ENCR
European Network
of Cancer Registries

Call for Data Protocol for European Population-Based Cancer Registries

Organised by

the European Commission's Joint Research Centre (JRC) and
the European Network of Cancer Registries (ENCR)

to update the **European Cancer Information System (ECIS)**

March 2022

European Commission | Joint Research Centre

ENCR is hosted by the European Commission's Joint Research Centre
Via E. Fermi 2749 | 21027 Ispra | Italy | jrc-encr@ec.europa.eu | www.encr.eu

TABLE OF CONTENTS

1. INTRODUCTION

The Joint Research Centre (JRC) is one of the Directorate Generals of the European Commission. The JRC employs scientists to carry out research in order to provide independent scientific advice and support to EU policy. In this role, the JRC is working closely with the ENCR Steering Committee to agree the priorities for enhancing the value and utilisation of cancer registries data at European level.

The European Network of Cancer Registries (ENCR) was established in 1989, in the framework of the Europe Against Cancer Programme of the European Commission. The ENCR is a professional, non-profit society dedicated to promoting collaboration between cancer registries (CRs), defining data collection standards and providing training to CR personnel. It aims to strengthen the basis for monitoring cancer burden in the EU and the rest of Europe, through the provision of regular and timely information from European CRs.

Since 2012 the JRC hosts the ENCR secretariat and supports ENCR activities aimed at standardisation and harmonisation of cancer data across Europe. Specifically, the following priorities have been identified and pursued:

1. The availability of accurate, reliable, comparable and up-to-date cancer indicators (incidence, mortality, survival, prevalence) across Europe;
2. The development and maintenance of the European Cancer Information System (ECIS) as a comprehensive cancer-information resource for hosting and processing the European cancer data.

ECIS is a comprehensive health and research infrastructure harmonising CRs' data, producing and disseminating cancer burden indicators to assess and facilitate the interpretation of the dynamics of cancer in Europe. A key component of the system is the ECIS web application, launched in February 2018 and populated with indicators computed from data submitted by the ENCR-affiliated registries.

To update the indicators in ECIS, the ENCR Secretariat at JRC coordinates calls for data to the ENCR registries. The present document details the guidelines for the data submission to the JRC, providing detailed instructions on the content and process of the submission through the

dedicated secure online JRC-ENCR portal. Submitted datasets are then processed by the JRC to check for compliance with ENCR agreed standards. Processing of CR data is carried out compliant to the data protection safeguards described in the record for the ECIS database DPR-EC-00417.

Data submitted by CRs is used to update the information available online in the ECIS web application, thus supporting information needs in the framework of European Commission's (EC) initiatives under Europe's Beating Cancer Plan. Information on use of the submitted data is specified in Annex 2.

In order to comply with the Regulation 2018/1725 on processing of personal data by European Union institutions, bodies, offices and agencies, prior to publication of indicators in ECIS a risk assessment for discosure of statistical indicators and scientific results deriving from the data analysis is carried out. This procedure is described in Annex 3.

To facilitate the submission and improve the quality of CRs data, the ENCR-JRC recommend to check beforehand the format of required files and the internal consistency of the incidence data using the JRC-ENCR Quality Check Software (QCS) (see Section 3.1.6). The checks performed by the JRC-ENCR QCS are based on the JRC Technical report "A proposal on cancer data quality checks: one common procedure for European cancer registries" .

Besides ECIS, subject to the consent by CRs data submissions might also serve other projects based on CR data (e.g. IARC's Cancer Incidence in Five Continents (CI5), the EUROCARE study, and possibly others). In such situations, the JRC may disclose CR data by forwarding cleaned data to the respective studies, validated according to pre-defined protocols and according to the conditions detailed in the record for the ECIS database DPR-EC-00417. To account for this possibility, the present protocol has been drafted in alignment with the requirements of CI5 and EUROCARE projects.

2. Timeline for data submission

Data submitted by CRs will be used for annual updates of the historical data available online in the ECIS web application. Moreover, they are used, complemented by additional sources, as a baseline for the computation of cancer incidence and mortality predictions.
For up to speed contribution to the ECIS project, submissions received by 31 January of each year will be taken into account for the annual update of the ECIS indicators.

3. REQUIRED DATA FILES

The following files should be submitted:

3.1 Cancer incidence file

3.2 Population file

3.3 Mortality file

3.4 Life tables

3.5 Questionnaire

3.1. Cancer incidence file

The cancer incidence file to be submitted shall include only pseudonymised data. The process of pseudonymisation is responsibility of the registries, and must be carried out at the registry level prior to submission of the incidence file.

Inclusion criteria for reportable incident cases:

3.1.1. Time period

All available registration years which are considered complete should be submitted.

3.1.2. Reporting cases

- All primary malignant tumours (ICD-O behaviour=3), including squamous cell carcinomas of skin.
- In situ tumours (behaviour=2): breast (ICD-O C50), Urothelial tumours (C65-C68), Ovary (C56), and skin melanoma.
- Uncertain behaviour tumours (behaviour=1): Thymoma (8580/1-8585/1), urothelial tumours (C65-C68), ovary (C56), central nervous system (C70-C72, C751-C753), gastrointestinal stromal tumours

(GIST) (8936/1) and gastroenteropancreatic neuroendocrine tumours (C15-C26, 8150/1-8153/1, 8155/1-8158/1, 8240/1-8242/1, 8248/1) .

- Benign tumours (behaviour=0): Central nervous system (C70-C72, C751-C753) and gastrointestinal stromal tumours (GIST) (8936/0).

3.1.3. Multiple primary tumours

All multiple primary tumours (MPT) are to be retained in the file. Due to differences on MPT definition among the European CRs, it is recommended to submit all primary tumours included in the CR dataset to assure data comparability.

3.1.4. Age

- All ages are eligible in the data submission.
- In age-restricted CRs, such as childhood CRs, all subsequent primary tumours of the registered patients should be included, if available, irrespective of age at diagnosis.

3.1.5. File format

The file should be formatted as follows:

- One record per tumour.
- The file should be a text file (.csv or .txt) with **semi colon (;) separators** and should include a header, with variables' names and order as specified in the text below.

  PAT; MoB; YoB; Age; Sex; Geo_Code; Geo_Label; TUM; MoI; YoI; BoD; Topo; Morpho; Beh; Grade; Autopsy; Vit_stat; MoF; YoF; Surv_time; ICD; CoD; TNM_ed; cT; cN; cM; pT; pN; pM; ToS; Stage; Surgery; Rt; Cht ; Tt; It; Ht; Ot; SCT
- To assist users in the creation of incidence files respecting this structure, the JRC CSV Data layout converter is available as a protocol data adapter. The software is downloadable from the ENCR website (https://www.encr.eu/download).

3.1.6. Data quality

Data should be verified and corrected before submission using the latest version of the JRC-ENCR Quality Check Software (QCS) downloadable from the ENCR website (https://www.encr.eu/download).

3.1.7. Requested variables and coding

The name, description and format, with corresponding missing/unknown values and coding schema, are summarised in Table 1.

Variables related to the patient

- Patient identification code (PAT)

Alphanumeric variable, maximum fifty characters.

Definition: The patient identification is a unique code assigned by the registry, or generated at the time of data submission, to refer to each registered cancer patient. For data protection reasons, it should not be the official personal identification number and should not to be used elsewhere (e.g. in a hospital).

Coding: Missing values are not allowed.

This code will be used for identifying patients with more than one primary tumour and for sending queries and logs to the CRs during the data cleaning process.

CRs need to keep a record of the correspondence between the patient identification number used in the registry and the code provided in their data submission.

- Date of birth

It consists of two separate variables: month and year of birth. The **date of birth will be used to compute/check the age at diagnosis**, therefore after validation of the dataset the two variables recording month and date of birth will be deleted from the final dataset, as no longer needed.

- Month of birth (MoB)

  Numeric variable, maximum two digits.

  Definition: Month of birth of the patient.

