TECHNICAL REPORT

HIV Estimates Accuracy Tool manual

www.ecdc.europa.eu

ECDC TECHNICAL REPORT

HIV Estimates Accuracy Tool manual

This manual was commissioned by ECDC to the National Institute of Public Health – National Institute of Hygiene and the Department of Hygiene, Epidemiology and Medical Statistics, National and Kapodistran University of Athens in 2016 and reviewed and approved by ECDC.

Project team: Magdalena Rosińska (National Institute of Public Health – National Institute of Hygiene Warsaw), Nikos Pantazis (National and Kapodistrian University of Athens)

Tool developer: Daniel Lewandowski (NextPage Software)

ECDC project manager: Chantal Quinten

Acknowledgments for help and advice on the project: Anastasia Pharris, Andrew Amato, Signe Gilbro, Emiliano Farinella and Vivian Tse.

Suggested citation: European Centre for Disease Prevention and Control. HIV Estimates Accuracy Tool manual. Stockholm: ECDC; 2019.

Stockholm, January 2019

ISBN 978-92-9498-302-2
DOI 10.2900/629999
Catalogue number TQ-03-19-045-EN-N

Contents

Figures

1 Introduction

The HIV Estimates Accuracy Tool is an R-based application that uses statistical methods to allow for adjusted estimates from HIV surveillance data taking into account the issues of missing data and reporting delay. While it does not replace the knowledge of data analysis with adjustments, it is intended for routine application in surveillance as no complex programming skills are needed.

The tool accepts case-based surveillance data for HIV prepared in the format specified for the European Surveillance System (TESSy) uploads in RecordType=HIV or HIVAIDS. Case-based surveillance data containing the required set of variables and consistent with the TESSy format in coding of variables may also be used.

The tool performs multiple imputations for missing values using joint multivariate normal models (and extensions) or full conditional specification (also known as multiple imputation by chained equations – MICE). Additionally, the tool allows for correction of delays in reporting through reverse time hazard estimation. The adjustments may be used in combination or separately.

The outputs include results from a set of pre-defined analyses in the form of a report containing tables and graphs and data sets in various formats in which the corrections have been incorporated and are ready for further analysis.

This manual guides through the HIV Estimates Accuracy Tool, explains why each step of the tool may be needed and how to interpret the output and what actions may be needed to be taken based on the output.

1.1 Methodological background

Missing data occur when values for certain variables are not recorded. If cases with missing values are excluded from analysis, it may lead to biased and potentially less precise estimates.

Missing data arise from one of the following mechanisms:

- data missing completely at random (MCAR) – value is missing independently of the value itself and of any other factors including observable covariates.
- data missing at random (MAR) – value is missing independently of the value itself, but the fact that it is missing may depend on other covariates.
- data missing not at random (MNAR) – the fact that a value is missing may depend on the value that is not observed, e.g. transmission category is not recorded as sex between men due to possible stigma.

MCAR mechanism is rarely encountered, but in this case, even simple analysis excluding cases with missing values provides unbiased estimates. Furthermore, it is impossible to discriminate between MAR and MNAR based on observed data alone. Expert opinion regarding the details of the data collection process is needed. Typically, the analysis begins with an assumption of MAR and this is the focus of the tool.

It is also useful to check if data follow a monotone missingness pattern. In this pattern, incomplete variables can be ordered so that if the value of the first variable is missing, then the value of the second variable is as well, along with the values of all the following variables. In addition, regardless of the first variable, if the value of the second variable is missing, then the value of the third and all subsequent variables are also missing.

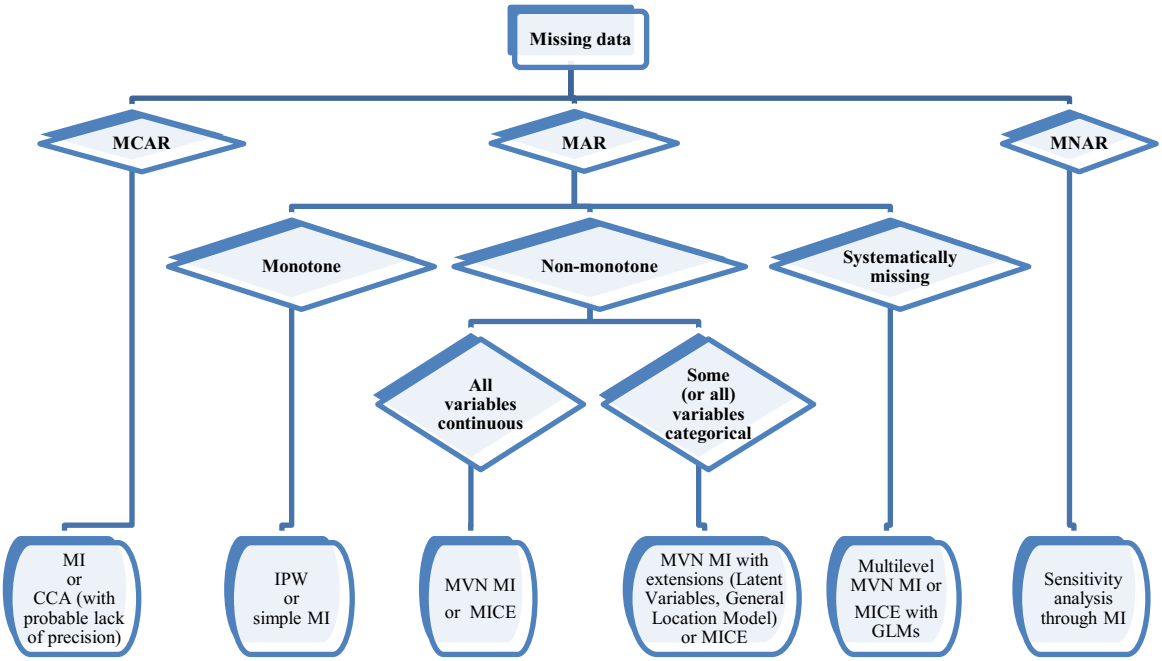
The most popular and flexible method of dealing with missing data (MCAR or MAR) involves multiple imputations (MI), firstly introduced by Rubin in 1987. The MI method involves filling each of the missing values with values randomly sampled from an appropriate distribution. The imputation is performed M times (typically 5–10) and in effect M so called pseudo-complete data sets are obtained. The model of interest (also called 'substantive model') can be fitted to each of the imputed data sets in order to estimate the parameter of interest and its variance M times. These can be combined using Rubin's rules to obtain an overall (average over M) estimator and its associated variance. This variance is enlarged to account for the uncertainty about the missing values.

The appropriate distribution to sample from is estimated from an imputation model. The main approaches of MI are based on joint modelling (multivariate normal model) or full conditional specification (MICE).

The multivariate normal imputation relies on the assumption that the joint distribution of all variables under consideration is multivariate normal. If data contain a mixture of continuous and categorical variables, multivariate normal MI can be extended to the latent normal or general location models. Alternatively, multiple imputations can be performed with the full conditional specification method (MICE). With MICE, separate specific models are constructed for each of the variables to be imputed depending on their type. These univariate models are fitted iteratively for each partially observed variable using both observed and previously imputed data of the remaining variables until the procedure converges.

Both the joint modelling and full conditional specification approaches can be extended to data sets combining data from different national surveillance systems through multilevel multiple imputation. The suggested approach to missing data is presented in Figure 1.

Figure 1. Appropriate methods to deal with missing data depending on missing data characteristics



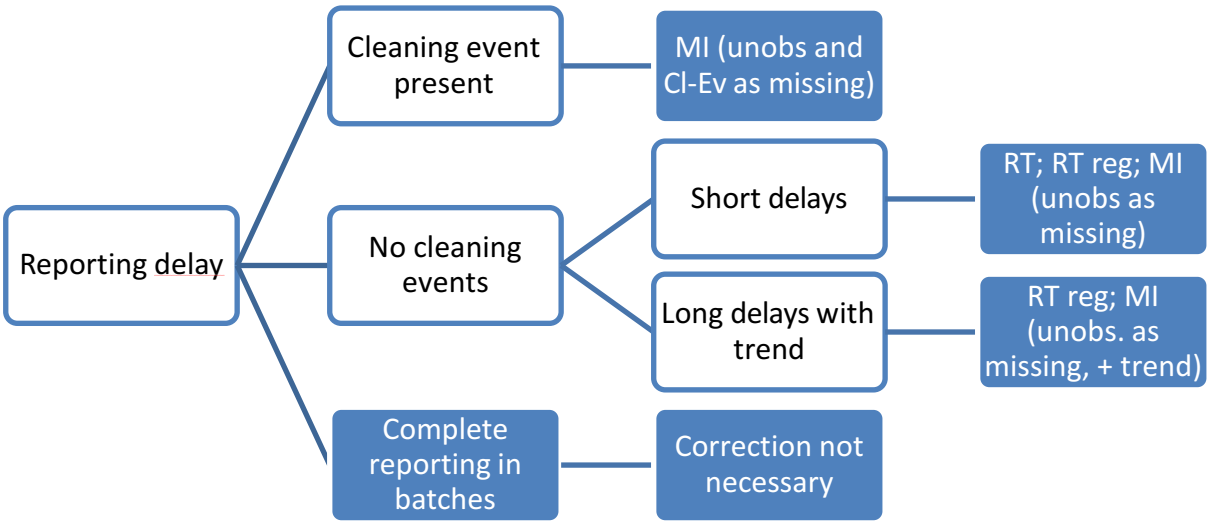
MCAR: missing completely at random
MAR: missing-at-random
MNAR: missing not at random
MI: multiple imputations
CCA: complete-case analysis
IPW: inverse probability weighting
MVN MI: multivariate normal MI
MICE: MI by chained equations.

Reporting delay is the time from case diagnosis to notification and it causes an artificial drop in the number of cases during the last data collection year. The majority of modern adjustment techniques rely on estimation of the delay distribution independently of the diagnosis rate. Once the estimate for the delay distribution is obtained, it is used to estimate the proportion of cases already reported given the diagnosis date and end date of data collection.

The reporting delay distribution can be estimated in a non-parametric way using a multinomial model (assuming there is a maximum delay) or reverse time transform and estimating the survivorship function with left-truncated data. In practice, both confidence intervals and point estimates for delay probabilities are equivalent for the two approaches. Both models allow incorporation of covariates that may impact the reporting delay, including the time of diagnosis. Alternatively, missing data techniques as discussed above could be applied. In this method, the counts of the cases, which will be reported with delay, are treated as missing and imputed. This technique also allows to remove data from the time periods, when specific activities were undertaken in surveillance system, which could alter the usual reporting delay patters. In e.g. this could refer to control activities that result in reports of old cases ('cleaning event').

Increasingly, HIV surveillance systems rely on cyclic uploading of complete data on new diagnoses during a predefined period of time from laboratory databases. In case of such batch reporting, delay may still be calculated, but using adjustment methods is not necessary. The suggested approach to reporting delays is presented in Figure 2.

Figure 2. Appropriate methods to deal with reporting delays



MI: multiple imputations of yet unobserved counts and artificially removed counts recorded during cleaning events
Cl-Ev: cleaning event
RT: reverse time estimation of reporting delay distribution
RT reg: reverse time estimation based on Cox proportional hazard regression.

1.2 Methods used in the HIV Estimates Accuracy Tool

The tool offers a possibility to perform both joint modelling (through multivariate normal model) and full conditional specification MI. Joint modelling is implemented with the 'jomo' R-package, full conditional specification through the 'mice' R-package and application of Rubin's rules through the 'mitools' R-package.

The tool first imputes missing values for gender (single imputation). Since other variables are imputed separately for males and females and gender is missing only for a small proportion of case, this simplifies the procedure. Gender 'Other' is imputed as either male or female. This is a simplification for statistical procedures, but for inference, it is recommended to go back to the original code for these cases.

The imputation model for males and females includes variables to be imputed (transmission category, migrant status, CD4 count – unless missing completely, age at diagnosis) and variables considered to always be known (AIDS at diagnosis, diagnosis year). The flexibility of this model includes the possibility to exclude CD4 count, transmission and/or migrant status (done automatically if the variable is systematically missing) and modelling of the time trend. A flexible model of the time trend is included in the form of cubic spline. The number of knots of the spline may be selected by the user in the 3–5 range.

Obtaining appropriate imputation requires a procedure that allows estimation of joint distribution. This is an iterative procedure that has to converge before the samples may be drawn to impute the missing values. The number of iterations needed for the procedure to converge is called burn-in. In addition, a number of iterations is necessary between subsequent imputations in order to avoid autocorrelation of these imputations.

Basic estimates before and after MI adjustments obtained using Rubin's rules and appropriate models are implemented within the interactive report. The report supports estimates obtained with the spline model of the trend, i.e. a congenial model with the imputation model, and also a discrete model for the diagnosis year. The first one provides smoothed estimates that may be quite different from the actual case counts observed in surveillance.

When adjusting for reporting delay in surveillance, the time units used vary from one day to one year. HIV data in Europe are traditionally collected quarterly. In addition, data are usually presented annually, so only longer delays of several months can lead to underestimating the number of diagnoses in the most recent years. Accordingly, a quarter was selected as an appropriate unit for measurement of the reporting delay.

The reporting delay is calculated only if both the quarter of the diagnosis and notification are available. In case the calculated value is less than 0, it is set to missing. The estimation of the reporting delay distribution is performed using the records, which contain a valid value for the reporting delay variable, unless imputation of the reporting delay is selected. In the latter case, reporting delays are imputed along other variables containing missing values based on other covariates as well as available information about the dates (maximum plausible reporting delay).

Truncation time is assumed to be the latest notification quarter that occurs in the data set. However, truncation time may be manually changed by the user in the reporting delay parameter window, e.g. if data do not entirely cover the last quarter. In addition, the user may choose to limit data only to cases diagnosed recently.

The reporting delay distribution is estimated based on survival techniques. Firstly, reverse time transform is applied, subtracting the reporting delay from truncation time and taking the diagnosis quarter as entry time. Next, standard survival techniques for right truncated data are applied, including stratified estimation of survival curves or proportional hazard regression model. The stratification covariates may be selected from transmission category, migration status and sex. If missing values in the covariates are encountered, they are treated as a separate category. The proportional hazard regression model contains the year of diagnosis by default as predictor in addition to other selected variables.

Individual weight is assigned for each case based on covariate pattern and the number of quarters between diagnosis and truncation time. Next, the adjustment formula, which makes use of both the weight and the case count by covariate pattern, is applied in order to obtain adjusted counts and respective standard errors, for each distinct covariate pattern. Further, these adjusted counts are combined under the assumption of independence into an overall estimate.

Reporting delay estimation models do not account for possible differences in reporting during the year. If cases are uploaded in batches, e.g. once per year, the estimates provided by the tool will not be valid.

If both adjustments are selected, the tool will first perform the imputation, then calculation of reporting delay weight. Reporting delay distribution estimation is performed separately for each imputed data set. Weighted (adjusted for reporting delay) estimates are produced for each imputed data set, which are then combined using Rubin's rules.

The report can be produced with both adjustments or with only one of them.

1.3 Specific issues

This section collects information on issues that may be encountered on different stages of data preparation, interpretation and outputting in one place.

Issue	Impact	Suggested solutions
Acceptable level of missingness	There are no clear guidelines on acceptable levels of missingness. However, any violation of the imputation model's assumptions will have more pronounced consequences with high proportions of missing data.	In EU/EEA HIV surveillance data, missingness in most of the key covariates is below 20% with the exception of CD4 count. The tool uses methods to minimise the impact of non-normally distributed CD4 count. In case of a high percentage of missing values, consider increasing the number of imputations beyond the typically used number of 5–10, as otherwise the estimates can be inaccurate.
Systematically missing CD4 count	It has impact on imputation of missing values. If detected, the tool will proceed with reduced imputation models that do not contain CD4 count.	The imputation is still valid, except that no outputs are produced with CD4 counts. CD4 counts will not be imputed in this case.
Negative values in imputations	Imputations use normal-based models. On rare occasions, the values in one imputed set may be not plausible (e.g. negative CD4 counts).	This is a correct value, as the estimations are based on multiply imputed sets.
Incomplete information for reporting delay variables	The reporting delay weights are not calculated (set to 1) if the reporting delay variables are missing. In addition, in case of regression method (reporting delay with trend), the weights are not calculated for cases with missing predictors.	In case the level of missingness is substantial for reporting delay, the adjustment may not be appropriate. In case of moderate missingness level, including the imputation of reporting delay is suggested.

1.4 Purpose of manual

This manual goes through the tool step by step and explains the functionalities of each part. Section 2 covers the data preparation side of the tool.

For each section on tabs and functionalities that the tool provides (section 3 onwards), the following items are included:

- Description – provides a short description on what the corresponding elements of the tool are and what type of output they provide.
- Process – what to do with the output provided by Stata.
- Interpretation – meaning of output is described.
- Further actions – what to do if there are any output issues. This may mean carrying out another analysis or modifying the data.

Disclaimer: the dummy data set based on the TESSy HIV data set is used as a model for this manual. This data set was developed solely for training purposes. Data do not refer directly to any country, has not been validated by ECDC experts and results produced in this manual cannot be interpreted and used for any reliable inferences.

