

APPENDIX A. FORMULAE USED IN THE CALCULATIONS

This Appendix explains the calculation of the various indices used in this study.

Crude Death Rates

The crude death rate (CDR) is simply the total number of deaths in an area of interest (e.g. Louth or the entire country) in a specified time period (normally one year) divided by the total population at risk within that area. The resulting quotient is multiplied by a constant (100,000 in this study) to make the figures easier to comprehend (in much the same way as percentages are calculated by multiplying the quotient by 100). The formula is:

$$CDR = \frac{d}{n} \times 100,000$$

where d is the number of deaths in the area and n is the total population of the area.

Age Specific Death Rates

The main problem with crude death rates when used to compare the risk of mortality in different areas is that they can produce very misleading results if the areas have different age compositions. The risk of dying increases with advancing age, consequently if one area has a predominantly elderly population then the number of deaths, and hence the crude death rate, will tend to be higher than in another area with a younger population. However, it is possible that the risk of death amongst people of similar age might actually be higher in the second area (i.e. the one with the lower crude death rate). To establish whether this is the case, it is preferable to compare like with like. The simplest way to do this is to calculate age specific death rates (ASDRs).

To calculate ASDRs, the population in each area is subdivided into a number of age groups (e.g. 0-4 years, 5-9 years, etc.). Five or 10 year age bands are normally chosen. For each age group the ASDR is calculated as:

$$ASDR_i = \frac{d_i}{n_i} \times 100,000$$

where $ASDR_i$ is the age specific death rate for age group i (e.g. 55-64), d_i is the number of deaths in age group i in the area of interest and n_i is the population of age group i in the area.

The main problem with ASDRs is that you end up with the same number of indices for each area as there are age groups (maybe 10 or 20, depending upon the size of the age bands). If you are comparing more than two or three areas, there may consequently be a surfeit of information making it more difficult to 'see the wood for the trees'.

Standardised Mortality Ratios

Ideally what we want is a single index for each area which takes account of variations in age composition. This may be achieved by calculating standardised mortality ratios (SMRs). These facilitate the direct comparison of areas with different age compositions by calculating the number of deaths that would have occurred if each area had the same age composition as a specified standard population. When comparing counties within a country, the standard population is often defined as the population of the whole country. There are two methods of standardisation, referred to as the *direct* and the *indirect* methods.

In the direct approach, an age specific death rate is calculated for each age group in the area of interest. The resulting ASDRs are then each multiplied by the population in the corresponding age groups in the standard population to calculate the number of deaths which would have occurred in the standard population if it had the same death rate at each age as the area of interest. These estimates are summed to give the total deaths which would have occurred in the standard population. This figure is then divided by the actual number of deaths in the standard population to produce a ratio of deaths in the area of interest to the number in the standard population. The resulting quotient is generally multiplied by 100 to create a 'percentage'. Values greater than 100 indicate that the area of interest has more deaths than would be expected, values less than 100 indicate fewer deaths than would be expected. More formally, the formula is:

$$SMR (Direct) = \frac{\sum_i \left(N_i \frac{d_i}{n_i} \right)}{D} \times 100$$

where d_i is the number of deaths in age group i in the area of interest, n_i is the population of age group i in the area of interest, N_i is the population of age group i in the standard population, and D is the total number of deaths of all ages in the standard population. The sigma sign indicates that the term within the brackets should be calculated for each value of i (i.e. for every age group) and the answers should then be summed.

In the indirect approach, an age specific death rate is calculated for each age group in the standard population. These ASDRs are then multiplied by the populations in the corresponding age groups in the area of interest and summed to give the number of deaths which would be expected in the area of interest if it had the same age specific death rates as the standard population. The actual number of deaths in the area of interest is then divided by the expected number of deaths and multiplied by 100 to produce a 'percentage'. As with the direct method, SMR values greater than 100 indicate that the area of interest has more deaths than would be expected, whilst values less than 100 indicate fewer deaths than would be expected. The formula may be written as:

$$SMR (Indirect) = \frac{d}{\sum_i \left(n_i \frac{D_i}{N_i} \right)} \times 100$$

where d is the total number of deaths in the area of interest, n_i is the population in age group i in the area of interest, D_i is the number of deaths in age group i in the standard population and N_i is the population of age group i in the standard population.

The direct and indirect methods should give similar but not necessarily identical answers. This raises questions as to which method is best. If one does not have information on the number of deaths in each age group in each of the areas of interest, then one has no option but to use the indirect method. The indirect method may also be preferable if the numbers of deaths in each area are very small, as it will tend to produce more stable results. However, if the number of deaths in each area are reasonably large, and

information is available on the age of death, then the direct method is preferable. The direct method is used for the comparisons between Louth and the rest of the country in this study, but the indirect method was used to compare mortality in Drogheda, Dundalk and the rest of the county as no information was available on the age at death for the three areas.