  Coding: The range of valid values is 1-12. **If the month of birth cannot be provided** for some reason, it should be **coded as 99**.

  It is preferable to have the data without imputation of the missing values (value=99). If, however, the month of birth for some tumours has already been imputed, please detail the imputation rule in the data call questionnaire (question 1.13).

- Year of birth (YoB)

  Numeric variable, four digits.

  Definition: Year of birth of the patient.

  Coding: Provide the full 4-digit year (for example 1942). The year of birth should not be less than 1842 (>1842). If the year of birth cannot be provided, it should be coded as 9999.

- Sex at birth (Sex)

Numeric variable, one digit.

Definition: This variable refers to the biological and physiological characteristics that define men and women.

Coding: It should be coded as 1 (male), 2 (female) or 3 (other). If sex cannot be provided, it should be coded as 9.

Tumour variables

- Code of the geographical area of residence at diagnosis (Geo_code)

Alphanumeric variable, max 10 digits.

Definition: Code for the geographical area of residence of the patient at the time of diagnosis for each tumour.

For harmonisation purposes, the Geo_code should follow the NUTS (Nomenclature of Territorial Units for Statistics) classification level 2 (NUTS2).

NUTS codes are the geolocation variables used by EUROSTAT in all their data sets, subdividing the economic territory of the European Union (EU) into regions at three different levels (NUTS 1, 2 and 3 respectively, moving from larger to smaller territorial units). NUTS geolocation codes are available for all EU-27 countries plus UK, as well as countries belonging to the European Free Trade Association (EFTA), candidate countries awaiting accession to the EU or potential candidates.

The NUTS codes will be available for each cancer registry in the JRC portal. More information can be found in Annex 1.

For other European countries where the NUTS classification is not available, the Geo_code variable corresponds to the highest level of administrative sub-divison in the area covered by the cancer registry which can be provided.

If a valid value cannot be provided for some cases, it should be coded as XX99.

If it is not applicable, it should be left blank.

Note: Due to the peculiarity of French site-specific CRs, whose area overlaps with the area of general registries, the CRs of France are requested to code the geographical area according to the NUTS3-***Départements.***

- Name of the geographical area of residence at diagnosis (Geo_label)

Alphanumeric variable, max 50 digits.

Definition: Description (name) for the geographical area of residence of the patient at the time of diagnosis for each tumour.

The Geo_label is the name of the geographical area corresponding to the Geo_code.

If a valid description cannot be provided for some cases, it should be coded as 9.

If it is not applicable, it should be left blank.

- Tumour identification (TUM)

Alphanumeric variable, maximum fifty characters.

Definition: This variable is assigned by the registry. It allows the identification of two or more tumours for the same patient. It can be (but does not need to) a sequence number.

For data protection reasons, it should not be the official personal identification number and should not be used elsewhere.

Please refer to the "Submission instructions" in section 4 of this document for additional details.

Coding: Missing values are not allowed.

The combination of the patient identification variable and the tumour identification variable should be unique for each tumour.

- Age at diagnosis (Age)

Numeric variable, maximum three digits.

The exact age is important because it is used to calculate age-specific and age-standardised rates. If available, cancer registries should use **day**, **month** and **year** of the date of birth and the date of incidence to calculate the exact age at diagnosis for each tumour.

Definition: **Latest completed year of age** at the time of diagnosis.

Coding: The range of valid values is 0-120.

- Date of incidence

It consists of two separate variables: month and year of incidence.

The updated ENCR DoI recommendation should be followed to record the date of incidence.

- Month of incidence (MoI)

Numeric variable, maximum two digits.

Definition: Month of incidence recoded according to the updated ENCR DoI recommendation.

Coding: The range of valid values is 1-12. **If the month of incidence cannot be provided** for some reasons, it should be **coded as 99**.

It is preferable to have the data without imputation of the missing values. If, however, the month of incidence for some tumours has already been imputed, please detail the imputation rule in the data call questionnaire (questions 1.15).

- Year of incidence (YoI)

Numeric variable, four digits.

Definition: Year of incidence recorded according to the updated ENCR DoI recommendation.

Coding: Missing values are not allowed. The range of valid values is from 1941 to present.

- Basis of diagnosis (BoD)

Numeric variable, one digit.

Definition: This variable indicates the degree of certainty with which the diagnosis of cancer has been established. This information will be used for computing quality indicators, such as the percentage of cancer cases identified through the Death Certificate Only (DCO), or the percentage of cases with histological verification of the diagnosis.

Coding: This variable should be coded, according to the ENCR BoD recommendation:

0 → Death certificate only (DCO)

1 → Clinical

2 → Clinical investigation

4 → Specific tumour markers

5 → Cytology

6 → Histology of a metastasis

7 → Histology of a primary tumour

9 → Unknown

Note: Cases registered as DCO are cancers for which no information could be obtained, other than a death certificate mentioning cancer. These cases are included in cancer incidence statistics for the year of death. Nevertheless, the true date of diagnosis and the duration of the survival are unknown, and these data cannot normally be included in survival analyses.

- Topography (Topo)

Alphanumeric variable, four characters.

Definition: This variable indicates the anatomic site of the primary tumours.

Coding: It should be coded according to the third revision of the International Classification of Diseases for Oncology (ICD-O-3.1).

CRs who use other International Classifications of Diseases (ICD) versions should convert their codes to ICD-O-3 prior to submission, using any appropriate software (for instance, the IARCcrgTools).

The full four-digit characters' ICD-O-3 code should be provided, including the initial letter, but without the decimal point (“.”). For example, supraglottis should be coded as C321.

When the primary site of the tumour is unknown, the topography should be coded as C809. The topography of the metastasis should not be attributed to the primary tumour.

- Morphology (Morpho)

Numeric variable, four digits.

Definition: This variable records the type of cell that has become neoplastic, the specific histological term.

Coding: It should be coded according to any version of the ICD-O-3 (ICD-O-3.1 or ICD-O-3.2). CRs who use other ICD versions should convert their codes to ICD-O-3 prior to submission, using any appropriate software (for instance, the IARCcrgTools).

The valid range of morphology codes is 8000-9993. Missing values are not allowed. **Malignant tumour, NOS, should be coded as 8000, leukaemia, NOS should be coded as 9800 and malignant lymphoma, NOS as 9590 (with behaviour code 3, see below).**

- Behaviour (Beh)

Numeric variable, one digit.

Definition: This variable indicates whether a tumour is malignant, benign, in situ, or of uncertain behaviour.

Coding: It should be coded according to any version of the ICD-O-3 (ICD-O-3.1 or ICD-O-3.2) as follows:

0 → Benign neoplasms

1 → Neoplasms of uncertain or unknown behaviour

2 → In situ neoplasms

3 → Malignant neoplasms stated or presumed to be primary

Note: codes 6 (malignant, metastatic site/malignant, secondary site) and 9 (malignant, uncertain whether primary or metastatic site) should not be used by CRs: the correct behaviour code in these case is 3.

- Grade (Grade)

Numeric variable, one digit.

Definition: This variable describes how much a tumour resembles the normal tissues from which it arose and is also used to denote cell lineage for leukaemias and lymphomas.

Coding: Except for tumours of the central nervous system and urothelial tumours, **solid malignant tumour** should be coded according to the ICD-O-3.1 as follows:

1 → Well differentiated

2 → Moderately differentiated

3 → Poorly differentiated

4 → Undifferentiated, anaplastic

When a diagnosis indicates two different degrees of grading or differentiation, the higher number should be used as the grade code. For example: if the diagnosis is moderately differentiated squamous cell carcinoma with poorly differentiated areas, the grade should be coded as 3.

For **leukaemias and lymphomas**:

5 → T-cell; T-precursor

6 → B-cell; Pre-B; B-precursor

7 → Null cell; Non T-non B

8 → NK cell (natural killer cell)

For **all**:

9 →Unknown

The grade of the central nervous system tumours should be coded according to table 27 of the ICD-O-3.1 (pages 28 and 29).

