Using the Longitudinal Study cancer data for research

Angela Donkin and Lin Hattersley

Contents

1. Introduction

This paper in the ONS Longitudinal Study (LS) User Guides series aims to provide potential users of the cancer data within the LS with information on the reliability of the data and other issues that affect correct use of these data. In addition, it provides a brief overview of the types of analysis that can be done and cancer research that has previously been conducted using the LS.

Research into cancer has considerable public health significance; more than one in three people in England and Wales will develop a cancer at some time in their life, and cancer causes one in four deaths. The government has set a target (for England) of reducing the cancer death rate in people aged under 75 by 20% over the years to 2010 (Department of Health, 1999). In addition, a National Cancer Director, charged with overseeing the achievement of this target, has been appointed and a National Cancer Plan has been published (Department of Health, 2000).

Over the next ten years, trends in cancer incidence, survival and mortality will be closely monitored at the national, regional and local levels. However, because the clinical presentation of cancer may be many years after an original exposure or follow an accumulation of factors over time, cancer incidence and mortality generally exhibit only slowly increasing or decreasing trends. For example, lung cancer rates depend on smoking habits 20 or more years previously. So, with the exception of cancers for which there have been major public health interventions, such as screening for breast cancer, past trends are usually a good guide to the future (Quinn et al. 2001). With over 30 years of data, the LS is a valuable tool for observing differences in cancer trends between sub-groups of the population.

The LS is a representative 1% sample of the population of England and Wales. The initial sample was drawn from the 1971 census on the basis of four birth dates, and subsequently information on cancer registration, mortality, births to sample members, migration and further census information has been linked to these members and any new members who join the study through immigration or birth. The LS holds data on the date of each cancer diagnosis, the type of cancer registered, and whether it is a multiple cancer. More information has been available since 1993, for example treatment type, and grade and stage of cancers of the breast and cervix.

The LS an excellent data source for some types of population based research into cancer. It has both advantages and disadvantages. One of the advantages of using the LS for research into cancer incidence is that individuals are traced over time, and hence the LS has numerator and denominator information at any one point in time, or over any period of time. In addition, there is the capacity to link cancer registration data to a wide range of descriptive data. These data may be from: each of the censuses since 1971; other event files such as births data; or from files on small area statistics. In addition, the LS holds census data relating to other members of the individual's household. Another advantage is the fact that a researcher using the LS can link cancer registration to the (eventual) underlying cause of death from 1971. This facility has only been available on National Cancer Files for the 1990s.

The major limitation is that the LS is a 1% sample, and hence the numbers of cancer registrations, particularly for less common cancers, may be too small for meaningful analysis. Researchers using the LS often group five years of data together in order to look at cancer incidence, and indeed mortality.

2. The cancer registration system in England and Wales

Cancer registration is the process of maintaining a systematic collection of data on the occurrence and characteristics of malignant neoplasms and certain non-malignant tumours. The procedure is widely established and follows guidelines established by bodies such as the International Union Against Cancer (UICC), the International Agency for Research on Cancer (IARC), the International Association of Cancer Registries (IACR) and the World Health Organisation (WHO).

More information on the background of the National Cancer Registration System can be found in Appendix G of *Cancer Trends in England and Wales 1950-1999* (Quinn et al 2001).

There have been three main problems with the cancer registration process from a national perspective. First, the timeliness of national data based on the full set of individual records depends on the speed of the slowest regional registry in completing its submissions to the Office for National Statistics (ONS). Second, the database is 'live' or 'dynamic' in the sense that records my be modified or deleted if new information is obtained. Finally, cancer registration is not statutory, and ONS has no organisational, managerial or financial control over regional registries.

In April 1999, the Advisory Committee on Cancer Registration, on behalf of the Department of Health, commissioned Professor Charles Gillis, Director of the West of Scotland Cancer Surveillance Unit, to undertake a further review of cancer registration in England. (Gillis, 1999) The review made a number of recommendations for how cancer registries should be strengthened, so that they would be able to contribute fully to the cancer modernisation agenda by providing robust data to support the planning and monitoring of cancer service delivery and identifying the scope for NHS intervention in relation to deprivation and cancer.(Department of Health, 2000) The Department of Health are leading the implementation of the recommendations of the Gillis review. In parallel, there has been considerable discussion about the implications of section 60 of the recent Health & Social Care Act governing which data about NHS patients may be collected and used without informed consent. (HMS0, 2001) Cancer registration data have historically been collated by the regional cancer registries without informed consent.

Care is required in the interpretation of cancer registration statistics, particularly when addressing either trends over time or differences between regions or countries. As the registration of cases of cancer is a dynamic process the figures for registrations published by a regional cancer registry may be different from those published by ONS which will generally have been produced at a different (usually later) time. This dynamic process also has an impact on the LS figures as described in more detail in this User Guide.

For the purposes of the national cancer registration scheme the term 'cancer' includes all malignant neoplasms and the reticuloses, that is conditions listed under codes C00 to C97 inclusive of the Tenth Revision to the International Statistical Classification of Diseases and Related Health Problems (ICD10). In addition, all *in situ* carcinomas and neoplasms of uncertain behaviour are registered. Benign neoplasms and neoplasms of unspecified nature of bladder and brain, including the pineal and pituitary glands, are also registered, together with hydatidiform mole.

On one specific cancer, it is well known that non-melanoma skin cancer (ICD10 C44) is greatly under-registered. This is an international problem.

It should be noted that some cancer registries are not able to collect all information about benign, uncertain and unspecified neoplasms and therefore these registration rates are believed to be underestimates of the true incidence in those regions. This is important to note when interpreting regional differences.

Further issues to be considered can be found later in this User Guide, and in Appendix H of *Cancer Trends in England and Wales 1950-1999* (Quinn et al 2001). There are issues surrounding the following topics:

Geographic coverage:
For instance, the boundaries of the cancer registries have changed over time.

Methods of data collection:
These differ considerably between regional registries and over time. Any large increases year to year in the number of registrations from a registry are likely to be a result of a concerted effort to increase the number of cases captured rather than real trends. However, not all registries carry out these procedures at the same time.

Completeness:
When the cancer registries data are submitted to ONS, a large number of cross-checks and quality checks are performed. If a record fails any of the vital checks then it is given a quality status of '3'. The national core contract for cancer registries requires that when a registry's data for a particular year are complete, no more than 0.5% should have a quality status '3' (NHS Executive 1996).

Accuracy of data:
Various indirect measures, such as mortality to incidence ratios, suggest that there is considerable variation between regions. For example, variations among the registries have been found in diagnostic factors, incidence date, stage of disease, treatment information, and the use of death information. (Huggett 1995)

Late registrations:
The point in time at which ONS, in consultation with regional registries, decides to produce tables for reference volumes is a compromise between two principal considerations – the need to minimise delay between the relevant data year and the publication of the detailed results, and the requirement to obtain a very high level of completeness of the data and hence minimise the number of late registrations. Likewise in the LS, late registrations are added after the main stream of data for that year has been added. The number of late registrations will vary according to the speed at which registries send back data and thus researchers should make sure, especially when researching recent years of data, to enquire about any late registration issues which might influence their results.

Duplicate registrations:
These can artificially inflate the figures. Duplication may arise for instance if a patient is resident in one area but treated in another although strenuous efforts are made at ONS to eliminate such duplicates.

Changes in coding systems:
Changes in coding systems may cause discontinuities in data. For national data held by ONS for incidence years 1971 – 1978, site is coded to ICD8. For incidence years 1979-1994 site is coded to ICD9 and from 1995 onwards site is coded to ICD10.

Completeness of flagging of registrations on the National Health Service Central Register (NHSCR):
The proportion of cancer registrations received by ONS which were successfully linked to an NHSCR record was on average 96% between 1971 and 1989, and 99% subsequently.

3. Content of the LS cancer data

Full descriptions of the LS cancer data together with variable names, lengths and coding frames are available in the LS data dictionary. The variables available in the LS cancer file are taken from the England and Wales cancer registration files. Any changes to the source file are reflected in the LS file and additions of new variables occur as necessary. For example, from 1995 cancer registration data were classified according to ICD10.

At the time of writing (June 2001) the LS cancer file held the following data for LS members (full details are given in the LS data dictionary):

a) Data about the person

- Date of birth
- Sex
- Ethnic origin - first introduced in 1993 but remains optional and is poorly completed.
- Country of birth
- Place of residence - NHS and local government administrative areas.
- Occupation – LS member's or that of parent if the LS member is a minor (coded to the 1970, 1980 and 1990 occupational classifications as appropriate to date of cancer registration).
- Employment status – LS member's or that of parent if the LS member is a minor (coded to the 1970, 1980 and 1990 classifications).
- Industry – LS member's or that of parent if the LS member is a minor (coded to the 1970, 1980 and 1990 industry classifications).
- Retirement Indicator – if the LS member is retired
- Social Class – LS member's or that of parent or spouse if the LS member is a minor or a married woman who has never worked.
- Socio-Economic Group (SEG) – LS member's or that of that of a parent if the LS member is a minor.
- Dates of exit and re-entry to the NHS (this occurs on emigration and subsequent re-entry to England & Wales – also on joining and leaving the Armed Forces)

b) Data about tumour & treatment

- Registration details (unique identity number – tumour based)
- Registration at screening (1993+)
- Number of registrations
- Age at diagnosis
- Diagnosis date
- Dead on registration
- Death certificate only (1993+)
- Date of death (if appropriate)
- Basis of diagnosis
- Behaviour of cancer
- Duration of survival
- Grade of cancer (1993+ breast & cervix only)
- Stage of cancer (1993+ breast & cervix only)
- Site of growth – ICD codes
- Type of growth – ICD codes
- Multiple tumour indicator
- Treatment type – chemotherapy, hormonal, radiotherapy, surgery, other indicators (1993+)

4. Linkage and validation

This section includes a short overview of the history of LS cancer data linkage methods. This has been included primarily to explain why differences between the national cancer data incidence rates and LS incidence rates may occur.

The linkage of event data to LS members is performed by the LS Unit at the National Health Service Central Register (NHSCR) in Southport. Methods of collection, linkage and addition of cancer data to the LS changed during the 1990s when NHSCR was computerised. Prior to this date, the procedure was:

- The OPCS (now ONS) cancer section in Titchfield would notify the cancer section at NHSCR in Southport of new cancers, which would then be flagged in the alphabetical index books containing the NHSCR records.
- If an LS flag was present for the entry, the NHSCR cancer section would send a copy of the cancer abstract card to the LS unit in Southport.
- The LS unit in Southport would then update the LS index cards and note the LS number on the cancer abstract.
- These abstract cards would then be sent to the LS Unit in Titchfield who would run an extract of all cancers occurring to LS members from the national file.
- A listing from the extract and the cards from Southport would then be manually matched.
- Any entries on the listing that had no matching cancer record from Southport would be sent back for query to NHSCR.

Once the queries had been resolved the data would be entered in the LS database once a year. It is probable that some cancers were missed due to the manual methods of notification.

The computerised NHSCR system, known as the Central Health Register Inquiry System (CHRIS) was formed from 1991 Family Health Service Authority records and included only live patients who were then in the country. LS member records were added for those members who had been present in the LS from 1971 onwards. Flags were transferred from the manual records to the Central Health Register Inquiry System (CHRIS) as necessary. Once CHRIS had been populated, tapes containing cancer registration identity numbers, rather than the previous paper records, were sent from Southport to Titchfield biannually. As long as CHRIS was flagged correctly, new cancer registrations for LS members together with the relevant LS numbers should have been included in the events tape sent to Titchfield. The cancer registration identity numbers for LS members are stored in what is known as the LS cancer 'progress' file which is then used to pull off the detailed cancer data from the national cancer system database in readiness for loading it into the LS database.

It should be noted that the incidence or diagnosis date for a cancer may be many years earlier than the registration date. Therefore, until the detailed cancer data, containing the diagnosis date, are linked to the LS members registration details there is no way to identify which diagnosis years those data involve.

After the 1991 Census all the national vital statistics systems at ONS (then OPCS) were redeveloped. The redevelopment of the LS data capture systems was completed in 1996. Because of the delays associated with redevelopment, the last LS cancer file available for analysis until 1997 was that for 1989. No cancer processing was done between January 1995 and April 1997 when the 1990 cancer data were added to the LS database. All late registrations available for cancers up to 1989 had been included prior to the beginning of redevelopment in January 1995. After that, late registrations were stockpiled with the intention of updating the files after redevelopment was complete. This has now been done.