2 Prerequisites

2.1 Data set

- The file should contain case-based records of HIV diagnoses.
- There are 19 required attributes/variables by the tool to run the adjustments. They are presented in Table 1 with the description of values required for each of the attribute/variable.
- Upload file should contain all these attribute/variable names except empty columns and columns containing a single value (e.g. ReportingCountry), which can be created directly in the tool.
- Different variable names are accepted by the tool as long as they can be mapped directly to these required variables in the 'Attribute mapping' utility in the tool. However, the variables must be coded as specified.

Number	Attribute/variable name	Description (as in TESSY metadata set 36 HotFix5)	Required values
1	RecordId	Unique identifier for each record within and across the national surveillance system	
2	ReportingCountry	Country reporting the record, according to the ISO list	Annex 1
3	Age	Exact age at diagnosis of HIV. Age as a crude number is preferred - calculated from date of diagnosis	0–100
4	FirstCD4Count	Variable specifies the CD4 cells count at the time of HIV diagnosis. Enter numeric value of CD4 (0–6000) or unknown (UNK)	0–6000
5	FirstCD4DateYear	Year of first CD4 cell count at time of diagnosis	>1985
6	CountryOfBirth	Country of birth of the patient according to the ISO list. Certain additional values used in surveillance also included (Annex 1). CountryofBirth preferred variable for migration status. If unknown, code as UNK or Blank	Annex 1
7	CountryOfNationality	Country of nationality of patient, according to the ISO list. Some additional values used in surveillance are also included (Annex 1)	Annex 1
8	RegionOfOrigin	Region of origin of patient	Annex 1
9	DateOfAIDSDiagnosisYear	Year of AIDS diagnosis For HIV cases initially reported at a pre-AIDS stage, the date of AIDS diagnosis is 'follow-up' information, which necessitates updating the record.	≥1984
10	DateOfDeathYear	Year of death because of HIV/AIDS	
11	DateOfDiagnosisYear	Year of first HIV diagnosis; clinical or laboratory diagnosis. Missing values not allowed.	≥1985
12	DateOfDiagnosisQuarter	Quarter of first HIV diagnosis; clinical or laboratory diagnosis	1,2,3,4
13	DateOfNotificationYear	Year HIV case was notified for first time to reporting country	≥1985
14	DateOfNotificationQuarter	Quarter in which HIV case was notified for first time to reporting country.	1,2,3,4
15	Gender	Patient gender. Transsexual should be coded as O-Other.	F=Female M=Male O=Other UNK=Unknown
16	Outcome	Information on whether case is alive or deceased. Death should be due to reported disease.	A=Alive D=Died UNK=Unknown
17	PlaceOfNotification	Place of first notification of case to regional authority. Select the most detailed NUTS level possible.	
18	PlaceOfResidence	Place of residence of patient at disease onset. Select the most detailed NUTS level possible.	
19	Transmission	Describes most probable route of transmission Nosocomial infection includes patients infected in health care settings. Case of occupational exposure should be classified as UNK 'Unknown or undetermined'. Cases that are not fully documented should be coded as UNK.	Transmission: HAEMO=haemophiliac patient HETERO=heterosexual contact IDU=ever injected drugs MSM=MSM/homo or bisexual male MTCT=mother-to-child transmission NOSO=Nosocomial TRANSFU=transfusion recipient Unk=Unknown or undetermined

- Out of 19 required attributes/variables by the tool:
  - Outcome, PlaceOfNotification, PlaceOfResidence, DateOfDeathYear and FirstCD4DateYear are not used by the current version of the yool and may be replaced by a column of missing values.
  - DateOfDiagnosisYear and DateOfAIDSYear are considered fully observed.
  - Imputation variables. transmission, CD4 count and migration variables (CountryOfBirth, CountryOfNationality, RegionOfOrigin) may have missing values, but if they are entirely missing, they will be excluded from imputation models.
  - Reporting delay variables DateOfDiagnosisYear, DateOfDiagnosisQuarter, DateOfNotificationYear and DateOfNotificationQuarter may have missing values (with the exception of DateOfDiagnosisYear), but if any are missing, a reporting delay is available.

  If one of the variables is not present in the data set, it may be artificially created (see 'Default values' inthe attributes mapping widow description).
- If the file to upload to the tool was previously uploaded to the TESSy database and successfully passed TESSy validation, there should be no problem with using it in the tool unless all 19 required by the attributes/variables are present in the file.

2.2 Online version

For restricted users, the HIV Estimates Accuracy Tool is a web tool available online through Shinyapps at http://ecdc.shinyapps.io/hivestimatesaccuracyui.

To access shinyapp.io, the user needs to create an account on Shinyapps.io. Authentication is possible through one of three methods:

- Shinyapps.io authentication. A new Shinyapps.io account can be created during the authentication process and requires only an email address and new password for the account.
- Google Authentication; or
- GitHub Authorisation.

To access the link, send an email to HIV.Modelling@ecdc.europa.eu with the subject 'Registration for HIV Estimates Accuracy Tool + Full Name'. An invitation email will be sent back.

The online version requires no installation. An active Internet connection is required. It is advised to use relatively recent versions of web browsers such as Chrome, Firefox, Internet Explorer, Edge or Safari with support for JavaScript enabled.

2.3 Offline version

There are two ways to download the offline version.

For experienced R users, a CRAN-like repository is set up for installing the tool as an R package in R GUI or RStudio.

The repository of R packages is available here: http://www.nextpagesoft.net/hiv-estimates-accuracy/repo. The tool can be installed using standard R commands executed in R console.

- Type
  > install.packages("hivEstimatesAccuracy", repo="http://www.nextpagesoft.net/hiv-estimates-accuracy/repo")
  and press ENTER. This will download and install latest version of the tool and all its dependencies.
- Once R is done with installation the tool can be run with command:
  > hivEstimatesAccuracy::RunApp()
- Periodically, the user can update the tool with the following command:
  > update.packages(repo='http://www.nextpagesoft.net/hiv-estimates-accuracy/repo')

Offline Windows x64 deployment package with embedded R environment

The other option for local tool deployment is to use a deployment package that includes all required software and R packages.

- Download the deployment package through the following link:
  http://www.nextpagesoft.net/hiv-estimates-accuracy/win/x64/hivEstimatesAccuracy-1.0.0.x64.zip (201 MB download size)
- Unpack the file to an arbitrary folder.

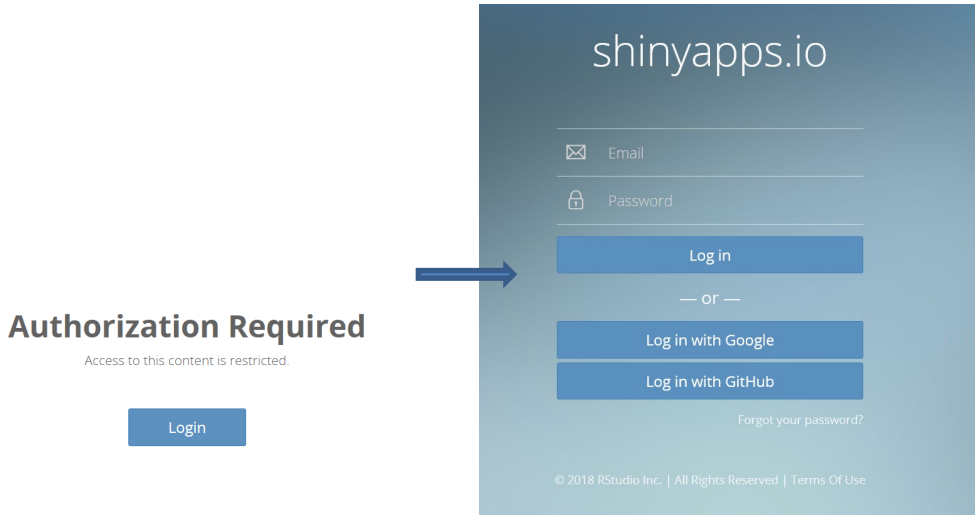
- After unpacking, a new folder will appear called 'hivEstimatesAccuracy'. Open the file 'hivEstimatesAccuracy.bat'. This will open the tool in the default web browser. When finished, close the browser window.

This offline package can be run only on 64-bit versions of Microsoft Windows (7, 8 and 10).

3 Using the HIV Estimates Accuracy Tool

3.1 How to open the tool

When accessing the online tool at http://ecdc.shinyapps.io/hivestimatesaccuracyui, the user will be taken to an authorisation screen.



To log in, the user needs to have an account on Shinyapps.io. Authentication is possible through one of three methods:

- Shinyapps.io authentication
- Google Authentication; and
- GitHub authorisation.

ECDC must register the user prior to first use Send an access request email to hiv.modelling@ecdc.europa.eu.

The offline version installed as R package can be opened by executing the following command in the R console:

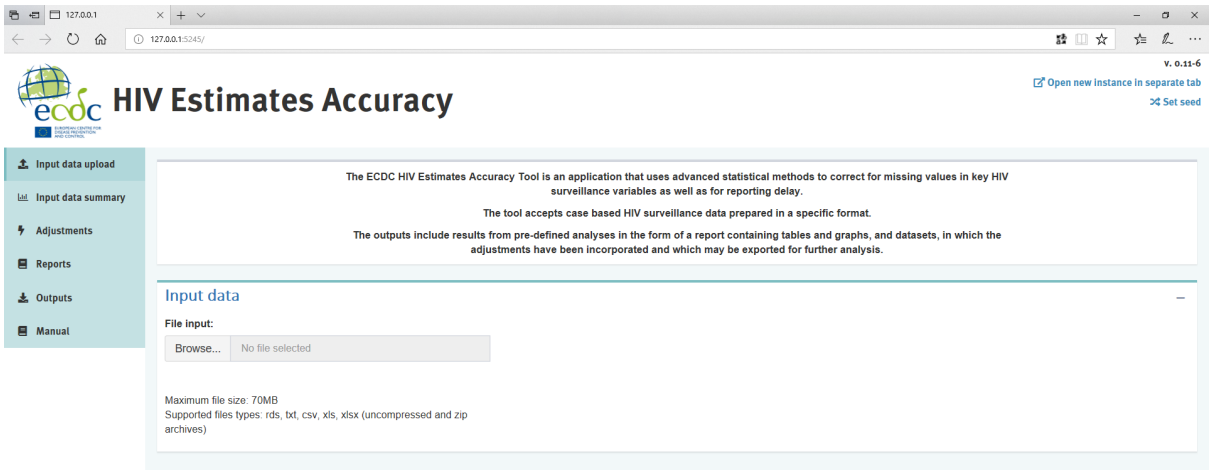
```
> hivEstimatesAccuracy::RunApp()
```

The offline version installed using the Windows x64 deployment package can be opened by clicking on 'hivEstimatesAccuracy.bat'.

In both cases, the tool will open as a new window in the default browser.

3.2 Construction of the tool

The tool is organised into tabs displayed in the left panel. It automatically opens at the first tab: Input data upload.



Tabs:

- Input data upload – further tabs are not active unless the data are uploaded and validated in this tab
- Input data summary – allows exploration of data and selection of filters for adjustments
- Adjustments – tab to specify adjustments and parameters for adjustments as well as to examine diagnostic charts
- Reports – allows creating and exporting a predefined report
- Outputs – contains output data sets that can be used for further analysis; and
- Manual.

The user can navigate freely between the tabs once the data set is uploaded.

4 Input data upload tab

This tab is always active and allows for uploading and preprocessing data.

4.1. Uploading data

Description

The tool allows case-based data sets corresponding to TESSy format. Supported files types include rds, txt, csv, xls and xlsx (uncompressed and zip archives). In the case of using the online version with larger data files, it is recommended to use Zip archives to speed up the data upload process.

Process

Select the 'Browse' button in the Input data section and navigate to the location of the data file.

Interpretation

The tool provides data summary (i.e. size of file, number of records, variable names) and opens new sections: 'Attributes mapping' and 'Migrant variables regrouping'.

Further actions

Check that the number of records and variables are uploaded correctly. Proceed to the 'Attributes mapping' section.

4.2 Uploading a saved application state

Description

The tool allows for uploading of a previously saved application state that contains uploaded and preprocessed data, as well as adjustments that were previously applied.

Process

Select the 'Browse' button in the Input data section and navigate to the location of the saved file. The file has the extension '.rds'. The default name starts with 'HIV_state_', followed by the date it was saved, but the file can be saved with the name specified by the user.

Interpretation

The previous work is uploaded. Mapped and preprocessed data are available for further analysis.

Further actions

Proceed to further tabs to continue the analysis.

4.3 Mapping and validating data

Description

The 'Attributes mapping' section offers the possibility to match between variable names used internally by the tool ('Attribute' column) and variables present in the input data ('Input data column'). The variable names used by the tool correspond to those used in the TESSy metadata set. If the variables in the input data have the same or similar names, they will be automatically identified by the tool and suggested in 'Input data column'. If the tool cannot identify the mapping, the field will be left blank.

Process

The mapping automatically proposes assigning the variables with names similar to or the same as the ones used by the tool. Other variables are manually mapped by selecting the appropriate variable (from the input data set) from the dropdown menu.

In case the variable has a single value and is not specified in the data set, it can be created directly in the tool by leaving 'Input data column' blank and specifying the variable value in the 'Default value' column.

If data are not available for a variable, the column in the tool can be created by entering 'NA' in the 'Default value' column.

When ready, click the 'Apply mapping' button at the top of the section.

Interpretation

Clicking on 'Apply mapping' will implement variable assignment and the validity checks. The tool automatically checks if the mapped variables contain valid values as required in a given covariate. A successful mapping process is indicated by the statements 'Mapping is valid' and 'Values are valid'.

In case of failed mapping, information is displayed as to which variable is problematic and the nature of the problem.

Valid mapping automatically triggers the preprocessing of data. During the preprocessing, a migrant status variable is created based on the following variables: CountryOfBirth, CountryOfNationality, RegionOfOrigin and AIDS at diagnosis based on DateOfAIDSDiagnosisYear and DateOfDiagnosis. Moreover, a single imputation of gender is performed.

Preprocessed data may be inspected at the bottom of the page under 'Input data records pre-processed'.

Further actions

Once the validity of mapping and values of the variables are confirmed, proceed to further tabs. Regrouping of the migrant variables is also possible.

4.4 Defining the migrant variable categorisation

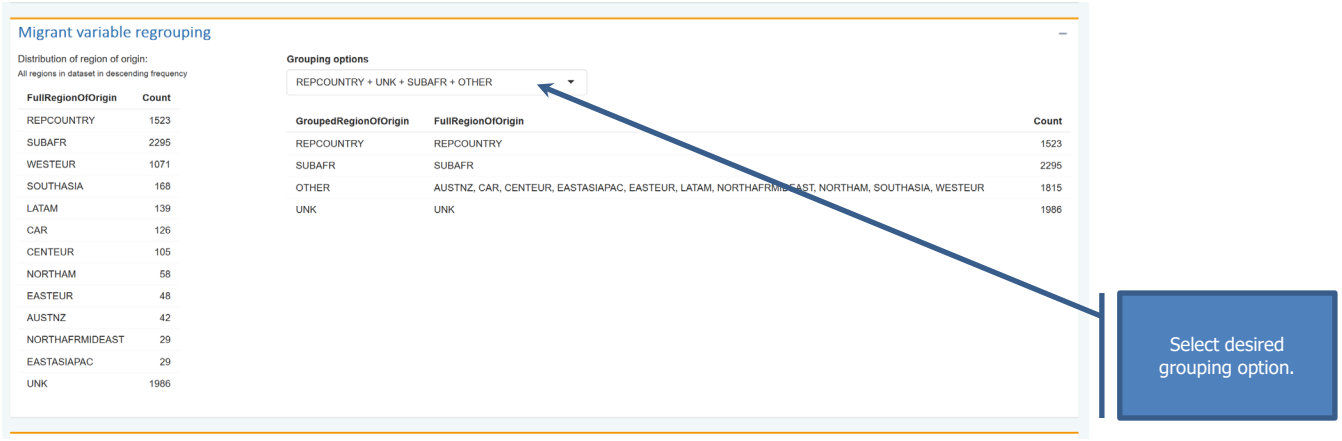
Description

The migrant status variable is created based on the following variables: CountryOfBirth, CountryOfNationality and RegionOfOrigin in combination with ReportingCountry. Based on this, regrouping the FullRegionOfOrigin variable is created based on categorisation used in TESSy (Annex 1). The FullRegionOfOrigin variable may be regrouped into categories that are the most relevant to the particular country.

Process

The following options are available:

- REPCOUNRTY+UNK+OTHER
- REPCOUNRTY+UNK+SUBAFR+OTHER
- REPCOUNRTY+UNK+3 most prevalent regions+OTHER; and
- Custom.



Custom regrouping may be created by selecting 'Custom option' and creating a group using the option 'Add group'.

Regions may be added to the group by clicking on the 'FullRegionOfOrigin' field and adding regions from the drop-down menu. Regions are added by clicking on appropriate names.

Unselected regions are automatically grouped into the 'OTH' group.

Interpretation

Distribution of cases by RegionOfOrigin is provided to guide grouping. After grouping, the number of cases in each group is automatically provided. Small numbers in particular groups can cause instability of adjustments and should be avoided.

Further actions

Select appropriate grouping and proceed to further tabs.