Variables related to the follow-up

- Incidental finding of cancer at autopsy (Autopsy)

Numeric variable, one digit.

Definition: It marks the cases discovered only at autopsy, that are included in cancer incidence statistics. For these cases, the date of incidence is the same as the date of death. This variable is extremely important in survival analysis because cases incidentally discovered at autopsy, as well as DCO cases, must be excluded from survival statistics.

Coding: It should be coded as:

0 → No (not found at autopsy)

1 → Yes (found at autopsy)

9 → Unknown

Note: For the cases discovered only at autopsy the vital status is always 2 (dead) and the survival time is 0 days.

- Last known vital status (Vit_sta)

Numeric variable, one digit.

Definition: This variable describes the patient's vital status as last known to the CR. This information may be collected using either 'active' or 'passive' methods of follow up.

Coding: It should be coded as:

1 → Alive

2 → Dead

9 → Unknown

If the CR adopts a passive follow-up, patients who are not known to be deceased would normally be assumed to be alive at the date of the most recent linkage between the registry data and a death index or other vital status records. The vital status of those patients should be coded as "1" (alive).

If patients cannot be traced by any active follow-up procedure, their vital status may remain undetermined and should then be coded as "9" (unknown).

- Date of last known vital status

This variable consists of two separate variables: month and year of last known vital status.

It corresponds to the most recent date for which the patient's last known vital status was available. If the patient is deceased, the date of last known vital status should be the date of death.

For registries using passive follow-up and the patient is alive, it corresponds to the most recent date for which death certificates have been linked to registrations.

If the patient has emigrated or has been lost to follow-up, the last date at which he/she was known to be alive should be reported.

The date of last known vital status will be used to compute/check the duration of survival, therefore after validation of the dataset the two variables month and year of last known vital status will be deleted from the final dataset, as no longer needed.

- Month of last known vital status (MoF)

  Numeric variable, maximum two digits.

  Definition: The month of the most recent date for which the patient's last known vital status was available.

  Coding: The range of valid values is 1-12. When the month of the last known vital status **cannot be provided**, it should be coded as 99.

It is preferable to have the data without imputation of the missing values. If, however, the month of the last known vital status for some tumours has already been imputed, detail the imputation rule in the data call questionnaire (questions 1.20.2).

• Year of the last known vital status (YoF)

Numeric variable, four digits.

Definition: Year of the most recent date for which the patient's last known vital status was available.

Coding: The range of valid values is from 1941 to present.

When the year of the last known vital status cannot be provided, it should be coded as 9999.

EUROCARE project: vital status should be updated at least one year after the latest complete incidence year (i.e. to at least 31/12/2019 if incidence is provided to 2018).

- Duration of survival in days (Surv_time)

Numeric variable, maximum five digits.

The exact duration of survival is essential. The CRs should use day, month and year of the date of incidence and the date of last known vital status to calculate the duration of the survival in days.

Definition: It is the number of days between full dates (including days) of the last known vital status and the date of incidence.

Coding: The values should be ≥ 0. When the duration of survival cannot be provided, it should be coded as 99999.

- Official underlying cause of death (CoD)

Alphanumeric variable, maximum four characters.

Definition: This variable is used to estimate cause-specific survival. It records the official underlying cause of death according to standard international coding rules.

Coding: It should be coded according to the International Classification of Diseases (ICD). The dot (.) between the third and the fourth digits should not be included. For example, if the underlying cause of death is malignant neoplasm of laryngeal cartilage, this should be coded as C323 (according to ICD-10). If the underlying cause of death is acute myocardial infarction unspecified, the CoD should be coded as 4109 (according to ICD-9).

When this variable is not applicable, it should be left blank.

Note: If vital status is 1 (alive) the CoD and ICD should be left blank; if vital status is 2 (dead) and the cause of death is unknown, CoD should be coded as R99 (ICD-10)/7999 (ICD-9) or 9999 if ICD is different from ICD-9 or ICD-10.

- ICD edition used for coding cause of death (ICD)

  Numeric variable, maximum two digits.

  Coding: This variable, coded as a number lower than 12, should be provided if the underlying cause of death has been reported.

  When this variable is not applicable, it should be left blank.

  Note: if vital status is 2 (dead) and ICD is unknown, it should be coded as 99.

Variables related to the tumour stage at diagnosis

The stage at diagnosis is particularly useful information for the interpretation of international survival comparisons, for the evaluation of screening programs, and other studies.

Notes:

- When TNM and/or TNM stage grouping is/are available, they should be reported in preference to any other coding system.
- If cTNM (clinical TNM) is available and the primary tumour was not resected, the pTNM (pT, pN, pM) should be left blank.
- If the CR does not know whether the TNM is pathological or clinical, it should be recorded as clinical and be specified in the data call questionnaire.
- When Toronto Childhood Cancer Stage Guidelines recommend TNM classification for a specific tumour (Tier 2), it should be recorded in the specific fields for the T-, N- and M categories (cT, cN, cM and pT, pN, pM)
- If TNM is not available or not applicable, cTNM and pTNM should be left blank and (if possible) type of stage (stage system) and stage should be reported.

- TNM edition (TNM_ed)

  Numeric variable, maximum two digits.

  This variable should be provided when any TNM and /or TNM stage grouping have being reported.

  Coding: Valid values are numbers ≤ 8, or 99 when the information is not available.

- TNM: clinical T-category (cT)

  Alphanumeric variable, maximum twelve characters.

  Definition: The variable encodes information on the extent of the primary tumour, based on clinical evidence.

Coding: It should be coded according to any edition of the TNM classification without the “T” - for example: 1a, not T1a. When the information cannot be provided, it should be coded as 9. Since TNM ed. 7, MX (distant metases cannot be assessed) is no longer a valid option.

- TNM: clinical N-category (cN)

Alphanumeric variable, maximum twelve characters.

Definition: This variable provides information on the absence or presence and extent of the regional lymph node metastasis, based on clinical evidence.

Coding: It should be coded according to any edition of the TNM classification, without the “N” - for example: 0, not N0; 3a, not N3a. When the information cannot be provided, it should be coded as 9.

- TNM: clinical M-category (cM)

Alphanumeric variable, maximum ten characters.

Definition: This variable describes the absence or presence of distant metastasis, based on clinical evidence.

Coding: It should be coded according to any edition of the TNM classification, without the “M" - for example: 0, not M0; 1a, not M1a. When the information cannot be provided, it should be coded as 9.

- TNM: pathological T-category (pT)

Alphanumeric variable, maximum twelve characters.

Definition: This variable encodes information on the extent of the primary tumour based on pathological evidence.

Coding: It should be coded according to any edition of the TNM classification, without the “T” - for example: 1a, not T1a.

When the information cannot be provided, it should be coded as 9.

- TNM, pathological N-category (pN)

Alphanumeric variable, maximum twelve characters.

Definition: This variable provides information on the absence or presence and extent of regional lymph node metastasis, based on pathological evidence.

Coding: It should be coded according to any edition of the TNM classification, without the “N”- for example: 0, not N0; 3a, not N3a. When the information cannot be provided, it should be coded as 9.

- TNM, pathological M-category (pM)

Alphanumeric variable, maximum twelve characters.

Definition: This variable describes the absence or presence of distant metastasis, based on pathological evidence.

Coding: It should be coded according to any edition of the TNM classification without the "M" - for example 1a, not M1a. When the information cannot be provide, it should be coded as 9.

- Staging system (ToS)

Alphanumeric variable, maximum three characters.

Definition: This variable describes the system used by the CR for coding stage.

