Control of data quality for population-based cancer survival analysis

Ruoran Li[1], Louise Abela[1], Jonathan Moore[1], Laura M Woods[1], Ula Nur[1], Bernard Rachet[1], Michel P Coleman[1]

[1] Cancer Research UK Cancer Survival Group, London School of Hygiene and Tropical Medicine, London, UK

Correspondence to: Ms Ruoran Li, Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT UK

Email address: ruoran.li@lshtm.ac.uk Tel: +44 (0)20 7927 2855

This work was funded by a Cancer Research UK programme grant (C1336/A11700).

Article type: Original research articles

Abstract word count: 192

Word count: 3073

References: 29

Tables: 5

Figures: 2

Abstract

Background

Population-based cancer survival is one of the important measures of the overall effectiveness of cancer care in a population. Population-based cancer registries collect data that enable the estimation of cancer survival. To ensure accurate, consistent and comparable survival estimates, strict control of data quality is required before the survival analyses are carried out. In this paper, we present a basis for data quality control for cancer survival.

Methods

We propose three distinct phases for the quality control. Firstly, each individual variable within a given record is examined to identify departures from the study protocol; secondly, each record is checked and excluded if it is ineligible or logically incoherent for analysis; lastly, the distributions of key characteristics in the whole dataset are examined for their plausibility.

Results

Data for patients diagnosed with bladder cancer in England between 1991 and 2010 are used as an example to aid the interpretation of the differences in data quality. The effect of different aspects of data quality on survival estimates is discussed.

Conclusions

We recommend that the results of data quality procedures should be reported together with the findings from survival analysis, to facilitate their interpretation.

Keywords: *Cancer, Survival, Research design, Population register, Epidemiology*

1 Introduction

Population-based cancer survival is one of the important measures of the overall effectiveness of cancer care and control in a population, alongside incidence and mortality. Trends in cancer survival provide further indication of improvements in diagnosis and treatment.[1]

Standard checks required for cancer incidence data have been described[2-4] and are embodied in the widely used IARC Check program.[5] However, additional quality checks are required for survival analysis, as the completeness and validity of data on vital status (alive, dead or lost to follow up) and follow-up time of the patients become crucial.

The interpretation of survival comparisons between countries or populations (defined by calendar period, socio-economic status, race or ethnicity) relies on the thoroughness of quality control procedures, which ensure that incomplete, ineligible or incoherent tumour records are flagged and excluded. We describe a set of quality control procedures that have been applied to population-based data for several recent national and international studies of cancer survival.[6-9] This set of procedures can form a basis for data quality control in cancer survival analysis.

2 Materials and Methods

2.1 Cancer registry data

Cancer registries collate data from sources such as hospitals, general practitioners, pathology departments, cancer referral units and screening programmes, and obtain one record for each tumours including patient demographic (date of birth, sex, residence or postcode, ethnicity, patient identifier), tumour (date of diagnosis, topography, morphology, behaviour, microscopic confirmation, stage at diagnosis), treatment (surgical procedure, chemotherapy, radiotherapy) and outcome (date and place of death) data. [10,11] This process may not be completed for six to nine months, until a patient's course of treatment has finished.

Information on the patient's vital status is later added from sources such as the regional or national death indexes, social security, health insurance, death certificates, physician or hospital contacts and/or home visits. The key concern is that the eventual deaths of all registered cancer patients are recorded. The quality and completeness of this information is essential for accurate estimation of survival.

2.2 Defining the cancers

Cancers are defined by their anatomic location (site) and microscopic appearance (morphology), and whether they are benign, *in situ*, malignant or of uncertain behaviour (behaviour), under the International Classification of Diseases[12] or the International Classification of Disease for Oncology.[13] Various utilities exist to convert ICD codes between the various revisions.[14,15]

In what follows, we write from the perspective of a general cancer registry, with data on all cancers.