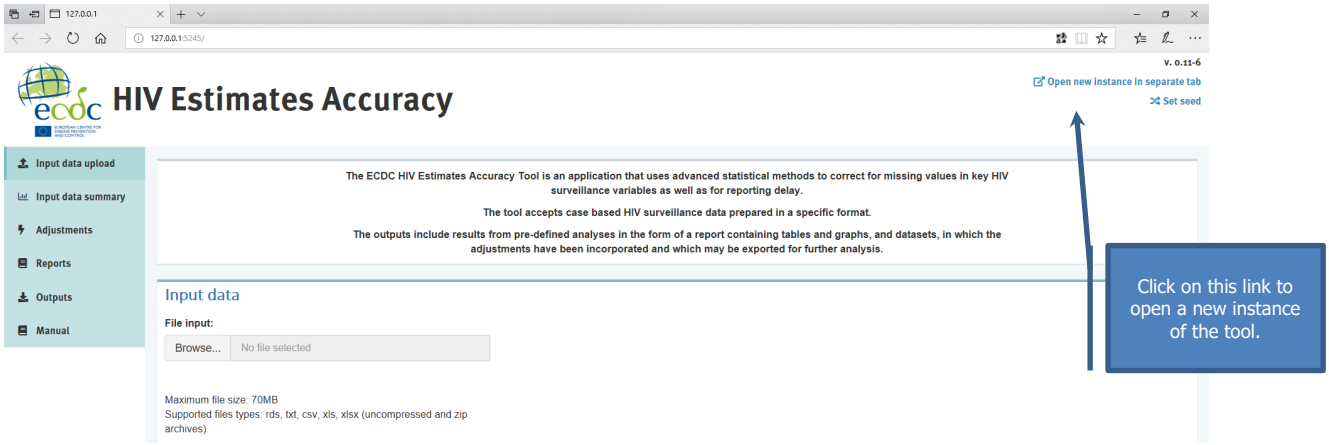
4.5 Opening a new instance of the tool

Description

It is possible to work with more than one window (instance of the tool) open. All instances will operate separately and independently. Data/saved workspaces mustt be uploaded independently to each instance of the tool.

Process

In order to open a new instance of the tool, select the 'Open new instance in separate tab' button in the top right corner. The new instance may be opened at any time of the analysis by going back to the 'Input data' upload tab.



Interpretation

The tool will open a new empty tab, requiring a new data upload. The user can avoid duplicating the mapping process by first saving the workspace with preprocessed data for further upload in the new instance of the tool.

Further actions

Proceed with adjustments.

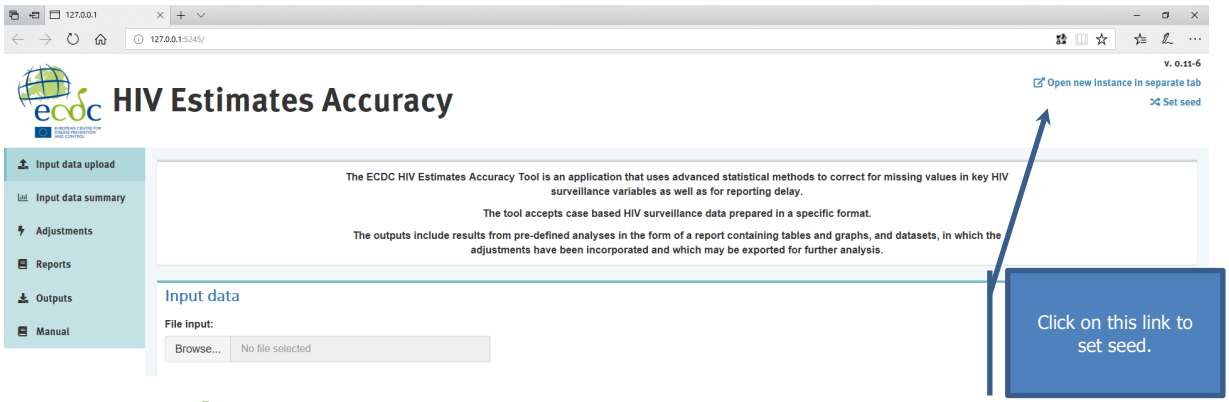
4.6 Setting the seed for the random processes used by the tool

Description

The tool uses a random number generator when imputing missing values. This means that each time the adjustments are run, the results could be slightly different. In order to receive exactly the same results, the random number generator should be initialised with the same number (seed).

Process

To set up the seed, select the 'Set seed' button in the 'Input data' upload section. A pop-up window will allow the user to enter the number to become a seed. Enter an empty value or type "default' to remove the fixed seed.



Interpretation

The seed is set that will be used in the further analysis.

Further actions

Proceed with adjustments.

5 Input data summary tab

This tab allows for inspecting data quality issues present in the input data.

5.1 Inspecting missing data patterns

Description

Section 1 provides a summary of the missing values for key epidemiological variables: age, CD4 count, transmission and migration status, overall and separately for each gender. There are no missing values for gender as these are imputed at the data preprocessing step.

Section 2 provides a summary of the trends for the proportion of missing values in key variables by diagnosis year.

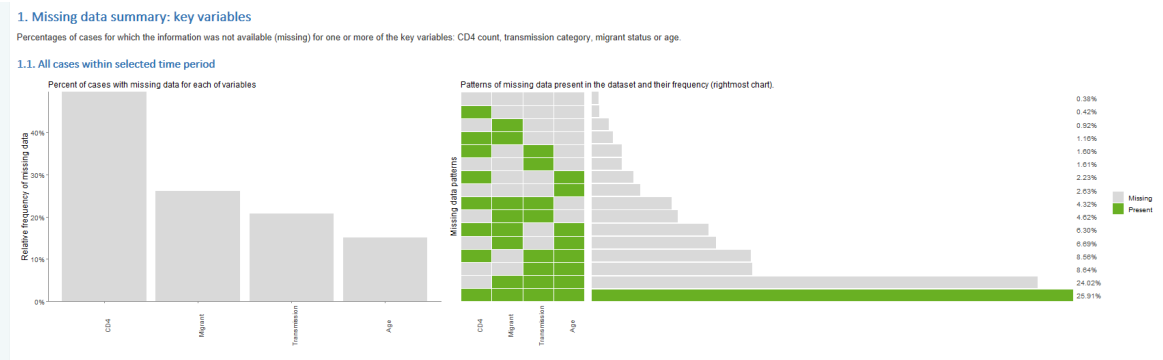
Process

The output is generated automatically when moving to the Input data summary tab. The user can select time periods for which data are summarised by selecting filters.

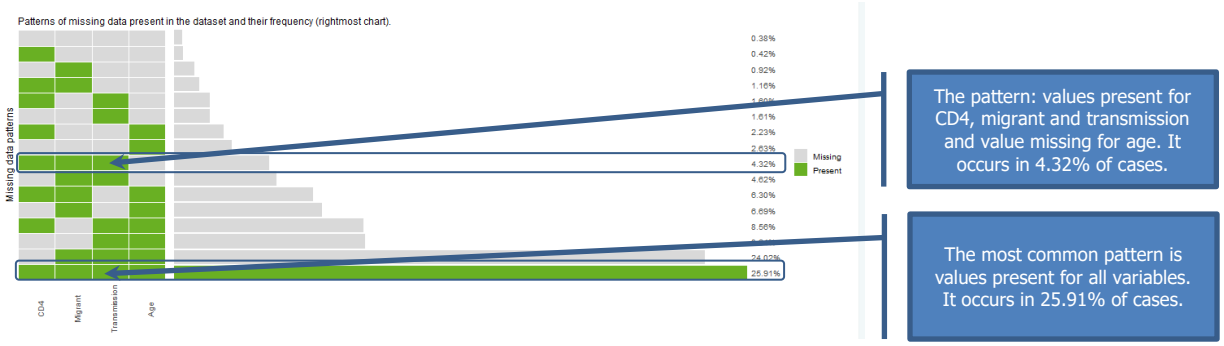
Interpretation

In section 1, for all cases, as well as separately for males and females, two graphs are presented:

- percent of cases with missing data for each of variables; and
- patterns of missing data present in the data set and their frequency (right chart).



The graph showing patterns of missing data at the right side displays which patterns of missing/present values are present in the data. A pattern is defined by which of the four variables considered are present (green) and missing (grey). It is displayed on the graph as green or grey boxes in columns corresponding to the particular variables. The left side of the chart shows the distribution of missing values patterns in the data. This indicates in what proportion of cases a values-specific pattern of missing data occurs. The pattern for which values are present for all considered variables is displayed in green. Patterns are sorted by the frequency in which they occur in data.



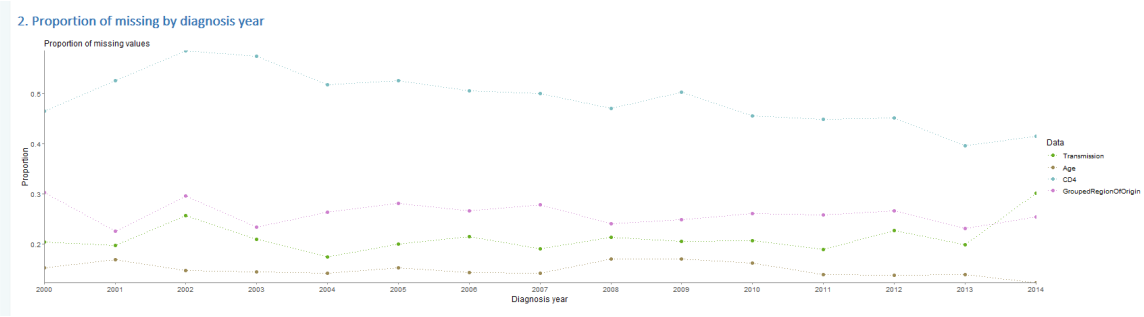
The graphs allow for checking if the reported levels of missing values are correct for the input data and may help in deciding whether the data should be restricted for further analysis.

Additionally:

- If a variable is completely missing, it will not be used in the imputation models and it will not be imputed.

- If specific variables tend to miss together, it indicates that the variables are not missing at random and analysing only the complete cases may lead to bias.
- The pattern of missing values may be monotonous or heterogeneous (non-monotone). Monotone missing patterns are represented as 'grey triangles' without green cells within. Patchy patterns indicate heterogeneity. At the moment, the adjustment methods implemented in the tool assume non-monotone missing patterns. They are also valid (although less efficient) for the monotone missing pattern.

In section 2, the proportion of missing values in each of the considered variables is provided.



When looking at the trends, it is particularly important to look for time periods when variables were entirely missing. These may occur if a variable was introduced to surveillance at one point in time and it is not available for cases reported before that date.

Including such historical data in imputations will result to a certain degree of extrapolation of available data to periods with no available data. If periods with no available data are long, the imputations may be less accurate.

Further actions

Select the data period for adjustments and proceed to further tabs.

5.2 Inspecting reporting delay patterns

Description

Section 3 provides information on the availability of data necessary to calculate the reporting delay, i.e. year and quarter of diagnosis year of notification and quarter of identification.

Section 4 displays the observed distribution of the reporting delay. The distribution is smoothed to provide the overall picture of the reporting delay. The picture can be also generated for a subset of data through filtering data by the year of diagnosis and/or notification at the top of the tab.

Section 5 displays the average delay by notification quarter. It also provides a line for trends in reporting delay by notification quarter and the upper limit of typical delays given the variability of the average reporting delay by quarter.

Process

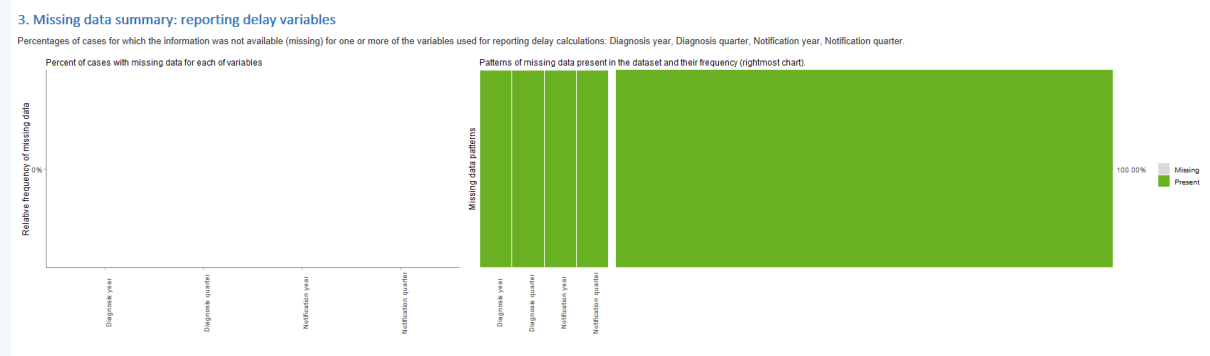
The output is generated automatically when moving to the Input data summary tab. The user may select time periods for which the data are summarised by applying filters.

Interpretation

The two graphs displaying the patterns of missing values are similar to graphs for the other variables:

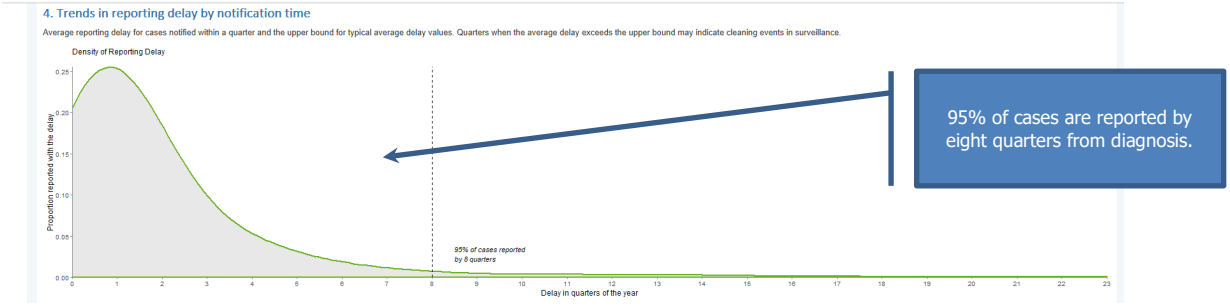
- percent of cases with missing data for each of variables; and
- patterns of missing data present in the data set and their frequency.

In the example below, the required variables are 100% complete, so only one pattern of missing/available values is present.

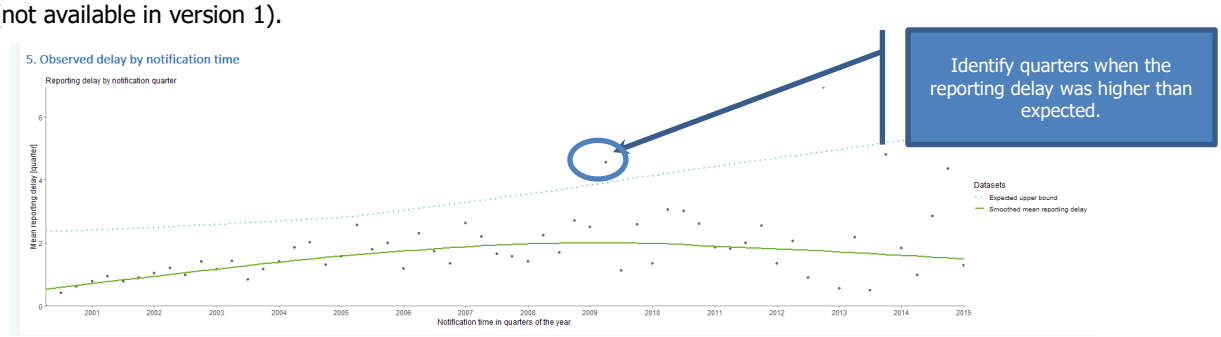


The reporting delay weight is estimated based on cases for which all four variables are available (100% complete pattern). If missing data are present, the results may be less accurate, especially in case of substantial missingness for reporting delay variables.

The next section provides observed distribution of reporting delay. It provides an overview of how important the reporting delay is in the input data set. This distribution does not represent the real distribution of reporting delay as cases not yet reported will have a longer delay, so the observed distribution underestimates the true distribution. The vertical line represents the quarter in which 95% of cases were reported. Since data are also usually analysed with some delay, if 95% of cases are reported within two quarters, the delay adjusting for reporting delay will not make much difference. In the case of the sample data, this is eight quarters, indicating important delays.



In the last summary, average by notification quarter allows for identifying quarters during which the average reporting delay exceeded expected values, indicating a 'cleaning event'. The blue line indicates the threshold for a possible cleaning event. If there is a cleaning event, especially if it lasts longer than just one quarter or takes place in more recent years, one should consider applying reporting delay correction by missing values imputation method (not available in version 1).



Further actions

Decide whether the reporting delay correction is necessary. In case of large proportions of missing values in the variables required for calculation of the reporting delay, consider using the 'Impute reporting delays' option in the multiple imputation parameters.