During the 1980s late registrations of cancer were common and were added to the national cancer files as soon as they became available (see *Cancer statistics, registrations 1995-1997*, Series MB1 no. 28, page 84, Figure 1A). At this point the national cancer data were held in a flat file system but a new dynamic database was developed in the 1990s and went live in 1996. The cancer database is constantly updated with late registrations as soon as they become available, and duplicates and other records are deleted. Consequently the cancer data for the LS do not remain static. Although there are some late registrations for births and deaths, these are few due to the legal requirement to register these events within a short time. This is not the case for cancer where there has been no legal requirement for registration.

When the cancer registration identity numbers received by tape from Southport and loaded into the progress file are run against the national cancer data this results in the linkage of what are referred to as 'accepted' records to be run against the LS database. 'Accepted' records are genuine records on the Vital Statistics Cancer Database, however they still need to be put through a validation run for entry onto the LS. 'Accepted' records also include duplicates and records which will fail LS database consistency checks and be rejected on loading. Not all

consistency check failures are rejected: records with date of birth discrepancies and sex discrepancies are loaded but the database holds variables which will allow their identification. Other consistency check failures such as 'date of event before date of entry' or 'date of event after date of death' are investigated further and may eventually be loaded onto the database. The records that fail validation are queried with Southport, and sometimes with the Cancer section at Titchfield. Records which fail consistency checks can still be added to the database if after further investigation it is agreed that the event and LS member have been correctly matched.

5. Data quality

The basic measurement of quality for LS Cancer data is done by examining linkage rates, sampling fractions and incidence rates. Ideally, if the LS were an unbiased representative 1 per cent sample of the population of England and Wales, then it would be expected that 1 per cent of the occurrences of an event to the national population would occur to LS members. However, this is dependent on a number of factors including the accuracy of the linkage method, changes in denominators and sensitivity to small numbers of events.

The linkage rate is calculated as follows:

$$\frac{\text{the actual number of cancers occurring to LS members in a year}}{\text{the expected number of cancers occurring to LS members in a year}} \times 100$$

Given that four birth dates in any year are used to select the LS sample, the calculation of the expected number of cancers in the LS is based on the expected sampling fraction multiplied by the number of cancers recorded as occurring in England and Wales for the year in question. This calculation is:

$$\left(\frac{4}{365.25} \times 100\right) \times \text{the number of cancers occurring in a year in England \& Wales}$$

The numbers of cancers in the national system increased in the 1980s and 1990s compared with the published national results due to late registrations being received and entered onto the database. The LS has also been updated with late registrations and any duplicate records have been deleted. As a result both the denominators and the numerators for calculating linkage changed. These changes affected the years 1981 to 1992 and brought the linkage rates down in most cases but there is still a degree of variation. Because the LS is a sample, some variation in linkage rates will be due to small sample numbers. Other sources of variation are duplicate records and late registrations.

Although some of the difference between observed and expected numbers of cancers in a year falls within normal sampling variation, some of it is due to differences between the three computer systems (Cancer system, LS system and CHRIS system) together with a small proportion due to human error. Any tracing done prior to 1991 for LS cancer data at Southport relied on the experience of both the NHSCR cancer staff and the LS Unit staff operating a complex manual system which allowed more opportunity for error than the current system.

Validation is done against the latest set of national cancer figures . As the cancer registration system is dynamic (see above) every time the LS cancer data are updated, validation of the new years data and the re-validation of previous years are required. Only when this is complete are the data ready for release to users.

The scope of quality checks on LS cancer data has been widened to include information on certain specific cancers as well as all cancers. A list of the validation tables available to users of the LS cancer data is shown below. These tables can be found in *Annex C* .

- Table C1 - *All cancer registrations 1971 – 1994* by year of diagnosis
- Table C2 - *All cancer registrations 1981 – 1994* by year of diagnosis, age group and sex
- Table C3 - *All malignant cancer registrations 1981 – 1994* (excluding ICD9 173 – non-malignant skin cancer) by year of diagnosis, age and sex
- Table C4 - *Specific cancers registered 1981 – 1994* by year of diagnosis and sex. These tables cover the following major cancers: stomach, colorectal, lung, breast and prostate.

Tables C1, C2 and C3 include linkage rates and sampling fractions; Table C3 also includes incidence rates.

95 per cent confidence intervals (based on the proportion expected given the number in the population) are shown for the expected numbers of cancers occurring to the LS sample. In those years where the observed number of cancers is outside the confidence interval, the rows in the tables are marked in grey. Where the tables are based on occurrences of 'all cancers' (benign as well as malignant) LS cancer registrations for later years (from the mid 1980s on) show a tendency to be over-sampled. The sampling fraction should ideally be 1.09% but frequently the number of cancers identified and linked to LS members may be above the number expected (over-sampling) or below it (under-sampling).

Among malignant cancers – ICD9 140-208 excluding 173 (Table C3) over-sampling has been more common for malignancies occurring to males than those occurring to females. Tables C4a-e show a number of major

malignancies which are of particular interest to researchers. These are stomach cancer, colorectal cancer, lung cancer, female breast cancer and prostate cancer. Over-sampling has been found for occurrences of colorectal cancer among males in 1983 and females in 1989; for lung cancer among males in 1989 and 1990 and females in 1991 and 1992; and for female breast cancer in 1994. Under-sampling is less common, but has been found for occurrences of female breast cancer in 1985 and prostate cancer in 1986 and 1987.

The inclusion of more late registrations has reduced under-sampling. However, for some specific cancers, such as breast and prostate cancer, although there has been a general improvement there are still appreciable differences in incidence rates between England and Wales and the LS for certain years.

Over-sampling as well as under-sampling has occurred and this suggests that either duplicate records are still being received for some years, or that incorrect notifications of cancers have been added to the database. Notifications that a deletion is required often occur long after the cancer was put on the database.

The differences that have been found between the LS and the England and Wales cancer data do not appear to be systematic. However, it must be noted that because the national system is dynamic the two sets of cancer data will never be completely in line. Cancer data in the 1980s were particularly prone to late registrations and during the period that these registrations were received major changes have occurred to the methods of processing both in Southport and in Titchfield. This has had an effect on the quality of LS cancer data.

Prior to this exercise, the source of national cancer data used by the LS for validation was the cancer annual reference volumes (Series MB1). The volume first used for validating 1992 cancers was published in 1998, but in the same year an updated series of national cancer figures for 1971 – 1992 was published on CD. Each set of figures is superseded annually when the latest year's national figures are published. Each replacement of national figures made changes to the LS sampling fractions, linkage and incidence rates, and in some cases these changes were substantial. Validation has now been done utilising the latest available national figures for 1982 to 1994 but the current LS quality figures will only remain in effect until the national cancer figures are updated again. However, it is not expected that there will be such major changes in the historic national data as seen previously and the annual reference volumes are expected to be a more reliable data source from 1990 onwards.

6. The effect on cancer incidence and mortality of changing from ICD9 to ICD10

Cancer registration data from 1971 to 1978 inclusive were coded using the Eighth Revision of the International Classification of Diseases (ICD8); data from 1979 to 1994 inclusive using the Ninth Revision (ICD9); and from 1995 onwards these data were coded to the Tenth Revision (ICD10). The coding of deaths changed from ICD9 to ICD10 in 2001.

There are marked differences between ICD9 and ICD10 coding frames overall (see Rooney & Smith, 2000). The new classification has 21 chapters compared with 17, and although the full ICD codes are still 4 characters long they are now alpha-numeric rather than numeric. There are some new cause codes and some existing cause codes have either been transferred into new chapters or from one existing chapter to another. The codes for cancer have remained relatively unchanged, although some cancers have been moved from their current ICD9 groupings into other cancer groupings.

The ICD9 site codes for malignant neoplasms were in the series 140–208; this changed in ICD10 to C00–C97. Changes have been made that affect the coding of cause when using both ICD9 and ICD10 in the same analysis. This will affect any incidence data occurring from January 1995 onwards and cancer deaths occurring from January 2001 onwards.

These changes are as follows with details of code changes being given in Table 1 below.

- Under ICD9 colorectal cancer included cancer of the anus. This is now has in a category on its own under ICD10.
- Under ICD9 cancer of the trachea was included with carcinoma of the bronchus and lung. It is now separate.
- Codes for carcinoma of the breast differentiated between male and female breast under ICD9. This is no longer the case under ICD10.
- Under ICD9 ovarian cancer was included as part of malignant neoplasms of the ovary and uterine adnexa. Under ICD10 ovarian cancer now has its own code.
- A new code for cancers with independent primary multiple sites has been introduced under ICD10.

Table 1: Changes in coding of certain malignant cancers between ICD9 and ICD10

Site	***ICD9 code***	***ICD10 code***
All	140 – 208 excluding 173	C00 – C97 excluding C44
Colo-rectal Of which anus	153 – 154 154.2, 154.3, 154.8	C18, C19, C20 C21
Trachea, bronchus & lung Of which trachea	162.0 – 162.9 162.0	C34 C33
Breast (female) Breast (male)	174 175	C50 C50
Ovary and uterine adnexa Of which ovary	183.0 – 183.9 183.0	C57 (excludes ovary) C56
Independent primary multiple	N/A	C97

These changes mean that care will have to be taken when 1995 cancer incidence data come on stream in the LS, and when 2001 mortality data are added to the LS in late 2002. 2001 deaths will be coded to both ICD9 and ICD10. However, any researcher considering using cause of death data in survival analysis from 2001 onwards must be aware that it is not only the classification of cause that has changed. The instructions (in particular Rule 3) governing the certification of death and the identification of the main underlying cause have also changed (Rooney & Smith 2000).

7. Research into cancer incidence

Researchers may wish to use the cancer registration information held in the LS for a number of reasons, for example to further understand national trends by making use of the descriptive data that the LS holds on individuals. Alternatively they may wish to link individual data to data about areas in which the study members have lived, or about their households. The following section provides an overview of methods commonly used to report on cancer incidence, issues which should be considered when conducting this type of analysis, and an overview of past cancer research using the LS.

7.1 Calculation of incidence rates, and direct and indirect standardisation

Crude rate

The crude rate per 100,000 person-years at risk for a cancer is calculated as the total number of cases registered per time period as a proportion of the total available (at risk) person years in that time period, multiplied by 100,000.

Cancer rates vary greatly with age and the crude rate is heavily influenced by the demographic structure of the population. Hence, if the population structure changes over time the crude rates over that period will not be directly comparable. Furthermore, it is not appropriate to compare crude rates across geographical areas (e.g. between cancer registries) where the age structure of the population differs. Therefore, in order to assess time trends in registration data or to compare incidence across geographical areas or between registries, it is desirable to standardise the rates with respect to age.

There are two main approaches to standardisation, the direct and indirect methods (Boyle & Parkin 1991).

Direct age standardisation

Direct age standardisation uses the age structure of a real (e.g. England and Wales) or artificial (e.g. European Standard Population) 'standard' population. The age-specific rates in the study population are calculated and applied to the standard population in order to calculate the number of events or cases which would have occurred in the standard population, if the observed rates in the study population had occurred in the standard population. These expected events (or cases) are then summed and divided by the total of the standard population to obtain the directly age standardised rate. Thus direct age standardisation can be described as calculating a weighted mean of the age-specific rates in the study population, using as weights the distribution of the standard population.

Five-year age groups (0-4, 5-9, ... 85+) are often used, but ten-year age groups or other variations can also be used depending on data availability and distribution. Age standardisation can be performed over a limited age range - for example, 15-64 years - and age groups can be truncated at either end of the age spectrum. However, the larger the age brackets employed, the less precise is the age adjustment.

The choice of the standard population is usually determined by the most important comparisons which are to be made, keeping in mind that the 'best' standard from the point of view of precision is that which is closest to all the populations being compared - and indeed may be derived from them (e.g. their sum or average). National figures may be adjusted to the world standard population to facilitate international comparisons, although rates from industrialised countries so adjusted will often appear quite low compared to their crude rate because the world population standard incorporates age structures of developing countries which have a much 'younger' age distribution. The European standard population is often used to report on UK data and this has a comparable age structure to that of the UK. In the analysis of trends over time, the most recent census population age structure is often used as the standard.

