Office for National Statistics

Quality assurance of administrative data used in cancer registrations and cancer survival statistics

Investigation of the administrative data sources used in the production of cancer survival statistics for England. In partnership with Public Health England, we publish several statistical bulletins on cancer survival each year.

Contact:
Sophie John, John Broggio
cancer.newport@ons.gov.uk
+44 (0)1633 436935

Release date:
3 July 2019

Next release:
To be announced

Table of contents

1 . Introduction

The UK Statistics Authority, in accordance with the Statistics and Registration Service Act 2007 and signifying compliance with the Code of Practice for Statistics, has designated the following statistics as National Statistics:

- Cancer registration statistics, England
- Adult cancer survival in England
- Geographic patterns of cancer survival
- Index of cancer survival for Clinical Commissioning Groups in England

In addition, two Experimental Statistics also rely on the data covered in this report:

- Adult cancer survival by stage at diagnosis for England
- Cancer survival for children in England

Cancer statistics receive more than limited media interest but are not highly politically sensitive, so have been assessed as being of medium public interest and value. The data collection is more complex, coming from the NHS, Public Health England and the Office for National Statistics; potential areas of risks are monitored and checks that are detailed below (and in the cancer QMIs) are developed to reduce the level of risk, resulting in a medium level of data quality concern.

Applying these ratings to the risk matrix produced by the Office for Statistical Regulation for evaluating quality assurance and audit arrangements, the cancer statistics are considered to be of medium risk.

An informal discussion of the contents and uses of these publications is available in a user guide as well as more technical Quality and Methodology Information reports on cancer registrations, the index of cancer survival and the remaining cancer survival bulletins.

2 . Operational context and administrative data collection

For the cancer registrations publication, there are two sources of data used: mid-year population estimates, published by the Office for National Statistics (ONS) and cancer registration data, which are collected by Public Health England (PHE). A Quality and Methodology Information (QMI) report on the population estimates and a peer-reviewed Data Resource Profile (DRP) of the cancer registration data are available.

The cancer registration data are also used in every cancer survival publication. For every cancer survival publication, a check on the vital status of each cancer patient is made through NHS Digital's Personal Demographics Service (PDS). National population data, that combine mortality data and mid-year population estimates into life tables, are used in every cancer survival publication for adults. There is quality information published in the mortality QMI and life tables QMI and a briefing document on the PDS is available on request from NHS Digital.

Figure 1 shows a summary of the administrative data flows that are captured and utilised in the statistical outputs on cancer registrations and survival. Cancer registration data is used in all cancer publications. National population data is used in the cancer registration publication and all cancer survival publication for adults. For all cancer survival publications, the PDS is used to confirm the vital status of each cancer patient.

Figure 1: Summary administrative data flows for patient-level data

Classifications and definitions used in all cancer publications can be found in the Concepts and definitions sections of the cancer registrations QMI, cancer survival QMI and the index of cancer survival QMI.

3 . Accuracy and quality of data and estimates

In every data source, there are potential sources of bias and error that could materially impact on the estimates produced. These are discussed in more detail in their respective Quality and Methodology Information (QMI) reports (population estimates QMI, mortality QMI and life tables QMI), data resource profile (cancer registration data) and briefing document (NHS Digital's Personal Demographics Service PDS)).

Table 1 identifies the main potential sources of bias and error together with actions taken to minimise the risks to data quality.

Table 1: Potential sources of bias and error in the administrative data and mitigating steps taken to minimise the risks to data quality

