

CRATE user manual

Version 2.0

Table of Contents

- 1. Overview
- 2. Installation
- 3. Usage
- 4. Aggregation levels
- 5. Output reports
- 6. Advanced configuration
 - 6.1. Metadata of input file
 - 6.2. Execution flags
 - 6.3. Output reports
- 7. Insights
- 8. Known issues

1. Overview

CRATE (Cancer Registries Aggregation Tool for ECIS) is the R routine which aggregates record-level data from cancer registries' archives according to the dimensions considered in the [ECIS web application](#). It can be used to prepare data for different studies: [ECIS Historical data](#), [ECIS Childhood data](#), [ECIS Estimates](#), [ECIS Survival](#) and [ECIS Prevalence](#).

The input file of CRATE is the record-level incidence data extracted from the registry archive according to the requirements of the [ECIS data-call protocol](#). Incidence data for CRATE needs to be cleaned and validated. For ECIS Survival and Prevalence studies, other input files are listed in section "3. Usage".

2. Installation

Before running CRATE, the R environment needs to be configured according to the following preparatory steps:

1. Install the R software (<https://www.r-project.org/>) on your computer (at least version 4.5.x)
2. Unzip the "crate-<version>.zip" package in any directory on your computer
3. Take note of the path where you unzip the ZIP package, since this will be your **working directory**
4. Launch the R console

Only for Historical, Childhood and Estimates studies, if not already installed before, install the following three packages:

- a. `install.packages("dplyr")`
- b. `install.packages("stringr")`
- c. `install.packages("lifecycle")`

Only for the Estimates study it is necessary to install also following components:

1. `install.packages("data.table")`
2. the external module executing ICD-10 conversion (see instructions received by email for the study on Estimates).

Only for the Survival, Prevalence and Survival-Childhood studies it is necessary to run the command `crate-install.R` or to install also all packages from the *dependencies* subfolder.

Once you unzip the "crate-<version>.zip" file, the following files will appear in your working directory:

- **crate-historical.R**: Launcher script for the Historical study
- **crate-childhood.R**: Launcher script for the Childhood study
- **crate-estimates.R**: Launcher script for the Estimates study
- **crate-install.R**: Launcher script to install all the packages included in the *dependencies* subfolder for Survival, Prevalence and Survival-Childhood studies
- **crate-survival-all-ages.R**: Launcher script for the Survival and Prevalence studies
- **crate-survival-childhood.R**: Launcher script for the Survival Childhood study
- **config/age_groups_18.csv**: Configuration file containing definitions for the age breakdown in 18 age groups
- **config/age_groups_20.csv**: Configuration file containing definitions for the age breakdown in 20 age groups
- **config/cancer_entities_childhood_2025.csv**: Configuration file containing the cancer-entities definitions for Childhood and Survival-Childhood studies
- **config/cancer_entities_estimates_2025.csv**: Configuration file containing the Estimates cancer-entities definitions
- **config/cancer_entities_historical_2026.csv**: Configuration file containing the cancer-entities definitions for Historical, Survival, and Prevalence studies
- **config/crate_global_settings.csv**: Configuration file containing the global settings for the studies of Survival, Prevalence and Survival Childhood
- **config/crate_user_settings.csv**: Configuration file containing the settings for the studies of Survival, Prevalence and Survival Childhood to be review and filled in by the user
- **docs/CRATE-aggregation-with-R-user-manual.pdf**: User manual (this document)
- **docs/Notice.txt**: Policy and information about CRATE usage licence
- **docs/Licence.txt**: User license for using the CRATE software
- **lib**: The sub-folder containing the R source files used as library by the main script (that is: *main.R*)

These files should never be edited or modified by the user, since this could trigger unpredictable behavior or wrong aggregation results.

Optional: to verify that all the required packages have been correctly installed, please execute in the R console the three commands below:

```
system.file(package='dplyr');  
system.file(package='stringr');  
system.file(package='lifecycle');
```

If the results are different from an empty string, then the three packages are correctly installed.