2.3 Quality control

Quality control procedures are designed to ensure that survival analyses include only patients resident in the defined territory who were diagnosed with a primary, invasive, malignant neoplasm during a defined calendar period, and whose tumour record is valid and logically coherent.[16]

We propose three distinct phases for the quality control of cancer data for survival analysis (Figure 1). In the following sections, we will describe the rationale and process for each of these phases with accompanying examples. As any data quality control process, feedback is provided to the data sources, i.e. the registries, which will result to data checks and may lead to modifications. In a study involving several registries, quality control would entail discussion between the analytic centre and the registry concerned.

Phase 1: Protocol adherence (variables)

Are the individual variables within a given record compliant with protocol? A protocol specifies all permissible values for each variable,[9] such as last known vital status: alive=1, dead=2, lost to follow-up=3, unknown=9 (Table 1), or that the month of a date is in the range 1-12. Protocol adherence involves checking each variable in each record to confirm that its value falls within the specified range, and tabulating the number and proportion of variables that meet the protocol definitions (Table 2). Records containing variables that are not compliant with protocol should be reviewed for correction or re-coding. Data sets with substantial proportions of error will require further detailed checks by the cancer registry concerned.

Phase 2: Eligibility and exclusion (records)

Are the variables in each record eligible and logically coherent for analysis? We recommend a two-stage selection process (Table 3).

Tumour records in the raw data are first selected as eligible for analysis only if they are for an invasive, primary, malignant neoplasm diagnosed during the period defined for analysis in a patient who was resident in the territory covered by the registry. *In situ*, non-malignant or secondary tumours should be excluded.

Next, each eligible tumour record is checked for internal logical coherence and validity for inclusion in survival analyses. Such checks include that the day, month and year of each date are coherent, the sequence of dates is plausible (e.g. diagnosis precedes or is equal to the date of death), that the vital status and sex are both known, and that the cancer was not registered only from a death certificate (death-certificate-only, or DCO) or from an autopsy. Duplicate registrations, synchronous tumours and second (third, etc.) primary cancers (often referred to as multiple primary tumours) at the same anatomic site are also excluded. However, we recommend retention in the analyses of eligible multiple primary tumours which are not at the same anatomic site as an earlier tumour.[17,18]

From a practical perspective, automated programs embodying the criteria mentioned above are applied. Each tumour record is checked against both the ineligibility and exclusion criteria, and assigned one or more error flags, as applicable. All records that fail one or more criteria are then excluded from the data in a defined sequence of descending severity, applying the most basic reasons for exclusion first. Counts are made of the number of tumour records that fail each criterion, and a separate count is made of the number and proportion of patients excluded from the data on the basis of each criterion. The results can be presented in standard tables to facilitate examination of data quality. The tables show the total number of records in the raw data, the number that remain after removal of ineligible records, the number (and proportion of eligible patients) excluded because the tumour record failed one or more criteria, and finally the number of patients whose data can be included in survival analysis. The results of this process would lead to review and revision of the data if errors are confirmed. An example is shown in Table 4, based on the data preparation for a recent analysis to produce the official National Statistics for cancer survival in England.[19]

It is particularly important to check the dates and their sequence in each tumour record. Complete dates (day, month, year) should be used in survival analysis, because estimates of survival are otherwise biased, particularly for short-term survival.[20] Individual dates should be checked for validity (e.g. 31 February is invalid). The sequence of the dates of birth, diagnosis and last known vital status must also be logically coherent (see Table 3). Records with the date of diagnosis outside the predefined range should be excluded. Similarly, records with the date of last known vital status after the predefined end of follow-up (and before the date of data extraction) should be censored as alive at the date of end of follow-up. The distribution of the day and month of each date should also be examined. For example, peaks of distribution of certain values (e.g. 15 for days) reflect high proportion of imputed dates.

Phase 3: Distribution of key characteristics – editorial tables (data sets)

Editorial tables are used to examine aspects of data quality in the data file as a whole. For example, one should examine the number and proportion of DCO registrations over time and the distribution of cancers by deprivation or ethnicity over time. These tables are useful in examining the data for a single registry, but also in comparing the data sets for many registries (Table 5).