Potential source of bias and error	Safeguards taken to minimise the risks to data quality
The Personal Demographics Service (PDS) of NHS Digital may not be notified in a timely manner of patients having embarked England or having died.	Each survival publication that uses the vital status returns from the PDS measures up to the end of a calendar year. To reduce the likelihood of a delayed notification of death or embarkation, the patient files are sent to the PDS at least three weeks into the new calendar year.
Corrupted identifiers may prevent the PDS data and the cancer registration data from linking.	Internal validation procedures in NHS Digital and Public Health England (PHE) are applied to identifiers and the number of patients returned unmatched are monitored. Cancer patients in the cancer registry are also routinely matched to other third-party datasets, including historical ONS cancer incidence records to test the validity of historical identifiers.
Cancer registrations are known not to be complete until five years following the diagnosis date in up to 2% of cancers.	The completeness of cancer registrations is monitored for both the level of completeness and for any patterns in the types of cancer that are not registered in a timely manner.
Survival statistics can only consider people who have sought medical help with their cancer and so have a diagnosis before death.	Trends in the levels of patients whose cancer is only found on their death are monitored.
Cancer registration data items that define the cancer of a patient may contain errors, for example, uncertain dates that define the age at diagnosis of patients.	There are extensive internal quality assurance (QA) procedures applied in the registration processes. A further set of survival QA tests are applied to each cohort of cancer patients that are being considered for survival analysis – see QA principles, standards and checks applied by data suppliers for a list of these checks.
Coding errors in extracting and linking the data sources; coding errors in producing the survival estimates.	The SQL and statistical code is written and executed independently by two teams to ensure the data and outputs agree. Further comparisons to check for consistency with previous estimates are made.
Life tables used in adult cancer survival publications reflect registered deaths of people in England and Wales; the populations used are estimates and may need adjusting if net migration patterns alter significantly from predictions.	Patients who were diagnosed in England but move and die abroad are censored from cancer survival calculations to prevent biasing of the estimates. When population estimates require revision, a revised back series of survival estimates is produced for users.

For patients whose death or embarkation from England is not notified in a sufficiently timely manner to PDS, they may inflate survival estimates. If corrupted identifiers prevent cancer registration data from matching PDS data, then deaths or embarkations that should be captured in the survival estimates may be missed; each patient with corrupted identifiers will have impact the survival estimates in the same way as late notifications of deaths or embarkations.

Each individual instance of late notification or corrupted identifiers will have a very small impact on the cancer survival estimates produced; monitoring of the returns from PDS over successive years shows that there are very few late registrations of deaths or leaving England. The cancer survival estimates reported at each geographical breakdown include a sufficient volume of diagnoses that these increases will not be statistically significant and will be undetectable at the level of reported precision of the estimates.

The delay in arriving at a confirmed cancer registration is assessed in the cancer registrations bulletin and shows similar levels, slightly decreasing over time, in each reported diagnosis year. Analysis of the cancer registrations that are untimely does not reveal biases in the cancers, age at diagnosis (and death), sex or residential area of patients; this lack of bias means it is unlikely that published estimates are significantly impacted.

Cancer registrations are over 98% complete when used in the cancer bulletins; the potential accuracy gained from waiting for 100% completeness (another four years) is less than the need for timely reporting of cancer statistics.

The cancer survival estimates are designed to reflect the effectiveness of the healthcare system in treating diagnosed patients. Including people in survival estimates who die with cancer that is only detected after death (called Death Certificate Only registrations or DCOs) would artificially depress survival estimates who may never have needed or sought treatment for cancer.

For people with undiagnosed cancer at death, they may be reflected in the cancer registration bulletin if their cancer is detectable after death in a post mortem. The rates of people in this position are monitored and are reported as part of the UK and Ireland Association of Cancer Registries Key Performance Indicators.

Errors or inconsistencies in the cancer registration data can cause bias in the estimates produced that alter the age distributions observed and because of this, the rates of cancer registrations reported. The procedures in registering cancer diagnoses discussed in the Data Resource Profile address how these errors are mitigated.

The procedures followed by PHE are the result of international agreements in how to code and record cancer, with coding systems and guidance being issued by the International Agency for Research on Cancer (the cancer agency of the World Health Organisation) and bodies specialising in classifying aspects of a cancer diagnosis (for example, the Union for International Cancer Control's cancer staging system). To promote consistent reporting of cancer registration data to PHE, PHE has authored a Cancer Outcomes and Services Dataset (COSD) which has been the standard for reporting of cancer in the NHS since January 2013.

Errors or inconsistencies in the data may mean that the time a patient with cancer is recorded as living following a diagnosis is either too long (will inflate survival estimates) or too short (will decrease survival estimates).