Remark: The maximum number of records that can be handled by CRATE depends on the amount of the internal memory of the user's computer (a minimum of 4GB of RAM is suggested).

3. Usage

CRATE reads the input incidence dataset, which is the record-level incidence file compliant with the [ECIS data-call protocol](#) requirements. This must be a text file using the **semicolon** (i.e. ";") to separate the different variables, correctly named and coded. For the ECIS Childhood data study only records with Age < 20 years are considered.

Only for the Survival, Prevalence and Survival Childhood studies, CRATE needs as input also the **lifetable.csv** and **population.csv** files that must be compliant with the [ECIS data-call protocol](#) requirements.

Only for the Survival, Prevalence and Survival Childhood studies, the user must fill in the following information in the corresponding **config** **lcrate_user_settings.csv**, inside the column **setting_value**:

- **qcs_dir_path** : the path where QCS has been installed by the user
- **incidence_first_year** : First year of incidence data used for survival analysis
- **incidence_last_year** : Last year of incidence data used for survival analysis
- **survival_follow_up_last_year** : Last year of complete survival follow-up. E.g. with 2020, this means that follow-up reaches 2020-12-31
- **dataset_population_value_type** : Indicates whether the population file contains the 'mid-year' (mid), start of the year (start), or end of the year population (end).

The user should also read through the other settings, which do have defaults, but they may not be right in the case of the user:

- **dataset_cancer_record_path** : Absolute or relative path to .csv containing cancer records as rows, as per the call for data for the 'incidence' dataset
- **java_command** : name of command or absolute path to Java executable. If Java has been installed and the java executable is included via the system PATH environment variable, then the simple command line command `java` will be available and the default will work directly. If you need to e.g. use a different version of Java, you can supply the path to its executable here
- **dataset_lifetable_path** : Relative or absolute path to lifetable file. Must be csv
- **dataset_lifetable_value_type** : Indicates whether the lifetable contains probabilities (p) or hazard rates (h)
- **dataset_lifetable_age_group_type** : Indicates whether the lifetable contains values for 18 age groups of width 5 (except last 85+; 18of5) or 101 age groups of width 1 (last 100+; 101of1)
- **survival_follow_up_last_year** : Last year of survival follow-up. E.g. with 2020 this means that follow-up reaches 2020-12-31
- **dataset_population_path** : Relative or absolute path to population file. Must be csv.

The setting_example and the setting_default for each setting are reported in the following table:

setting_name	setting_example	setting_default
dataset_cancer_record_path	dataset.csv	dataset.csv
qcs_dir_path	C:/Users/jrc-qcs-2.4.1/jrc-qcs-2.4.1	
java_command	Files/OpenJDK/jdk-22.0.2/bin/java.exe	java
dataset_lifetable_path	my_lifetable.csv	lifetable.csv
dataset_lifetable_value_type	h	p
incidence_first_year	1985	
incidence_last_year	2020	
survival_follow_up_last_year	2020	
dataset_population_path	my_population.csv	population.csv
dataset_population_value_type	end	end

Further details for the Survival and Prevalence studies' settings can be found in the CRATE Technical Details document in the *ldoc* subdirectory.

To run CRATE, the user can choose two different options:

1. Run CRATE using the default configuration, without performing any setup of the R environment
2. Setup a custom run time environment, in order to have more control over the script execution

Advanced users can change also other configuration parameters (see section [6. Advanced Configuration](#) for details).

Option 1: default configuration

Using the default configuration, the record-level incidence file to be aggregated MUST be copied to the **working directory**, and the file MUST be renamed as **dataset.csv**. Then, to run CRATE the user must open an R console and copy-paste one of the following commands, depending by the target study type:

```
source("<absolute path of your working directory>/create-historical.R")
source("<absolute path of your working directory>/create-childhood.R")
source("<absolute path of your working directory>/create-estimates.R")
source("<absolute path of your working directory>/crate-survival-all-ages.R")
source("<absolute path of your working directory>/crate-survival-childhood.R")
```

Remark: only for the Survival and Prevalence studies, the files named **lifetable.csv** and **population.csv** MUST be placed in the same path of the input dataset. This file must respect the current structure required by the ECIS data call protocol, and must use comma as separator (not semicolon).