For the purposes of presentation it is often desirable to calculate rate ratios such that a rate is presented in relation to an average overall rate for England and Wales, or in relation to 'desirable' rates from populations which may have low incidence. If the age-standardised rate for each study population is divided by the crude rate of the standard population a directly age standardised incidence ratio is produced. This is often multiplied by 100 to dispense with decimals and for presentation as a commonly understood percentage. Calculation of rate ratios requires additional information on the crude rate of incidence in the standard population, not merely the age distribution.

The main advantage of direct standardisation is precision in the adjustment of the effects of age, and, since various reports from the UK use the European Standard Population for instance, comparable rates can be produced. A

further advantage is that legitimate comparisons can be made between all population groups which have had their rates directly age standardised to the same standard.

The disadvantage of direct standardisation is that instability in age specific rates from the study population could occur if these are based on small number of events or cases and/or small populations. Although this can be allowed for by the calculation of confidence intervals, there is no simple method for calculating confidence intervals of directly standardised rates based on small numbers; it has been suggested that one should use the Poisson method for fewer than 30 cases (Morell 1998).

Indirect age standardisation

Indirect age standardisation uses a series of age-specific incidence rates as the standard (i.e. age-specific event rates in a standard population). These standard rates are applied to the age-specific denominator populations in the study population to produce the number of cases or events which would be expected in the study population had the rates in the standard population prevailed. The observed cases in the study population are divided by this expected number, to produce the indirect age standardised incidence ratio, often called the standardised incidence ratio (SIR). This is usually multiplied by 100 to produce a percentage.

For indirect standardisation, the standard rate is normally the overall rate of the population from which the smaller units are derived. For example, the male SIR for various occupations would be based on the overall incidence rate for all males. When an overall rate is used as a standard it is essential that the study populations are a small part of the total population such that they can be assumed to be independent. For example, it would be problematic to compare one group with a standard if that group made up 30% of the standard population. The indirect age standardised ratio for a particular group can be converted to a rate by multiplication with the crude rate of the standard population.

The main advantage of the indirect method is that it is not necessary to know the age distribution of (observed) cases in the study population. Other advantages of the indirect method include the ease of calculation of confidence intervals using the Poisson method, and the production of a ratio statistic in a one step procedure.

A disadvantage of the indirect method is that it is generally considered to be less precise in adjusting for age than the direct method, particularly when the age structure of the study population is radically different from that of the standard (for example the armed forces). In this case it will often be found that the directly and indirectly standardised mortality or incidence ratios are quite different, with the direct method being more accurate provided the number of events in the study population are sufficient. Another disadvantage is that it is usually considered that SIRs from study populations can only be legitimately compared with the standard (1.00 or 100) and not with each other. Armitage and Berry (1994), however, provide a method of calculation of confidence intervals which would permit such inter-group comparisons given knowledge of age-specific events in the study population.

Use of the indirect method is cumbersome when a series of comparisons of various cancer sites are required, since a standard set of incidence rates must be derived for each disease.

Previous research on cancer incidence using the LS has been conducted using both directly (e.g. Brown et al. 1997) and indirectly (e.g. Leon, 1988) standardised incidence ratios. If data are available, both direct and indirect methods can be used on the same material and the results compared. Both methods should provide similar rates or ratios; if there are differences, then the strengths and weaknesses of both methods should be assessed in relation to the data to determine which approach is likely to be most appropriate in the circumstances.

It should also be noted that standardised ratios (both direct and indirect) should not be used in calculations or displayed in diagrams without appropriate logarithmic transformation. That is, if the baseline value is taken as 100, a doubling of a SIR to 200 is equivalent to a halving of the SIR to 50 in the other direction.

Notes on calculating directly and indirectly standardised rates and confidence intervals are found in *Annex A* below. These can be applied to both cancer incidence and mortality data (substituting incidence figures with mortality figures). The notes in this section are taken almost exclusively from Morell (1998*) Quantitative Methods in Demography*. Those requiring further statistical advice are also encouraged to read Boyle and Parkin (1991).

7.2 General issues affecting research into cancer incidence

a) Issues specific to the LS

Small sample size

As the LS is an approximately 1% sample, the number of registrations for specific less common cancers may be too small to conduct meaningful analysis. Care must be taken to ensure that the number of registrations is sufficient to answer the question posed. While figures from the LS can be expected to match those in the national population within certain confidence intervals, the size of the confidence interval will depend upon the type of cancer under study and may be very large for rarer cancers.

Individuals with more than one cancer

Due to the longitudinal nature of the data, the LS provides information on whether a cancer was the first or subsequent registration for that individual. The new person based cancer registration database also does this, but in the published figures for cancer registrations each primary cancer is counted separately. Much research on cancer using the LS has the aim of identifying and/or describing causal pathways and groups of the population who are more at risk, and when conducting this sort of research it is typical to look only at first primary cancers. This is because those who have had a first primary cancer have an increased chance of developing a second primary cancer. One of the reasons why rates calculated using the LS might appear smaller than those expected from a 1% sample of the population is that researchers are typically looking to explain variations in incidence and would use only first primary cancers. Depending on the cancer, between 4-7% of all cases registered can be expected *not* be first primary cancers. In addition, the number of subsequent cancers is increasing as survival for first cancers improves and with increases in longevity; this may also have a slight influence on the rates of change in incidence between time periods.

Cohort versus period analysis

Much research on cancer involves following a group of people to determine which type of people develop cancer and which do not. For instance, we might look at those who are aged 40-60 at the 1981 census and the incidence of cancer between 1986-1991. Obviously these figures will be different from those for people aged 40-60 at a midpoint between 1986 and 1991; they are, however, also likely to be substantially different from figures for 45-65 year olds in 1986-1991. Some of the original sample from 1981 will have exited from the study, some people may have re-entered: for instance they may have died or have emigrated. The cohort who were present at census will necessarily decrease over time.

b) Issues not specific to the LS

Incomplete registration 1971-75

It is known that the registration of cancers in England and Wales for 1971-75 was incomplete. This is covered in more detail in Leon (1988). Less than 90% of cancers were registered during this period and there was some geographical bias to this due to variations in completeness between cancer registries. The bias does not however appear to be very substantial, as regional variations in cancer registration rates in the LS are broadly similar to those seen in national mortality data. The later addition of late registrations will not have managed to clear this bias. If these data are required then the Leon report provides further details on controlling for variation between regions to ensure that any apparent socio-demographic/socio-economic variation observed is not an artefact.

7.3 Examples of research into cancer incidence using the LS

The following section provides brief summaries of papers based on analysis using the LS cancer incidence data. This section is intended merely to provide an overview of the versatility of the LS and the range of research questions that can be answered from it. It is not a comprehensive review of LS cancer research. A list of publications using LS cancer data is given in *Annex B* below.

The majority of studies using the cancer registration data in the LS have examined socio-economic difference in cancer incidence. LS Series No.3, *Social distribution of cancer, 1971-1975* (Leon, 1988) is one of the most comprehensive pieces of work using cancer incidence data from the LS. In addition to providing an overview of the LS, of cancer registration and of validity issues Leon reports extensively on variations in incidence by a number of social variables, particularly marital status, reproductive history and socio-economic factors.

The results from the LS are generally in accordance with other studies. For example, single women have higher rates of cancers of the breast and 'other uterus' than married women. Conversely rates of cervical cancer are lower for the single than the married or divorced. However, the report showed a 'J-shaped' relationship of breast cancer with age at first birth. Those who had their first child between 16-19 appeared to be at greater risk of breast cancer than those having their first child between the ages 20-34, after which risk increased.

More recently Brown et al. (1997) examined social class patterns in the incidence of breast, lung and cervical cancer in women, and lung cancer in men, for the period 1976-89. The work was primarily based on the 1971 cohort as a long follow-up period was required. At working ages, there was very little difference in breast cancer incidence between women in non-manual and manual classes. At older ages the incidence was higher in women in non-manual classes than in those in manual classes. Cervical cancer incidence was considerably higher among younger women in manual than in non-manual classes and these differences were greatest in 1986-89, indicating a growing divide. Among both younger and older men and women, strong class differentials in the incidence of lung cancer were evident in 1986-89, with those in manual classes much more likely to have lung cancer.

In another study by Brown et al. (1998) the incidence of stomach, colorectal and pancreatic cancers from 1976-90 was examined for men and women aged 30 years and over by their housing tenure and social class. Large socio-economic differences in the incidence of stomach cancer for both men and women were found. The pattern of colorectal cancer was less clear, with women in more advantaged social groups experiencing higher incidence while for men there was no significant association. Pancreatic cancer showed no association with socio-economic status. Consistent findings with each indicator strengthen the interpretation of the results. Risk factors for these cancers are known to vary by socio-economic status, and the authors concluded that this study demonstrates the importance of continued monitoring of the distribution of cancer incidence.

Harding and Rosato (1999) calculated standardised incidence ratios for commonly occurring cancers and all cancers using the age-sex-specific rates for first generation Scottish, Irish, West Indian and South Asian male and female migrants in the LS. The incidence of all malignant neoplasms among West Indians and Indians was low. Among South Asians the pattern was consistent for Hindus, Sikhs and Moslems (identified using names analysis (Webster, 1989)). Scottish females showed raised incidence of lung cancer.

Regidor et al [1] used data from the LS to evaluate the validity of the theory of fundamental social causes, which postulates that the persistence of the association between socio-demographic factors and disease is predictable, and perhaps inevitable, due to the existence of a series of social conditions which they term the fundamental causes of disease. These fundamental causes involve resources like knowledge, money, power, prestige, and social connections, which strongly influence people's ability to avoid risk and to minimise the consequences of disease once it occurs. The results indicate that, except in the case of lung cancer, where the results observed can be attributed to differential changes in smoking behaviour by social class, the trend in the association between social class and the incidence of other types of breast cancer between 1976-81 and 1986-91 cannot be explained by the fundamental social causes theory. The authors postulate that it may be that this theory can only explain the association between socio-economic status and the incidence of health problems when the risk factors and/or preventative practices related with the health problem are well known, and if, in addition its etiologic fraction and/or preventive fraction is very high.

[1] Regidor E, Donkin A, Elisa Calle M, Dominguez V. Evaluation of the fundamental social causes theory by examining trends in the incidence of cancers with positive and negative social gradients. *Journal of Epidemiology and Community Health* (Submitted for publication 2001)

8. Research into cancer mortality

8.1 Calculation of mortality rates, and direct and indirect standardisation

The methods for calculating standardised mortality rates and ratios are similar to those used for cancer incidence rates and ratios described above, replacing incidence data with those for mortality. The analysis of survival following cancer registration is described in the next section.

8.2 General issues affecting research into cancer mortality

a) Issues specific to the LS

Timeliness

The mortality data are more timely and comprehensive than those for cancer registration. Mortality data are published within a year of the death occurring and are linked to LS records within two years.

b) Issues not specific to the LS

Accuracy of death certification

As death registration is compulsory there is 100% coverage. However, there are known imperfections in the data from death certification. The mortality data are more timely and comprehensive, usually the LS holds mortality data up to two years prior to the current date, and as death registration is compulsory there is 100% coverage. However there are still known imperfections in data from death certificates (Abramson *et al.* 1974, Cameron & McGoogan 1981). The way in which cause of death is coded can vary between individual practioners and cultures. In addition one study of the level of agreement between autopsy and clinical diagnosis at cause of death examined the influence of the socio-economic position of the deceased. Compared to pathologists, clinicians tended to over diagnose neoplasms as the underlying cause of death in the non-manual group and to under diagnose them in the manual group. This would tend towards underestimation of the degree to which lower socio-economic groups experience higher overall cancer mortality (Samphier *et al.* 1988). The numbers were too small to see if there were any disagreements for individual sites by socio-economic position. However a larger study from the US suggested that there, misclassification of cancer sites is not great enough to account for the heterogeneity of associations with socio-economic position (Percy *et al.* 1981

Grulich et al. (1995) considered whether the rise in cancer mortality in older people was real, or due to changes in certification and coding of cause of death between 1970 and 1990. They found that death coding and certification artefacts were much larger in older persons and that in those aged 75-84 change in the position of recording cancer on the death certificate could potentially account for 46% of the recorded increase in deaths from prostate cancer and 28% of the increase in deaths from breast cancer. The same effect would *not* be seen in relation to cancers with high or rapid mortality such as lung or stomach because secondary conditions would never have been as likely. Part but not all of the increase in cancer mortality rates may have been explained by the implementation of ICD coding Rule 3 in 1984. Researchers should therefore acknowledge the possibility that increasing trends in slower cancers may be artefactual. Any large increase between 1994 and 1995 may be an indicator of such an influence and methods could be employed to 'smooth' the trend.