In adult cancer survival estimates, the cancer registration data can be correct and consistent but the age of the patient at diagnosis falls outside the range for which the life tables are reliable. Because of the statistical techniques applied in producing survival estimates, these estimates are more sensitive than rates of registration and an enhanced set of quality controls on the data are applied to reduce the bias of the survival estimates (see Producer's quality assurance investigations & documentation).

Each individual error will have a low impact on the estimates produced; the checks on the data are designed to prevent many individual errors from being included, which may manifest in an error with a larger impact.

Coding errors can have a significant impact on the accuracy and quality of the estimates produced. The PHE team that extract and link the data used in the production of the cancer bulletins are trained in SQL, and follow documented extraction procedures. The extraction team works closely with their colleagues that develop the databases used and the cancer registration protocols so that changes to internal and external data recording is reflected in the data extracts produced. Two extracts are independently coded and produced; the value of each data item for each cancer diagnosis must match before this step is signed off.

The PHE team is also trained in applying and developing survival analysis techniques. Two sets of independently written code are run to produce the estimates in the bulletins and follow documented procedures; each estimate produced from the two sets of runs must match before the estimates can be considered for consistency checking against previous publications and international comparisons.

After applying consistency checks (for example, rarer cancers have a lower number of diagnoses than common cancers, cancers with generally poor prognosis have lower survival estimates than ones with good prognosis), the estimates are then shared with senior members of the PHE and ONS teams for independent scrutiny against international and previous estimates, before jointly agreeing the new estimates produced are fit for publication.

For the cancer survival in adult bulletins, even small changes to the life tables have a very large and significant impact on all the estimates produced because of the statistical techniques used. The number of people registered as having died at a given age and location in a year may be incorrectly recorded or the population estimates may need significant revision.

If these errors cause the mortality rates in the life tables to increase, then the survival estimates produced will also increase; conversely, if errors cause the mortality rates in the life tables to fall, then the survival estimates produced will also fall. These effects will be larger for 5-year survival estimates than 1-year survival estimates and larger for cancers with good prognosis than cancers with poor prognosis.

When the data underlying the life tables need revising, the team produces a revised back series of survival estimates to allow users to assess trends on a consistent basis.

4 . Communication with data supply partners

The provisions of Section 251 of the NHS Act 2006 provide the legal basis to PHE for collecting patient-level data on cancer patients for specified purposes, without consent. This is reviewed annually by the Confidentiality Advisory Group of the Health Research Authority. Strict technical and contractual controls are put in place to prevent unauthorised access and use of the data, with staff undergoing regular training on data protection and information governance.

Secure file transfer systems are used to send and receive patient-level data between the NHS and PHE. Within PHE, all patient-level data are stored on secure servers with role-based access. Only aggregated data are shared between the ONS and PHE to produce the bulletins and are also transferred via a secure file transfer system.

All statistical outputs transferred between the ONS and PHE are produced by two teams working independently to produce the output. When there is internal agreement, the output is shared between the ONS and PHE. Senior members of the ONS and PHE teams then meet to discuss the outputs and agree that they are fit for publication.

During this process, weekly meetings are held between the ONS and PHE to discuss progress towards reaching agreed statistical outputs. These are supplemented by meetings internal to each team and joint workshops to review previous outputs and plan future outputs.

The combined ONS and PHE team has responded to concerns raised about the statistical outputs created and have issued corrections to both cancer registrations and cancer survival outputs. They have also self-identified concerns in the appropriateness of older data sources before publication, which drove the development and publication of life tables by the ONS to support the calculation of cancer survival estimates in adults.

In the publication of the most recent Index of cancer survival bulletin, the combined team issued a survey for users to complete. The survey responses will help inform a user event to be held in autumn 2019 to discuss how the cancer outputs can remain relevant to users.

The data items shared by PHE to the NHS PDS to facilitate the checking of patients' vital status are:

- NHS number
- date of birth
- postcode of residence

The data items returned by the NHS PDS are:

- local patient identifier
- trace result NHS number
- date of death
- old NHS number
- new NHS number
- returned current posting
- date field last modified
- returned date of current posting or date of death, if deceased

The data used by the PDS are updated by trained NHS staff who work in healthcare settings and the database (the Spine) it stores the data in applies validation checks when records are attempted to be updated by a user.