For example, to perform the **Historical** study on the default *dataset.csv* file :

```
source("C:/Home/Temp/crate/crate-historical.R")
```

Option 2: custom configuration

This option allows running CRATE without renaming the record-level incidence file. To enable this option, the name and extension of the input dataset MUST be setup using the **dataset_meta** variable in the R console, as follow:

```
# General syntax:
dataset_metadata <- c(<dataset file's name>, <dataset file's extension>, <working directory>);
```

Example:

```
# Usage example:
dataset_metadata <- c("my_file_name", "txt", "C:\\Temp\\ENCR\\data");
```

Warning

The custom-configuration approach implies that the output reports will be produced in the new working directory (in the example above: "C:\\Temp\\ENCR\\data"), and the name of ALL output files will be changed accordingly. For instance, in the example above with "my_file_name.txt", the custom-configuration approach would produce the file "my_file_name_aggregated.txt" instead of "dataset_aggregated.txt" (see section [5. Output reports](#) for details).

This approach can be useful when the user wishes to aggregate different datasets with different names, namely data separated by incidence period or other stratifications inside the same registry. By choosing a custom configuration, the user could:

1. Copy all input datasets in the custom working folder
2. Set the *dataset_meta* variable (in the R console) in order to match the name of the first dataset to be processed
3. Run CRATE
4. Go back to step 2, assign to the *dataset_meta* variable the name of the next input dataset to be processed, then repeat the step 3.

4. Aggregation levels

The current version of CRATE aggregates records according to the following variables (not done by Survival and Prevalence studies):

- Sex (SEX): male or female
- Year of incidence (YOI): the whole incidence time period included in the input record-level data file
- Age group (AG) : 5-year age groups, from minimum "0-4 years" to maximum "95+ years"
- Cancer entity (ENTITY_ID): cancer entity definitions as described in the configuration file (for the target study)
- Geographical area of residence at diagnosis (GEO_CODE): the geographical codes provided by the cancer registry (optional)

Then, after this basic aggregation, some post processing tasks are executed, depending by the target study.

Example: for the Historical study, the *ce_inall* and *ecis_web* columns (see Historical configuration file) are used as follows:

- The cancer entities contributing to the definition of "All entities excluding keratinocytic skin cancers" are those where the column **ce_inall** is different from "NO"
- The cancer entities that will be published in the ECIS web application are those where the column **ecis_web** is different from "NO"

5. Output reports

CRATE produces in the output files in the output sub directory defined for each study:

- The **historical** sub directory contains the output files of aggregation for the Historical study
- The **childhood** sub directory contains the output files of aggregation for the Childhood study
- The **estimates** sub directory contains the output files of aggregation for the Estimates study
- The **survival** sub directory contains the output files of the estimates for the Survival study
- The **survival-childhood** sub directory contains the output files of the estimates for the Survival-Childhood study
- The **prevalence** sub directory contains the output files of the estimates for the Prevalence study.

The output files for incidence Historical, Childhood, and Estimates studies are:

- **dataset_checklog.txt**: report documenting the results of the aggregation process
- **dataset_discarded.csv**: dataset containing all records discarded because of missing or invalid values of the main variables (topography, morphology, behaviour, grade, sex, year of incidence, age/year of birth) or because of excluded values (valid values, but excluded from the aggregation)
- **dataset_notinall.csv**: dataset containing records matching the definition of at least one cancer entity, but not matching the definition of "All entities excluding keratinocytic skin cancers"
- **dataset_unmatched.csv**: dataset containing all valid records not matching any cancer entity definitions
- **dataset_selected.csv**: dataset containing a copy of the records actually used for the aggregation (i.e. records not discarded, nor excluded and therefore matching at least one cancer entity definition)
- **dataset_aggregated.csv**: dataset containing the aggregated number of cases by sex, age group, year of incidence and cancer entity
- **dataset_aggregated_by_area.csv**: similar to "dataset_aggregated.csv", with aggregation also over the GEO_CODE variable.