The basic distributions of the key variables in the records are analysed, such as the counts of cases by year of diagnosis, distribution of DCOs by age or time, or the proportion of tumours with morphological verification and the actual distribution of morphology. Editorial tables permit greater visual scrutiny of the data, such as differences in the proportion of DCO by age, time or socio-economic status. This enables comparisons between different deprivation categories, years and cancer registries.

Exclusion and editorial tables are shared with the registry, both to help identify improbable distributions of variables, and to document trends in data quality over time. For studies with more than one cancer registry, such tables provide valuable comparative information.

3 Results

Data quality control processes help to shed light on observed survival differences between geographical regions and over time periods. Differences in the proportion of tumour records that were eligible for survival analyses could reflect differences in data quality or in diagnostic and coding practices.

3.1 Interpretation of differences in data quality

An English dataset including bladder cancer patients diagnosed between 1991 and 2010 is used as an example. Recommendations to exclude some papillary tumours of the bladder that would previously have been classified as invasive were implemented by UK cancer registries, for tumours registered from 2000.[21,22] Patients with urothelial papillary tumours with high survival who would have been included in survival analyses prior to 2000 may then have been excluded for analyses if diagnosed after 2000.

This change would not be reflected at the protocol adherence phase, as there is no change in the range of morphology codes for bladder cancer. However, the change in coding practice can be faithfully represented in the exclusion table (Table 4).; the proportion of registered patients with benign or uncertain tumours of the bladder increased dramatically from 2.8% for patients diagnosed in 1991-1995 to 29.3% in 2006-2010, which leads to an increase in the proportion of ineligible patients from 10.6% to 45.1%.

This was also clear in the tabulation of new cases by morphology and year of diagnosis in editorial tables. A drastic drop in number of patients diagnosed with invasive bladder cancer with morphology codes 8120 and 8130 in the year 2000 was observed, as the result of the change in the coding practice (Figure 2).

This change would produce an artificial downward trend in bladder cancer survival, without any real change in survival times for patients with genuinely invasive malignancy. It would produce a downward trend in bladder cancer incidence, but would not influence observed mortality. Regional variations in survival and incidence within England may also be explained by this change in practice; the change in coding practice happened gradually in the different regional cancer registries between 1986 and 1999.[23] Data quality control processes can thus highlight potential artefacts in the data that can inform the interpretation of survival estimates.

Potential differences and changes in the quality of cancer registration can also be evaluated. For bladder cancer in England, the number of DCO registrations decreased from 4.3% in 1991 to 1.8% in 2010 (Table 4), which reflects an improvement in the quality of cancer registration in England. Similarly, data quality difference between registries can be assessed by the completeness and validity of vital statistics information and by the proportion of records with DCOs.

Other factors can also influence the comparability and continuity of registration data for survival analyses. For example, introduction of screening programmes (*e.g.* for breast cancer) allows the detection of a high proportion of low stage cancers, which would result in an increase in cancer survival estimates. The exclusion and editorial output from the data quality control process would offer an insight to these changes. Survival trends should also be interpreted alongside trends in incidence and mortality.

4 Discussion

This paper provides an overview of the data quality control methods currently used by the Cancer Research UK Cancer Survival Group to prepare population-based cancer registry data for the estimation of cancer survival. We recommend application of strict data quality control procedures to ensure internally valid and externally comparable survival estimates. This monitoring of quality control methods is a continuous process.[5] It involves routine checking for validity and consistency, and maintenance and updating of the cleaning programmes that are used to identify and flag inconsistencies or possible errors, and to present the results in suitable tables and graphics.

