some records this was changed from the code supplied by each contributing registry in order to maintain consistency across different countries. These changes were based upon the ICD-O-2 or ICD-O-3 topography and morphology codes supplied by the registry and were made primarily in situations where these codes were inconsistent with the supplied ICD10 code as derived by IARC²⁴⁴.

Due to size constraints only a few of the countries included in the compendium have been included in this report. The number of cases diagnosed by sex and five-year age group, available online as a companion to the publication, was used to reproduce the published rates as this approach also allowed aggregated data for the European Union (EU) to be derived. Data for this political entity was derived in two ways:

- Using the 15 countries making up the EU from 1995-2004 (EU-15). This was derived using data from Republic of Ireland, UK (including Northern Ireland), France, Germany, Italy, Spain, Denmark, Belgium, Netherlands, Sweden, Finland and Austria. The remaining three countries, Luxembourg, Greece and Portugal, are not included in the compendium while some of the data for individual countries does not cover 100% of the population.
- Using the 27 countries making up the EU from 2007 to present (EU-27). This was derived using data from the twelve countries listed above plus data from Bulgaria, Czech Republic, Estonia, Latvia, Lithuania, Malta, Poland, Slovakia and Slovenia. However data from Romania, Cyprus and Hungary is not available.

Data from the USA has also been included for comparative purposes. This data comes from SEER, which collates data from 14 different cancer registries in the USA but does not represent 100% of the population.

Data quality differs between the various countries. An overview of these differences is supplied in table A2.4 and refers to the data used by IARC in the production of the compendium.

A2.5.2: Survival

Survival data for 20 countries within Europe are available from the EURO CARE-IV study that was conducted in 2007 and investigated patients diagnosed with cancer in 1995-1999 with follow up to the end of 2003²³⁸. Both Northern Ireland and Republic of Ireland provided data for this study with results for most cancer sites available. The participating European countries are listed in table A2.5, however not all participating cancer registries (e.g. France, Germany and Spain) cover 100% of the population of their country. Consequently, despite the large sample size, caution should be exercised when interpreting survival differences between Ireland and these countries.

Table A2.4: Data quality of cancer incidence (excluding NMSC) in countries contributing to the IARC publication "Cancer incidence in five continents: Volume IX"

	% MV	% DCO	% other & unspecified		% MV	% DCO	% other & unspecified
Republic of Ireland	84.4%	2.8%	5.5%	Finland	92.8%	7.9%	1.7%
Northern Ireland	80.5%	1.4%	5.5%	Sweden	98.3%	0.0%	4.7%
England & Wales	78.6%	4.3%	5.5%				
Scotland	84.8%	0.6%	5.2%	Ireland (NI+ROI)	83.2%	2.4%	5.5%
France	95.7%	3.1%	3.0%	UK (Inc. NI)	79.3%	3.9%	5.5%
Spain	88.1%	3.6%	4.1%	EU-15 countries	85.8%	3.9%	4.2%
Italy	84.5%	1.4%	2.5%	EU-27 countries	84.9%	4.2%	4.1%
Germany	81.7%	12.4%	3.0%				
Belgium	93.5%	4.2%	5.5%	Australia	90.8%	1.4%	3.7%
Netherlands	95.2%	0.0%	4.5%	Canada	87.0%	1.5%	3.6%
Denmark	89.6%	0.4%	4.7%	USA (SEER)	94.3%	1.1%	2.6%
Austria	90.4%	7.9%	1.7%				

Source: IARC²⁴³; MV: Microscopically verified, DCO: Death certificate only
 Note: Other & unspecified refers to cancers coded as C26, C39, C48, C76 & C80

The methodology used in the EUROCARE-IV study is similar to the survival methodology used throughout this report with the exception that the Hakulinen method as opposed to the Ederer II method was used to derive expected survival. The differences between five-year age standardised relative survival results using these two methods is minor, with typical differences of less than 0.2%.

Table A2.5: European countries covered by the EUROCARE-IV study and population coverage by participating cancer registries

Country	Population coverage	Country	Population coverage	Country	Population coverage
Northern Ireland	100%	Denmark	100%	Norway	100%
Republic of Ireland	100%	Finland	100%	Poland	9%
England	100%	France	17%	Portugal	43%
Wales	100%	Germany	1%	Slovenia	100%
Scotland	100%	Iceland	100%	Spain	16%
Austria	100%	Italy	28%	Sweden	100%
Belgium	58%	Malta	100%	Switzerland	17%
Czech Rep.	8%	Netherlands	34%		

Source: EUROCARE-IV²³⁸

A2.6: Statistical software

The SPSS statistical software package was used to develop the All-Ireland datasets and produce all incidence, mortality, treatment and prevalence data, including the logistic regression models for treatment application. The STATA software package was used to produce all survival data including modelling of excess mortality. Dr. Paul Dickman²⁴⁵ original developed the syntax used for survival analysis with some modifications made by Dr. Paul Walsh (National Cancer Registry, Ireland).

A2.7: Accuracy and rounding

The majority of values presented in this report are rounded to one decimal place with the exception of average numbers of cases/deaths that are rounded to the nearest whole number. All percentage values, calculations of differences, significance tests etc however are calculated using the maximum number of decimal places available rather than the rounded figures available in tables. Totals and percentages presented in the body of the text may thus occasionally differ from those calculated directly from values presented in tables.