If comparing counties, the standard population would normally be defined as the population of the whole country. One problem with using the national population as the standard when making comparisons over a 30 year period is that the composition of the national population changes slightly from year to year, resulting in a shifting 'standard'. However, using the direct method, it is possible to use any population you wish as the standard (with only minor adjustments to the formula above). This allows you to use a fixed population as the standard, thereby providing a firmer basis for comparisons over time. The European Standard Population, a static but totally fictitious population defined by the World Health Organization, was therefore used as the standard population in this study.²⁵

Finally, given that everyone must die eventually, some authorities argue that the inclusion of the elderly could distort the calculations. For example, if an area had a very low death rate at all ages, then a higher percentage of its population would survive to old age. This in turn could result in it having an above average age specific death rate in the oldest age group, which in turn could give the impression that the area has excess deaths overall, whereas in fact the opposite may be the case. It is by no means inevitable that its age specific death rate for the oldest age group will be above average, but given that distortions are possible it is recommended that the deaths amongst the elderly should be omitted from the analysis. The SMRs in this study are therefore calculated using only deaths under the age of 75.

Standardised Death Rates

Standardised Mortality Ratios provide a ready means for comparing the age standardised mortality in different areas at a given point in time. However, because the average of the SMRs for all the areas in a given year is always approximately 100, they do not provide a useful means for gauging whether the situation in an area is improving or disimproving in absolute terms over time. Age standardised rates, by expressing mortality as a rate (e.g. the number of deaths per 100,000 persons), provide a more suitable measure for gauging absolute changes over time provided one uses a fixed standard population (such as the WHO's European Standard Population).

The formula for the Directly Standardised Rate (DSR) is similar to that for a Directly Standardised Mortality Ratio, except that the standardised number of deaths in the numerator is divided by the total number at risk in the standard population. The formula is:

$$DSR = \frac{\sum_i \left(N_i \frac{d_i}{n_i} \right)}{N} \times 100000$$

where d_i is the number of deaths in age group i in the area of interest, n_i is the population of age group i in the area of interest, N_i is the population of age group i in the standard population, and N is the total number of people in the standard population. The quotient is multiplied by 100,000 to express the rate as deaths per 100,000, but any figure could be used.

²⁵ Details of the WHO European Standard Population are provided by Waterhouse, J., Correa, P. Muir, C. and Powell, J. eds. (1976) *Cancer Incidence in Five Continents*, Volume III. IARC, Lyon. (Cited in Information Management Unit, Department of Health and Children (2003) *Public Health Information System*, Version 6, Appendix B.)

As with SMRs, an argument can be made in favour of omitting deaths amongst the elderly. Deaths of those aged 75 or more are therefore omitted in the rates reported in this study.

Morbidity Rates

Rates similar to each of the above may be calculated using morbidity data – i.e. information on the number of people diagnosed as having cancer, irrespective of whether it resulted in death. A distinction should be made between **incidence** rates (which measure the number of new cases in a given time period, usually 12 months) and **prevalence** rates (which measure the number of people as having the disease). Prevalence rates may refer either to a specific point in time (**point prevalence**) or to all people who were ill during a particular time period (**period prevalence**). Period prevalence rates are generally higher than incidence rates for the same period because in addition to the new cases in the time period (i.e. incidence) they also include cases first diagnosed in a previous time period but still receiving treatment. It is therefore important to know whether rates refer to incidence or prevalence, and also the time period covered. This study uses only incidence rates.

Simple Regression

Simple regression models are used to identify the linear trend in a time series. The formula for a regression line is:

$$y = a + bx$$

where y is the dependent variable (in our case a mortality rate of some sort), x is the independent variable (in our case the year), and a and b are two parameters, referred to as the intercept and the gradient, which need to be calculated.

The intercept is the value y would have when x is zero (in our case the mortality rate in the year 0 AD). It may be estimated using the formula:

$$a = \bar{y} - b\bar{x}$$

where \bar{x} and \bar{y} are the mean (i.e. average) of all the x values and all the y values respectively and b is the gradient (calculated using the formula below).

The gradient (or slope) is a measure of how ‘steep’ the line is. Large gradients (representing a steep line) indicate that the mortality rate changes by a large amount each year. The gradient may be either positive (indicating the mortality rate tends to increase each year) or negative (indicating a decrease). The gradient is given by the formula:

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

where n is the number of observations (in our case years). The summations, indicated by the capital sigma signs, are for all n cases.

The intercept is only used in this study to help determine the location of the regression line in the figures. The gradient is also required for these calculations, but it is also used in the text as a measure of the average

change. To make the numbers easier to interpret, the gradient is typically multiplied by 10 to calculate the average change over a decade.

Correlation Coefficients

Correlation coefficients measure the extent to which the values of two variables (e.g. cancer rates and urbanisation) increase or decrease together. When the values of the variables refer to areas (e.g. counties), the correlation coefficient can be regarded as a measure of the extent to which the maps of the two variables are similar.

The formula for a correlation coefficient is:

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}}$$

where r is the correlation coefficient, n is the number of observations (i.e. areas) and x and y are the values of the two variables.