It is impossible to be completely prescriptive about all the quality-control tests that should be conducted before analysis of survival in a given data set. For example, if the analyses will involve examination of survival by stage at diagnosis, it will be necessary to perform tests of the completeness and validity of the data on stage, and perhaps to perform multiple imputation for missing values of stage. The range of tests to be performed will also depend on the variables that are collected by the cancer registry concerned; and in the case of a comparative analysis involving several registries, on the variables included in the study protocol.

However, several key variables required for population-based survival analyses must be completely and accurately recorded in the registry for accurate estimation of survival. For example, if follow-up for patient's vital status is not complete and deaths are not all recorded properly (error in the 'vital status' variable), patients may become 'immortals' (Table 5) and over-estimation of survival would occur. In this case, a cleaning process which allows the identification of probable 'immortals' would be essential for estimating the scale of the problem. Ideally, full dates of birth, dates of diagnosis and dates of follow-up should always be used to ensure complete data assessment and unbiased survival estimation.[20]

By contrast, we recommend inclusion in survival analyses of patients who died on the same day as the diagnosis of their cancer. It may be necessary to assume for these patients that death occurred one day after diagnosis, if the statistical software cannot deal with zero survival time. Excluding such observations would artificially over-estimate survival.

Cancer registrations based solely on a death certificate (death-certificate-only or DCO registrations) are assigned to the date of death for the purposes of cancer incidence, but they cannot be included in survival analyses[24] because the duration of survival is unknown. If DCOs represent a high proportion of all registered cases, this suggests under ascertainment of incident cases. If the true (but unknown) duration of survival for patients registered as a DCO is shorter than the average, a high proportion of DCO cases may also lead to over-estimation of survival.[25] The proportion of DCOs will be zero in countries where death certificates are not used to initiate a new cancer registration or where access to the cause of death is not legal; this may give rise to some under registration of incident cases.

The choice of whether to include the second, third (etc.) tumour in a given person (multiple primary) in survival analyses will affect the interpretation of results. We recommend excluding multiple primaries at the same anatomic site. Including multiple primaries at the same site would permit inclusion of two deaths for a single person in the same survival analysis of a type of malignancy (typically define by ICD-O topography codes). Multiples with different morphology within the same organ remain rare. It is statistically feasible to allow the same person to contribute two events (e.g. episodes of influenza) in a cohort analysis of endpoints. However, in practice, because morphology remains missing or is too general in high proportions in many registries, it seems safer to exclude all multiples at the same site for the ease of comparison between registries and over time.

By contrast, it is generally advisable to include a person with two malignancies that have occurred at different anatomic sites in the analysis of survival for each of the sites. Including multiple primaries at different sites reduces the bias in comparison of survival between registries due to different observation periods, age, registration quality and completeness of registration.[18] For example, if we were to exclude them, a subsequent cancer of a patient would be excluded in a registry with records of the first cancer, while it would be treated as the first cancer in a younger registry and included in survival analyses; this would bias survival comparisons. Including multiple primary tumours also avoids the conceptual difficulties that arise from the definition of multiple primary malignancy, which differs widely between the two main sets of international rules (SEER and IARC).[26,27] The general effect of inclusion is to reduce survival estimates by a variable amount depending on the proportion of multiple primaries, the cancer site, and the extent to which survival for subsequent tumours is shorter than for first primary tumours.[17] One final caveat, which is that if we are to analyse survival from all cancers combined, in which we pool the data from more than one anatomic site as typically defined, then it would again become inappropriate to include a single person more than once in the analyses, which should then be confined to first primary malignancies.

We recommend consistent application of data quality checks in preparing population-based data for the estimation of cancer survival, in order to ensure accuracy, consistency and comparability of the estimates. The data quality assurance procedures should be reported when presenting the results of survival analyses, in order to facilitate their interpretation.

Conflict of Interest Statement

None declared.

Ethical Approval

The study protocol and the Cancer Research UK Cancer Survival Group's system-level security policy at LSHTM were approved by the Ethics and Confidentiality Committee of the National Information Governance Board on 21 April 2010 (extension of PIAG 1-05(c)2007).