5.3 Applying filters

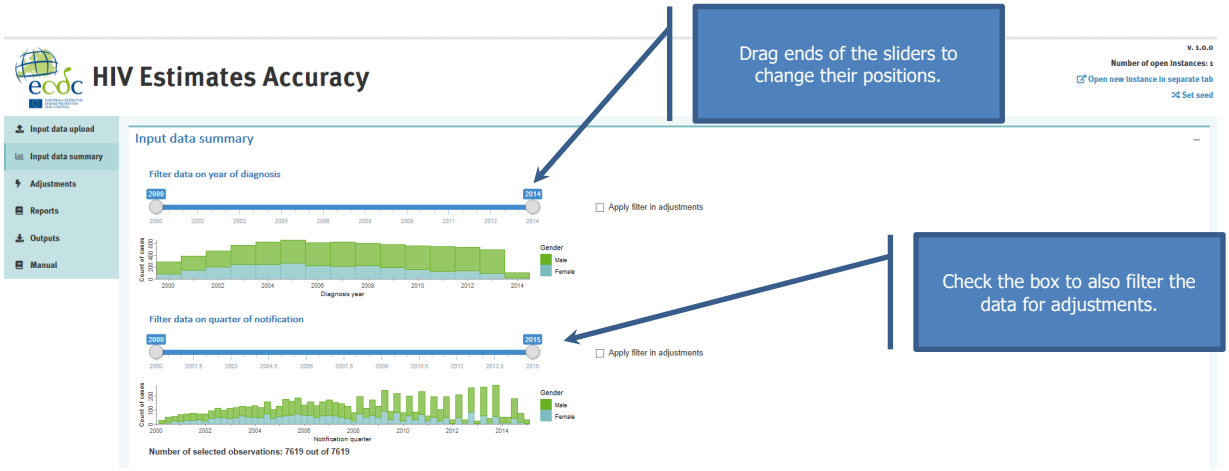
Description

The Input data summary tab allows for applying filters on the year of diagnosis and the year and quarter of notification. These filters may be applied to inspect the data in the Input data summary tab, but can be also

passed onto the adjustments. When passed onto the adjustments, the filtering will also have an effect on the output data sets.

Process

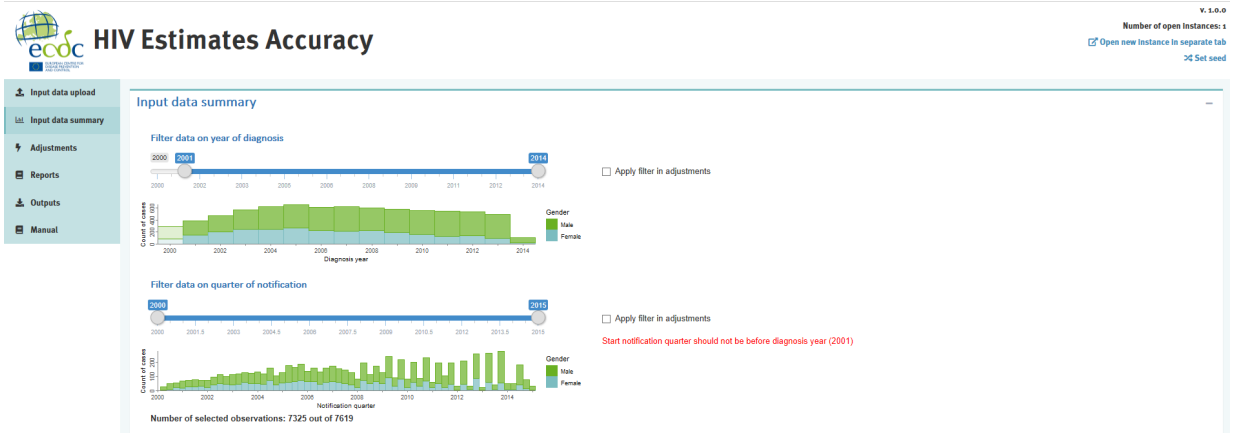
The filters may be applied by using sliders. Both the start and end times may be changed for both the year of diagnosis and time of notification. The chart below each slider shows the distribution of cases by gender among the included and excluded cases. The application of filters has an immediate effect on the graphs in the same tabs. The selected filters may be also applied to data that will be used for adjustments by checking an appropriate box.



Interpretation

Filtering used for adjustments will also have an effect on output data. Only cases meeting filtering criteria will be included in the output data set.

Both the slider for the diagnosis year and notification time may be changed freely. However, it is not recommended for adjustments to apply a set of filters for which the earliest year of diagnosis is before the earliest year of notification. This may lead to overestimation of the reporting delay, as among cases diagnosed in the period prior to the earliest notification time, only those reported with delay will be included. In case such a filter is included, the tool issues a warning.



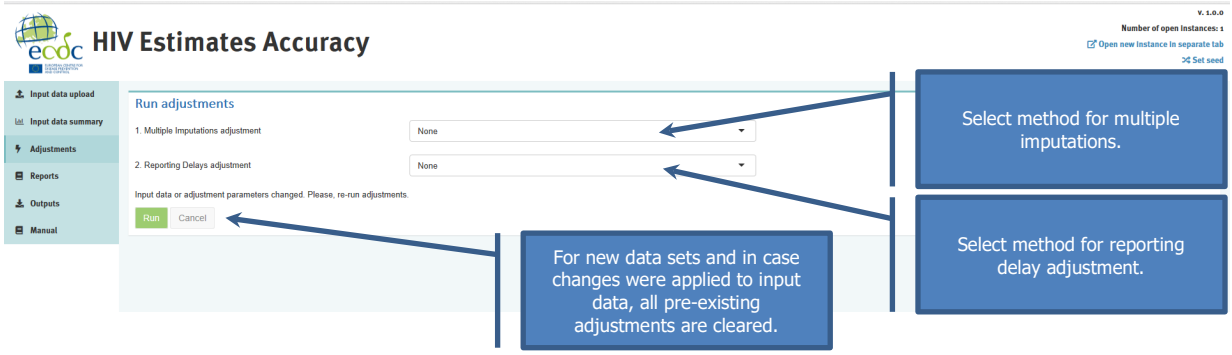
Further actions

Inspect filtered data. Decide on filtering to be used for adjustments. Proceed to further tabs.

6 Adjustments tab

The adjustment tab allows the user to specify adjustments and their parameters, apply them and look at the diagnostics output.

In case a new data set is uploaded or the uploaded workspace is changed, e.g. through application of (different) filters, the remaining part of this tab and any pre-existing adjustments are automatically cleared.



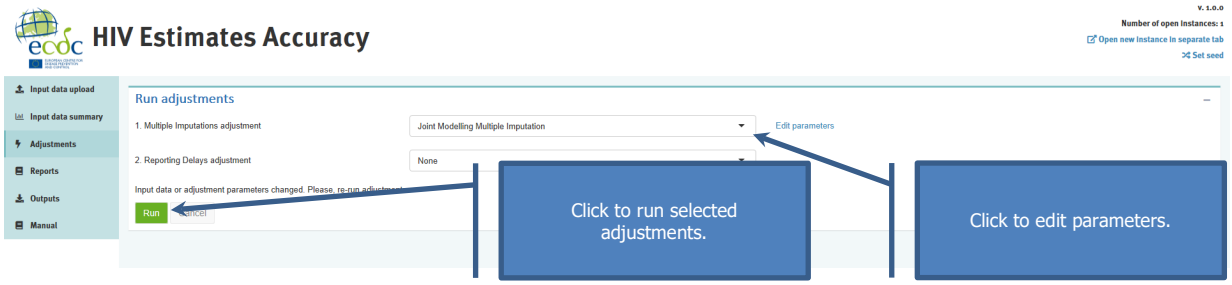
6.1 Joint modelling multiple imputation

Description

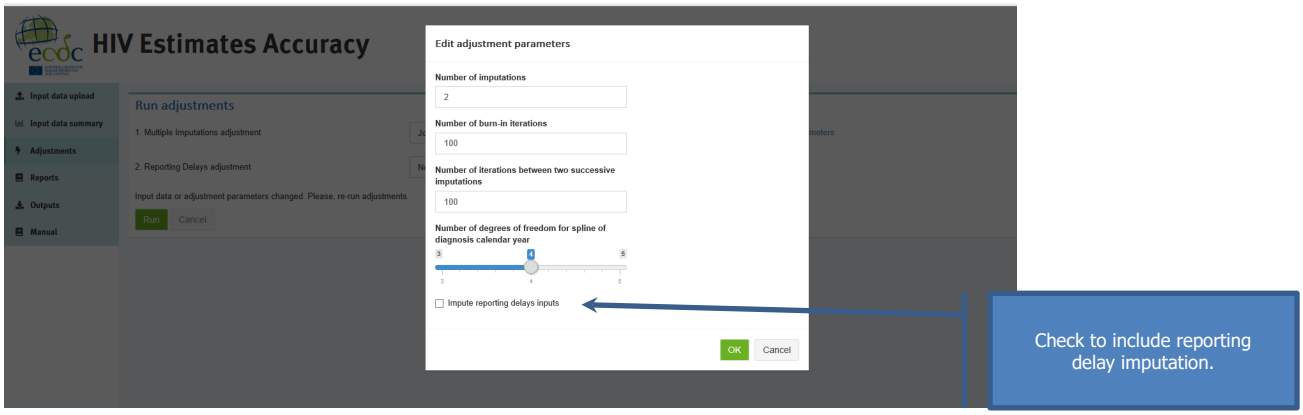
This option performs multiple imputations with joint multivariate normal modelling. This is an iterative procedure that can be time-consuming. The amount of the time needed depends on the parameters set. It is recommended to start with lower numbers to look at the outputs and allow up to several hours for the final runs.

Process

Select 'Joint Modelling Multiple Imputation' from the drop-down menu in the 'Multiple Imputations adjustment' field. The option to edit parameters will appear on the right.



A pop-up window is displayed with preset parameter values can be edited. The interpretation and selection of proper values is provided below. If imputation of the reporting delay is intended, it should be specified by checking the box at the bottom.



Interpretation

Parameters relating to imputation procedure and imputation model are displayed below. Refer to the diagnostics section in order to select proper values.

Parameter	Description	What to select
Number of imputations	Number of imputed data sets that will be produced	For test runs, select 2. For the final adjustments, at least 5–10 imputations.
Number of burn-in iterations	Number of iterations after which method should converge	For test runs select 100. Generally higher numbers (order of thousands) are needed and this can be decided based on the adjustment diagnostics
Number of iterations between 2 successive imputations	Number of iterations between outputting the successive imputed data set, which should limit autocorrelation of imputed data sets	For test runs, select 100. Usually this is sufficient or too high. Refer to the adjustment diagnostics.
Number of degrees of freedom for splines of diagnosis calendar year	Parameter used to determine the degree of flexibility of the time trend in data (number of cases per year or median CD4 count per year)	Select between 3 and 5. Choose one that results in a best fitting model. Choose higher numbers if expecting fast-changing trends and highly non-linear trends of CD4 levels, transmission group, migration status and age over time. Usually 3 will be enough.
Impute reporting delay inputs	Imputes reporting delay in case quarter of diagnosis, notification year or quarter of notification are missing	Should be applied in case of substantial proportion of missing values in reporting delay variables.

Further actions

Test run the selected adjustments with smaller values for 'Number of imputations', 'Number of burn-in iterations' and 'Number of iterations between 2 successive imputations'. Inspect the diagnostics section. Rerun with improved parameters so that the diagnostics are satisfactory.

6.2 Multiple imputation by chained equations (MICE)

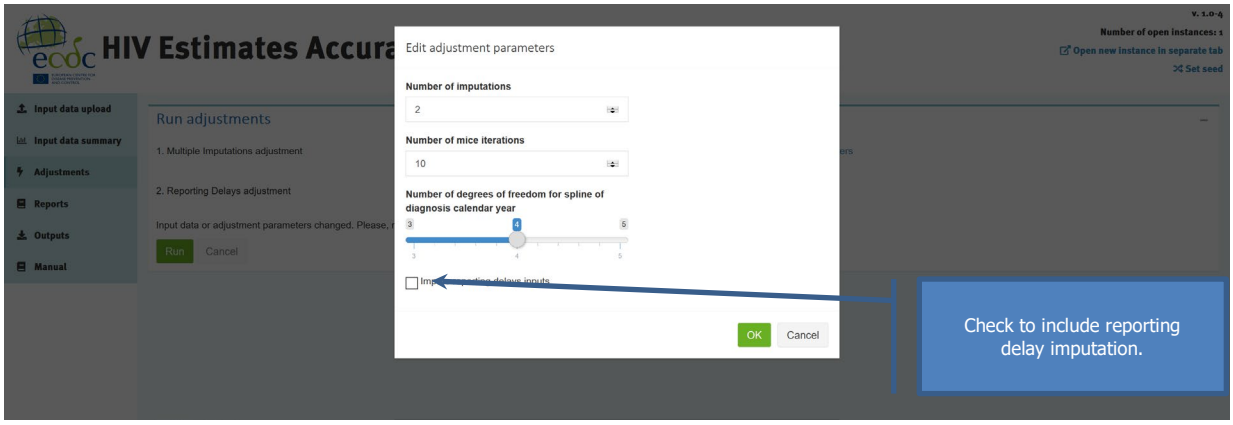
Description

This option performs multiple imputation by chained equations (MICE). This is an iterative procedure that can be time-consuming. The amount of time needed depends on the parameters set. It is recommended to start with lower numbers to look at the outputs and allow up to several hours for the final runs.

Process

Select 'Multiple Imputation using Chained Equations' from the dropdown menu in the 'Multiple Imputations adjustment' field. It is possible to edit parameters on the right.

A pop-up window is displayed with preset values of the parameters that can be edited. The interpretation and selection of proper values is provided below. If imputation of reporting delay is intended, it should be specified by checking the box at the bottom.



Interpretation

Parameters relating to imputation procedure and model are displayed below. Refer to the diagnostics section in order to select the proper values.

Parameter	Description	What to select
Number of imputations	Number of imputed data sets that will be produced	For test runs, select 2. For final adjustments, at least 5–10 imputations.
Number of MICE iterations	Number of iterations after which method should converge	For test runs, select 10. Generally higher numbers (usually 50 will be enough) are needed and this can be decided based on the adjustment diagnostics.
Number of degrees of freedom for splines of diagnosis calendar year	Parameter used to determine the degree of flexibility of the time trend in data (number of cases per year or median CD4 count per year)	Select between 3 and 5. Choose one that results in a best fitting model. Choose higher numbers if you expect fast changing trends and highly nonlinear trends of CD4 levels, transmission group, migration status and age over time. Usually 3 will be enough.
Impute reporting delay inputs	Imputes reporting delay in case either quarter of diagnosis, notification year or quarter of notification are missing	Should be applied in case of substantial proportion of missing values in reporting delay variables

Further actions

Test run the selected adjustments with smaller values for the 'Number of imputations' and 'Number of MICE iterations'. Inspect the diagnostics section. Rerun with improved parameters so that the diagnostics are satisfactory.

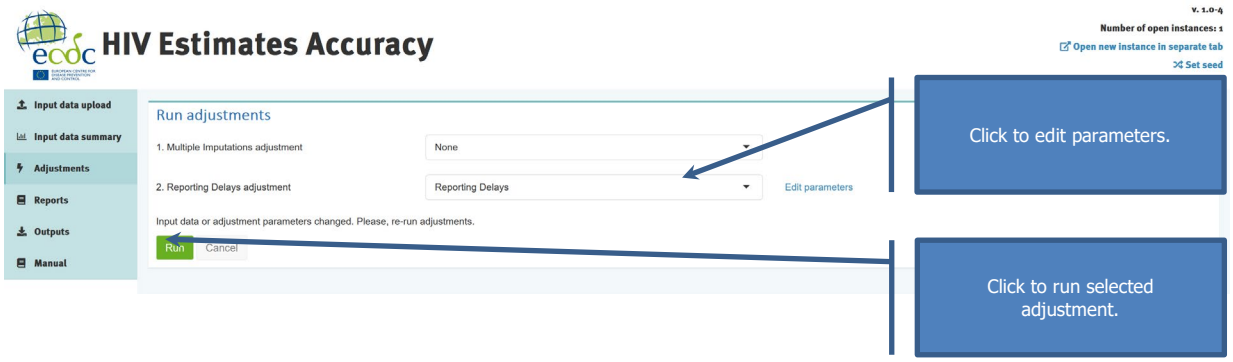
6.3 Simple reporting delay

Description

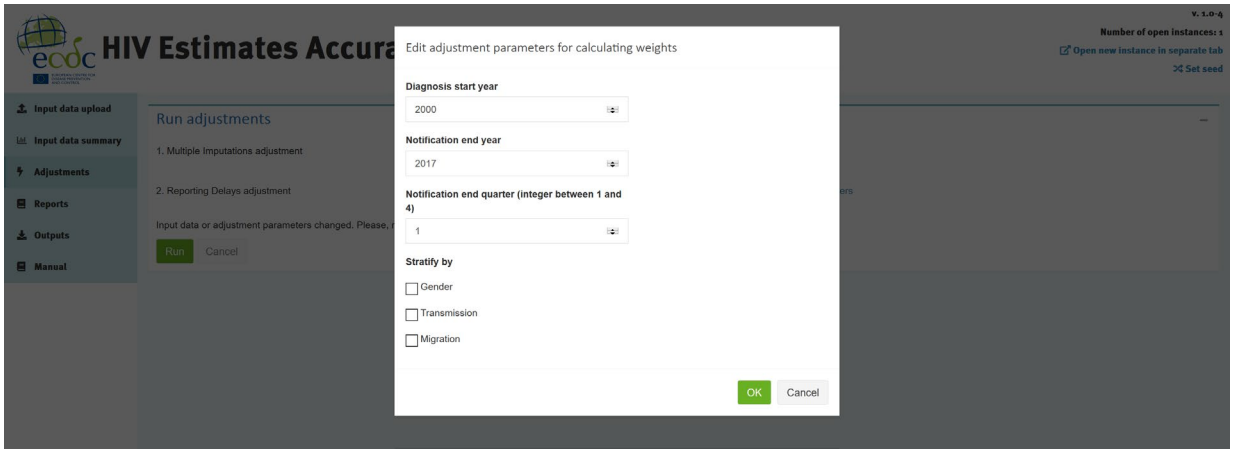
This option performs estimation of reporting delay distribution without regression modelling. An overall or stratum-specific distribution is estimated depending on the parameters selected.