The value of the correlation coefficient always lies within the range -1.0 and $+1.0$. Positive correlation coefficients indicate that the values of the variables tend to increase together (i.e. the higher values of both variables tend to be found in the same areas). Negative correlation coefficients indicate that high values of one variable tend to be associated with low values of the other (i.e. the maps will be mirror images of each other).

The numeric part of the coefficient measures the strength of the association. Values close to $+1.0$ or -1.0 indicate a very close association, whereas values close to zero indicate little or no association. When the coefficient is so large that it is unlikely to have occurred by chance unless the two variables are related in some way, the correlation is said to be significant.

APPENDIX B. AN ASSESSMENT OF THE IMPLICATIONS OF GEOCODING UNCERTAINTY

The address information provided to the National Cancer Registry is used to assign a code identifying the DED where the patient normally resides. This enables counts to be made of the number of cases in each area, thereby making it possible to map the data. In some instances it is not possible to geocode an address, in which case the record cannot be used in spatial analysis. Approximately 2.4 per cent of the patients for 1994-1997 were not geocoded. In other instances, the address may be geocoded but a degree of uncertainty may exist because the address information is ambiguous or incomplete. These records are identified in the registry. A total of 66,455 patients registered between 1994 and 1997 were geocoded, but only 61,436 were geocoded with confidence. This smaller set was used in Chapter 4 in the analysis of variations within County Louth, but the larger set was used in the comparisons between counties. This appendix examines the implications of using the smaller set instead of the larger set for inter-county comparisons.

Table 19 shows the percentage of patients who were geocoded with confidence in each county. The percentages are broadly similar, although those for some of the cities tend to be noticeably lower. There is also possibly a tendency for some of the more rural counties to have higher percentages.

County	Males	Females	County	Males	Females
Carlow	89.0	90.6	Kerry	97.4	97.8
Dublin C.B.	90.7	89.2	Limerick C.B.	84.5	76.9
Dublin Co.	92.3	92.9	Limerick Co	97.4	97.9
Kildare	94.7	93.2	Tipperary N.R.	95.3	97.2
Kilkenny	93.3	93.3	Tipperary S.R	87.5	85.6
Laois	91.4	91.8	Waterford C.B.	73.9	76.5
Longford	95.0	94.3	Waterford Co.	98.1	96.4
Louth	94.6	94.8	Galway	97.1	96.7
Meath	92.5	91.4	Leitrim	98.3	95.5
Offaly	95.4	95.7	Mayo	91.0	90.0
Westmeath	88.6	90.6	Roscommon	96.7	97.0
Wicklow	93.2	91.5	Sligo	95.8	92.4
Wexford	88.6	88.1	Cavan	98.1	97.7
Clare	94.5	93.5	Donegal	97.0	95.4
Cork CB	84.5	85.0	Monaghan	91.1	89.3
Cork Co.	94.4	93.8	National	92.8	92.0

Table 19. Percentage Of Patients Geocoded With Confidence In each County

Table 20 shows the mean crude incidence rate for each county 1994 to 1997 calculated using the smaller data set. This table is directly comparable with Table 8 which was calculated using the larger data set. The estimated rates in Table 20 are obviously smaller. Given that the cities have the lowest percentages of cases confidently coded, their rates tend to decline more than the more rural counties. This in turn results in some changes to the rank order of the counties. Despite being one of the more urban counties, Louth has an above average percentage of cases confidently coded and therefore it tends to move up the rank order slightly (from 17th to 14th for males, and from 12th to 11th for females). However, the main conclusions arising from Table 8 remain unchanged. The crude incidence rate in Louth is still close to the national average for males, and slightly above it for females.

Using the smaller data set alters the results for the age standardised and age specific rates in a similar manner, but it does not significantly alter the main conclusions.

	Males		Females	
	Rate	Rank	Rate	Rank
Carlow	394	26	342	24
Dublin C.B.	556	5	524	1
Dublin Belgard	267	34	286	33
Dublin Fingal	292	32	290	32
Dun Laoghaire / Rathdown	509	11	481	5
Kildare	322	30	315	28
Kilkenny	394	25	328	27
Laois	428	20	381	19
Longford	493	12	447	6
Louth	456	14	428	11
Meath	370	27	343	23
Offaly	418	22	381	18
Westmeath	451	15	411	13
Wexford	416	23	344	22
Wicklow	432	18	360	20
Clare	414	24	313	30
Cork C.B.	479	13	428	10
Cork Co.	451	16	398	16
Kerry	574	3	489	3
Limerick C.B.	278	33	244	34
Limerick Co.	368	28	332	26
Tipperary N.R.	447	17	389	17
Tipperary S.R.	423	21	347	21
Waterford C.B.	323	29	341	25
Waterford Co.	523	10	444	9
Galway C.B.	300	31	294	31
Galway Co.	553	7	420	12
Leitrim	650	1	485	4
Mayo	525	9	404	15
Roscommon	604	2	446	7
Sligo	553	6	445	8
Cavan	573	4	491	2
Donegal	526	8	409	14
Monaghan	429	19	315	29
National	449		398	

Table 20. Mean Crude Incidence Rate By County Using Only Confidently Coded Records