Acknowledgements

This work was funded by a Cancer Research UK programme grant (C1336/A11700). We thank the Office for National Statistics and the regional cancer registries in England for providing the data.

Role of the Funding Source

None.

References

1. Rachet B, Maringe C, Nur U, Quaresma M, Shah A, Woods LM, Ellis L, Walters S, Forman D, Steward JA, Coleman MP. Population-based cancer survival trends in England and Wales up to 2007: an assessment of the NHS cancer plan for England. *Lancet Oncol* 2009; **10**: 351-69
2. Parkin DM, Chen VW, Ferlay J, Galceran J, Storm HH, Whelan SL. *Comparability and quality control in cancer registration. IARC Technical Report No. 19*. Lyon: IARC; 1994
3. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. *Eur J Cancer* 2009; **45**: 747-55
4. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods Part II. Completeness. *Eur J Cancer* 2009; **45**: 756-64
5. Ferlay J, Burkhard C, Whelan S, Parkin DM. *Check and conversions programs for cancer registries IARC / IACR Tools for Cancer Registries*. Lyon, International Association for Research on Cancer, 2005.
6. Coleman MP, Rachet B, Woods LM, Mitry E, Riga M, Cooper N, Quinn MJ, Brenner H, Estève J. Trends and socio-economic inequalities in cancer survival in England and Wales up to 2001. *Br J Cancer* 2004; **90**: 1367-73
7. Woods LM, Rachet B, Shack LG, Catney D, Walsh PC, Cooper N, White C, Mak V, Steward JA, Comber H, Gavin AT, Brewster DH, Quinn MJ, Coleman MP, UK Association of Cancer Registries. Survival from twenty common adult cancers in the United Kingdom and the Republic of Ireland during the late twentieth century. *Health Stat Quart* 2010; **46**: 7-26
8. Coleman MP, Quaresma M, Berrino F, Lutz J-M, De Angelis R, Capocaccia R, Baili P, Rachet B, Gatta G, Hakulinen T, Micheli A, Sant M, Weir HK, Elwood JM, Tsukuma H, Koifman S, Azevedo e Silva G, Francisci S, Santaquilani M, Verdecchia A, Storm HH, Young JL, CONCORD Working Group. Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncol* 2008; **9**: 730-56
9. Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Maringe C, Nur U, Tracey E, Coory M, Hatcher J, McGahan CE, Turner D, Marrett L, Gjerstorff ML, Johannesen TB, Adolfsson J, Lambe M, Lawrence G, Meechan D, Morris EJ, Middleton R, Steward J, Richards MA, ICBP Module 1 Working Group. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. *Lancet* 2011; **377**: 127-38
10. Office for National Statistics. *Cancer statistics: registrations of cancer diagnosed in 2009, England. Series MB1 no. 40*. Newport, Office for National Statistics, 2011.
11. Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, eds. *Cancer registration: principles and methods. (IARC Scientific Publications No. 95)*. Lyon: International Agency for Research on Cancer; 1991
12. World Health Organisation. *International statistical classification of diseases and related health problems. Tenth revision*. Geneva: WHO; 1994
13. Fritz AG, Percy C, Jack A, Shanmugaratnam K, Sobin LH, Parkin DM, Whelan SL, eds. *International Classification of Diseases for Oncology (ICD-O)*. 3rd edn. Geneva: World Health Organisation; 2000
14. ICD Conversion Programs. *Surveillance Epidemiology and End Results (SEER)* 5 Jun 2013 http://seer.cancer.gov/tools/conversion/
15. IARCcrgTools Conversion programs. *IARC* 5 Jun 2013 http://www.iacr.com.fr/iarccrgtools.htm
16. Coleman MP, Babb P, Damiecki P, Grosclaude PC, Honjo S, Jones J, Knerer G, Pitard A, Quinn MJ, Sloggett A, De Stavola BL. *Cancer survival trends in England and Wales 1971-1995: deprivation and NHS Region. (Studies on Medical and Population Subjects No. 61)*. London: The Stationery Office; 1999, pp1-695.