If any of these files is missing, then the corresponding set of records is empty (*Example*: if *dataset_discarded.csv* is missing, then CRATE did not discard any record).

The output files for Survival, Survival-Childhood, and Prevalence studies are:

- **log.txt**: report documenting the results of the estimation process
- **survival_by_POI.csv** dataset containing age-adjusted Pohar Perme survival estimates
- **survival_by_POI_et_ECIS_AG_SURV.csv** dataset containing age-adjusted Pohar Perme survival estimates not adjusted by age
- **survival_latest_period.csv** dataset containing age-adjusted Pohar Perme survival estimates by period
- **survival_latest_period_by_ECIS_AG_SURV.csv** dataset containing Pohar Perme survival estimates by period and age group
- **survival_by_ECIS_POI.csv** dataset containing age-adjusted Pohar Perme survival estimates for the period 2000-2007
- **survival/quality_indicators_latest_period.csv** dataset containing the quality indicators for the same period analysis as the corresponding table of survival estimates
- **survival/quality_indicators_by_POI.csv** dataset containing the quality indicators for the same period analysis as the corresponding table of survival estimates, stratified by period of incidence
- **survival/quality_indicators_by_YOI.csv** dataset containing the quality indicators for the same period analysis as the corresponding table of survival estimates, stratified by year of incidence
- **survival-childhood/survival_childhood_latest.csv** dataset containing the observed-survival estimates
- **survival-childhood/survival_childhood_latest_by_AG_SURV_CHILDHOOD.csv** dataset containing the observed-survival estimates stratified by age group
- **prevalence_latest.csv** dataset containing the prevalence age-adjusted estimates for the last allowed date
- **prevalence_latest_by_ts_age.csv** dataset containing the prevalence estimates for the last allowed date stratified by age group (based on age at prevalence observation time point)

The user can find further information on the structure of these Survival, Prevalence and Survival-Childhood output files and on the content of each variable in the pdf file `\docs\crate_technical_details.pdf`.

As reported in the section "3.Usage", the names of these output files could be different if the user set a custom configuration.

The main output file of CRATE for the studies incidence Historical, Childhood and Estimates is the **dataset_aggregated.csv** output report containing the results of the aggregation process on all valid records that did match a valid cancer entity as defined in the configuration file (for the target study). If opened using a text editor, this file should look like the following:

```
SEX;YOI;AG;ENTITY_ID;NUMBERS
```

```
1;2006;14;68;1
```

```
1;2006;15;21;2
```

```
1;2006;15;88;1
```

```
1;2006;16;21;1
```

```
1;2006;16;60;1
```

```

1;2006;17;21;3
1;2006;17;22;1
1;2006;17;25;1
1;2006;17;30;1
1;2006;17;65;2
1;2006;17;69;1

```

Figure 1: example of results in the "dataset_aggregated.csv" output file

- The ENTITY_ID column refers to the ID of the ECIS cancer entities, as defined in the **ECIS_ID** column of the configuration file referred by the target study
- The NUMBERS column reports the number of cases corresponding to the same combination of SEX, YOI, AG and ENTITY_ID

The list of files produced depend by the target study and the user's configuration. For example:

- The **dataset_aggregated_by_area.csv** is optional, because it is produced only if the GEO_CODE variable contains at least **2** different values: if all GEO_CODE values are missing (or set to the same identical NUTS code), this report is **not** produced
- The **dataset_intermediate.csv** and **dataset_intermediate_by_area.csv** output files are produced only when running the Historical study, and are indented only for advanced users (i.e. they can be ignored)

6. Advanced configuration

CRATE can be run using the default configuration, as explained in section "3. Usage". However, advanced users can exploit the flexibility of the R script by tweaking some parameters to match specific needs.