Process

Select 'Reporting delay' from the drop-down menu in the 'Reporting delay adjustment' field. The option to edit parameters will appear on the right.



A pop-up window is displayed with preset values of the parameters that can be edited. The interpretation and selection of proper values is provided below.



Interpretation

The parameters relating to the reporting delay estimation are displayed. Filtering by diagnosis and notification year and quarter as part of the reporting delay parameter only affects estimation of the reporting delay weights. The output data will not be filtered as the estimated reporting delay weight will also be applied to the data outside of the filtered period specified as part of the reporting delay parameters.

Parameter	Description	What to select
Diagnosis start year	Only diagnoses made during this year or later will be included in the estimation.	If older data are unreliable or there was an important change in surveillance system, estimation could be performed on later data.
Notification end year and quarter	Only cases notified until this quarter will be included in the estimation.	This can be used to exclude the latest data if a cleaning event was performed at this time.
Stratification variables	For each of the cross sections of the values of the selected variables, a separate curve is created.	Important predictors of the reporting delay should be included. The method may be unstable if the stratification results in small numbers of cases in certain strata.

Further actions

Test run the selected adjustments. Inspect the diagnostics section. Rerun with improved parameters so that the diagnostics are satisfactory.

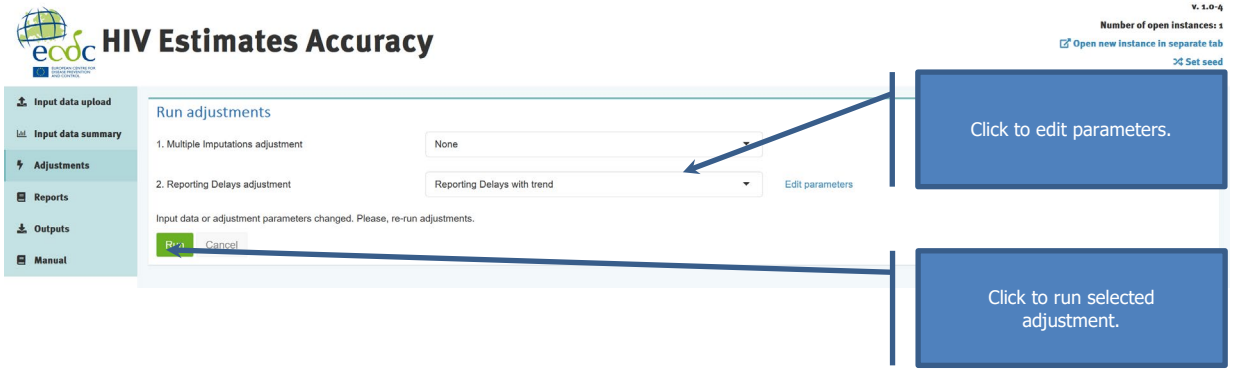
6.4 Reporting delay with trend

Description

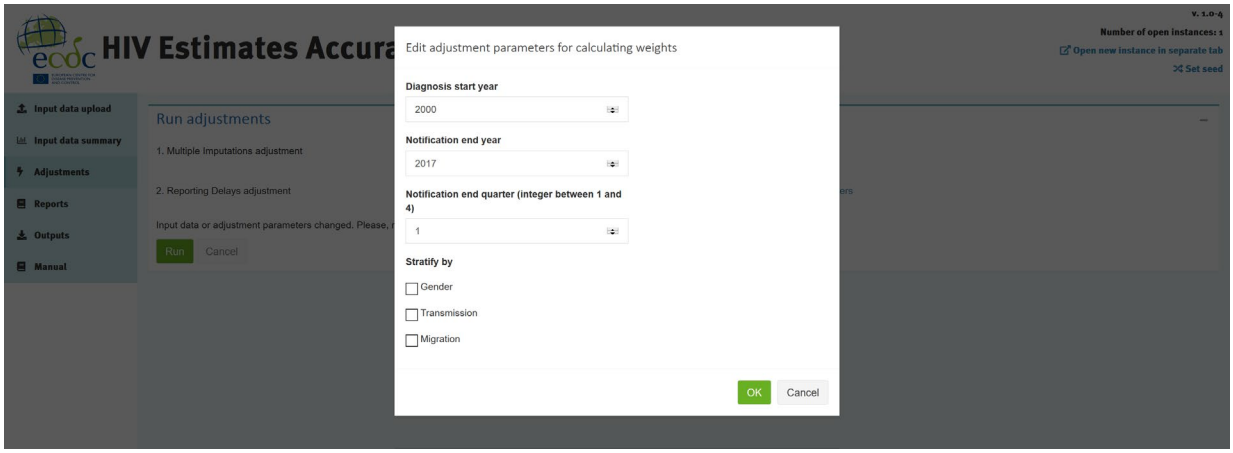
This option performs estimation of reporting delay distribution based on regression modelling of hazards in reverse time. Year of diagnosis is included by default. Additional covariates in the model are specified as stratification variables. An overall or stratum-specific distribution is estimated depending on the parameters selected.

Process

Select 'Reporting delay with trend' from the drop-down menu in the 'Reporting delay adjustment' field. The option to edit parameters will appear on the right.



A pop-up window is displayed with preset values of the parameters that can be edited. The interpretation and selection of proper values is provided below.



Interpretation

The parameters relating to reporting delay estimation are displayed. Filtering by diagnosis year and notification year and quarter as part of the reporting delay parameter only affects estimation of the reporting delay weights. Output data will not be filtered as the estimated reporting delay weight will also be applied to data outside the filtered period specified as part of the reporting delay parameters.

Parameter	Description	What to select
Diagnosis start year	Only diagnoses made during this year or later will be included in estimation.	If older data are unreliable or there was an important change in surveillance system, the estimation could be performed on the later data.
Notification end year and quarter	Only cases notified until this quarter will be included in estimation.	This can be used to exclude the latest data if a cleaning event was performed at this time.
Stratification variables	For each of the cross sections of the values of the selected variables, a separate curve is created.	Important predictors of the reporting delay should be included. The method may be unstable if the stratification results in small numbers of cases in certain strata.

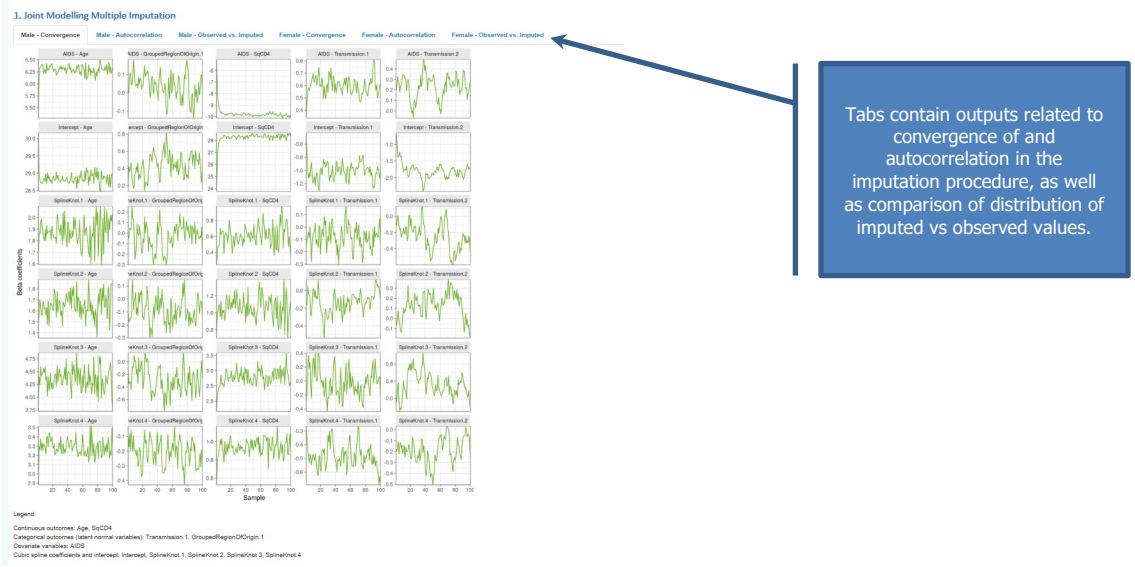
Further actions

Test run the selected adjustments. Inspect the diagnostics section. Rerun with improved parameters so that diagnostics are satisfactory.

6.5 Intermediate outputs of adjustments and diagnostics – joint modelling multiple imputations

Description

After running the joint modelling adjustment, the following section will present the intermediate outputs that can be used as diagnostics for the adjustments. The output is organised in tabs.



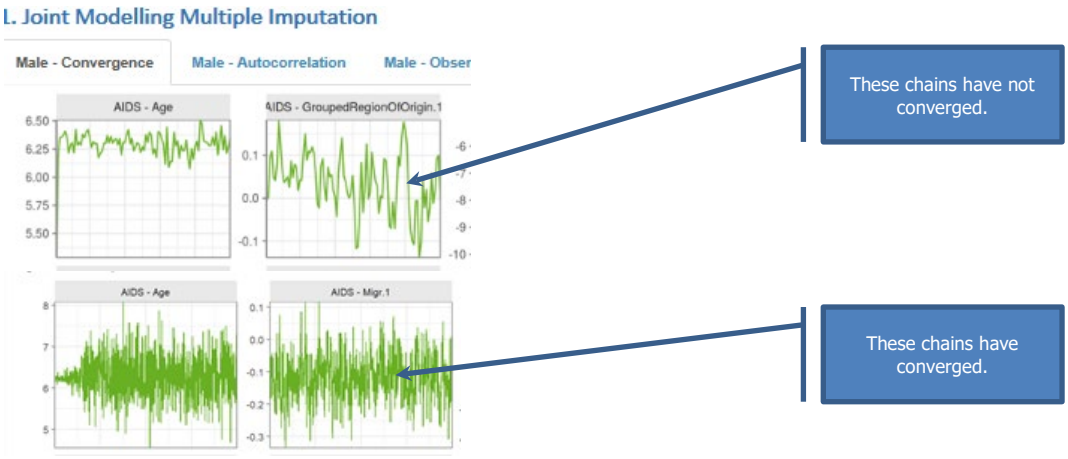
Process

Intermediate outputs are generated automatically when running adjustments.

Interpretation

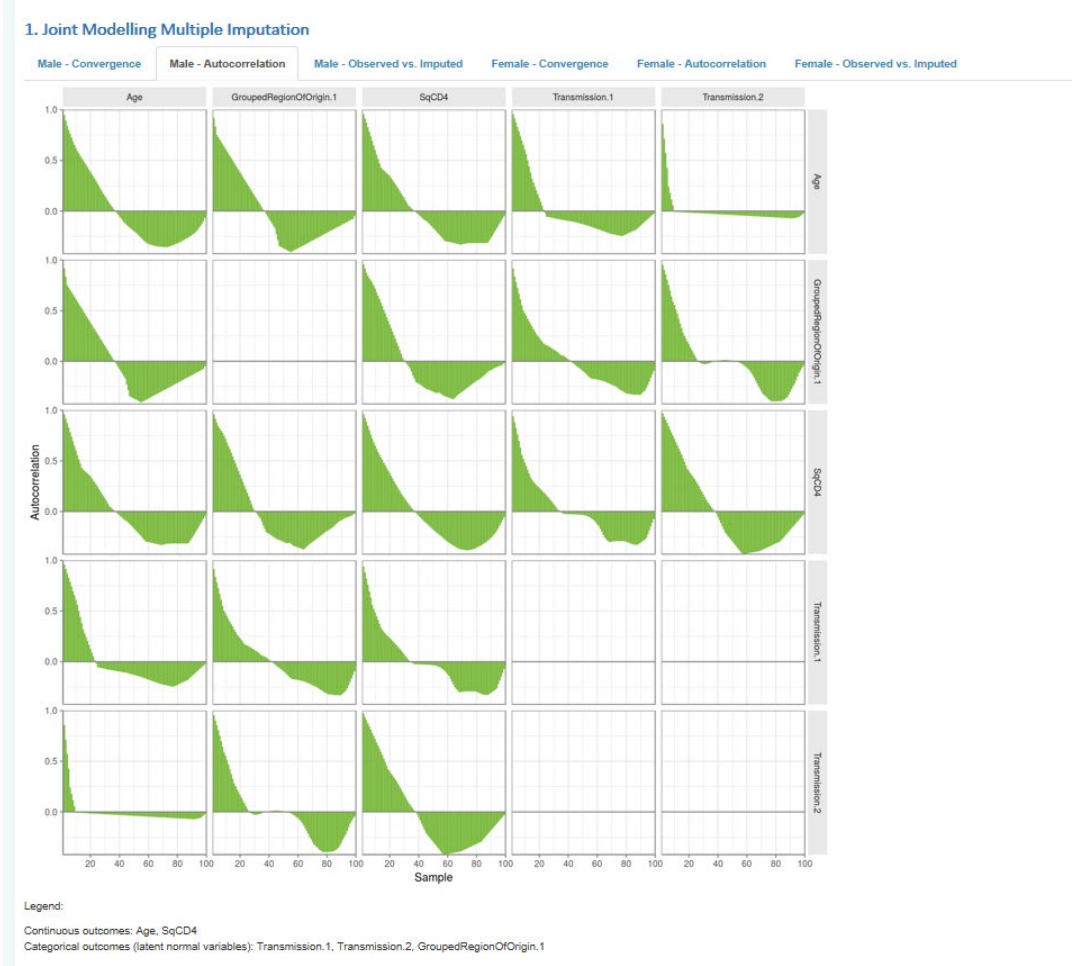
The output related to the convergence contains trace plots.

The use of the trace plots determines whether the procedure converged, assuring that the missing values are imputed from the correct distribution. In case of convergence, the trace plot for every parameter does not display any pattern. More iterations are needed in case certain parameters display certain trends that do not level off at the right of the graph. In case more iterations are needed, this can be controlled with 'Number of burn-in iterations'.

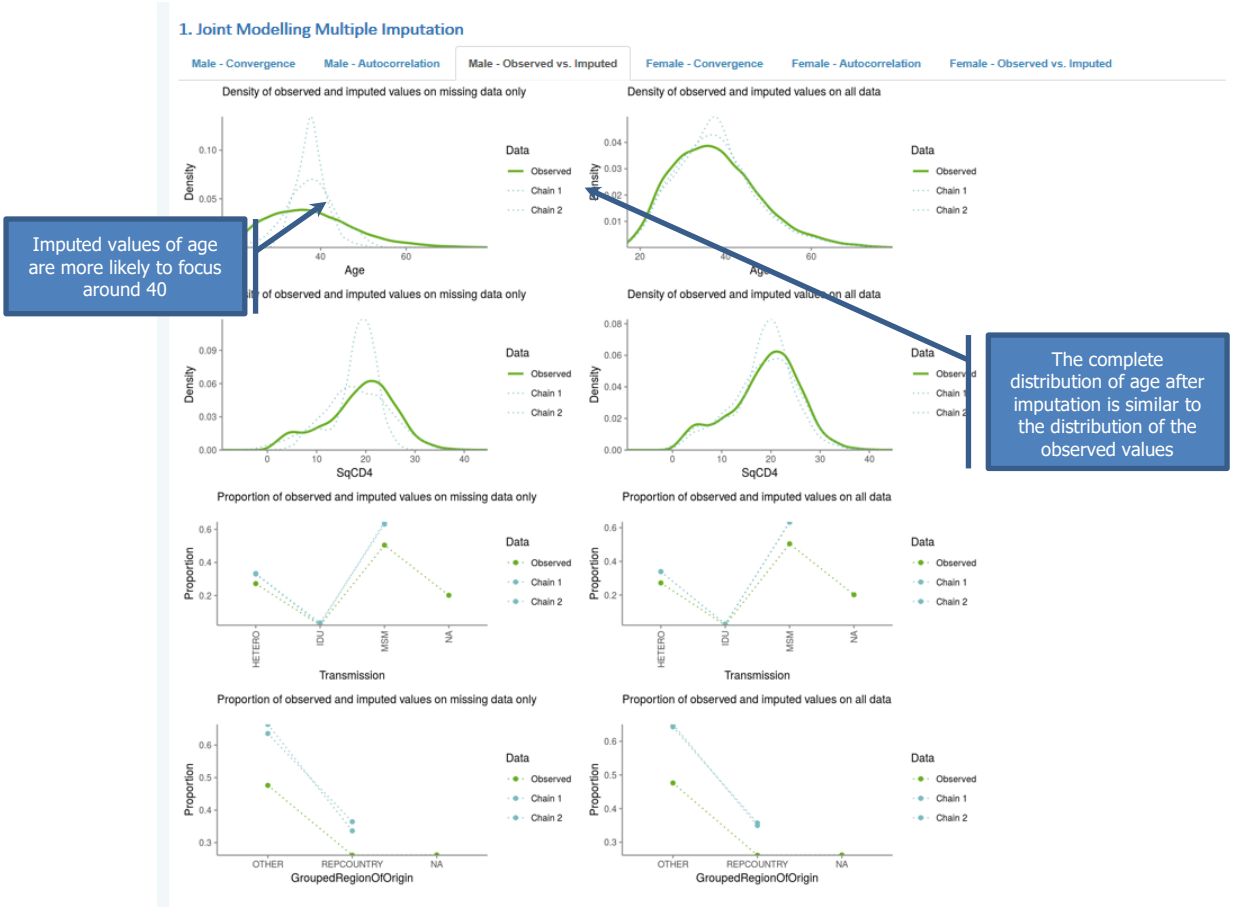


The autocorrelation plot informs about the number of iterations that should be performed between the subsequent imputations in order to ensure independence of these imputations. The aim should be that the autocorrelation should be insignificant.