17. Brenner H, Hakulinen T. Patients with previous cancer should not be excluded in international comparative cancer survival studies. *Int J Cancer* 2007; **121**: 2274-8
18. Rosso S, De Angelis R, Ciccolallo L, Carrani E, Soerjomataram I, Grande E, Zigon G, Brenner H, EUROCARE Working Group. Multiple tumours in survival estimates. *Eur J Cancer* 2009; **45 (Suppl. 6)**: 1080-94
19. Abela L, Rachet B, Whitehead S, Messer J, Coleman MP. Cancer Survival in England: Patients Diagnosed, 2006–2010 and Followed up to 2011. *Office for National Statistics* 23 Oct 2012
20. Woods LM, Rachet B, Ellis L, Coleman MP. Full dates (day, month, year) should be used in population-based cancer survival studies. *Int J Cancer* 2012; **131**: E1120-E1124
21. Pheby DFH, Martinez-Garcia C, Roumagnac M, Schouten LJ. Recommendations for coding bladder tumours. *European Network of Cancer Registries* 1995, last accessed 14 February 2008. www.encr.com.fr/workgr1C.htm
22. UK Association of Cancer Registries. Library of recommendations on cancer coding and classification policy and practice: bladder cancer. *UKACR* 2004, Cambridge, last accessed 14 February 2008. www.ukacr.org/codingpolicy/po9901.asp
23. Shah A, Rachet B, Mitry E, Cooper N, Brown CM, Coleman MP. Survival from bladder cancer in England and Wales up to 2001. *Br J Cancer* 2008; **99 (Suppl. 1)**: 86-9
24. Capocaccia R, Gatta G, Roazzi P, Carrani E, Santaquilani M, De Angelis R, Tavilla A, EUROCARE Working Group. The EUROCARE-3 database: methodology of data collection, standardisation, quality control and statistical analysis. *Ann Oncol* 2003; **14 (Suppl. 5)**: 14-27
25. Brenner H, Hakulinen T. Implications of incomplete registration of deaths on long-term survival estimates from population-based cancer registries. *Int J Cancer* 2009; **125**: 432-7
26. Multiple Primary and Histology Coding Rules Manual. *Surveillance Epidemiology and End Results (SEER)* 5 Jun 2013 http://seer.cancer.gov/tools/mphrules/download.html
27. IARC. International Rules for Multiple Primary Cancers (ICD-O Third Edition). *IARC* 2004 http://www.iacr.com.fr/iarccrgtools.htm
28. Rachet B, Woods LM, Mitry E, Riga M, Cooper C, Quinn MJ, Steward JA, Brenner H, Estève J, Sullivan R, Coleman MP. Cancer survival in England and Wales at the end of the 20th century. *Br J Cancer* 2008; **99 (Suppl. 1)**: 2-10
29. Berg JW. Morphologic classification of human cancer. In: Schottenfeld D, Fraumeni JF, eds. *Cancer epidemiology and prevention*. 2 ed. New York: Oxford University Press; 1996, pp28-44.

List of Tables and Figures

Table 1

Short description	Type[a]	No. of digits or characters	Valid values (or range of valid values)	Value to be used when valid data are missing
Unique ID[b]	A	Depending on the source cancer registry		Not allowed
Sex	N	1	1,2	9
Day of birth	N	1 or 2	1-31	99
Month of birth	N	1 or 2	1-12	99
Year of birth[c]	N	4	1895-2010	9999
Day of diagnosis	N	1 or 2	1-31	99
Month of diagnosis	N	1 or 2	1-12	99
Year of diagnosis[c]	N	4	1995-2010	9999
Last known vital status[d]	N	1	1,2,3	9
Day of last known vital status	N	1 or 2	1-31	99
Month of last known vital status	N	1 or 2	1-12	99
Year of last known vital status[c]	N	4	1995-2010	9999
ICD-O-3[e] Topography	A	4	C00.0-C80.9	Not allowed
ICD-O-3[e] Morphology	N	4	8000-9989	9999
Behaviour[f]	N	1	0,1,2,3,6,9	Not allowed

[a] A - Alphanumeric; N - Numeric

[b] Recognised only within the cancer registry (to enable correction of errors).