Increased case-fatality from cancers might explain the rest of the increase, but there was no evidence for this. Other possible explanations are that changes in diagnostic information available at death certification, or changes in certification practice of doctors (for instance, because of increased social acceptability of certifying cancer as a cause of death), may have increased the frequency with which doctors record cancer as an underlying cause.

Within the LS, 845 people with a cancer registration died in 1970-1985 with cancer of unspecified site given as the underlying cause. However, looking at the cancer registrations, only 74% of these had cancer of unspecified site registered; the other 26% had cancer of a specific site registered. In males the most common sites affected were lung and liver cancer (34% and 11% respectively of the specified sites) and in females, ovarian and stomach cancer (14% and 11% of specified sites). The percentage of persons with a specified cancer registration did not vary materially between 1971-78 and 1979-1985, or by age. Grulich et al. (1995)

Thus it is important for the researcher to investigate any potential changes in classification, coding or practices which may have occurred in the time period in which they are interested. A full set of notes and definitions for mortality data has been published (ONS 1999). This includes: base populations; occurrences and registrations;

areal coverage, death rates and standardisation; certification of cause of death; coding the underlying cause of death; analysis of the conditions mentioned on the death certificate; amended cause of death; accelerated registrations; legislation on the registration of deaths and the processing, reporting and analysis of mortality data including the introduction of the Ninth Revision of the International Classification of Diseases in 1979, industrial action taken by registration officers in 1981-92, and the amendment by OPCS in 1984 of WHO Rule 3 (one of the rules used to select the underlying cause of death).

The main change in introducing automated cause of death coding was in the interpretation of Rule 3. The death certificate is set out in two parts; part I gives the condition or sequence of conditions leading to death, while part II gives details of any associated conditions. Rule 3 states that 'if the condition selected by the General rules or Rules 1 and 2 can be considered a direct sequel of another reported condition, whether in part I or part II, select this primary condition'. The interpretation of Rule 3 was broadened by OPCS in 1984 so that certain conditions which were often terminal, such as bronchopneumonia or pulmonary embolism, could be considered a direct sequel of any more specific condition reported. The more specific condition would then be regarded as the underlying cause. This change in interpretation meant that the numbers of deaths from certain conditions such as pneumonia fell suddenly in 1984, while deaths from other conditions rose (Rooney & Devis, 1996). The change in 1993 was thus a move back to the internationally accepted interpretation of Rule 3 operating in England and Wales before 1984.

8.3 Examples of research into cancer mortality using the LS

Dolin, (1992) compared the percentage of workers in each occupation in areas with high bladder cancer mortality with the average for England and Wales. The sex-specific occupational makeup of each district was determined from the 1971 census information in the LS. The percentage of workers in 220 separate occupations in the high-risk areas was compared to the corresponding percentages for England and Wales. Ninety five per cent confidence intervals were based on the upper and lower Poisson expectations of the number of workers in each occupation in the high-risk areas. For men, occupations associated with high bladder cancer mortality areas largely fell into four categories: chemical, glass, engineering and textile-related occupations. The corresponding areas for females had a higher percentage of workers mainly in textile related industries.

Dolin and Cook-Mozaffari (1992) also used the LS to provide information on the number of males employees in each occupation and industry according to age and district of residence at 1971. This information was used in conjunction with occupational statements on the death certificates of 2,457 men aged 25-64 who had died from bladder cancer in selected coastal and estuarine regions of England and Wales during 1965-80. Excess mortality was found for deck and engine crew of ships, railway workers, electrical and electronic workers, shoemakers and repairers and tobacco workers. An excess of cases also occurred among food workers, particularly those employed in the bread and flour confectionery industry or involved in the extraction of animal and vegetable oils and fats. Use of a job-exposure matrix revealed elevated risk for occupations in which most workers were exposed to paints and pigments, benzene and cutting oils.

Harding and Allen (1996) identified the LS as a possible tool for ethnic minority cancer research, along with death certificates and the General Household Survey for risk factor information. Ethnicity was recorded in census data in 1991 and 2001. The LS, however, also has information on the country of birth of the parents for those members of the study present at the 1971 census and this, with name coding, can be used as a proxy for ethnic origin. The authors described those with higher or lower than expected mortality from selected cancers by country of birth. People born in India, the Caribbean and Africa had lower mortality from lung, skin and breast cancer compared to all men and women in England and Wales. The Irish however had higher mortality from lung cancer and both Irish and Caribbean women showed higher mortality from cervical cancer. Mortality from liver cancer was higher among Indian, Caribbean and Irish men.

9. Research into cancer survival

The data available in the cancer registration database – and hence within the LS - make it possible to look at the time which elapsed between cancer registration and death (survival). Such research can inform on socio-economic or geographical variations in cancer survival, differences in survival according to treatment type or the stage at which the cancer was diagnosed. The length of survival after a diagnosis of cancer is known to vary between types of cancer, with some cancers (e.g. lung) having a much shorter prognosis than others.

9.1 Calculation of survival rates and ratios

Crude survival is the proportion of individuals who survived for a defined period of time, for example one year or five years. This does not take account of the actual cause of death, which may or may not be cancer, or of risk of death (increasingly termed *background risk*) found among those in the population of the same age and sex as the person with a cancer registration.

Corrected or net survival can be calculated so that the mortality risk due to cancer can be differentiated from that which is due to background risk. This type of analysis requires information on the cause of death of both cancer patients and others who die in the relevant period, which the LS can offer. This measure gives the probability of survival from cancer in the absence of other causes of death, but does not take account of age.

Relative survival takes into account background risk, but does not require information about the cause of death in cancer patients. As with net survival, the method assumes that the risks of death from cancer and background causes can be considered to act independently. This method relies on estimates of death rates from other causes of death using routine vital statistics. The measure gives a ratio which shows the additional mortality risk from the cancer compared to the level of risk in the general population.

Age standardisation of relative survival rates

A relative survival rate is not age–standardised, and while it allows for age-specific mortality from other causes, an excess risk of death may be evident from the cancer itself as cancer is often age-dependent. For many cancers, although not all, relative survival declines with age (female breast and prostate are the principal exceptions, with lower survival in younger patients than in the middle aged).

It is advisable to age standardise to take into account the possibility of variation in age distributions between groups. Age standardisation is especially important if looking at either time trends or geographical differences in survival. As survival varies with age, changes in the age distribution of cancer patients over time, or differences between areas, might affect the results.

Age-standardised survival is the overall survival rate that would occur, if the age distribution of the group with cancer had been the same as the standard population.

Formulae for all these rates can be found in chapter 3, 'Methods' of *Cancer Survival Trends* (Coleman et al. 1999). More information can also be found in:

Parkin DM and Hakulinen, T. Analysis of Survival. In: *Cancer Registration Principles and Methods*, Eds. Jensen, O.M., Parkin D.M., Maclennan, R., Muir, C.S. and Skeet, R.G. IARC Scientific Publications 1991: Lyon.

9.2 Benefits of the LS for cancer survival research

Relative survival is a common method for analysing the survival of cancer patients. One of the reasons for its popularity is that it does not need cause of death information, which is not available through the national cancer files before the 1990s. One advantage of the LS is that net survival can also be calculated, since due to the linked nature of the data there is information both on cancer incidence and mortality by cause. The major problem with net survival is the poor quality (including a high proportion of 'non-specific' cancer) of the mortality data. (Quinn *et al* 2001)

It is also possible using the LS to investigate trends in survival due to the longitudinal nature of the data, and to look at these trends in terms of individual characteristics. It is not possible to do the latter with the national cancer files, although they can be linked to area-based measures (e.g. ward deprivation scores). The ability to follow cohorts through time also enables researchers to look at the survival of cancer patients given characteristics at census time points, registration or death.

9.3 Examples of research into cancer survival using the LS

There has been comparatively little research on cancer survival using the LS.

Murphy et al. (1990) conducted some research using the LS to determine whether or not the survival of women with cancer of the uterine cervix was associated with their marital status and social class. Apparent differences in crude survival by marital status and social class were examined. However these apparent differences were found to be accounted for by adjustment for age and stage of cancer.

Kogevinas et al. (1991) investigated the relationship between socio-economic status and cancer survival. Socio-economic status was assessed in terms of housing tenure. Council tenants, the low socio-economic group, had poorer survival than owner occupiers, the high socio-economic group, for the combined group of all cancers and for 11 out of 13 cancers examined in males and 12 out of 15 cancers examined in females. Differences were found irrespective of death and prognosis of cancer. Survival analysis by length of follow up for cancers of varying prognosis indicated that council tenants were more likely to present at a later stage than owner occupiers. The authors concluded that differences in survival of cancers of poor prognosis (e.g. oesophagus, pancreas and lung) where treatment has little effect could not be attributed to differences in treatment, but the survival differences for cancers of good prognosis (e.g. corpus uteri, bladder, skin) could, in part, be due to differences in treatment.

10. References

Abramson JH, Sacks MI, Cahana E. Death certification data as an indication of the presence of certain common conditions at death. J Chron Dis 1974: 24: 417-31

Armitage P, Berry G. In: Statistical Methods in Medical Practice. Oxford: Blackwell Scientific Publications, 1994: 477-81.

Boyle P, Parkin DM. Statistical methods for registries. In: Cancer Registration Principles and Methods, Eds. Jensen OM, Parkin DM, Maclennan R, Muir CS, Skeet RG. IARC Scientific Publications 1991: Lyon.

Cameron HM, McGoogan E. A prospective study of 1152 hospital autopsies: II. Analysis of inaccuracies in clinical diagnosis and their significance. J Pathol. 1981: 133: 285-300.

Coleman MP, Babb P, Damiecki P, et al . Cancer Survival Trends in England and Wales 1971-1995. Deprivation and NHS Region. 1999. The Stationery Office: London.

Department of Health. Saving Lives: Our Healthier Nation. 1999. London: Department of Health.

Department of Health. National Cancer Plan. 2000. London: Department of Health..

Department of Health. Action Programme on Cancer Registration. 2000. London: Department of Health http://www.doh.gov.uk/cancer/pdfs/actionprogramme.pdf

Gillis Report. 2000 http://www/doh.gov.uk/pdfs/gillis.pdf

HMSO. (2001) "Health and Social Care Act" http://www.hmso.gov.uk/acts/acts2001/20010015.htm

Huggett C. *Review of the Quality and Comparability of Data held by Regional Cancer Registries.* Bristol: Bristol Cancer Epidemiology Unit incorporating the South West Cancer Registry, 1995.

Morrell S. Quantitative methods in demography. University of Sydney: http://www.health.usyd.adu.au/demog98/course/course5.htm#Introduction

NHS Executive. *Core contract for purchasing Cancer Registration.* EL(96)7. 1996. London: NHS Executive.

Parkin DM, Hakulinen T. Analysis of Survival. In: Cancer Registration Principles and Methods, Eds. Jensen OM, Parkin DM, Maclennan R., Muir CS, Skeet RG. IARC Scientific Publications 1991: Lyon.

Percy C, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. Am J Public Health 1981: 71: 242-50.

Quinn M, Babb P, Brock A, Kirby L, Jones J. Cancer trends in England and Wales 1950-1999. 2001. The Stationery Office: London.

Rooney CIF, Smith SK. Implementation of ICD10 for mortality data in England and Wales from January 2001. HSQ No. 08, Winter 2000, The Stationery Office: London.

Samphier ML, Robertson C, Bloor MJ. A possible artefactual component in specific cause mortality gradients. J Epidemiol Comm Health 1988: 42: 138-43.

Swerdlow AJ. "Cancer registration in England and Wales: some aspects of the interpretation of the data", Journal of the Royal Statistical Society, Series A, 1986: 149:2, pp146-160

Vandenbroucke JP. Letter to the editor: A shortcut method for calculating the 95% confidence interval for the standardised mortality ratio. Am J Epidem. 1982: 115: 303-304.

Webster J. Using the OPCS Longitudinal Study to classify ethnic origin. Longitudinal Study Working Paper 41. City University, 1989.