Coding: It should be coded according to the following categories (Table 1):

A → Ann Arbor/ Lugano stage

D → Dukes' stage

E → Summary extent of disease

F → FIGO stage

S → TNM stage, unknown whether clinical or pathological

clS → clinical TNM stage

paS →pathological TNM stage

ypS → pathological TNM stage after neoadjuvant therapy

cpS → combination of clinical & pathological TNM stage

coS → condensed TNM stage

esS → essential TNM stage

Ti1 → Toronto Childhood Cancer Stage Tier 1 for paediatric tumours

Ti2 → Toronto Childhood Cancer Stage Tier 2 for paediatric tumours

COG → COG stage, Toronto Tier 2 for Wilms tumours,: findings at surgery when NO chemotherapy prior to surgery (see also Annex 4)

SIO → SIOP stage, Toronto Tier 2 for Wilms tumours: findings at surgery when chemotherapy prior to surgery (see also Annex 4)

8 → Other system

When the information cannot be provided, it should be coded as 9.

- Stage

Alphanumeric variable, maximum three characters.

Definition: This variable is defining the extent of disease at diagnosis.

When a CR reports TNM they should not submit stage, which will be coded centrally.

Coding: it should be coded according to the following categories (Table 1), see also for coding Toronto Childhood Cancer Stage Annex 4.

0 → Stage 0, stage 0a, stage 0is, carcinoma in situ, non-invasive

1 → Stage I, FIGO I, localized, localized limited (L), limited, Dukes A

1A -> Stage IA, FIGO IA, Ann Arbor

1B -> Stage IB, FIGO IB

1B1 -> FIGO IB1

…

2 → Stage II, FIGO II, localized advanced (A), locally advanced, advanced, direct extension, Dukes B

2A -> Sage IIA, FIGO IIA

2B -> Stage IIB, FIGO IIB

…

3 → Stage III, FIGO III, regional (with or without direct extension), R+, N+, Dukes C

4 → Stage IV, FIGO IV, metastatic, distant, M+, Dukes D

When the information cannot be provided, it should be coded as 9.

Variables related to treatment

Treatment variables refer to the curative **first course of anticancer therapy after diagnosis**. Purely symptomatic therapy (e.g. bypass surgery, pain relief) should not be considered.

- Surgery (Surgery)

Numeric variable, one digit.

Coding: It should be coded according to the following categories (Table 1):

0→ No

1 → Yes, without specification

2 → Yes, local surgery only

The following procedures should be considered as local surgery: polypectomy (mainly gastro-intestinal tract), transurethral resection (TUR; bladder & other urinary tract), cone biopsy/loop excision (cervix), as well as all other procedures which leave the organ in situ, such as cryosurgery, laser coagulation, thermoablation, radiofrequency ablation (RFA), etc.

3 → Yes, 'operative' surgery

'Operative' surgery includes all resections of the tumour which require the removal of an organ or a part of that organ, such as a lobectomy, hemicolectomy, hysterectomy, cystectomy, prostatectomy, etc.

9 → Missing/Unknown

Notes:

- If available, type of surgery (*local surgery* versus *operative surgery*) should be recorded for solid cancers of the following topographies: C01-C06, C16-C20, C30-C34, C53-C55, C61 and C65-C68. For other tumours, code 1 (surgery without specification) suffices.
- If both *local surgery* and *operative surgery* were performed for the same tumour, *operative surgery* should be registered.

- Radiotherapy (Rt)

Numeric variable, one digit.

Coding: It should be coded according to the following categories (Table 1):

0 → No

1 → Yes, without other specification

2 → Yes, neoadjuvant (pre-operative) radiotherapy

3 → Yes, adjuvant (post-operative) radiotherapy

9 → Unknown/missing

- Chemotherapy (Cht)

Numeric variable, one digit.

Coding: It should be coded according to the following categories (Table 1):

0 → No

1 → Yes, without other specification

2 → Yes, neoadjuvant (pre-operative) chemotherapy

3 → Yes, adjuvant (post-operative) chemotherapy

4 → Yes, both neoadjuvant and adjuvant chemotherapy

9 → Unknown/missing

- Targeted therapy, including monoclonal antibodies (Tt)

  Numeric variable, one digit.

  Definition: Targeted therapy comprises all drugs that block the growth of cancer cells by inhibition of certain pathways in the cancer cell. Traditional chemotherapy also affects other cells in the body that divide quickly. The main categories of targeted therapy are small molecules (mostly tyrosine kinase inhibitors such as imatinib and many other -nibs) and monoclonal antibodies (such as rituximab and many other -mabs). Monoclonal antibodies are considered a form of immunotherapy but should be coded as targeted therapy.

  Coding: It should be coded according to the following categories (Table 1):

  0 → No

  1 → Yes

  9 → Unknown/missing

- Immunotherapy, excluding monoclonal antibodies (It)

  Numeric variable, one digit.

  Coding: It should be coded according to the following categories (Table 1):

  0 → No

  1 → Yes

  9 → Unknown/missing

- Hormone therapy (Ht)

  Numeric variable, one digit.

  Coding: It should be coded according to the following categories (Table 1):

  0 → No

  1 → Yes

  9 → Unknown/missing

- Other or unspecified systemic therapy (Ot)

  Numeric variable, one digit.

  Coding: It should be coded according to the following categories (Table 1):

  0 → No

  1 → Yes, without other specification

  2 → Yes, neoadjuvant (pre-operative)

  3 → Yes, adjuvant (post-operative)

  9 → Unknown/missing

- Stem cell transplantation (SCT)

  Numeric variable, one digit.

  Coding: It should be coded according to the following categories (Table 1):

  0 → No

  1 → Yes

  9 → Unknown/missing.

Table 1. Variable name, description, format, missing/unknown values and coding schema

Variables should be separated by a semi-colon

Patient variables					
Variable name	**Variable description**	**Format**	**Maximum length**	**Missing/ unknown**	**Coding**
PAT[1]	Patient identification code	A	50	Not allowed	According to registry coding
MoB	Month of birth	F	2	99	Range of allowed values: 1 - 12
YoB	Year of birth	F	4	9999	Range of allowed values: > 1842 and ≤ the current year
Sex	Sex at birth	F	1	9	1 →Male 2 →Female 3 →Other
Tumour variables					
Geo_code	Code for the geographical area of residence at Diagnosis	A	10	XX99	NUTS2 when available or the highest level of administrative sub-divison that can be provided[2]. Blank → not applicable
Geo_label	Name of the geographical area of residence at Diagnosis	A	50	9	Blank → not applicable
TUM	Tumour identification	A	50	Not allowed	According to registry coding
Age	Age at diagnosis (incidence date) in years	F	3	999	Range of allowed values: ≥ 0 and < 121
MoI	Month of incidence	F	2	99	Range of allowed values: 1 - 12
YoI	Year of incidence	F	4	Not allowed	Range of allowed values: From 1941 to present
BoD	Basis of diagnosis	F	1	9	0→Death certificate only 1→Clinical 2→Clinical investigation 4→Specific tumour markers 5→Cytology 6→Histology of a metastasis 7→Histology of a primary tumour
Topo	ICD-O-3 topography code	A	4	Not allowed	Valid code in ICD-O-3
Morpho	ICD-O-3 morphology code	F	4	Not allowed	Valid code in any ICD-O-3 version
Beh	ICD-O-3 behaviour	F	1	Not allowed	0→ Benign neoplasm 1→ Neoplasm of uncertain and unknown behaviour 2→ In situ neoplasm 3→ Malignant neoplasm
Grade[3]	ICD-O-3 grade of differentiation / immunophenotype	F	1	9	1→Grade I, Well differentiated 2→ Grade II, Moderately differentiated 3→ Grade III, Poorly differentiated 4→Grade IV, Undifferentiated, anaplastic 5→ T-cell; T-precursor 6→ B-Cell; Pre-B; B-precursor 7→ Null cell; Non T-non B 8→ NK cell (natural killer cell) 9→ Not applicable

[1] *PAT* should be a code assigned by the registry that is not to be used elsewhere (e.g. in a hospital). So, it cannot be an official personal number. It may be an encrypted personal number as long as this specific encryption is not used by any other organisation. The JRC will provide the tool to the CRs to do it.

[2] The geographical area for French CRs should be coded according to NUTS3 (see Note on page 8).

[3] The *grade* of tumours of the central nervous system should be coded according to table 27 of ICD-O-3.1.