[c] The valid values for these variables depends on the eligible years of diagnosis, years of follow-up and age of patient in a predefined study protocol. Here, the date of diagnosis and follow-up are defined to be between 1st of January 1995 and 31st December 2010 and that patients under 100 years of age are eligible for the study.

[d] Last known vital status: 1 - alive, 2 - dead, 3 - lost to follow-up, 9 - unknown

[e] Anatomic site and morphological type of neoplasm, coded to: World Health Organisation. International Classification of Diseases for Oncology (ICD-O). In: Fritz AG, Percy C, Jack A, Shanmugaratnam K, Sobin LH, Parkin DM, Whelan SL, eds., 3rd ed. Geneva: World Health Organisation, 2000.

[f] Behaviour of neoplasm: 0 - Benign; 1 - Uncertain whether benign or malignant; 2 - Malignant carcinoma in situ; 3 - Malignant, primary site; 6 - Malignant, metastatic site; 9 - Malignant, uncertain whether primary or metastatic site

Table 2

			Colorectal		Lung		Breast	
			Compliant		Compliant		Compliant	
Variable name	**Short description**	**Coding allowed**	**No.**	**%**	**No.**	**%**	**No.**	**%**
VAR1	Unique ID	Up to 15 alphanumeric	383,895	**100**	251,274	**100**	341,769	**100**
VAR2	Sex	1-digit = 1, 2, 9	383,895	**100**	251,274	**100**	341,769	**100**
VAR3	Day of birth	1-2 digit 1-31, 99	383,887	**>99**	251,269	**>99**	341,659	**>99**
VAR4	Month of birth	1-2 digit 1-12, 99	383,887	**>99**	251,269	**>99**	341,659	**>99**
VAR5	Year of birth	4-digit	383,887	**>99**	251,274	**100**	341,769	**100**

Table 3

Definition	Comment
Ineligible records	
Incomplete data	Incomplete data item(s) such as sex, date of birth, date of diagnosis (for non-death-certificate-only records), date of last known vital status, postcode, site, morphology, and behaviour [28]
Not belonging to the population of interest	Checked when the postcode information is added
In situ neoplasm	Behaviour code 2
Benign, or uncertain if benign or malignant	Behaviour code 0 or 1
Metastatic	Behaviour code 6 or 9
Otherwise ineligible	Tumour specific checks on ICD codes such as anatomic location, morphology or behaviour, specific to a particular malignancy
Lymphoma [a] in a solid organ	Morphology for lymphoma in a solid organ
Leukaemia [a] or myeloma in a solid organ	Leukaemia or myeloma in a solid organ
Exclusion criteria	
Aged 100+	If cases are aged 100 years or more at diagnosis
Vital status unknown	If vital status is not known by the 'freeze date' [b]
Sex not known	Sex code 9
Sex-site incompatibility	Sex-specific tumours not compatible with recorded sex
Invalid dates or invalid sequence of dates	Dates of birth, diagnosis, death or censoring do not correspond to a real date; or sequence of dates is impossible
Death certificate only (DCO)	Case identified only by death certificate or case identifies only by autopsy
Duplicate registration	Identified if records have the same site code, sex, personal identification number (or cancer registry number), and cancer registry as another registration
Synchronous tumours	Synchronous tumours at a single site are considered one cancer[24]. Further synchronous records can be identified and excluded if two records are associated with the same site code, sex, date of birth, date of diagnosis and/or other combinations of identifiable information
Multiple primary at the same site	Multiples may be identified if two records have the same personal identification number and are of the same site, but with different dates of diagnosis

[a] ICD-O-3 morphology: 9590-9999

[b] The freeze date of a database is the date after which the database no longer updates with new information.