Annex A - Directly and indirectly standardised rates and confidence intervals

Direct standardisation

1. Calculate the number of deaths/incidences (c) that occur in each five year age group (a) in the sample of interest in the time period of interest for each social class or group of interest.

2. Calculate the sum of the Person Years at Risk (n) for those in each five year age band (a) for each social class or group of interest.

3. Divide 1 by 2 and multiply by 100, 000 giving crude rates (r_i) per 100,000 person years for each group. The crude rate across a number of groups is given by the total number of cases/deaths in those groups C divided by the total number of person years at risk N.

4. Apply these rates to the standard population (w) in 5 year age bands. Standard populations can be found in Appendix H of Quinn et al. *Cancer Trends in England and Wales* 1950-1999. 2001. The Stationery Office: London.

5. Add the 'expected' number of people across all age bands in each age category (e.g. add 0-4, 5-9, 10-15 results up for 0-15 age band).

6. Divide this by the number of people in total standard population in the age categories used.

7. Multiply by 100,000 to get the standardised rate.

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

$$C = \sum_{i=1}^{a} c_i = c_1 + c_2 + c_3 + ...c_a$$

$$N = \sum_{i=1}^{a} n_i = n_1 + n_2 + n_3 + ...n_a$$

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

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

To calculate the confidence intervals

In general the (100(1-α))% confidence interval of an age standardised rate (ASR) with standard error se(ASR) can be expressed as ASR +/- $Z_{\alpha/2}$ * (se(ASR))

Where $Z_{\alpha/2}$ is the standardised normal deviate, and thus 1.96 for a 95% CI and 2.58 for a 99% CI.

The standard error of an age standardised rate is: Se(ASR) = √Var(ASR)

Where the variance (var) = ((Age specific rate (a) * standard population for that age group^2 (100,000 – a))/person years at risk for that age group (n))/sum of the standard populations for all age groups

Small numbers

An alternative expression can be obtained, as outlined in Armitage and Berry (1987) when the numbers are small by making a Poisson approximation to the binomial variance of the age-specific rates (a). This results in an expression of the variance of the age-standardised rate (Var(ASR))

$$\int Var(ASR) = \frac{\sum_{i=1}^{A}(a_i w_i^2 \times 100000/n_i)}{(\sum_{i=1}^{A} w_i)^2}$$

(Where w = standard population in each age band, n= person years at risk and the standard error of the age standardised rate (s.e. (ASR)) is the square root of the variance.

Indirectly age–standardised rates

This is a comparison between observed and expected numbers of cases. The expected number of cases is calculated by applying a standard set of age-specific rates (a_i) to the population of interest:

$$\sum_{i=1}^{A} e_i = \sum_{i=1}^{A} a_i n_i \div 100000$$

where e_i, the number of cases expected in age class i , is the product of the 'standard rate' and the number of persons in age class i in the population of interest.

The standardised ratio (M) is now calculated by comparing the observed number of cases (r_i) with that expected

$$M = \frac{\sum_{i=1}^{A} r_i}{\sum_{i=1}^{A} e_i} = \frac{\sum_{i=1}^{A} r_i}{\sum_{i=1}^{A} a_i n_i \div 100000}$$

This is normally expressed as a percentage by multiplying by 100. When applied to incidence data it is commonly known as the standardised incidence ratio (SIR); when applied to mortality data it is known as the standardised mortality ratio (SMR).

Standard error of the standardised ratio

The variance of the above standardised ratio (M) is given by

$$Var(M) = \frac{\sum_{i=1}^{A} r_i}{(\sum_{i=i}^{A} a_i n_i \div 100000)^2}$$

and the standard error of the indirect ratio, s.e. (M) is the square root of the variance

$$s.e.(M) = \frac{\sqrt{\sum_{i=1}^{A} r_i}}{\sum_{i=i}^{A} a_i n_i \div 100000}$$

Vandenbroucke (1982) has proposed a short-cut method for calculating the (100(1-))% confidence interval of a standardised ratio, involving a two step procedure. First, the lower and upper limits for the observed events are calculated:

$$\text{Lower limit} = [\sqrt{\text{observed events}} - (z_{\alpha/2} \times 0.5)]^2$$

$$\text{Upper limit} = [\sqrt{\text{observed events}} + (z_{\alpha/2} \times 0.5)]^2$$

Division of these limits for the observed number of the expected number of events gives the approximate 95% confidence interval for the SIR.

Annex B - Papers using LS Cancer Data

Brown J, Harding S, Bethune B, Rosato M. (1997) "Incidence of Health of the Nation cancers by social class", Population Trends, 90 (Winter 1997), pp 40-47

Brown J, Harding S, Bethune A, Rosato M. (1998) "Longitudinal Study of socio-economic differences in the incidence of stomach, colorectal and pancreatic cancers", Population Trends, 94 (Winter 1998), pp 35-41

Davey-Smith G, Leon D, Shipley M, Rose G. (1991) "Socio-economic differentials in cancer among men", International Journal of Epidemiology, Vol 20: 2, pp 339-345

Dolin P. (1992) "A descriptive study of occupation and bladder cancer in England and Wales", British Journal of Cancer, Vol. 65:3 , pp. 476-478

Dolin PJ, Cook-Mozaffari P. (1992) "Occupation and bladder cancer: a death-certificate study", British Journal of Cancer, Vol. 66: 3, pp. 568-579

Grulich A, Swerdlow A, Dos Santos Silva I, Beral V. (1995) "Is the apparent rise in cancer mortality in the elderly real? Analysis of changes in certification and coding of cause of death in England and Wales, 1970-1990", International Journal of Cancer, 63, pp 164-168

Harding S, Allen EJ. (1996) "Sources and uses of data on cancer among ethnic groups", British Journal of Cancer, 74: Supplement XXIX, S17-S21

Harding S. (1998) "The incidence of cancers among second generation Irish living in England and Wales", British Journal of Cancer, 78(7), pp 958-961

Harding S, Brown J, Rosato M, Hattersley L. (1999) "Socio-economic differentials in health: illustrations from the Office for National Statistics Longitudinal Study. Health Statistics Quartely, 1, 5-15

Harding S, Rosato M. (1999) "Cancer incidence among first generation Scottish, Irish, West Indian and South Asian migrants living in England and Wales", Ethnicity and Health, 4 (1/2), pp 83-92

Jones DR. (1988a) "Cancer mortality following widow(er)hood in Office of Population Censuses and Surveys' Longitudinal Study", In Eylenbosch, W J., Depoorter A-M., van Larebeke N. (Eds.) Proceedings of the First International Symposium on Primary Prevention of Cancer, Antwerp, EORTC Monograph Vol. 19, New York: Raven Press

Jones DR. (1988b) "Cancer mortality following widow(er)hood in the Office of Population Censuses and Surveys' Longitudinal Study", In Watson M., Greer S., Thomas C. (Eds) Psychological oncology: proceedings of conference on bereavement and cancer, Leicester, 1987, Oxford: Pergamon Press

Jones DR, Goldblatt PO. (1986) "Cancer and mortality following widow(er)hood. Some further results from the Office of Population Censuses and Surveys Longitudinal Study", Stress Medicine, Vol. 2, pp 129-140

Jones DR, Goldblatt PO, Leon DA. (1984) "Bereavement and cancer: some data on deaths of spouses from the Longitudinal Study of the Office of Population Censuses and Surveys", British Medical Journal, Vol 289, pp 461-464

Kogevinas M. (1990) Longitudinal Study: Socio-demographic differences in cancer survival, OPCS Series LS No.5, London: The Stationery Office (formerly HMSO)

Kogevinas M. (1992) "Social Inequalities and Cancers", In: Preventing Cancers, Heller, T., Bailey, L., and Patterson, S. (eds), Milton Keynes: The Open University Press, pp 5-17

Kogevinas M, Marmot MG, Fox AJ, Goldblatt PO. (1991) "Socio-economic differences in cancer survival", Journal of Epidemiology and Community Health, Vol. 45, pp 216-219

Leon D. (1988a) The Social Distribution of Cancer, OPCS Series LS No 3, London: The Stationery Office (formerly HMSO)

Leon D. (1988b) "Socio-economic factors and the primary prevention of cancer", In Eylenbosch, W.J. and Depoorter, A-M van Larebeke N. (Eds.), Proceedings of the First International Symposium on Primary Prevention of Cancer, Antwerp, EORTC Monograph Vol. 115, p 19, New York: Raven Press

Leon D. (1989) "A prospective study of the independent effects of parity and age at first birth on breast cancer incidence in England and Wales", International Journal of Cancer, Vol.43, pp 986-991

Leon D, Wilkinson RG. (1989) "Inequalities in prognosis: socio-economic differences in cancer and heart diseases survival", Health Inequalities in European Countries, J. Fox (Ed.) Proceedings of the European Science Foundation Workshops, held in London 1984-86, Aldershot: Gower Press, pp. 280-300

Murphy M, Goldblatt P, Thornton-Jones H, Silcocks P. (1990) "Survival amongst women with cancer of the uterine cervix; the influence of marital status and social class", Journal of Epidemiology and Community Health, Vol. 44, pp. 293-296

Pugh H, Power C, Goldblatt P, Arber S. (1991) "Women's Lung Cancer Mortality, Socio-Economic Status and Changing Smoking Patterns", Social Science and Medicine, Vol. 32: 10, pp. 1105-1110

Annex C - Validation tables and incidence rates

Table 1: All Cancer Registrations 1971 - 1994 giving linkage rates and sampling fractions

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	95% C.I	Linkage rate	Observed Sampling Fraction **
1971*	109287	1260	1198	1131 - 1265	106.1	1.16
1972	162551	1778	1781	1699 - 1863	100.4	1.09
1973	168504	1946	1847	1763 - 1933	106.4	1.16
1974	182090	1944	1996	1909 - 2083	98.0	1.07
1975	181290	1928	1987	1900 - 2074	97.8	1.06
1976	185243	2024	2030	1942 - 2118	97.9	1.09
1977	189817	2025	2080	1991 - 2169	98.3	1.07
1978	188490	1969	2066	1977 - 2155	96.3	1.05
1979	199947	2085	2191	2100 - 2282	96.2	1.05
1980	201533	2173	2209	2117 - 2301	99.3	1.08
1981	220579	2338	2417	2321 - 2513	96.7	1.06
1982	222542	2307	2439	2343 - 2535	94.6	1.04
1983	226854	2528	2486	2389 - 2583	101.7	1.11
1984	231057	2399	2532	2434 - 2630	94.7	1.04
1985	250241	2689	2742	2640 - 2844	98.1	1.07
1986	249467	2743	2734	2632 - 2836	100.3	1.10
1987	259243	2958	2841	2737 - 2945	104.1	1.14
1988	271602	2992	2976	2870 - 3082	100.5	1.10
1989	271047	3043	2970	2864 - 3076	102.5	1.12
1990	275218	3159	3016	2909 - 3123	104.7	1.15
1991	281412	3399	3084	2976 - 3192	110.2	1.21
1992	292459	3423	3205	3095 - 3315	106.8	1.17
1993	287679	3161	3153	3044 - 3262	100.3	1.10
1994	296698	3489	3251	3140 - 3362	107.3	1.18

* 1971 contains numbers of cancers diagnosed from Census day 1971 to 31st December 1971

** The observed sampling fraction is calculated as ((LS actual number of cancers/ E&W actual number of cancers) * 100)

Rows marked in grey indicate where the numbers of LS actual cancers fall outside the 95% CI for expected numbers of cancers in the LS