The fields extracted by the PHE analysis team for producing cancer registration or cancer survival estimates are:

- NHS number
- tumour number
- sex
- date of birth
- date of diagnosis
- date of death, if deceased
- vital status
- vital status date
- ICD-10 topographical code
- ICD-O-2 morphological code
- ICD-O-2 behaviour code
- DCO flag
- stage at diagnosis
- deprivation quintile
- country of residence code
- region of residence code
- NHS England Region of residence code
- Cancer Alliance of residence code
- Sustainability and Transformation Partnership of residence code
- Clinical Commissioning Group of residence code

The geographical codes are obtained from the tumour table in the cancer registration tables that are linked to the National Statistics Postcode Lookup file in the creation of the cancer registration tables.

The life table data transferred from the ONS to PHE comprises:

- year
- age
- sex
- region
- deprivation quintile
- mortality rate
- probability of surviving one-year

The statistical output data transferred from PHE to the ONS is comprised of:

- diagnosis years
- age group
- sex
- method(s) of compiling the output
- regional codes appropriate to the publication
- cancer sites and their definition
- stage at diagnosis (survival by stage at diagnosis output only)
- standardised and un-standardised output
- smoothed output (childhood cancer survival output only)
- measures of uncertainty – confidence intervals (all survival output) or precision (Index of cancer survival only)

5 . Quality assurance principles, standards and checks applied by data suppliers

The system NHS Digital uses for its PDS contains data entry validations that are enforced every time a user of their system updates a record. The returns provided to users of the PDS contains coding information to indicate where (and why) matches could not be found, so that users can check the data they hold and enable a dialogue to resolve any concerns that the users may have.

PHE monitors the returns of cases to investigate and correct the records of cancer registrations where the PDS has indicated there may be a problem; almost all of these are from cancer registrations that occurred more than 20 years ago and, consequently, would be not included for reporting upon in the statistical bulletins.

The life tables published by the ONS and all the statistical outputs on cancer published by the ONS in partnership with PHE have been run by two independent teams, the first step of which is to ensure the raw data extracts agree. Further checks – including sensitivity testing and studying potential outliers – are made at appropriate points during the analytical process before the finalised, agreed output is shared between the ONS and PHE.

Cancer registrations are assessed by two cancer registration officers and any inconsistencies resolved. Every quarter, further data quality checks at the level of each record are applied, as well as checking trends and population-level metrics before the data registered during the quarter are released for statistical analysis. These higher-level checks are conducted by the cancer registration data quality team within PHE.

PHE has an information governance policy and training schedule that all its employees must follow. The cancer registration team has an enhanced level of training to complete because they are working with sensitive, patient-level data. PHE operates a data governance team to establish and monitor the use of safe methods of working with sensitive data and has a team dedicated to ensuring that applicants for use of PHE's data do so with a legitimate legal basis, current information governance training and ethical approvals.

The cancer registration team in PHE also has a data quality group that meets to maintain or improve the quality of cancer registration data. One is a joint registration officer, data quality and analysis meeting where unusual data patterns are discussed to establish if there are epidemiological reasons for the pattern, or if some data need reviewing, as well as prioritising future developments of the cancer registration data tables.

If the quality of some parts of the cancer registration data are found to be falling, PHE has another data quality group of data liaison officers. The role of PHE's data liaison officers is to go out to data providers and support providers in the use of their systems, so the data that PHE needs to make a cancer registration are available and of high quality.

This combined auditing and quality improvement work has ensured that serious errors (for example. an invalid date) in the main data items that make up a cancer registration have been found in fewer than 0.1% of cancer registrations for more than 10 years. To further ensure the survival estimates use cancer registration data that are self-consistent, a set of enhanced quality assurance tests are applied to survival analysis cohorts (these are discussed under Producer's quality assurance investigations & documentation.

The implications of important parts of the data collection and supply processes failing are discussed under Operational context and administrative data collection.

6 . Producer's quality assurance investigations and documentation

The extraction of data for the cancer registration bulletin is simple. The extraction follows a Standard Operating Procedure (SOP) written for the cancer tables, which ensures that there are no duplicate or unfinalised registrations included. After the SQL script has been independently checked, the extract is then loaded into a statistical package to check the data meets the criteria of SQL script. Aggregated tabulations are created so that the data can be transferred to the Office for National Statistics (ONS).