The following plot suggests a number of iterations between the imputations of more than 100 but graphs should be judged only if convergence is suggested by the previous type of graphs.



Finally, the 'Observed vs imputed' tab presents how the distribution of the imputed variables changes after imputation for each imputed chain. Two types of graphs are available: one comparing the distribution of observed and imputed values, the other comparing the distribution of observed and all values observed and imputed.



Generally, if the proportion of missing values is less, even if the distribution of the imputed values is very different, the final complete distribution is not impacted much by the imputed values. Conversely, with a large proportion of missing values, the distribution of imputed values becomes important and a faulty model may lead to bias. The imputed distribution is expected to be somewhat different than that observed distribution. However, the main trends are normally preserved. In any case, graphs should be judged only if convergence is suggested by the previous type of graphs.

Further actions

In case of lack of convergence, increase the number of iterations under 'Number of burn-in iterations'.

Increase the number of 'Iterations between subsequent imputations' if needed.

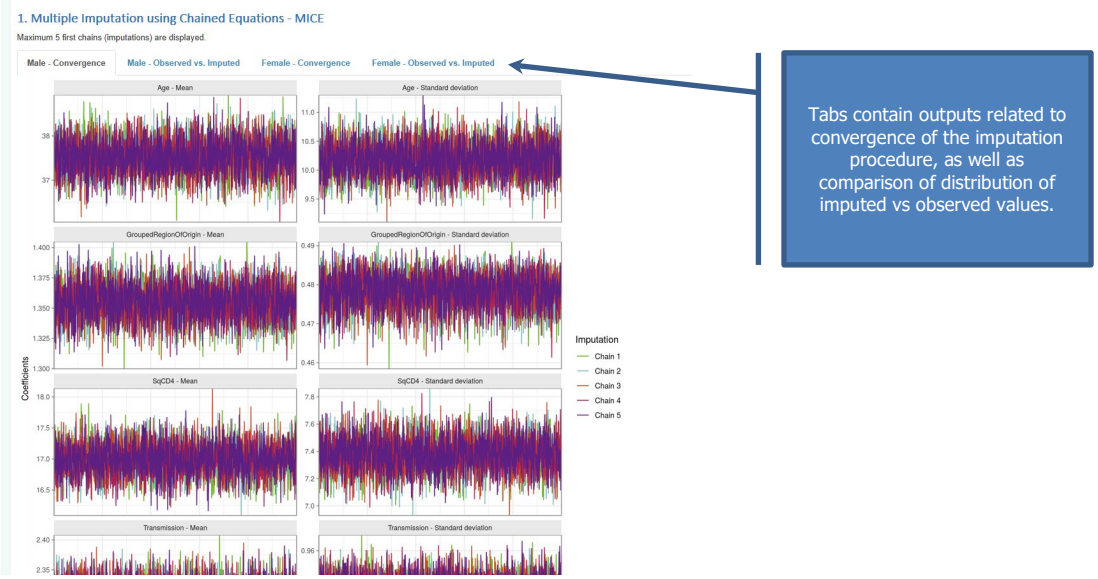
Rerun the analysis.

In case the distributions of the imputed values are very different from the observed values, rerun the analysis with MICE.

6.6. Intermediate outputs of adjustments and diagnostics – multiple imputation by chained equations (MICE)

Description

After running the joint modelling adjustment, the following section will present the intermediate outputs that can be used as diagnostics for the adjustments. The output is organised in tabs.



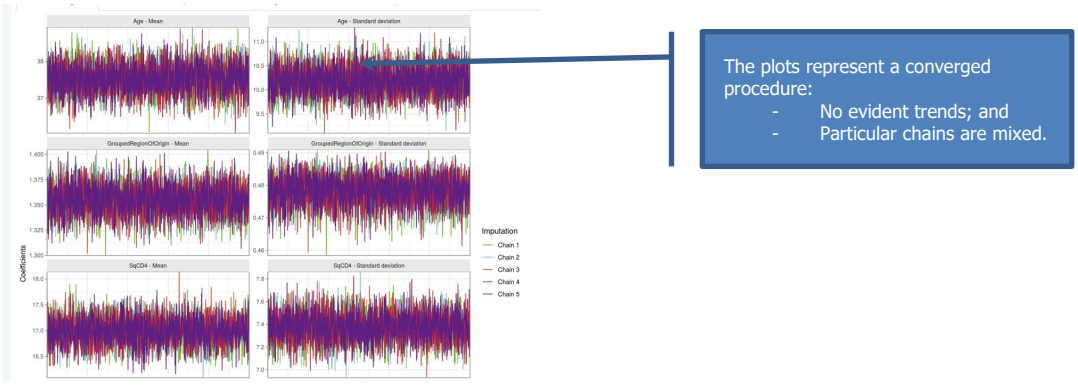
Process

Intermediate outputs are generated automatically when running the adjustments.

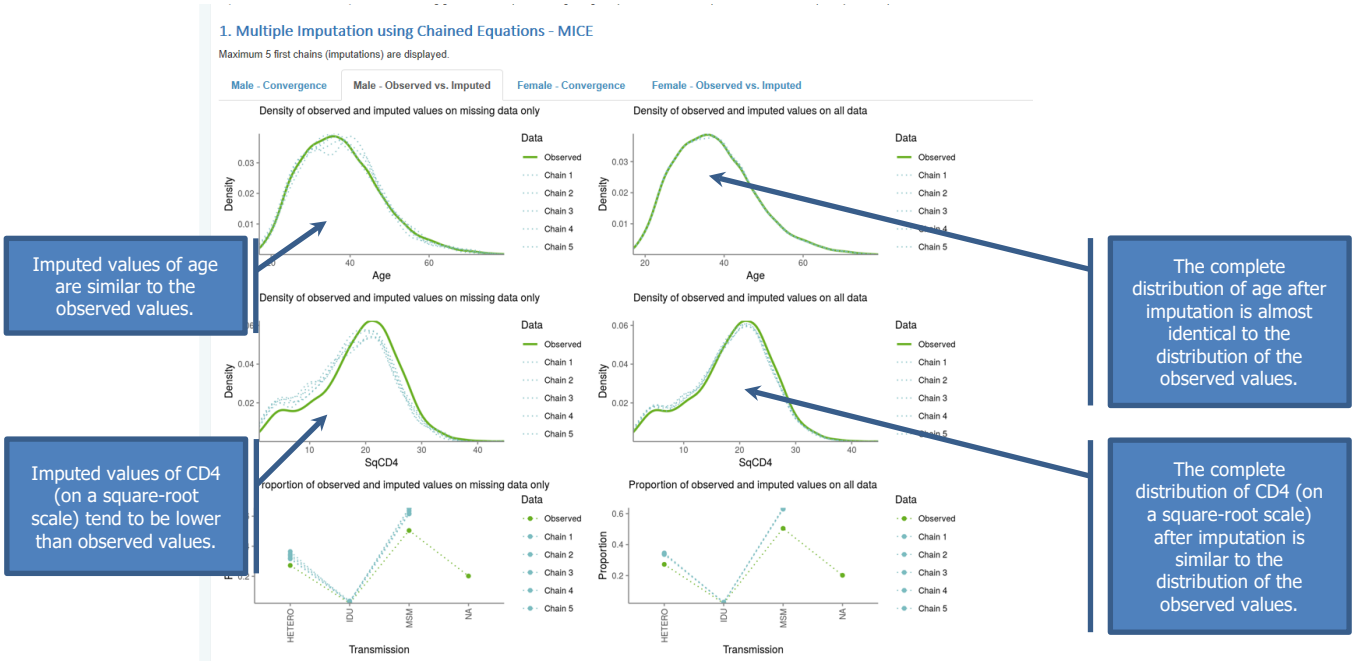
Interpretation

The output related to the convergence contains trace plots.

The use of the trace plots determines whether the procedure converged, assuring that the missing values are imputed from the correct distribution. In case of convergence, the trace plot for every parameter does not display any pattern. More iterations are needed in case certain parameters display ertain trends that do not level off at the right of the graph. In case more iterations are needed, this can be controlled under 'Number of mice iterations'. The picture below represents a converged procedure.



Finally, the 'Observed vs imputed' tab presents how the distribution of the imputed variables changes after imputation for each imputed chain. Two types of graphs are available: one comparing the distribution of observed and imputed values and the other comparing the distribution of observed and all values observed and imputed.



Generally, if the proportion of missing values is less, even if the distribution of the imputed values is very different, the final complete distribution is not impacted much by the imputed values. Conversely, with a large proportion of missing values, the distribution of imputed values becomes important and a faulty model may lead to bias. The imputed distribution is expected to be somewhat different than that observed distribution. However, the main trends are normally preserved. In any case, graphs should be judged only if convergence is suggested by the previous type of graphs.

Further actions

In case of lack of convergence, increase the number of iteration under 'Number of mice iterations'. Rerun the analysis.

In case the distributions of the imputed values are very different from the observed values, rerun the analysis with joint modelling.

6.7. Intermediate outputs of adjustments and diagnostics – reporting delay

Description

In case the reporting delay adjustment was selected, the intermediate output contains a visual representation of the reporting delay adjustment and results in univariable analysis of the selected predictors of the reporting delay adjustments.

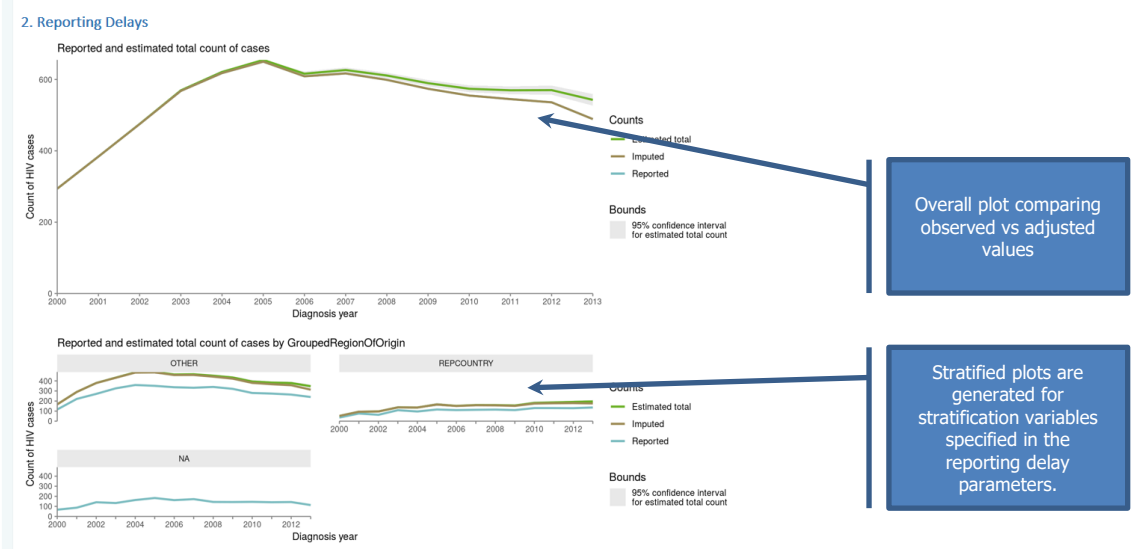
Process

Intermediate outputs are generated automatically when running the adjustments.

Interpretation

The plots show the observed and adjusted values. The overall plot is always generated to show how the overall count changes after adjusting for reporting delay. This allows for visual inspection if the adjusted trend looks plausible.

If stratification was introduced, trends by stratification variables are also displayed. In case the reporting delay adjustment was run together with the imputations, the graphs will display the observed trend, trends after imputations and trend after both imputations and the reporting delay adjustment. The imputation curve will be different from the observed curve only in case of plots stratified by a variable that has been imputed.



At the bottom of the page a table is displayed with univariable analysis of predictors of the reporting delay. As the reporting delay is modelled on the reverse time scale the interpretation of regression parameters (hazard ratio, HR) is not meaningful. Most importantly the p-value should be checked. Non important predictors could be excluded from the stratification variables.

The proportionality assumption test is provided for information only. For many countries' data this assumption was not met and the model used includes already stratification to deal with this problem.

Further actions

If the outputs are satisfactory proceed to further tabs. Otherwise, change parameters and rerun the analysis.

7 Reports

7.1 Creating report

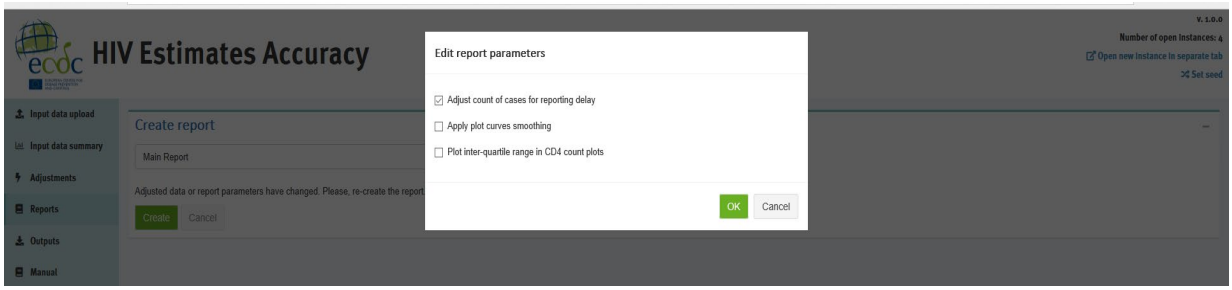
Description

In this tab, a predefined report with main findings is provided. In version 1 of the tool, only 'Main Report' is available. However, certain parameters may be set for the report.

Process

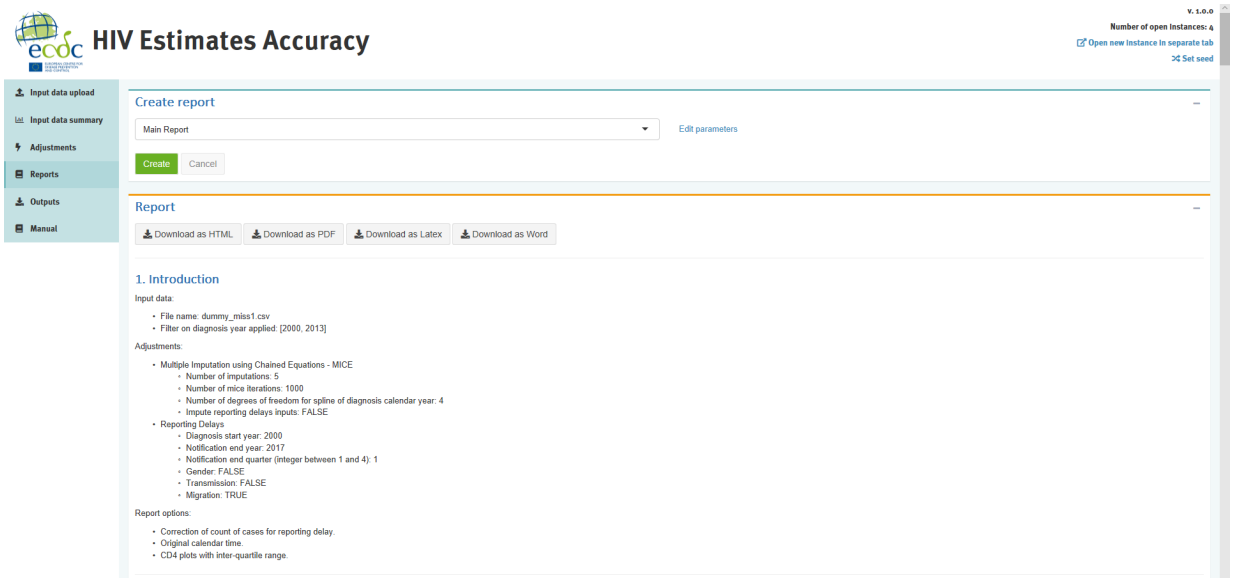
In order to control the output, three parameters should be selected:

- Adjust case counts for reporting delay. This option is selected by default if the reporting delay adjustment is applied. It can be unchecked to produce a report on imputed data excluding the reporting delay correction.
- Apply plot curves smoothing. This option refers to the way imputations are dealt with when producing plots. If no smoothing is selected, – treating each year separately and not taking into account any potential trends over time, the report will contain simple counts for the number of cases and means for CD4 counts. If smoothing is applied, both the counts and the CD4 counts are estimated from a regression model with year as continuous predictor. While this is more methodologically appropriate, the counts generated may be different than the observed ones.
- Plot inter-quantile range in CD4 count plot. This option affects the graphs presenting trends in CD4 counts. Inter-quartile ranges are presented if this option is selected.