Table 2: All cancer registrations by broad age group, sex and year of occurrence of cancer

a) Males Year of diagnosis	Age at cancer occurence 0-19						20-49					
	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS	Exp. in LS (95% CI)	Linkage rate	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS	Exp. in LS (95% CI)	Linkage rate
1981	6	1,063	0.56	12	5 - 19	50.0	96	8,079	1.19	89	71 - 107	107.9
1982	11	1,084	1.01	12	5 - 19	91.7	74	7,852	0.94	86	68 - 104	86.0
1983	12	1,067	1.12	12	5 - 19	100.0	93	8,182	1.14	90	72 - 108	103.3
1984	16	1,062	1.51	12	5 - 19	133.3	85	8,149	1.04	89	71 - 107	95.5
1985	11	1,031	1.07	11	5 - 17	100.0	98	8,682	1.13	95	76 - 114	103.2
1986	11	1,044	1.05	11	5 - 17	100.0	106	8,652	1.23	95	76 - 114	111.6
1987	9	1,142	0.79	13	6 - 20	69.2	98	9,201	1.07	101	81 - 121	97.0
1988	8	1,132	0.71	12	5 - 19	66.7	104	9,693	1.07	106	86 - 126	98.1
1989	9	1,036	0.87	11	5 - 17	81.8	115	9,684	1.19	106	86 - 126	108.5
1990	12	1,129	1.06	12	5 - 19	100.0	120	9,818	1.22	108	88 - 128	111.1
1991	11	1,050	1.05	12	5 - 19	91.7	126	10,127	1.24	111	90 - 132	113.5
1992	13	1,103	1.18	12	5 - 19	108.3	119	10,808	1.10	118	97 - 139	100.8
1993	17	1,131	1.50	12	5 - 19	141.7	129	10,558	1.22	116	95 - 137	111.2
1994	10	1,064	0.94	12	5 - 19	83.3	109	10,473	1.04	115	94 - 136	94.8

a) Males Year of diagnosis	50-69						70+						Total					
	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS	Exp. in LS (95% CI)	Linkage rate	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS	Exp. in LS (95% CI)	Linkage rate	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS	Exp. in LS (95% CI)	Linkage rate
1981	481	48,479	0.99	531	486 - 576	90.6	539	50,654	1.06	555	509 - 601	97.1	1,122	108,275	1.04	1,187	1120 - 1254	94.5
1982	532	47,872	1.11	525	480 - 570	101.3	526	52,238	1.01	572	525 - 619	92.0	1,143	109,046	1.05	1,195	1128 - 1262	95.6
1983	571	47,093	1.21	516	472 - 560	110.7	608	54,132	1.12	593	546 - 640	102.5	1,284	110,474	1.16	1,211	1143 - 1279	106.0
1984	506	46,949	1.08	515	471 - 559	98.3	538	55,047	0.98	603	555 - 651	89.2	1,145	111,207	1.03	1,219	1151 - 1287	93.9
1985	516	48,802	1.06	535	490 - 580	96.4	676	60,103	1.12	659	609 - 709	102.6	1,301	118,618	1.10	1,300	1230 - 1370	100.1
1986	508	48,193	1.05	528	483 - 573	96.2	634	59,071	1.07	647	597 - 697	98.0	1,259	116,960	1.08	1,282	1212 - 1352	98.2
1987	586	48,441	1.21	531	486 - 576	110.4	644	60,530	1.06	663	613 - 713	97.1	1,337	119,314	1.12	1,308	1238 - 1378	102.2
1988	579	51,245	1.13	562	516 - 608	103.0	672	63,161	1.06	692	641 - 743	97.1	1,363	125,231	1.09	1,372	1300 - 1444	99.3
1989	564	50,314	1.12	551	505 - 597	102.4	708	62,925	1.13	690	639 - 741	102.6	1,396	123,959	1.13	1,358	1286 - 1430	102.8
1990	554	49,593	1.12	543	498 - 588	102.0	787	64,391	1.22	706	654 - 758	111.5	1,473	124,931	1.18	1,369	1297 - 1441	107.6
1991	557	49,755	1.12	545	499 - 591	102.2	810	67,294	1.20	737	684 - 790	109.9	1,504	128,226	1.17	1,405	1332 - 1478	107.0
1992	617	50,525	1.22	554	508 - 600	111.4	849	71,711	1.18	786	731 - 841	108.0	1,598	134,147	1.19	1,470	1395 - 1545	108.7
1993	573	49,878	1.15	547	501 - 593	104.8	768	72,026	1.07	789	734 - 844	97.3	1,487	133,593	1.11	1,464	1389 - 1539	101.6
1994	598	50,586	1.18	554	508 - 600	107.9	921	75,470	1.22	827	771 - 883	97.3	1,638	137,593	1.19	1,508	1432 - 1584	108.6

Rows marked in grey indicate where the numbers of LS actual cancers fall outside the 95% CI for expected numbers of cancers in the LS

Table 2: All cancer registrations by age, sex and year of occurrence of cancer

b) Females Year of diagnosis	Age at cancer occurence											
	0-19						20-49					
	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS	Exp. in LS (95% CI)	Linkage rate	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS	Exp. in LS (95% CI)	Linkage rate
1981	9	886	1.02	10	4 - 16	90.0	194	19,097	1.02	209	181 - 237	92.8
1982	7	898	0.78	10	4 - 16	70.0	197	19,531	1.01	214	185 - 243	92.1
1983	14	950	1.47	10	4 - 16	140.0	234	20,732	1.13	227	198 - 256	103.1
1984	11	960	1.15	11	5 - 17	100.0	243	22,927	1.06	251	220 - 282	96.8
1985	13	1,014	1.28	11	5 - 17	118.2	298	27,775	1.07	304	270 - 338	98.0
1986	11	1,053	1.04	12	5 - 19	91.7	293	29,293	1.00	321	286 - 356	91.3
1987	12	1,237	0.97	14	7 - 21	85.7	354	31,988	1.11	351	314 - 388	100.9
1988	16	1,257	1.27	14	7 - 21	114.3	397	33,461	1.19	367	330 - 404	108.2
1989	17	1,200	1.42	13	6 - 20	130.8	369	32,933	1.12	361	324 - 398	102.2
1990	12	1,273	0.94	14	7 - 21	85.7	423	35,598	1.19	390	352 - 428	108.5
1991	16	1,190	1.34	13	6 - 20	123.1	419	35,309	1.19	387	349 - 425	108.3
1992	19	1,271	1.49	14	7 - 21	135.7	424	36,130	1.17	396	357 - 435	107.1
1993	9	1,254	0.72	14	7 - 21	64.3	364	35,214	1.03	386	348 - 424	94.3
1994	9	1,248	0.72	14	7 - 21	64.3	420	36,940	1.14	405	366 - 444	103.7

b) Females Year of diagnosis	50-69						70+						Total					
	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS	Exp. in LS (95% CI)	Linkage rate	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS (95% CI)	Exp. in LS (95% CI)	Linkage rate	LS cancers	E&W cancers	Observed Sampling Fraction	Exp. in LS	Exp. in LS (95% CI)	Linkage rate
1981	472	43,046	1.10	472	430 - 514	100.0	541	49,275	1.10	540	495 - 585	100.2	1,013	112,304	0.90	1,231	1163 - 1299	82.3
1982	468	42,931	1.09	470	428 - 512	99.6	492	50,136	0.98	549	503 - 595	89.6	1,163	113,496	1.02	1,244	1175 - 1313	93.5
1983	447	42,776	1.04	469	427 - 511	95.3	549	51,922	1.06	569	523 - 615	96.5	1,200	116,380	1.03	1,275	1205 - 1345	94.1
1984	444	42,717	1.04	468	426 - 510	94.9	556	53,246	1.04	584	537 - 631	95.2	1,248	119,850	1.04	1,313	1242 - 1384	95.0
1985	493	44,834	1.10	491	448 - 534	100.4	584	58,000	1.01	636	587 - 685	91.8	1,331	131,623	1.01	1,442	1368 - 1516	92.3
1986	528	44,517	1.19	488	445 - 531	108.2	652	57,644	1.13	632	583 - 681	103.2	1,491	132,507	1.13	1,452	1378 - 1526	102.7
1987	516	46,138	1.12	506	462 - 550	102.0	739	60,566	1.22	664	614 - 714	111.3	1,559	139,929	1.11	1,533	1457 - 1609	101.7
1988	539	48,989	1.10	537	492 - 582	100.4	677	62,664	1.08	687	636 - 738	98.5	1,582	146,371	1.08	1,604	1526 - 1682	98.6
1989	556	49,878	1.11	547	501 - 593	101.6	704	63,077	1.12	691	640 - 742	101.9	1,673	147,088	1.14	1,612	1534 - 1690	103.8
1990	542	50,525	1.07	554	508 - 600	97.8	830	62,891	1.32	689	638 - 740	120.5	1,758	150,287	1.17	1,647	1568 - 1726	106.7
1991	599	50,551	1.18	554	508 - 600	108.1	861	66,136	1.30	725	673 - 777	118.8	1,895	153,186	1.24	1,679	1599 - 1759	112.9
1992	552	51,133	1.08	560	514 - 606	98.6	829	69,778	1.19	765	711 - 819	108.4	1,816	158,312	1.15	1,735	1654 - 1816	104.7
1993	562	49,040	1.15	537	492 - 582	104.7	739	68,578	1.08	752	699 - 805	98.3	1,744	154,086	1.13	1,689	1609 - 1769	103.3
1994	579	49,493	1.17	542	497 - 587	106.8	843	71,424	1.18	783	728 - 838	107.7	1,795	154,087	1.16	1,689	1609 - 1769	106.3

Rows marked in grey indicate where the numbers of LS actual cancers fall outside the 95% CI for expected numbers of cancers in the LS

Table 3: All malignancies - ICD 140 - 208 (excluding 173) by sex, 1981 - 1994 showing linkage rates, sampling fractions and incidence rates per 100,000

Males

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected (95% CI)	Linkage rate	Sampling Fraction	England & Wales - incidence rates per 100,000	LS - incidence rates per 100,000
1981	92411	947	1013	951 - 1075	93.5	1.02	382	358
1982	93312	978	1023	961 - 1085	95.6	1.05	386	370
1983	94576	1091	1036	973 - 1099	105.3	1.15	391	411
1984	95092	978	1042	979 - 1105	93.9	1.03	392	368
1985	100357	1087	1100	1035 - 1165	98.8	1.08	412	407
1986	97756	1054	1071	1007 - 1135	98.4	1.08	400	393
1987	100478	1120	1101	1036 - 1166	101.7	1.11	409	417
1988	103661	1127	1136	1070 - 1202	99.2	1.09	421	418
1989	102711	1165	1126	1061 - 1191	103.5	1.13	415	430
1990	103706	1246	1137	1071 - 1203	109.6	1.20	417	456
1991	105637	1241	1158	1092 - 1224	107.2	1.17	423	454
1992	109331	1290	1198	1131 - 1265	107.7	1.18	436	468
1993	109429	1189	1199	1132 - 1266	99.2	1.09	434	429
1994	112145	1340	1229	1161 - 1297	109.0	1.19	434	429

Rows marked in grey indicate where the numbers of LS actual cancers fall outside the 95% CI for expected numbers of cancers in the LS

Females

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected (95% CI)	Linkage rate	Sampling Fraction	England & Wales - incidence rates per 100,000	LS - incidence rates per 100,000
1981	90460	999	991	930 - 1052	100.8	1.10	355	357
1982	91534	964	1003	941- 1065	96.1	1.05	359	346
1983	92628	972	1015	953 - 1077	95.8	1.05	363	348
1984	94025	989	1030	967 - 1093	96.0	1.05	368	354
1985	100288	1057	1099	1034 - 1164	96.2	1.05	391	376
1986	98145	1116	1076	1012 - 1140	103.7	1.14	382	396
1987	103026	1187	1129	1064 - 1194	105.1	1.15	400	420
1988	105091	1135	1152	1086 - 1218	98.5	1.08	407	401
1989	106393	1166	1166	1099 - 1233	100.0	1.10	410	410
1990	106003	1181	1162	1096 - 1228	101.6	1.11	408	414
1991	109160	1343	1196	1129 - 1263	112.3	1.23	418	470
1992	112239	1294	1230	1162 - 1298	105.2	1.15	429	451
1993	109890	1164	1204	1136 - 1272	96.7	1.06	419	404
1994	112175	1299	1229	1161 - 1297	105.7	1.16	419	404

Rows marked in grey indicate where the numbers of LS actual cancers fall outside the 95% CI for expected numbers of cancers in the LS

Table 4: Major cancers by sex, 1981 - 1994 showing linkage rates, sampling fractions and incidence rates per 100,000

Table 4a: Stomach cancer (ICD 151)

Males

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected 95% CI	Linkage rate	Sampling Fraction	England & Wales - incidence rates per	LS - incidence rates per 100,000
1981	7374	81	81	63 - 99	100.0	1.10	30	31
1982	7213	82	79	62 - 96	103.8	1.14	30	31
1983	7374	83	81	63 - 99	102.5	1.13	30	31
1984	7013	77	77	60 - 94	100.0	1.10	29	29
1985	7219	76	79	62 - 96	96.2	1.05	30	29
1986	7105	76	78	61 - 95	97.4	1.07	29	28
1987	6728	77	74	57 - 91	104.1	1.14	27	29
1988	6964	75	76	59 - 93	98.7	1.08	28	28
1989	6672	65	73	56 - 90	89.0	0.97	27	24
1990	6419	70	70	54- 86	100.0	1.09	26	26
1991	6224	76	68	52 - 84	111.8	1.22	25	28
1992	6354	70	70	54 - 86	100.0	1.10	25	25
1993	5982	62	66	50 - 82	93.9	1.04	24	22
1994	6115	83	67	51 - 83	123.9	1.36	24	30