At the point of creation of the tabulations, the current estimates are compared by the Public Health England (PHE) team to previously published cancer registration bulletins. The ONS team carries out further checks for unusual patterns in the data, unusual combinations of cancer type and sex and check the new data against trends from earlier data. PHE will seek and provide medical or coding advice where unusual patterns are found as a part of the checks that ONS apply.

The data required for survival analysis are more complex and require a longer time series of data. A longer time series of data is required to ensure that the first tumour of each type is used for each patient in survival analysis. More complex data are required because of the enhanced quality assurance checks needing to be applied before survival analyses can be undertaken. These data quality checks ensure that patients and their cancer(s) are uniquely identified and that the data about the patient and their cancer(s) are self-consistent.

The following criteria are used to identify the patients that are eligible to be included in the analysis (and the final number of eligible patients is provided as part of the publication release):

- patients should have a unique identifier; this is to make sure cancers for one patient are not assigned to another patient
- patients should have a complete date of birth, so their age can be calculated at various time points
- adults should be aged between 15 and 99 years at diagnosis; to match to the life tables (see the Quality and Methodology Information report for more information)
- children should be aged between 0 and 14 years at diagnosis
- patients should have a known sex; this is a data quality check and to match to life tables
- patients should have a complete date of cancer diagnosis; this is a data quality check, to match to life tables and to calculate survival time
- patients who have died should have a complete registered date of death; data quality check and to match to life tables
- patients should have a known date of being recorded as alive or dead; data quality check and to calculate their survival time after diagnosis
- patients should be resident in England and have a valid postcode for their usual place of residence at the time of diagnosis; match patients to life tables
- cancers should be (potentially) lethal, newly diagnosed in the studied cohort and a primary cancer (one that hasn't spread from another part of the body); this is so the date of original diagnosis is known
- cancers of the blood (e.g. lymphomas, leukaemia and myelomas) should not occur in a solid cancer; data quality check
- patients are included even if they have further new cancer diagnoses later in the period of interest; this ensures that a patient is only included once in each group of patients and survival time is counted from the earliest diagnosis of the cancer of interest in each period of interest
- patients are excluded if they have had a primary cancer in the same site diagnosed before the period of interest; if a patient has two or more cancers of the same type, it is not clear whether survival time from that type of cancer should be measured from the first or later diagnosis
- patients are included where the earliest diagnosis of a cancer of interest occurred within the period of interest even if they have a primary cancer of another site diagnosed at any time; this treats the patients where the cancer registry is unaware of previous cancer diagnoses in the same way as where this medical history is known
- cancers where the only confirmed record of the cancer is on the patients' death certificate are excluded; as cancer survival attempts to assess the effectiveness of the health system in treating patients with cancer, patients in which cancer is only found after death cannot contribute to this assessment.
- the sequence of dates should be consistent; data quality check, e.g. a patient should not be diagnosed before they are born

Other decisions applied include:

- where a patient dies on the date of diagnosis and have more records than those on a death certificate, then these patients should be included in the survival analyses but should have one day added to the recorded date of death to prevent Stata's stset command from excluding those patients

- when two or more tumours of the same type are diagnosed on the same day for a patient, the one with the worst prognosis is chosen for inclusion; this ensures that a patient is only included once in each group of patients

- coding the cancers with reference to ICD-10 to select similar groups of cancers; the details of the coding applied are included in each bulletin.

Applying these checks mitigates against some of the potential sources of error in the data (see Operational context and administrative data collection).

Having two teams independently extract and produce the statistical outputs identifies uncertainty in how methodologies should be implemented; these differences are used as a tool for learning across the teams. Once there are agreed estimates, sensitivity testing (e.g. varying the composition of a cohort or the censor date) is undertaken to ensure the results are stable. Further comparisons are made to previously published output and results of appropriate international studies.

By consistently applying these quality assurance measures and continually monitoring the data received and the estimates produced, the PHE team can assess the quality of the administrative data received. The data governance and quality assurance teams that PHE have in place and the relationships they have with their data providers means that the likely degree of risk to the quality of the administrative data is low.