Table 1. Variable name, description, format, missing/unknown values and coding schema

Variables should be separated by a semi-colon

Variable name	Variable description	Format	Maximum length	Missing/ unknown	Coding
Variables related to follow-up					
Autopsy[4]	Incidental finding of cancer at autopsy	F	1	9	0→No 1→Yes
Vit_stat	The last known vital status	F	1	9	1→ Alive 2→ Dead
MoF	Month of last known vital status	F	2	99	Range of allowed values: From 1 to 12
YoF	Year of last known vital status	F	4	9999	Range of allowed values: > 1941 and ≤ the current year
Surv_time	Duration of survival in days	F	5	99999	≥ 0
ICD[5,6]	ICD edition for coding cause of death	F	2	99	Range of allowed values: <12 Blank → Not applicable
CoD[5,6]	Official underlying cause of death	A	4	R99 (ICD-10) 7999 (ICD-9)	According to ICD Blank → Not applicable
Stage variables					
TNM_ed	TNM edition	F	2	99	Allowed values: ≤ 8
cT[7]	Clinical T-category	A	12	9	According to the TNM Classification of Malignant Tumours Blank → not applicable
cN[7]	Clinical N-category	A	12	9	
cM[7]	Clinical M-category	A	12	9	
pT[7,8]	Pathological T-category	A	12	9	
pN[7,8]	Pathological N-category	A	12	9	
pM[7,8]	Pathological M-category	A	12	9	
ToS	Staging system	A	3	9	A → Ann Arbor/ Lugano stage D → Dukes' stage E → Extent of disease F → FIGO stage S → TNM stage, unknown whether clinical or pathological clS → clinical TNM stage paS → pathological TNM stage ypS -> pathological TNM stage after neoadjuvant therapy cpS → combination of clinical & pathological TNM stage coS → condensed TNM stage esS → essential TNM stage Ti1 → Tier 1 stage for paediatric tumours Ti2 → Tier 2 stage for paediatric tumours COG → COG Tier 2 stage for Wilms tumours, findings at surgery when NO chemotherapy prior to surgery SIO → SIOP Tier 2 stage for Wilms tumours: findings at surgery when chemotherapy prior to surgery 8 → Other staging system

[4] In autopsy cases, incidentally found at autopsy, the *vital status* is always 2 (dead) and the *survival* time is 0 days.
[5] If the vital status is 1 (alive) the *CoD* and *ICD* should be left blank;
[6] if the vital status is 2 (dead) and the cause of death is unknown, CoD should be coded as R99 (ICD-10)/7999 (ICD-9) or 9999 and ICD should

be coded as 99

[7] If TNM is not available or not applicable, cTNM (*cT, cN, cM*) and pTNM (*cT, cN, cM*) should be coded as 9 and be left blank respectively and (if possible) *Staging system (ToS)* and *stage* should be coded.

[8] If cTNM is available and the primary tumour was not resected the pTNM (*pT, pN, pM*) should be left blank.

Table 1. Variable name, description, format, missing/unknown values and coding schema

Variables should be separated by a semi-colon

Stage variables					
Variable name	**Variable description**	**Format**	**Maximum length**	**Missing/ unknown**	**Coding**
Stage	Stage	A	3	9	0 → Stage 0, stage 0a, stage 0is, carcinoma in situ, non-invasive 1 → Stage I, FIGO I, localized, localized limited (L), limited, Dukes A 1A -> Stage IA, FIGO IA, Ann Arbor 1B -> Stage IB, FIGO IB 1B1 -> FIGO IB1 … 2 → Stage II, FIGO II, localized advanced (A), locally advanced, advanced, direct extension, Dukes B 2A -> Sage IIA, FIGO IIA 2B -> Stage IIB, FIGO IIB … 3 → Stage III, FIGO III, regional (with or without direct extension), R+, N+, Dukes C … 4 → Stage IV, FIGO IV, metastatic, distant, M+, Dukes D … See also Annex 4 for Toronto childhood cancer stage
Treatment variables					
Surgery[9,10]	Resection of the primary tumour	F	1	9	0 → No 1 → Yes, without specification 2 → Yes, local surgery only[a] 3 → Yes, 'operative' surgery[b]
Rt	Radiotherapy	F	1	9	0 → No 1 → Yes, without specification 2 → Yes, neoadjuvant (pre-operative) radiotherapy 3 → Yes, adjuvant (post-operative) radiotherapy

[9] If available, type of surgery (*local surgery* [2] versus *operative surgery* [3]) should be coded for solid cancers of the following cancer sites: C01-C06, C16-C20, C30-C34, C53-C55, C61 and C65-C68. For other cancers, code 1 (surgery without specification) suffices.

[10] If both *local surgery* and *operative surgery* were performed for the same tumour, *operative surgery* should be coded.

[a] The following procedures should be coded as local surgery: polypectomy (mainly gastro-intestinal tract), transurethral resection (TUR; bladder & other urinary tract), cone biopsy/loop excision (cervix), as well as all other procedures which leave the organ in situ, such as cryosurgery, laser coagulation, thermoablation, radiofrequency ablation (RFA), etc.

[b] This includes all resections of the tumor which require the removal of an organ or a major part of that organ, such as a lobectomy, hemicolectomy, hysterectomy, cystectomy, prostatectomy, etc.

Table 1. Variable name, description, format, missing/unknown values and coding schema

Variables should be separated by a semi-colon

Treatment variables					
Variable name	**Variable description**	**Format**	**Maximum length**	**Missing/ unknown**	**Coding**
Cht	Chemotherapy	F	1	9	0 → No 1 → Yes, without other specification 2 → Yes, neoadjuvant (pre-operative) 3 → Yes, adjuvant (post-operative) 4 → Yes, both neoadjuvant and adjuvant
Tt[11]	Targeted therapy (including monoclonal antibodies)	F	1	9	0 → No 1 → Yes
It	Immunotherapy (excl. monoclonal antibodies)	F	1	9	0 → No 1 → Yes
Ht	Hormone therapy	F	1	9	0 → No 1 → Yes
Ot	Other or unspecified systemic therapy	F	1	9	0 → No 1 → Yes, without other specification 2 → Yes, neoadjuvant (pre-operative) 3 → Yes, adjuvant (post-operative)
SCT	Stem cell transplantation	F	1	9	0 → No 1 → Yes

[11] Targeted therapy comprises all drugs that block the growth of cancer cells by inhibition of certain pathways in the cancer cell. Traditional chemotherapy also affects other cells in the body that divide quickly. The main categories of targeted therapy are small molecules (mostly tyrosine kinase inhibitors such as imatinib and many other -nibs) and monoclonal antibodies (such as rituximab and many other -mabs). Monoclonal antibodies are considered a form of immunotherapy but should be coded as targeted therapy.

3.2 Population file

Information on population data should be provided from official censuses, from intercensal/postcensal estimates provided by Vital Statistics Departments, or equivalent, or other official sources.

The population data should have the same geographical and temporal reference as for the cases of the incidence file.

Registries having cases belonging to more than one geographical area (Geo_code) need to send population data for each of the areas, specifying the reference area in the Geo_code variable as described next.

Scope

The population data should correspond to the cancer case file with respect to:

- registration area
- time period by year
- sex
- age-range
- Geo_code

Geo_code should refer to the same greographical areas included in the incidence file.

If possible, population figures should give the mid-year estimates for each sub-category.

File format

The population data should be submitted in the form of a text file with semi colon (;) separator, and should include headers with names as specified in examples 1 or 2.

The file should contain the following variables:

- Calendar year
- Sex
- Age: by single year of age, if possible, or otherwise 21 standard age ranges (see below)
- Geo_code
- Geo_label
- Number of residents

The variables in the population file should be in the same order as reported above.

If the population data are available by single year of age and, for example, the period for the cancer cases is 1992-2013, the population file should be provided as in EXAMPLE 1.