Females

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected 95% CI	Linkage rate	Sampling Fraction	England & Wales - incidence rates per	LS - incidence rates per 100,000
1981	4957	61	54	40 - 68	113.0	1.23	19	22
1982	4840	48	53	39 - 67	90.6	0.99	19	17
1983	4708	46	52	38 - 66	88.5	0.98	19	16
1984	4654	43	51	37 - 65	84.3	0.92	18	15
1985	4719	63	52	38 - 66	121.2	1.34	18	22
1986	4340	50	48	34 - 62	104.2	1.15	17	18
1987	4553	47	50	36 - 64	94.0	1.03	18	17
1988	4312	50	47	34 - 60	106.4	1.16	17	18
1989	4288	47	47	34 - 60	100.0	1.10	17	17
1990	3991	46	44	31 - 57	104.5	1.15	15	16
1991	3976	47	44	31 - 57	106.8	1.18	15	16
1992	3895	52	43	30 - 56	120.9	1.34	15	18
1993	3641	39	40	28 - 52	97.5	1.07	14	14
1994	3599	37	39	27 - 51	94.9	1.03	14	13

Table 4b: Colorectal cancer (ICD 153 - 154)

Males

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected 95% CI	Linkage rate	Sampling Fraction	England & Wales - incidence rates per	LS - incidence rates per 100,000
1981	12078	129	132	110 - 154	97.7	1.07	50	49
1982	12159	125	133	111 - 155	94.0	1.03	50	47
1983	12448	165	136	113 - 159	121.3	1.33	51	62
1984	12556	134	138	115 - 161	97.1	1.07	52	50
1985	13117	156	144	121 - 167	108.3	1.19	54	58
1986	12643	137	139	116 - 162	98.6	1.08	52	51
1987	13195	147	145	122 - 168	101.4	1.11	54	55
1988	13646	146	150	126 - 174	97.3	1.07	55	54
1989	13850	160	152	128 - 176	105.3	1.16	56	59
1990	13981	160	153	129 - 177	104.6	1.14	56	58
1991	14079	165	154	130 - 178	107.1	1.17	56	61
1992	14972	184	164	139 - 189	112.2	1.23	59	67
1993	14898	157	163	138 - 188	96.3	1.05	59	57
1994	14811	175	162	137 - 187	108.0	1.18	59	63

Females

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected 95% CI	Linkage rate	Sampling Fraction	England & Wales - incidence rates per	LS - incidence rates per 100,000
1981	13238	141	145	122 - 168	97.2	1.07	52	51
1982	13079	146	143	120 - 166	102.1	1.12	51	52
1983	13287	141	146	122 - 170	96.6	1.06	52	50
1984	13141	141	144	121 - 167	97.9	1.07	52	50
1985	13662	134	150	126 - 174	89.3	0.98	53	48
1986	13422	170	147	123 - 171	115.6	1.27	52	60
1987	13775	161	151	127 - 175	106.6	1.17	54	57
1988	14050	158	154	130 - 178	102.6	1.12	54	56
1989	14067	179	154	130 - 178	116.2	1.27	54	63
1990	13885	157	152	128 - 176	103.3	1.13	53	55
1991	14000	165	153	129 - 177	107.8	1.18	54	58
1992	14784	146	162	137 - 187	90.1	0.99	56	51
1993	13991	168	153	129 - 177	109.8	1.20	53	58
1994	14093	159	154	130 - 178	103.2	1.13	54	55

Rows marked in grey indicate where the numbers of LS actual cancers fall outside the 95% CI for expected numbers of cancers in the LS

Table 4c: Lung cancer (ICD 162)

Males

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected 95% CI	Linkage rate	Sampling Fraction	England & Wales - incidence rates per	LS - incidence rates per 100,000
1981	28142	298	308	274 - 342	96.8	1.06	116	113
1982	28021	279	307	273 - 341	90.9	1.00	116	105
1983	27688	286	303	269 - 337	94.4	1.03	114	108
1984	26911	264	295	262 - 328	89.5	0.98	111	99
1985	28271	295	310	276 - 344	95.2	1.04	116	110
1986	26539	299	291	258 - 324	102.7	1.13	108	112
1987	26254	298	288	255 - 321	103.5	1.14	107	111
1988	26383	277	289	256 - 322	95.8	1.05	107	103
1989	25416	313	279	246 - 312	112.2	1.23	103	115
1990	25037	312	274	242 - 306	113.9	1.25	101	114
1991	24890	299	273	241 - 305	109.5	1.20	100	109
1992	24998	294	274	242 - 306	107.3	1.18	99	106
1993	23504	253	258	227 - 289	98.1	1.08	93	91
1994	23314	285	255	224 - 286	111.8	1.22	92	103

Females

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected 95% CI	Linkage rate	Sampling Fraction	England & Wales - incidence rates per	LS - incidence rates per 100,000
1981	9458	115	104	84 - 124	110.6	1.22	37	41
1982	9698	102	106	86 - 126	96.2	1.05	38	37
1983	9871	97	108	88 - 128	89.8	0.98	39	35
1984	10157	110	111	90 - 132	99.1	1.08	40	39
1985	10987	115	120	99 - 141	95.8	1.05	43	41
1986	10826	130	119	98 - 140	109.2	1.20	42	46
1987	11422	147	125	103 - 147	117.6	1.29	45	52
1988	11737	113	129	107 - 151	87.6	0.96	45	40
1989	11601	125	127	105 - 149	98.4	1.08	45	44
1990	11567	118	127	105 - 149	92.9	1.02	44	41
1991	11808	153	129	107 - 151	118.6	1.30	45	53
1992	12352	160	135	112 - 158	118.5	1.30	47	56
1993	12105	143	133	111 - 155	107.5	1.18	46	49
1994	12297	158	135	112 - 158	117.0	1.28	47	55

Rows marked in grey indicate where the numbers of LS actual cancers fall outside the 95% CI for expected numbers of cancers in the LS

Table 4d: Female Breast cancer (ICD 174)

Females

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected 95% CI	Linkage rate	Sampling Fraction	England & Wales - incidence rates per	LS - incidence rates per 100,000
1981	22683	237	249	218 - 280	95.2	1.04	89	85
1982	23214	237	254	223 - 285	93.3	1.02	91	85
1983	23076	232	253	222 - 284	91.7	1.01	90	83
1984	23319	269	256	225 - 287	105.1	1.15	91	96
1985	25630	225	281	248 - 314	80.1	0.88	100	80
1986	25310	270	277	245 - 309	97.5	1.07	98	96
1987	26479	296	290	257 - 323	102.1	1.12	103	105
1988	27172	267	298	264 - 332	89.6	0.98	105	94
1989	28591	302	313	279 - 347	96.5	1.06	110	106
1990	29250	323	321	286 - 356	100.6	1.10	112	113
1991	31097	360	341	305 - 377	105.6	1.16	119	126
1992	31957	365	350	314 - 386	104.3	1.14	122	127
1993	30497	328	334	298 - 370	98.2	1.08	116	114
1994	31671	388	347	311 - 383	111.8	1.23	120	135

Rows marked in grey indicate where the numbers of LS actual cancers fall outside the 95% CI for expected numbers of cancers in the LS

Table 4e: Prostate cancer (ICD 185)

Males

Year of diagnosis	England & Wales n	LS Actual n	LS Expected n	LS Expected 95% CI	Linkage rate	Sampling Fraction	England & Wales - incidence rates per	LS - incidence rates per 100,000
1981	9365	89	103	83 - 123	86.4	0.95	39	34
1982	9720	95	107	87 - 127	88.8	0.98	40	36
1983	10098	124	111	90 - 132	111.7	1.23	42	47
1984	10347	99	113	92 - 134	87.6	0.96	43	37
1985	11213	124	123	101 - 145	100.8	1.11	46	47
1986	11406	91	125	103 - 147	72.8	0.80	47	34
1987	11874	99	130	108 - 152	76.2	0.83	48	37
1988	12617	144	138	115 - 161	104.3	1.14	51	53
1989	12799	134	140	117 - 163	95.7	1.05	52	49
1990	13399	158	147	123 - 171	107.5	1.18	54	58
1991	14313	176	157	133 - 181	112.1	1.23	57	64
1992	15823	178	173	147 - 199	102.9	1.12	63	65
1993	17210	189	189	162 - 216	100.0	1.10	68	68
1994	19399	234	213	185 - 241	109.9	1.21	77	84

Rows marked in grey indicate where the numbers of LS actual cancers fall outside the 95% CI for expected numbers of cancers in the LS