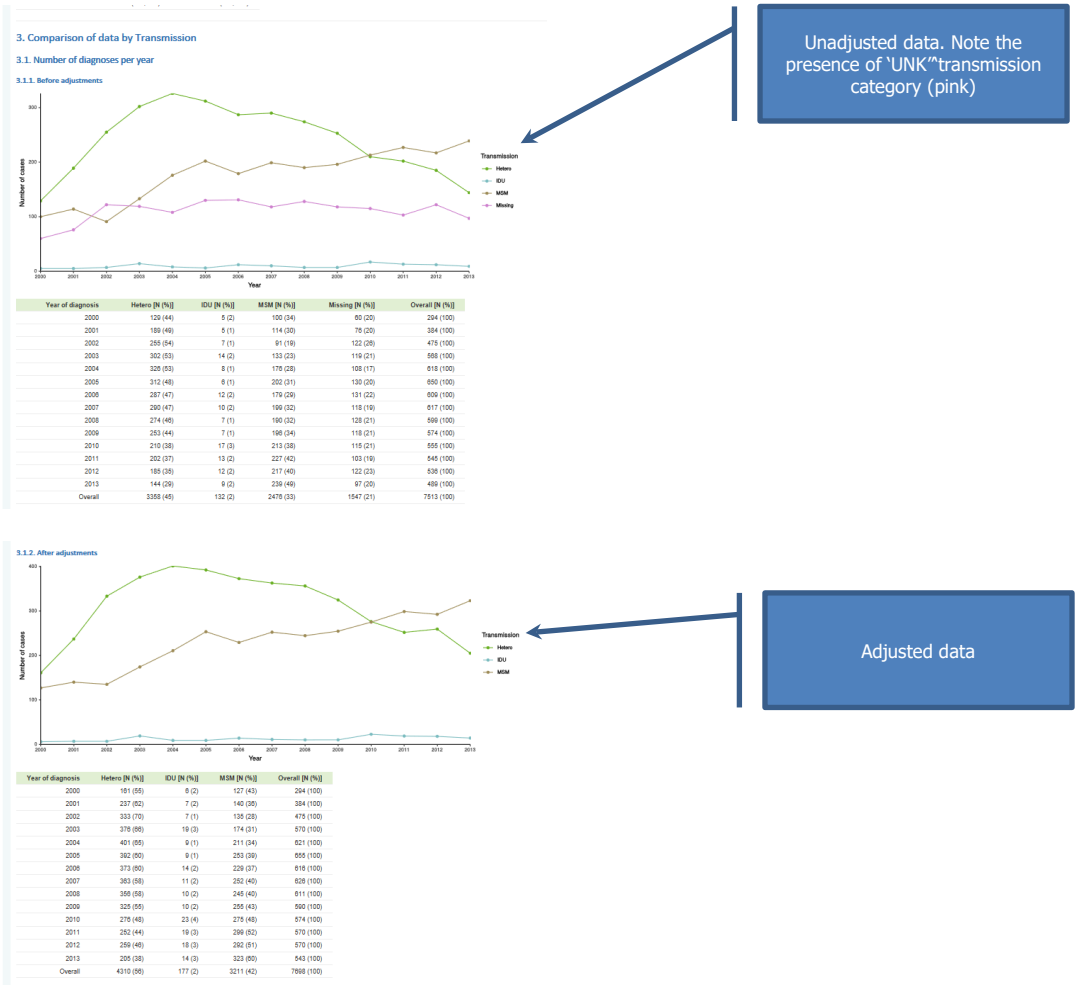


Interpretation

In the first section of the report, all selected options are summarised for convenience.



The following section contain comparisons of trends by covariates for unadjusted and adjusted data.



The last section provides an additional comparison of the overall counts observed and adjusted for reporting delay. The 'Weight not estimated' column provides information on the number of cases where it was not possible to estimate the reporting delay weight. The estimated number of yet unreported cases is also provided.

5. Comparison of the reported and estimated number of diagnoses per year

Diagnosis year	Reported	Weight estimated	Weight not estimated	Estimated unreported [N (95% CI)]	Estimated total [N (95% CI)]
2000	294	219	75	0 (0, 0)	294 (294, 294)
2001	384	384	0	0 (0, 0)	384 (384, 384)
2002	475	475	0	0 (-1, 2)	475 (474, 477)
2003	568	568	0	2 (-2, 5)	570 (566, 573)
2004	618	618	0	3 (-1, 7)	621 (617, 625)
2005	650	650	0	5 (-0, 10)	655 (650, 660)
2006	609	609	0	7 (1, 13)	616 (610, 622)
2007	617	617	0	10 (2, 17)	627 (619, 634)
2008	599	599	0	12 (4, 20)	611 (603, 619)
2009	574	574	0	16 (7, 25)	590 (581, 599)
2010	555	555	0	19 (9, 29)	574 (564, 584)
2011	545	545	0	25 (14, 36)	570 (559, 581)
2012	536	536	0	34 (21, 47)	570 (557, 583)
2013	489	489	0	54 (38, 70)	543 (527, 559)
Total	7513	7438	75	187 (93, 281)	7700 (7606, 7794)

Further actions

The report may be exported.

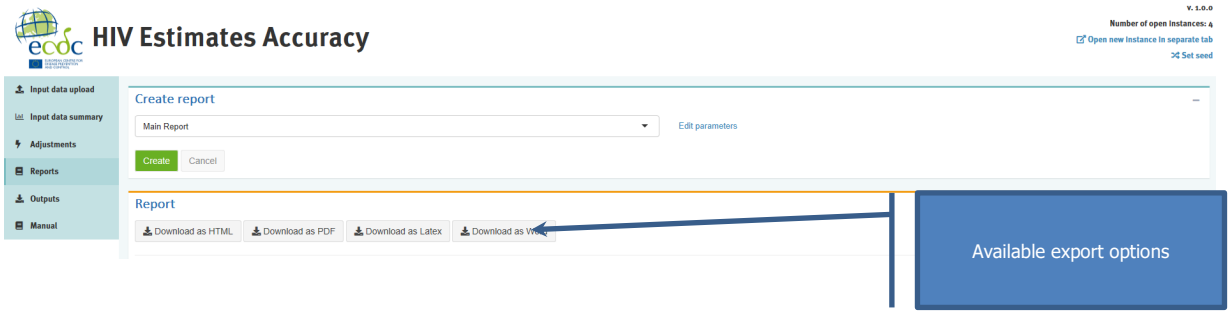
7.2 Exporting report

Description

The report may be exported to different formats: HTML, PDF, Word or LaTeX. If using the offline version, the user needs to have LaTeX installed in order to generate the PDF version of the report.

Process

Select the desired format from the buttons available at the top of the report.



8 Outputs

8.1 Adjusted data set

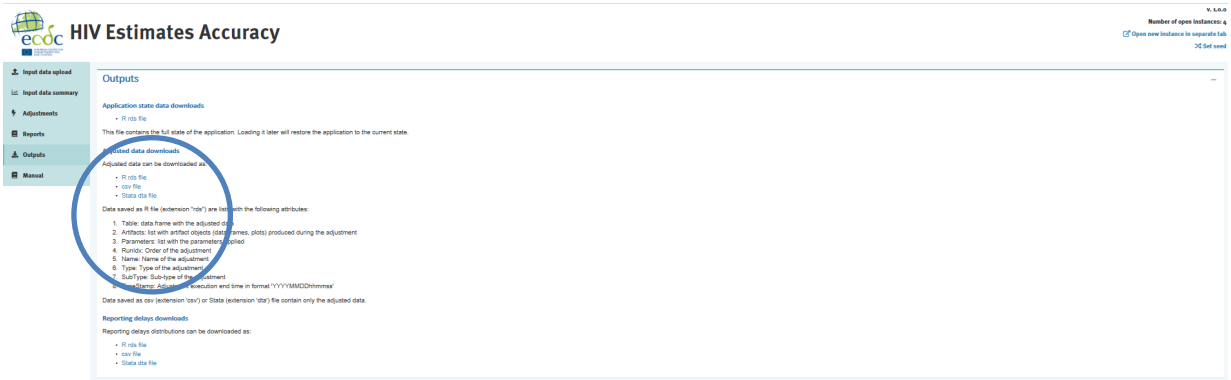
Description

The full data set with adjustments may be exported. If both imputation and reporting delay adjustments are run, the output data will be a multiply imputed data set with reporting delay weight. This data set contains the original data as uploaded to the tool, variables created during the preprocessing procedure, variable imputation and variable weight representing weight due to reporting delay. The data set contains original data (imputation=0) and subsequent copies of the data set with missing values imputed (pseudo-complete data sets, imputation=1,2).

It can be exported in multiple formats (R, CSV and Stata). Apart from the data, the R file contains additional information about the adjustment performed, as well as certain outputs such as graphs.

Process

Select the desired format from the 'Adjusted data downloads' section.



8.2 Reporting delay weights

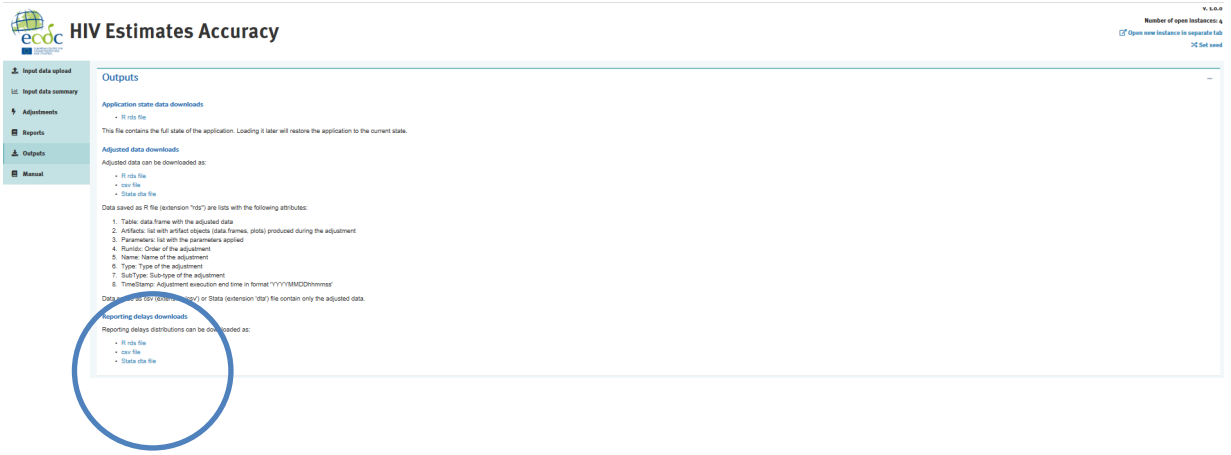
Description

The data set contains the reporting delay distribution (the probability of reporting within a certain number of quarters after the diagnosis) and the confidence intervals. If the stratification was included separate distribution for each stratification variable pattern are provided. This distribution may be used to adjust data for reporting delay outside of the tool.

It can be exported in multiple formats (R, csv and Stata file). The R file apart from the data contains additional information about the adjustment performed as well as some outputs such as graphs.

Process

Select the desired format from the 'Adjusted data downloads' section.



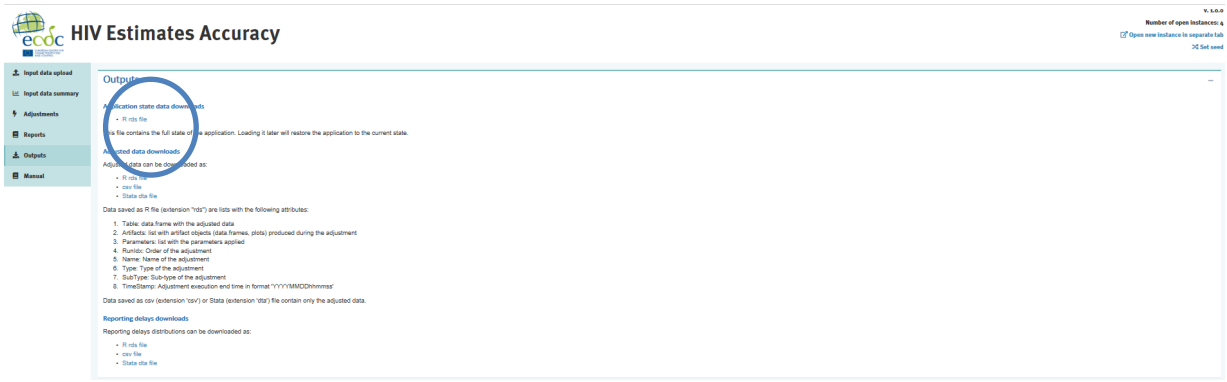
8.3 Application state data

Description

This .rds file contains the current status of the work, including preprocessed data and adjusted date if needed. This file can be uploaded the next time the tool is used or a new instance of the tool is launched in order to continue or modify adjustments. The default name starts with 'HIV_state_' followed by the date it was saved, but the file can be saved with a name specified by the user.

Process

Select the R .rds file in the 'Application state data downloads' section.



References

Rosinska M, Pantazis N, Janiec J, Pharris A, Amato-Gauci AJ, Quinten C, ECDC HIV/AIDS Surveillance Network. Potential adjustment methodology for missing data and reporting delay in the HIV Surveillance System, European Union/European Economic Area, 2015. Euro Surveill. 2018 Jun;23(23). doi: 10.2807/1560-7917.ES.2018.23.23.1700359.

Missing values

Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken: Wiley; 2002. p. 381.

Carpenter JR, Kenward MG. Missing data in randomised controlled trials: a practical guide. Birmingham: Health Technology Assessment Methodology Programme, p. 199. Available from: http://researchonline.lshtm.ac.uk/id/eprint/4018500.

Schafer JL. Analysis of incomplete multivariate data. 1st ed. Boca Raton: Chapman & Hall/CRC; 1997. p. 430.

Quartagno M, Carpenter JR. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. Stat Med. 2016 Jul 30;35(17):2938-54.

Jolani S, Debray TPA, Koffijberg H, van Buuren S, Moons KGM. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. Stat Med. 2015 May 20;34(11):1841–63.

Quartagno M.,Carpenter J. jomo: Multilevel Joint Modelling Multiple Imputation [Internet, software]. Vienna: R Foundation for Statistical Computing; 2018. Available from: http://cran.r-project.org/package=jomo.

van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011 Dec;45(3):1-67. Available from: http://www.jstatsoft.org/v45/i03.

Lumley T. mitools: Tools for multiple imputation of missing data [Internet, software]. Vienna: R Foundation for Statistical Computing; 2014. Available from: http://cran.r-project.org/package=mitools.

Reporting delay

Lawless JF. Adjustments for Reporting Delays and the Prediction of Occurred but Not Reported Events. Can J Stat. 1994 Mar;22(1):15.

Brookmeyer R, Liao JG. The analysis of delays in disease reporting: methods and results for the acquired immunodeficiency syndrome. Am J Epidemiol. 1990 Aug;132(2):355–65.

Lagakos SW, Barraj LM, Gruttola VD. Nonparametric analysis of truncated survival data, with application to AIDS. Biometrika. 1988 Sep 1;75(3):515-523.

Kalbfleisch JD, Lawess JF. Regression models for right truncated data with applications to AIDS incubation times and reporting lags. Stat Sin. 1991 Jan;1:19–32.

Pagano M, Tu XM, Gruttola VD, MaWhinney S. Regression Analysis of Censored and Truncated Data: Estimating Reporting- Delay Distributions and AIDS Incidence from Surveillance Data. Biometrics. 1994 Dec;50(4):1203.