If population data are not available by single year of age, 21 age ranges should be provided by age-group as following: **0** (under 1 year), **1-4** (age group 1-4), **5-9** (age group 5-9), **10-14** (age group 10-14), **15-19** (age group 15-19), **20-24** (age group 20-24), **25-29** (age group 25-29), **30-34** (age group 30-34), **35-39** (age group 35-39), **40-44** (age group 40-44), **45-49** (age group 45-49), **50-54** (age group 50-54), **55-59** (age group 55-59), **60-64** (age group 60-64), **65-69** (age group 65-69), **70-74** (age group 70-74), **75-79** (age group 75-79), **80-84** (age group 80-84), **85-89** (age group 85-89), **90-95** (age group 90-95), **95+** (age group 95 and over). In this case, the population file should be provided as in EXAMPLE 2.

EXAMPLE 1

Calendar year	Sex	Age unit	Geo_ code	Geo_label	Number of residents
1992	1	0	AT11	Burgenland	$N_{1992,1,0}$
1992	1	1	AT11	Burgenland	$N_{1992,1,1}$
1992	1	2	AT11	Burgenland	$N_{1992,1,2}$
1992	1	3	AT11	Burgenland	$N_{1992,1,3}$
1992	…	…	…		…
1992	2	0	AT11	Burgenland	$N_{1992,2,0}$
1992	2	1	AT11	Burgenl and	$N_{1992,2,1}$
1992	2	2	AT11	Burgenland	$N_{1992,2,2}$
1992	2	3	AT11	Burgenland	$N_{1992,2,3}$
…	…	…			
2013	1	100	AT34	Vorarlberg	$N_{2013,1,100}$
2013	2	0	AT34	Vorarlberg	$N_{2013,2,0}$
2013	…	…	AT34	Vorarlberg	
2013	…	100	AT34	Vorarlberg	$N_{2013,2,100}$

Sex = 1 → males; Sex=2 → females

EXAMPLE 2

Calendar year	Sex	Age range	Geo_ code	Geo_label	Number of residents
1992	1	0	AT11	Burgenland	$N_{1992,1,1}$
1992	1	1-4	AT11	Burgenland	$N_{1992,1,2}$
1992	1	5-9	AT11	Burgenland	$N_{1992,1,3}$
1992	1	10-14	AT11	Burgenland	$N_{1992,1,4}$
1992	…	…	…		…
1992	2	0	AT11	Burgenland	$N_{1992,2,1}$
1992	2	1-4	AT11	Burgenland	$N_{1992,2,2}$
1992	2	5-9	AT11	Burgenland	$N_{1992,2,3}$
1992	2	10-14	AT11	Burgenland	$N_{1992,2,4}$
…	…	…			
2013	1	95+	AT34	Vorarlberg	$N_{2013,1,21}$
2013	2	0	AT34	Vorarlberg	$N_{2013,2,1}$
2013	…		AT34	Vorarlberg	
2013	…	95+	AT34	Vorarlberg	$N_{2013,2,21}$

Sex = 1 → males; Sex=2 → females

Accompanying information required

- Any other coding than the recommended above should be documented in the data call questionnaire (section 2).
- The reference to the source of population data should be provided in the data call questionnaire (section 2).
- The reference calendar data (e.g. 1st January, 31st December, etc) must be indicated in the data call questionnaire (section 2).

3.3 Mortality file

National CRs are NOT required to submit national mortality statistics, as these can be retrieved directly from the EUROSTAT or WHO databases.

For sub-national registries, mortality statistics are partially available either from the previous submission, or from EUROSTAT (if the registry region corresponds to at least NUTS2). Sub-national registries in regions not overlapping with NUTS2 boundaries need to complement mortality statistics. The mortality data already available in the ECIS database for sub-national registries is listed here. Registries should submit additional years of mortality not yet available in ECIS, consistently with the incidence years submitted.

The mortality data should be the official cancer mortality data, as obtained from the Vital Statistics Department, or equivalent, and based on certificates/death records.

Mortality data will be published in the ECIS web application and used to compute quality indicators.

Scope

The mortality data for the area covered by the CR should include all residents whose underlying cause of death was cancer.

The mortality data should correspond to the cancer cases file with respect to:

- registration area
- time period by year
- sex
- age-range.

File format

The mortality data should be submitted in the form of a text file with semi colon (;) separator, and include headers with names as specified in examples 3 or 4.

The file should contain the following variables:

- Calendar year
- Sex
- Age: single year of age, if possible, or otherwise 21 age range (see below)
- Cause of death: 3 digits of the applicable International Classification of Diseases (ICD)
- Number of deaths

The variables in the mortality file should be in the same order as reported above.

Overall mortality data for all ages combined (total number of deaths) is acceptable only if no breakdown information by age-group is available to the registry.

If the number of deaths is available by single year of age and, for example, the period for cancer cases is 1992-2013, the mortality file should be provided as in EXAMPLE 3.

Alternatively, the number of deaths for the combination of calendar year, sex, age range and cause of death should be provided (EXAMPLE 4), using the following age range codes: **0** (under 1 year), **1-4** (age group 1-4), **5-9** (age group 5-9), **10-14** (age group 10-14), **15-19** (age group 15-19), **20-24** (age group 20-24), **25-29** (age group 25-29), **30-34** (age group 30-34), **35-39** (age group 35-39), **40-44** (age group 40-44), **45-49** (age group 45-49), **50-54** (age group 50-54), **55-59** (age group 55-59), **60-64** (age group 60-64), **65-69** (age group 65-69), **70-74** (age group 70-74), **75-79** (age group 75-79), **80-84** (age group 80-84), **85-89** (age group 85-89), **90-95** (age group 90-95), **95+** (age group 95 and over).

EXAMPLE 3

Calendar year	Sex	Age unit	Cause of death	Number of Deaths
1992	1	0	C00	$N_{1992,1,0,C00}$
1992	1	1	C00	$N_{1992,1,1,C00}$
1992	1	2	C00	$N_{1992,1,2,C00}$
1992	1	3	C00	$N_{1992,1,3,C00}$
1992	…	…	…	…
1992	2	0	C00	$N_{1992,2,0,C00}$
1992	2	1	C00	$N_{1992,2,1,C00}$
1992	2	2	C00	$N_{1992,2,2,C00}$
1992	2	3	C00	$N_{1992,2,3,C00}$
…	…	…		
2013	1	100	C97	$N_{2013,1,100,C97}$
2013	2	0	C00	$N_{2013,2,0,C00}$
2013	…	…		
2013	…	100	C97	$N_{2013,2,100,C97}$

Sex = 1 → males; Sex=2 → females

EXAMPLE 4

Calendar year	Sex	Age range	Cause of death	Number of Deaths
1992	1	0	140	$N_{1992,1,1,140}$
1992	1	1-4	140	$N_{1992,1,2,140}$
1992	1	5-9	140	$N_{1992,1,3,140}$
1992	1	10-14	140	$N_{1992,1,4,140}$
1992	…			
1992	2	0	140	$N_{1992,2,1140}$
1992	2	1-4	140	$N_{1992,2,2,140}$
1992	2	5-9	140	$N_{1992,2,3,140}$
1992	2	10-14	140	$N_{1992,2,4,140}$
…	…			
2013	1	95+	208	$N_{2013,1,21,208}$
2013	2	0	140	$N_{2013,2,1,140}$
2013	…			
2013	…	95+	208	$N_{2013,2,21,208}$

Sex = 1 → males; Sex=2 → females

Accompanying information required

- Any coding, other than that recommended, should be documented also in the data call questionnaire (section 3).
- A reference to the source of population data should be provided in the data call questionnaire (section 3).

3.4 Life tables – ONLY for cancer registries providing follow-up and survival data

Life tables, i.e. the background mortality in the general population of the administrative territory covered by the cancer registry, must be provided by registries covering their entire period of incidence or the period in which the follow-up is available.