Chart 1. Stomach cancer (ICD151) incidence rates per 100,000

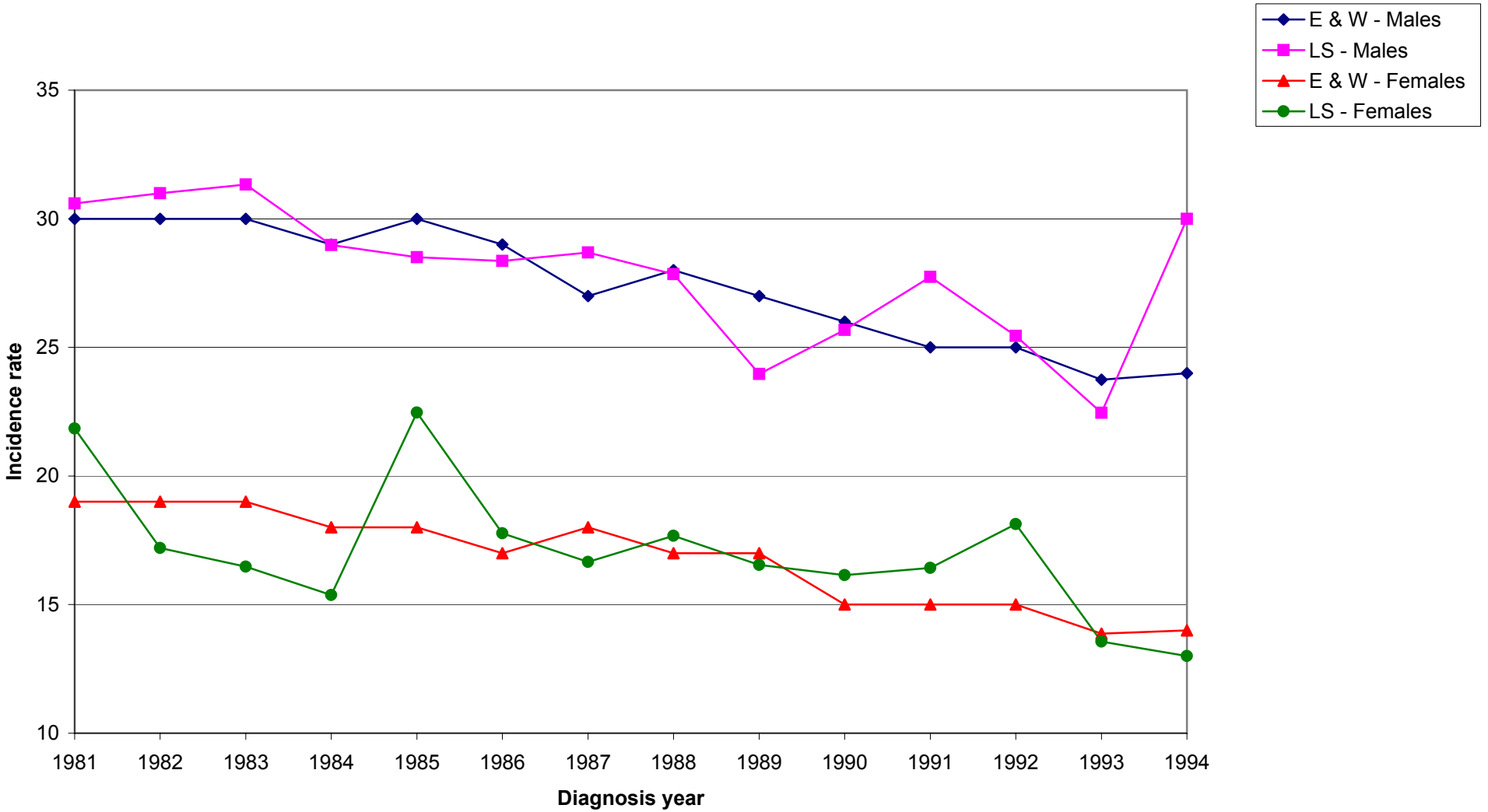

Chart 2. Colorectal cancer (ICD 153 - 154) incidence rates per 100,000

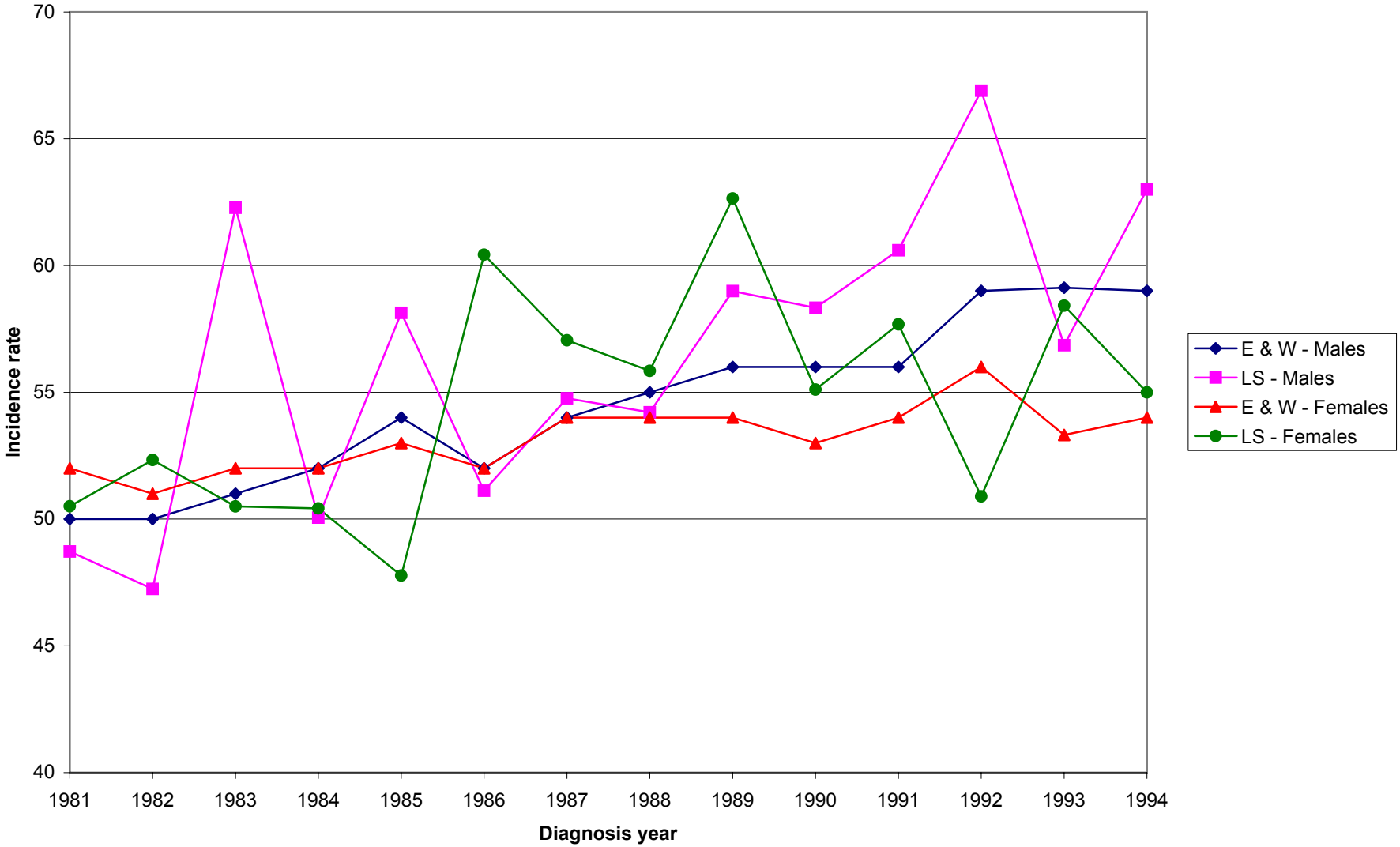

Chart 3. Lung cancer (ICD 162) incidence rates per 100,000

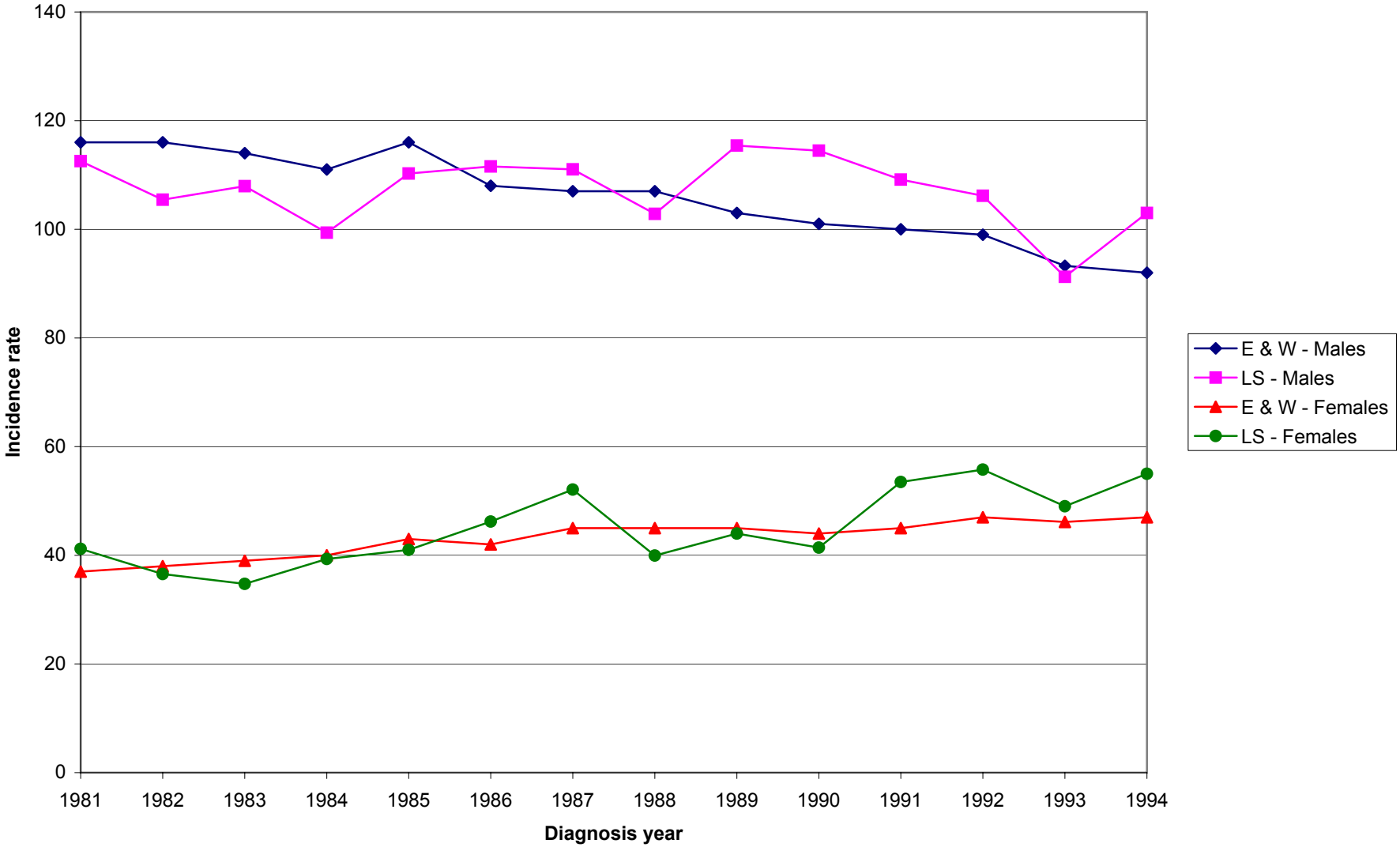

Chart 4. Female Breast cancer (ICD 174) incidence rates per 100,000

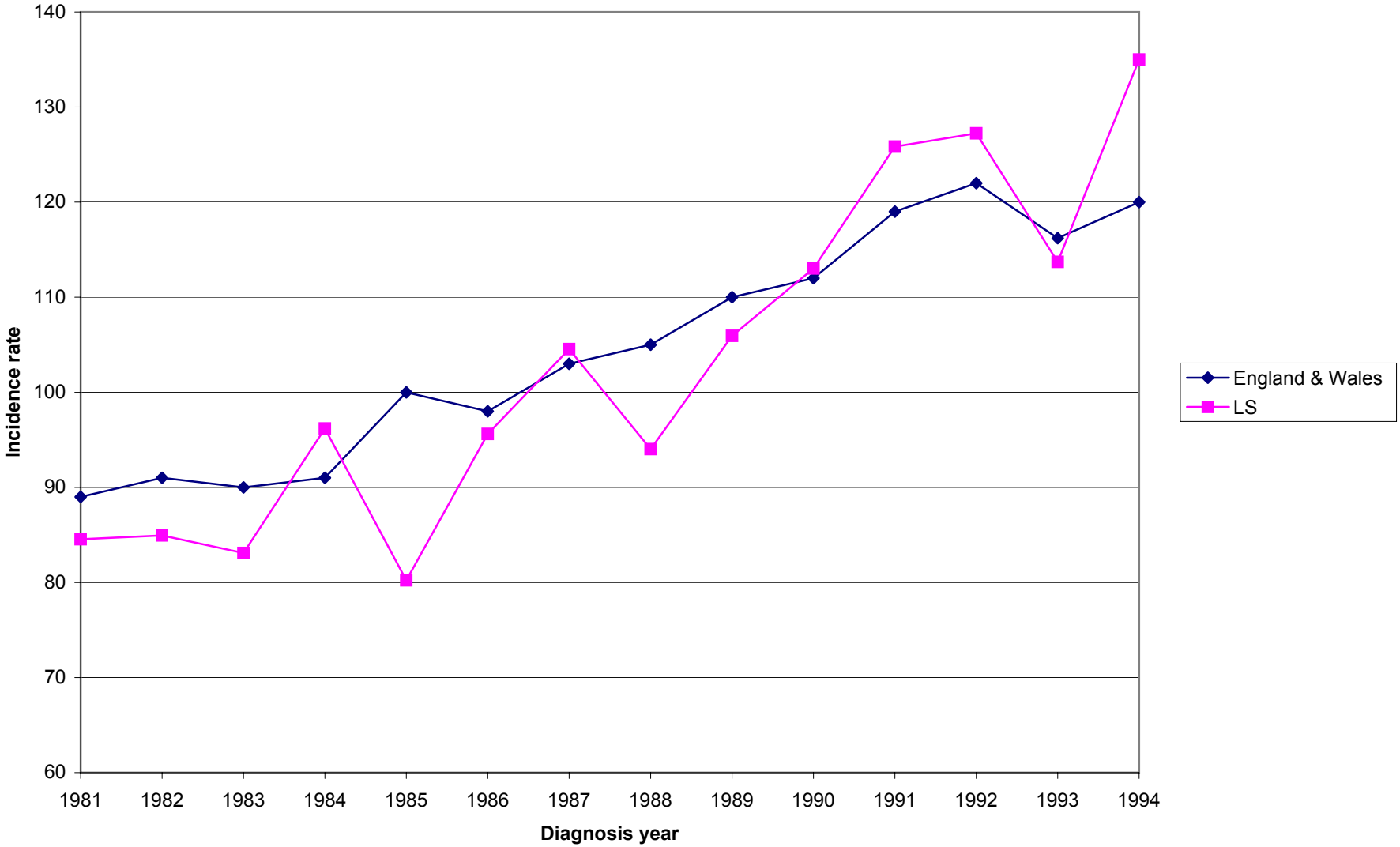

Chart 5. Prostate cancer (ICD185) incidence rates per 100,000