Annex 1. Codes used for countries and regions

	Name	Code	FormalName	RegionOrigin
1	Taiwan	TW	Republic of China	EASTASIAPAC
2	Afghanistan	AF	Islamic Republic of Afghanistan	SOUTHASIA
3	Albania	AL	Republic of Albania	CENTEUR
4	Algeria	DZ	People's Democratic Republic of Algeria	NORTHAFRMIDEAST
5	American Samoa	AS	NA	EASTASIAPAC
6	Andorra	AD	Principality of Andorra	WESTEUR
7	Angola	AO	Republic of Angola	SUBAFR
8	Anguilla	AI	NA	CAR
9	Antarctica	AQ	NA	UNK
10	Antigua and Barbuda	AG	Antigua and Barbuda	CAR
11	Argentina	AR	Argentine Republic	LATAM
12	Armenia	AM	Republic of Armenia	EASTEUR
13	Aruba	AW	NA	CAR
14	Australia	AU	Australia	AUSTNZ
15	Austria	AT	Republic of Austria	WESTEUR
16	Azerbaijan	AZ	Republic of Azerbaijan	EASTEUR
17	Bahamas	BS	Commonwealth of the Bahamas	CAR
18	Bahrain	BH	Kingdom of Bahrain	NORTHAFRMIDEAST
19	Bangladesh	BD	People's Republic of Bangladesh	SOUTHASIA
20	Barbados	BB	Barbados	CAR
21	Belarus	BY	Republic of Belarus	EASTEUR
22	Belgium	BE	Kingdom of Belgium	WESTEUR
23	Belize	BZ	Belize	LATAM
24	Benin	BJ	Republic of Benin	SUBAFR
25	Bermuda	BM	NA	CAR
26	Bhutan	BT	Kingdom of Bhutan	SOUTHASIA
27	Bolivia (Plurinational State of)	BO	Plurinational State of Bolivia	LATAM
28	Bonaire, Sint Eustatius and Saba	BQ	NA	CAR
29	Bosnia and Herzegovina	BA	Bosnia and Herzegovina	CENTEUR
30	Botswana	BW	Republic of Botswana	SUBAFR
31	Bouvet Island	BV	NA	CAR
32	Brazil	BR	Federative Republic of Brazil	LATAM
33	British Indian Ocean Territory	IO	NA	SUBAFR
34	British Virgin Islands	VG	NA	CAR
35	Brunei Darussalam	BN	Brunei Darussalam	SOUTHASIA
36	Bulgaria	BG	Republic of Bulgaria	CENTEUR
37	Burkina Faso	BF	Burkina Faso	SUBAFR
38	Burundi	BI	Republic of Burundi	SUBAFR
39	Cabo Verde	CV	Republic of Cabo Verde	SUBAFR
40	Cambodia	KH	Kingdom of Cambodia	SOUTHASIA
41	Cameroon	CM	Republic of Cameroon	SUBAFR
42	Canada	CA	Canada	NORTHAM
43	Cayman Islands	KY	NA	CAR
44	Central African Republic	CF	Central African Republic	SUBAFR
45	Chad	TD	Republic of Chad	SUBAFR
46	Chile	CL	Republic of Chile	LATAM
47	China	CN	People's Republic of China	EASTASIAPAC
48	China, Hong Kong Special Administrative Region	HK	NA	EASTASIAPAC
49	China, Macao Special Administrative Region	MO	NA	EASTASIAPAC
50	Christmas Island	CX	NA	AUSTNZ
51	Cocos (Keeling) Islands	CC	NA	AUSTNZ
52	Colombia	CO	Republic of Colombia	LATAM
53	Comoros	KM	Union of the Comoros	SUBAFR
54	Congo	CG	Republic of the Congo	SUBAFR
55	Cook Islands	CK	Cook Islands	EASTASIAPAC
56	Costa Rica	CR	Republic of Costa Rica	LATAM
57	Croatia	HR	Republic of Croatia	CENTEUR
58	Cuba	CU	Republic of Cuba	CAR
59	Curaçao	CW	NA	CAR
60	Cyprus	CY	Republic of Cyprus	CENTEUR
61	Czechia	CZ	Czech Republic	CENTEUR
62	Côte d'Ivoire	CI	Republic of Côte d'Ivoire	SUBAFR

	Name	Code	FormalName	RegionOrigin
63	Democratic People's Republic of Korea	KP	Democratic People's Republic of Korea	EASTASIAPAC
64	Democratic Republic of the Congo	CD	Democratic Republic of the Congo	SUBAFR
65	Denmark	DK	Kingdom of Denmark	WESTEUR
66	Djibouti	DJ	Republic of Djibouti	SUBAFR
67	Dominica	DM	Commonwealth of Dominica	CAR
68	Dominican Republic	DO	Dominican Republic	CAR
69	Ecuador	EC	Republic of Ecuador	LATAM
70	Egypt	EG	Arab Republic of Egypt	NORTHAFRMIDEAST
71	El Salvador	SV	Republic of El Salvador	LATAM
72	Equatorial Guinea	GQ	Republic of Equatorial Guinea	SUBAFR
73	Eritrea	ER	State of Eritrea	SUBAFR
74	Estonia	EE	Republic of Estonia	EASTEUR
75	Ethiopia	ET	Federal Democratic Republic of Ethiopia	SUBAFR
76	Falkland Islands (Malvinas)	FK	NA	LATAM
77	Faroe Islands	FO	NA	WESTEUR
78	Fiji	FJ	Republic of Fiji	EASTASIAPAC
79	Finland	FI	Republic of Finland	WESTEUR
80	France	FR	French Republic	WESTEUR
81	French Guiana	GF	NA	LATAM
82	French Polynesia	PF	NA	EASTASIAPAC
83	French Southern Territories	TF	NA	SUBAFR
84	Gabon	GA	Gabonese Republic	SUBAFR
85	Gambia	GM	Republic of the Gambia	SUBAFR
86	Georgia	GE	Georgia	EASTEUR
87	Germany	DE	Federal Republic of Germany	WESTEUR
88	Ghana	GH	Republic of Ghana	SUBAFR
89	Gibraltar	GI	NA	WESTEUR
90	Greece	EL	Hellenic Republic	WESTEUR
91	Greenland	GL	NA	WESTEUR
92	Grenada	GD	Grenada	CAR
93	Guadeloupe	GP	NA	CAR
94	Guam	GU	NA	SOUTHASIA
95	Guatemala	GT	Republic of Guatemala	LATAM
96	Guernsey	GG	NA	WESTEUR
97	Guinea	GN	Republic of Guinea	SUBAFR
98	Guinea-Bissau	GW	the Republic of Guinea-Bissau	SUBAFR
99	Guyana	GY	Republic of Guyana	LATAM
100	Haiti	HT	Republic of Haiti	CAR
101	Heard Island and McDonald Islands	HM	NA	AUSTNZ
102	Holy See	VA	Holy See	WESTEUR
103	Honduras	HN	Republic of Honduras	LATAM
104	Hungary	HU	Hungary	CENTEUR
105	Iceland	IS	Republic of Iceland	WESTEUR
106	India	IN	Republic of India	SOUTHASIA
107	Indonesia	ID	Republic of Indonesia	SOUTHASIA
108	Iran (Islamic Republic of)	IR	Islamic Republic of Iran	SOUTHASIA
109	Iraq	IQ	Republic of Iraq	NORTHAFRMIDEAST
110	Ireland	IE	Ireland	WESTEUR
111	Isle of Man	IM	NA	WESTEUR
112	Israel	IL	State of Israel	WESTEUR
113	Italy	IT	Republic of Italy	WESTEUR
114	Jamaica	JM	Jamaica	CAR
115	Japan	JP	Japan	EASTASIAPAC
116	Jersey	JE	NA	WESTEUR
117	Jordan	JO	Hashemite Kingdom of Jordan	NORTHAFRMIDEAST
118	Kazakhstan	KZ	Republic of Kazakhstan	EASTEUR
119	Kenya	KE	Republic of Kenya	SUBAFR
120	Kiribati	KI	Republic of Kiribati	SOUTHASIA
121	Kuwait	KW	State of Kuwait	NORTHAFRMIDEAST
122	Kyrgyzstan	KG	Kyrgyz Republic	EASTEUR
123	Lao People's Democratic Republic	LA	Lao People's Democratic Republic	SOUTHASIA
124	Latvia	LV	Republic of Latvia	EASTEUR
125	Lebanon	LB	Lebanese Republic	NORTHAFRMIDEAST
126	Lesotho	LS	Kingdom of Lesotho	SUBAFR
127	Liberia	LR	Republic of Liberia	SUBAFR
128	Libya	LY	Libya	NORTHAFRMIDEAST

	Name	Code	FormalName	RegionOrigin
129	Liechtenstein	LI	Principality of Liechtenstein	WESTEUR
130	Lithuania	LT	Republic of Lithuania	EASTEUR
131	Luxembourg	LU	Grand Duchy of Luxembourg	WESTEUR
132	Madagascar	MG	Republic of Madagascar	SUBAFR
133	Malawi	MW	Republic of Malawi	SUBAFR
134	Malaysia	MY	Malaysia	SOUTHASIA
135	Maldives	MV	Republic of Maldives	SOUTHASIA
136	Mali	ML	Republic of Mali	SUBAFR
137	Malta	MT	Republic of Malta	WESTEUR
138	Marshall Islands	MH	Republic of the Marshall Islands	SOUTHASIA
139	Martinique	MQ	NA	CAR
140	Mauritania	MR	Islamic Republic of Mauritania	SUBAFR
141	Mauritius	MU	Republic of Mauritius	SUBAFR
142	Mayotte	YT	NA	SUBAFR
143	Mexico	MX	United Mexican States	LATAM
144	Micronesia (Federated States of)	FM	Federated States of Micronesia	SOUTHASIA
145	Monaco	MC	Principality of Monaco	WESTEUR
146	Mongolia	MN	Mongolia	EASTASIAPAC
147	Montenegro	ME	Montenegro	CENTEUR
148	Montserrat	MS	NA	CAR
149	Morocco	MA	Kingdom of Morocco	NORTHAFRMIDEAST
150	Mozambique	MZ	Republic of Mozambique	SUBAFR
151	Myanmar	MM	Republic of the Union of Myanmar	SOUTHASIA
152	Namibia	NA	Republic of Namibia	SUBAFR
153	Nauru	NR	Republic of Nauru	SOUTHASIA
154	Nepal	NP	Federal Democratic Republic of Nepal	SOUTHASIA
155	Netherlands	NL	Kingdom of the Netherlands	WESTEUR
156	New Caledonia	NC	NA	SOUTHASIA
157	New Zealand	NZ	New Zealand	AUSTNZ
158	Nicaragua	NI	Republic of Nicaragua	LATAM
159	Niger	NE	Republic of the Niger	SUBAFR
160	Nigeria	NG	Federal Republic of Nigeria	SUBAFR
161	Niue	NU	Niue	EASTASIAPAC
162	Norfolk Island	NF	NA	AUSTNZ
163	Northern Mariana Islands	MP	NA	SOUTHASIA
164	Norway	NO	Kingdom of Norway	WESTEUR
165	Oman	OM	Sultanate of Oman	NORTHAFRMIDEAST
166	Pakistan	PK	Islamic Republic of Pakistan	SOUTHASIA
167	Palau	PW	Republic of Palau	SOUTHASIA
168	Panama	PA	Republic of Panama	LATAM
169	Papua New Guinea	PG	Independent State of Papua New Guinea	EASTASIAPAC
170	Paraguay	PY	Republic of Paraguay	LATAM
171	Peru	PE	Republic of Peru	LATAM
172	Philippines	PH	Republic of the Philippines	SOUTHASIA
173	Pitcairn	PN	NA	SOUTHASIA
174	Poland	PL	Republic of Poland	CENTEUR
175	Portugal	PT	Portuguese Republic	WESTEUR
176	Puerto Rico	PR	NA	CAR
177	Qatar	QA	State of Qatar	NORTHAFRMIDEAST
178	Republic of Korea	KR	Republic of Korea	EASTASIAPAC
179	Republic of Moldova	MD	Republic of Moldova	EASTEUR
180	Romania	RO	Romania	CENTEUR
181	Russian Federation	RU	Russian Federation	EASTEUR
182	Rwanda	RW	Republic of Rwanda	SUBAFR
183	Réunion	RE	NA	SUBAFR
184	Saint Barthélemy	BL	NA	CAR
185	Saint Helena	SH	NA	SUBAFR
186	Saint Kitts and Nevis	KN	Saint Kitts and Nevis	CAR
187	Saint Lucia	LC	Saint Lucia	CAR
188	Saint Martin (French Part)	MF	NA	CAR
189	Saint Pierre and Miquelon	PM	NA	NORTHAM
190	Saint Vincent and the Grenadines	VC	Saint Vincent and the Grenadines	CAR
191	Samoa	WS	Independent State of Samoa	EASTASIAPAC
192	San Marino	SM	Republic of San Marino	WESTEUR
193	Sao Tome and Principe	ST	Democratic Republic of Sao Tome and Principe	SUBAFR
194	Saudi Arabia	SA	Kingdom of Saudi Arabia	NORTHAFRMIDEAST
195	Senegal	SN	Republic of Senegal	SUBAFR

	Name	Code	FormalName	RegionOrigin
196	Serbia	RS	Republic of Serbia	CENTEUR
197	Seychelles	SC	Republic of Seychelles	SUBAFR
198	Sierra Leone	SL	Republic of Sierra Leone	SUBAFR
199	Singapore	SG	Republic of Singapore	SOUTHASIA
200	Sint Maarten (Dutch part)	SX	NA	CAR
201	Slovakia	SK	Slovak Republic	CENTEUR
202	Slovenia	SI	Republic of Slovenia	CENTEUR
203	Solomon Islands	SB	Solomon Islands	EASTASIAPAC
204	Somalia	SO	Federal Republic of Somalia	SUBAFR
205	South Africa	ZA	Republic of South Africa	SUBAFR
206	South Georgia and the South Sandwich Islands	GS	NA	LATAM
207	South Sudan	SS	Republic of South Sudan	NORTHAFRMIDEAST
208	Spain	ES	Kingdom of Spain	WESTEUR
209	Sri Lanka	LK	Democratic Socialist Republic of Sri Lanka	SOUTHASIA
210	State of Palestine	PS	State of Palestine	NORTHAFRMIDEAST
211	Sudan	SD	Republic of the Sudan	NORTHAFRMIDEAST
212	Suriname	SR	Republic of Suriname	LATAM
213	Svalbard and Jan Mayen Islands	SJ	NA	WESTEUR
214	Swaziland	SZ	Kingdom of Swaziland	SUBAFR
215	Sweden	SE	Kingdom of Sweden	WESTEUR
216	Switzerland	CH	Swiss Confederation	WESTEUR
217	Syrian Arab Republic	SY	Syrian Arab Republic	NORTHAFRMIDEAST
218	Tajikistan	TJ	Republic of Tajikistan	EASTEUR
219	Thailand	TH	Kingdom of Thailand	SOUTHASIA
220	Former Yugoslav Republic of Macedonia	MK	Former Yugoslav Republic of Macedonia	CENTEUR
221	Timor-Leste	TL	Democratic Republic of Timor-Leste	SOUTHASIA
222	Togo	TG	Togolese Republic	SUBAFR
223	Tokelau	TK	NA	EASTASIAPAC
224	Tonga	TO	Kingdom of Tonga	EASTASIAPAC
225	Trinidad and Tobago	TT	Republic of Trinidad and Tobago	CAR
226	Tunisia	TN	Republic of Tunisia	NORTHAFRMIDEAST
227	Turkey	TR	Republic of Turkey	CENTEUR
228	Turkmenistan	TM	Turkmenistan	EASTEUR
229	Turks and Caicos Islands	TC	NA	CAR
230	Tuvalu	TV	Tuvalu	EASTASIAPAC
231	Uganda	UG	Republic of Uganda	SUBAFR
232	Ukraine	UA	Ukraine	EASTEUR
233	United Arab Emirates	AE	United Arab Emirates	NORTHAFRMIDEAST
234	United Kingdom of Great Britain and Northern Ireland	UK	United Kingdom of Great Britain and Northern Ireland	WESTEUR
235	United Republic of Tanzania	TZ	United Republic of Tanzania	SUBAFR
236	United States Minor Outlying Islands	UM	NA	SOUTHASIA
237	United States Virgin Islands	VI	NA	CAR
238	United States of America	US	United States of America	NORTHAM
239	Uruguay	UY	Eastern Republic of Uruguay	LATAM
240	Uzbekistan	UZ	Republic of Uzbekistan	EASTEUR
241	Vanuatu	VU	Republic of Vanuatu	EASTASIAPAC
242	Venezuela (Bolivarian Republic of)	VE	Bolivarian Republic of Venezuela	LATAM
243	Viet Nam	VN	Socialist Republic of Viet Nam	SOUTHASIA
244	Wallis and Futuna Islands	WF	NA	EASTASIAPAC
245	Western Sahara	EH	NA	NORTHAFRMIDEAST
246	Yemen	YE	Republic of Yemen	NORTHAFRMIDEAST
247	Zambia	ZM	Republic of Zambia	SUBAFR
248	Zimbabwe	ZW	Republic of Zimbabwe	SUBAFR
249	Åland Islands	AX	NA	WESTEUR

European Centre for Disease Prevention and Control (ECDC)

Address:
Gustav III:s boulevard 40, SE-169 73 Solna, Sweden

Tel. +46 858601000
Fax +46 858601001
www.ecdc.europa.eu

An agency of the European Union
www.europa.eu

Subscribe to our publications
www.ecdc.europa.eu/en/publications

Contact us
publications@ecdc.europa.eu

Follow us on Twitter
@ECDC_EU

Like our Facebook page
www.facebook.com/ECDC.EU

ECDC is committed to ensuring the transparency and independence of its work

In accordance with the Staff Regulations for Officials and Conditions of Employment of Other Servants of the European Union and the ECDC Independence Policy, ECDC staff members shall not, in the performance of their duties, deal with a matter in which, directly or indirectly, they have any personal interest such as to impair their independence. Declarations of interest must be received from any prospective contractor(s) before any contract can be awarded.
www.ecdc.europa.eu/en/aboutus/transparency

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
  via EU Bookshop (http://bookshop.europa.eu);
- more than one copy or posters/maps:
  from the European Union's representations (http://ec.europa.eu/represent_en.htm);
  from the delegations in non-EU countries (http://eeas.europa.eu/delegations/index_en.htm);
  by contacting the Europe Direct service (http://europa.eu/europedirect/index_en.htm) or calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (*).

  (*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (http://bookshop.europa.eu).

Publications Office