All-causes of **death probabilities** in the general population, **by sex, age and calendar year**, should be provided to **6 decimal** places or an equivalent number of significant figures (e.g. 0.012345 for a rate of 1,234.5 per 100,000). The format of all-causes of death information must be specified in the Questionnaire, where you state whether you are providing **probabilities** or **rates**. Since all-causes death probabilities are highly dependent on age, values should be **preferably** given by **one-year age classes** (from 0 to 99 or more). If this is not possible, age should be grouped by **no more than five years**: in this case, please specify how the life tables were smoothed in the data call questionnaire (section 4).
It is essential to have accurate all-causes death probabilities for the elderly to accurately estimate relative survival in this age group. The oldest age class can be open ended (e.g. 90 years and over), but the lower boundary of **the oldest age class should not be less than 85 years**.

Life tables should have the same geographical and temporal reference as for the cases of the incidence file. Registries having cases belonging to more than one geographical area (Geo_code) are invited to send life tables for each of the areas, if available, specifying the reference area in the Geo_code and Geo_label variables as described next. If available, **NUTS2 or the same administrative regions** used for incidence file should be provided.

Life tables available from the previous submission are listed here. Registries should submit additional years of life tables not yet available, consistently with the incidence years submitted.

Please document the source of demographic data in the data call questionnaire (section 4).

EXAMPLE 5

Calendar year	Sex	Annual age (years)	Geo_code	Geo_label	All causes death probability
1990	1	0	AT11	Burgenland	0.003228
1990	1	1	AT11	Burgenland	0.000272
1990	1	2	AT11	Burgenland	0.000376
..	…	…	…	…	…
1990	1	99	AT11	Burgenland	0.414117
1990	2	0	AT11	Burgenland	0.000379
1990	2	1	AT11	Burgenland	0.000376
1990	2	2	AT11	Burgenland	0.000373
…	…	…	…	…	…
1990	2	99	AT11	Burgenland	0.389871
…	…	…	…	…	…
2013	1	0	AT34	Vorarlberg	0.002528
2013	...	…	…	…	…
2013	2	99	AT34	Vorarlberg	0.342862

Sex = 1 → males; Sex=2 → females
Geo_code: Code of the geographical area of residence at diagnosis
Geo_label: Name of the geographical area of residence at diagnosis (Geo_label)

3.5 Questionnaire

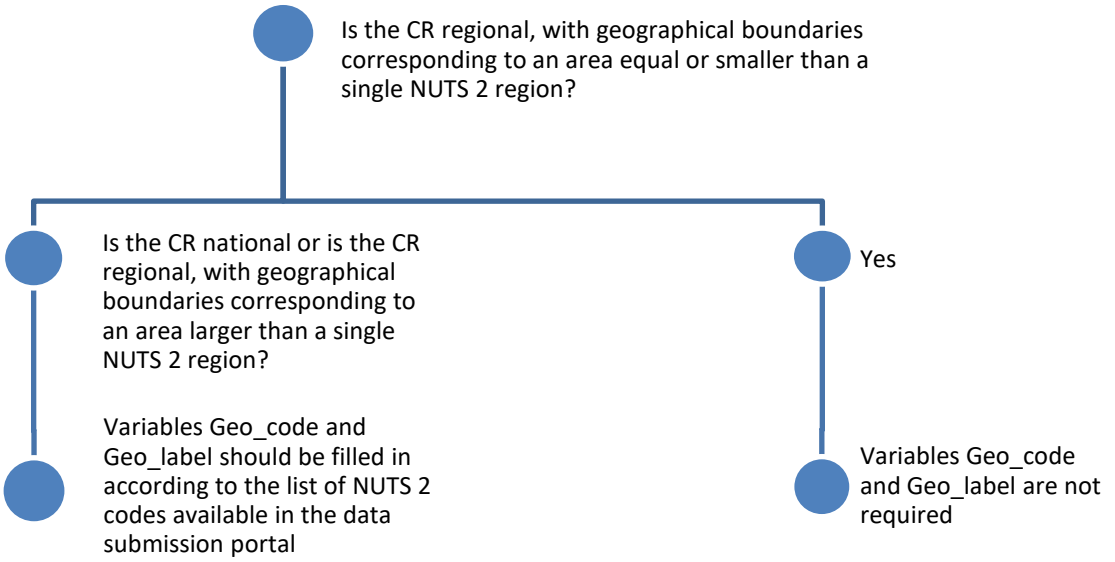
The questionnaire is an essential part of the data submission process and for data interpretation and comparability among registries.

CRs will be invited to fill in the questionnaire at the time of submission and this step will be a pre-requisite for the completion of the data submission process. The questionnaire is focused on the datasets submitted.

Annex 1 – Geographical variables

This section provides more information on how to fill in the Geo_code and Geo_label variables depending on whether the CR is national or regional.

The decision tree below relate to CR from EU-27 countries plus UK, as well as countries belonging to the European Free Trade Association (EFTA), candidate countries awaiting accession to the EU or potential candidates. A list of the regions and codes will be available in the CR-specific area of the Portal.



The scheme above refers to the incidence, population and life tables files. If Geo-code and Geo-label cannot follow the NUTS nomenclature as requested for the incidence file, please specify in the questionnaire which geographical areas the Geo-code and Geo-label refer to in the life table files.

For other European countries where the NUTS classification is not available, the Geo_code variable corresponds to the highest level of administrative sub-divison in the area covered by the cancer registry which can be provided.

Annex 2 – Data use of the files provided by the European population-based cancer registries

1. Update of the indicators published online in the ECIS web application:
   - National estimates of cancer incidence and mortality for the major cancer entities in European countries;
   - Cancer indicators by cancer registry/geographical area (historical data): number of cases, crude rates, age-standardized rates, age-specific rates and cumulative risk by year, sex, age range and cancer entity;
   - Childhood cancer incidence indicators by cancer registry/geographical area :
     - Number of cases by 5-years age group, diagnostic group and subgroup according to the International Childhood Cancer Classification, third edition (ICCC-3), grouped by at least 3 years of incidence period;
     - Crude and age-standardized rates and cumulative risk by diagnostic group and subgroup according to the ICCC-3, grouped by at least 3 years of incidence period.

2. New indicators planned for inclusion in the ECIS web application:
   - Data quality indicators by cancer entity, sex, age group and calendar period:

     Proportion of cases microscopically verified

     Proportion of cases with unspecified morphology

     Mortality-to-incidence (M:I) ratio

     Proportion of cases with death certificate only (DCO)

     Missing or not valid values of the following variables: age, sex, topography, morphology, behaviour, grade, basis of diagnosis and year of incidence, vital status and last known vital status date/survival.
   - 1-year/ 5-year survival by cancer entity, sex and incidence period.
   - Distribution of cases by stage and/or type of treatment, cancer entity, sex, age range and incidence period - depending on availaibility and quality of data submitted by the cancer registries.

Annex 3 – Risk assessment analysis for discosure of statistical indicators and scientific results deriving from the data analysis in ECIS

The JRC discloses statistical indicators through the publicly available ECIS website, (https://ecis.jrc.ec.europa.eu/) and publications.

A number of specific risk-reducing measures are taken, including aggregation of data, to prevent identification of natural persons from any of the published indicators. These risk-reducing measures take into account all the means reasonably likely to be used, and therefore ensure that the information is not personal data in accordance with the Regulation 2018/1725 on processing of personal data by European Union institutions, bodies, offices and agencies.

The publication of indicators in the form of aggregated data is allowed as far as the reported figures cannot be related to any identifiable natural person. This results both from ethical considerations and from the purpose limitation principle. Indeed, it is necessary to ensure that the collection and storage of personal data results in uses that fulfil the very purpose of the processing, namely, to contribute to scientific research on cancer for the development of public health policies, in particular effective cancer surveillance and proper monitoring of the cancer burden in Europe.

The following methodology is implemented to ensure that aggregated data do not relate to identifiable individuals.