If no parameter is set in the R console, the script will perform the aggregation using the **default** configuration producing output files with the name convention as described in the section "4. Output reports". Moreover, the aggregation is executed under the following assumptions:

- Any missing numeric value (**NA** in R) is replaced with an empty string
- Aggregation is performed using **20 Age Groups** (0-4, 5-9, ... 90-94, 95+)

Advanced users can run the R script overriding this configuration from the R console, as explained below. It's also possible to change some parameters directly in the R script by searching the "CUSTOM_SELECTION" keyword in the source file. However this is NOT suggested when preparing the dataset for the ECIS submission since tampering with the source code could lead to unpredictable results.

6.1. Metadata of input file

Name and extension of the input dataset can be changed setting the **dataset_meta** variable (in the R console). For details, see *Option 2* in the "3. Usage" section.

6.2. Execution flags

The execution of CRATE can be changed by setting the **flags** variable (in the R console) *before* running the script.

```

# General syntax:
flags <- c(<use_20_age_groups>,<fill_missing_with_empty_string>, <clean_all_resources>);

# Usage example:
flags <- c(TRUE, TRUE, TRUE);

```

The code in the example above enables all the *flags-related* features, which are (in the same order as the command above):

- *use_20_age_groups*: if set to FALSE, the definition of 18 Age Groups is used (instead of 20 Age Groups)
- *fill_missing_with_empty_string*: if set to FALSE, all produced output datasets contain the value "NA" where the input dataset has missing value (i.e. empty string)
- *clean_all_resources*: if set to FALSE, all final objects (intermediate data frames) are kept in the current workspace.

6.3. Output reports

Only for Historical, Childhood and Estimates studies, the **number of output files** produced during the aggregation process can be changed by setting the **output_reports** variable (in the R console) *before* running the script.

```
# General syntax:
output_reports <- c(<discarded_report>, <unmatched_report>);

# Usage example:
output_reports <- c(FALSE, FALSE);
```

The code in the example above disables the production of the following output reports:

- If the *discarded_report* flag is set to FALSE: the report named like "*dataset_discarded.csv*" will NOT be produced
- If the *unmatched_report* flag is set to FALSE: the report named like "*dataset_unmatched.csv*" will NOT be produced

7. Insights

CRATE applies few selected checks and replacements on missing and unknown values of aggregation variables before processing the input dataset (aka *data sanity*):

- **MOI (Month Of Incidence)** and **MOB (Month Of Birth)**: missing (i.e. empty) and *unknown* ("99") values are replaced with value **6** (June)
- **AGE**: if missing or unknown, AGE is calculated as the difference between the incidence date (MOI, YOI) and the birthdate (MOB, YOB) - if available for the calculation
- **GRADE**: if missing (i.e. empty), GRADE is replaced with the value **9** (Grade *unknown*)
- **BEH (behaviour)**: if BEH is **6** or **9**, BEH is replaced with value **3** (Malignant neoplasms stated or presumed to be primary)

8. Known issues

CRATE is currently in the **testing** stage. For this reason, the user experience could be not straightforward, since some manual tasks must be performed when installing and/or running the software.

The following improvements will be considered in the future:

1. Replace step 4 of the installation procedure with a task such as: "install the CRATE package using the R **install package** feature from local files menu option and the ZIP file provided by the JRC". This should allow users to install the script using something like "*library/crate*" and then running a command like "*df_aggregated <- crate.R(path, fromYear=1980, toYear=2022)*"
2. It is currently not possible to **run CRATE from a operating system (OS) console** (that is: using "RScript" executable of the user's operative system)
3. If more R installations are available (for instance: both R base and RStudio), then some local settings could be shared by these installations. In case of issues, try to run the tool **twice**: in the second run the settings (proper values in the *input_path* and *output_path* variables) from previous runs are no longer used
4. We are deliberately **not** removing all objects from the CRATE working memory to reduce the risk of deleting your previous work (if any). Therefore, in case of dirty R memory and/or status, you should save your data and then execute the command *rm(list=ls(all=TRUE))*

For **support** please contact JRC-ENCR@ec.europa.eu.