Table 4

	1991-95		1996-2000		2001-05		2006-2010	
Total registered	**62,331**	**100.0**	**67,993**	**100.0**	**72,517**	**100.0**	**80,165**	**100.0**
Ineligible	***Patients***	***%***	***Patients***	***%***	***Patients***	***%***	***Patients***	***%***
Incomplete data	130	0.2	75	0.1	101	0.1	10	<0.1
Not resident in England	96	0.2	70	0.1	5	<0.1	2	<0.1
In situ neoplasm	4,573	7.3	11,006	16.2	11,255	15.5	12,621	15.7
Benign or uncertain	1,760	2.8	5,013	7.4	17,895	24.7	23,460	29.3
Metastatic	49	<0.1	31	<0.1	53	<0.1	65	<0.1
Otherwise ineligible[a]	11	<0.1	8	<0.1	0	0.0	0	0.0
Lymphoma[b]	0	0.0	0	0.0	0	0.0	0	0.0
Leukaemia or myeloma	0	0.0	0	0.0	0	0.0	0	0.0
Total ineligible [c]	**6,619**	**10.6**	**16,203**	**23.8**	**29,309**	**40.4**	**36,158**	**45.1**
Total eligible	**55,712**	**100.0**	**51,790**	**100.0**	**43,208**	**100.0**	**44,007**	**100.0**
Aged 100+	23	<0.1	28	<0.1	28	<0.1	41	<0.1
Vital status unknown	222	0.4	128	0.2	81	0.2	87	0.2
Sex not known	0	0.0	0	0.0	0	0.0	0	0.0
Sex-site error	0	0.0	0	0.0	0	0.0	0	0.0
Invalid dates	16	<0.1	2	<0.1	9	<0.1	1	<0.1
Death certificate only	2,387	4.3	1,688	3.3	1,064	2.5	790	1.8
Duplicate registration	0	0.0	0	0.0	0	0.0	0	0.0
Synchronous tumours	166	0.3	262	0.5	68	0.2	59	0.1
Multiple primary same site	80	0.1	90	0.2	94	0.2	128	0.3
Total exclusions [d]	**2,894**	**5.2**	**2,198**	**4.2**	**1,344**	**3.1**	**1,106**	**2.5**
Patients available for analyses	**52,818**	**94.8**	**49,592**	**95.8**	**41,864**	**96.9**	**42,901**	**97.5**

[a] Other criteria of anatomic location, morphology or behaviour, specific to a particular malignancy. In general, they refer to secondary malignancy at relevant site.

[b] Morphology for lymphoma in a solid organ excluded for survival analysis at the solid organ site. These cases of lymphomas would be included in the lymphoma analysis.

[c] Of total registered patients.

[d] Of total eligible patients.

Table 5

Type1	**Data quality**
	1) The distribution of the day and month (and year) of the date of birth, date of diagnosis and date of death (or lost to follow-up), which should be, at least for day and month, roughly uniform
	2) No. (%) of records representing multiple primaries at different sites by calendar year
	3) No. (%) of Death Certificate Only (DCO) registrations by cancer and year of registration
	4) No. (%) of records with morphological verification, by cancer and year of diagnosis
	5) No. (%) of records with implausible age or duration of survival by cancer and year of diagnosis[a]
Type 2	**Descriptive counts and proportion: (No./%)**
	6) Cancer by sex and year of diagnosis
	7) Cancer, deprivation and/or ethnicity and year of diagnosis
	8) Morphology group[b] by cancer, year and period of diagnosis

[a] Referred to as ‘immortals’. Depending on the study design, ‘immortals’ may be defined as patients aged 105 years or over who are not known to have died, or who have apparently survived five or more years from a highly lethal cancer, e.g. brain, oesophagus, stomach or pancreas. They should be defined prior to the analysis, and may be systematically excluded from the data.

[b] *Berg et al 1996* [29]