In order to prevent identification from aggregated data, the dimension of the population size in relation to the number of cases is taken into account. This approach is based on the probability of identification of a cancer case in a specific registry's dataset, and is computed as the ratio between the number of cancer patients and the background population, for a given combination of the variables allowed as filters for dissemination (so far: sex/cancer entity/age/year). We need to consider not only the intrinsic identifiability of the data (the probability with which a person can be identified in each single cancer registry dataset), but also the risk of additional information being used to possibly disclose the identity of a person (re-identification).

Based on well-established statistical practice, we can consider 0.05 (1 in 20 or 5%) as a reasonable threshold to be on the safe side as for the re-identification probability.

For the ECIS data, table cells for a specific registry are defined by a Cartesian crossing of four independent variables:

1. Cancer entity (60 different sites for incidence)
2. Sex (male, female or "both sexes = male and female together")
3. Age (18 age groups)
4. Calendar year of incidence.

Let's consider the following scenario for a low-value frequency in a table cell given by a narrow selection of a specific age class of females only, in a specific year for a cancer incidence considered rare and in a small region/cancer registry (real data taken from ECIS):

1. Cancer entity: tongue (defined by the ICD-O-3 codes: C01-C02)
2. Sex: female
3. Age: age group 55 to 59
4. Calendar year of incidence: 2012

This combination identifies in the specific registry 1 case of cancer of the tongue on a background population of 9,568 women within the 55 to 59 age group. The probability of identifying a person from this cell is then 1/9,568 or 0.000104515 (0.01%,), where 1 is the number of persons in the cell and 9,568 is the background population. Since the value of the calculated probability is smaller than 0.05 (5%), the cell may be publicly released with negligible risk of re-identification of any individual person.

Additional variables, such as stage, can also be applied to the Cartesian crossing and define new combinations. The procedure to assess if the resulting frequencies should be published or not is the same as above, in which cell-specific probabilities are compared against the threshold value of 0.05 (5% probability). Should additional parameters be introduced in ECIS to filter the data, the same cell-specific probabilities approach would be applied.

The smallest background population in a specific region/cancer registry of the ECIS database concerns the following situation:

1. Cancer site: Oesophageal cancer (ICD-O-3 code C15)
2. Sex: male
3. Age: age group 80 to 84
4. Calendar year of incidence: 2007

This combination identifies 1 case of oesophageal cancer on a background population of 749 males. Following the procedure described above, the probability of identifying this person is given by 1/749 or 0.0013 (0.13%), which is still below the threshold of 0.05 (5%).

Whereas the procedure for ascertaining the probability threshold is generic, there is one notable restriction. If the background population is extremely small, in the order of a few hundred persons, it could be argued that despite fulfilling the statistical criteria, there is a large probability that a person can indeed be identified. An example would be the inhabitants of a small village who all know each other on a personal level. This situation is however very far away from the catchment areas of the European CRs (even the smallest of them corresponding to metropolitan areas).

The specific situation of reporting on childhood cancers deserves separate considerations. For childhood cancers, where the numbers are more unstable compared to adults given they are based from the lower occurrences encountered in children, not showing the aggregated numbers and maybe presenting only rates would hamper comparability (due to the instability of the rates when computed from small background populations) and the possibility to hint to the magnitude of the problem in the different areas.

Childhood cancers are reported in a specific section of ECIS, separately from the all-ages part. In this section (ECIS-Childhood) a less detailed visualisation in comparison to the adults is adopted, according to the following specifications:

• Diagnostic group – for children, tumors are defined according to the International Classification of Childhood Cancer (ICCC), which is a system that bases the malignancy classification on the histological traits of the tumor (type of tissue) rather than on the tumor

sites, as for the adults. This makes the reference to the cancer site much less immediate than for the adults. ECIS-childhood considers the 12 diagnostic groups and 47 subgroups of ICCC;

- Age – 5-yrs age groups are considered, the same as for the adults;

Therefore, depicting the most extreme situation of a "1" in a table cell for childhood cancers, this would mean that:

- in the specific area of the registry … – the smallest registries insist at NUTS3 level, i.e. areas with population between 15000 and 80000 units;
- … there was one child - not known if male or female;
- … aged within a specific 5 yrs interval … - for instance between 5 and 10 yrs old;
- … diagnosed of one cancer classified according to its histology (ICCC)

The smallest background population in a specific region/cancer registry of the childhood ECIS database concerns the following situation:

1. Cancer subgroup: 3b Astrocytoma
2. Sex: female
3. Age: age group 0 to 4
4. Calendar year of incidence: 2008

This combination identifies 1 case of astrocytoma on a background population of 2924 females. Following the procedure described above, the probability of identifying this person is given by 1/2924 or 0.0003 (0.03%), which is still below the threshold of 0.05 (5%). The risk of disclosure of an individual is therefore negligible.

Annex 4 – Reporting childhood tumour stage to the ENCR-JRC data call according to Toronto guidelines

Tumour	Tier 1	Stage value in the ENCR-JRC data call protocol	Tier 2	Stage value in the ENCR-JRC data call protocol
Acute lymphoblastic leukaemia	CNS negative CNS positive	0 1	CNS1 CNS2 CNS3	1 2 3
Acute myeloid leukaemia	CNS negative CNS positive	0 1	CNS negative CNS positive	0 1
Hodgkin's lymphoma	**Ann Arbor/Lugano** stage IA/B stage IIA/B stage IIIA/B stage IVA/B	 1, 1A, 1B 2, 2A, 2B 3, 3A, 3B 4, 4A, 4B	**Ann Arbor/Lugano** stage IA/B/IE stage IIA/B/ stage IIIA/B/IIIS stage IVA/B	 1, 1A, 1B, 1E 2, 2A, 2B, … 3, 3A, 3B, 3S 4, 4A, 4B
Non-Hodgkin lymphoma	Limited Advanced	1 4	St Jude/Murphy - stage I St Jude/Murphy - stage II St Jude/Murphy - stage III St Jude/Murphy - stage IV	1 2 3 4
Neuroblastoma	Localised Locoregional INRGSS—MS disease Metastatic	1 2 3 4	INRGSS—localised L1 INRGSS—locoregional L2 INRGSS—MS disease INRGSS - metastatic	1 2 3 4
Wilms tumour*	Localised Metastatic	1 4	Stage I Stage II Stage III Stage IV	1 2 3 4
Wilms tumour**	Localised Metastatic	1 4	y-stage I y-Stage II y-stage III y-stage IV	1 2 3 4
Rhabdomyosarcoma	Localised Metastatic	1 4	TNM stage 1 TNM stage 2 TNM stage 3 TNM stage 4	1 2 3 4

*COG stage (no chemo prior to surgery)
**SIOP stage (chemo prior surgery)

Tumour	Tier 1	Stage value in the ENCR-JRC data call protocol	Tier 2	Stage value in the ENCR-JRC data call protocol
Non-rhabdomyosarcoma soft-tissue sarcomas	Localised Metastatic	1 4	TNM stage 1 TNM stage 2 TNM stage 3 TNM stage 4	1 2 3 4
Osteosarcoma	Localised Metastatic	1 4	Localised Metastatic	1 4
Ewing's sarcoma	Localised Metastatic	1 4	Localised Metastatic	1 4
Retinoblastoma	Localised (intraocular) Regional (orbital or regional lymph nodes) Distant (extra-orbital)	1 3 4	IRSS stage 0 IRSS stage I IRSS stage II IRSS stage III IRSS stage IV	0 1 2 3 4
Hepatoblastoma	Localised Metastatic	1 4	Pretext Stage I Pretext Stage II Pretext Stage III Pretext Stage IV	1 2 3 4
Testicular	Localised Regional Metastatic	1 2 4	TNM stage I TNM stage II TNM stage III	1 2 3
Ovarian	Localised Regional Metastatic	1 2 4	FIGO stage I FIGO stage II FIGO stage III FIGO stage IV	1 2 3 4
Medulloblastoma and other CNS embryonal tumour	M0 or localised M+ or metastatic	1 4	M0 M1 M2 M3 M4	0 1 2 3 4
Ependymoma	M0 M+	1 4	M0 M1 M2 M3 M4	0 1 2 3 4