

ECDC TECHNICAL DOCUMENT

Documentation for use of the HIV Estimates Accuracy Tool



This documentation 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 Kapodistrian University of Athens in 2016 and reviewed and approved by the 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 and Emiliano Farinella

Suggested citation for the manual:

European Centre for Disease Prevention and Control. ECDC TECHNICAL DOCUMENT
Documentation for use of the HIV Estimates Accuracy Tool. Stockholm: ECDC; 2018

Stockholm, November 2018

© European Centre for Disease Prevention and Control, 2018

Reproduction is authorised, provided the source is acknowledged.

Contents

1	Introduction	1
1.1	Methodological background.....	1
1.2	Methods used in the HIV Estimates Accuracy Tool	3
1.3	Specific issues	4
1.4	How this document works.....	5
2	Prerequisites.....	5
2.1	Dataset	5
2.2	Online version.....	7
2.3	Offline version	8
3	Using the HIV Estimates Accuracy tool	9
3.1	How to open the tool.....	9
3.2	Construction of the tool	10
4	Input data upload tab.....	10
4.1	Uploading data.....	10
4.2	Uploading a saved application state	10
4.3	Mapping and validating data.....	11
4.4	Defining the migrant variable categorisation	12
4.5	Opening a new instance of the tool.....	13
4.6	Setting the seed for the random processes used by the tool.....	13
5	Input data summary tab	14
5.1	Inspecting missing data patterns	14
5.2	Inspecting reporting delay patterns.....	16
5.3	Applying filters	17
6	Adjustments tab.....	18
6.1	Joint Modelling Multiple Imputation.....	18
6.2	Multiple Imputation using Chained Equations - MICE	20
6.3	Simple reporting delay	21
6.4	Reporting delay with trend.....	22
6.5	Intermediate outputs of adjustments and diagnostics – joint modelling multiple imputations.....	24
6.6	Intermediate outputs of adjustments and diagnostics – multiple imputation with chained equations, MICE	26
6.6	Intermediate outputs of adjustments and diagnostics – reporting delay.....	28
7	Reports	29
7.1	Creating report	29
7.2	Exporting report.....	31
8	Outputs.....	31
8.1	Adjusted dataset	31
8.2	Reporting delay weights.....	32
8.3	Application state data	32
	References	34
	Annex 1. Codes used for countries and regions.....	35

1 Introduction

The HIV Estimates Accuracy Tool is an R - based application that uses statistical methods to allow for adjusted estimates from the 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 the following formats: RecordType = HIV or RecordType = HIVAIDS. Case-based surveillance data containing the required set of variables and consistent with the TESSy format in coding of the variables may also be used.

The tool performs multiple imputations for the 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 to correct for delays in reporting through reverse time hazard estimation. The adjustments may be used separately or in combination.

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

This document will guide you through the HIV Estimates Accuracy Tool, explain why each step of the tool may be needed, explain 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 some 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:

- a) **data missing completely at random (MCAR)** - a value is missing independently of the value itself and of any other factors including observable covariates;
- b) **data missing at random (MAR)** - a value is missing independently of the value itself but the fact that it is missing may depend on other covariates;
- c) **data missing not at random (MNAR)** – the fact that a value is missing may depend on the value, which 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. Further, it is not possible to discriminate between MAR and MNAR based on the 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 the data follow a **monotone missingness pattern**. In this pattern, the incomplete variables can be ordered so that if the value of the first variable is missing then the value of the second variable is missing, as well as the values of all the following variables. Further, regardless of the first variable, if the value of the second variable is missing, then the value of the third one and all the following variables are also missing and so on.

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 we obtain M so called pseudo-complete datasets. The model of

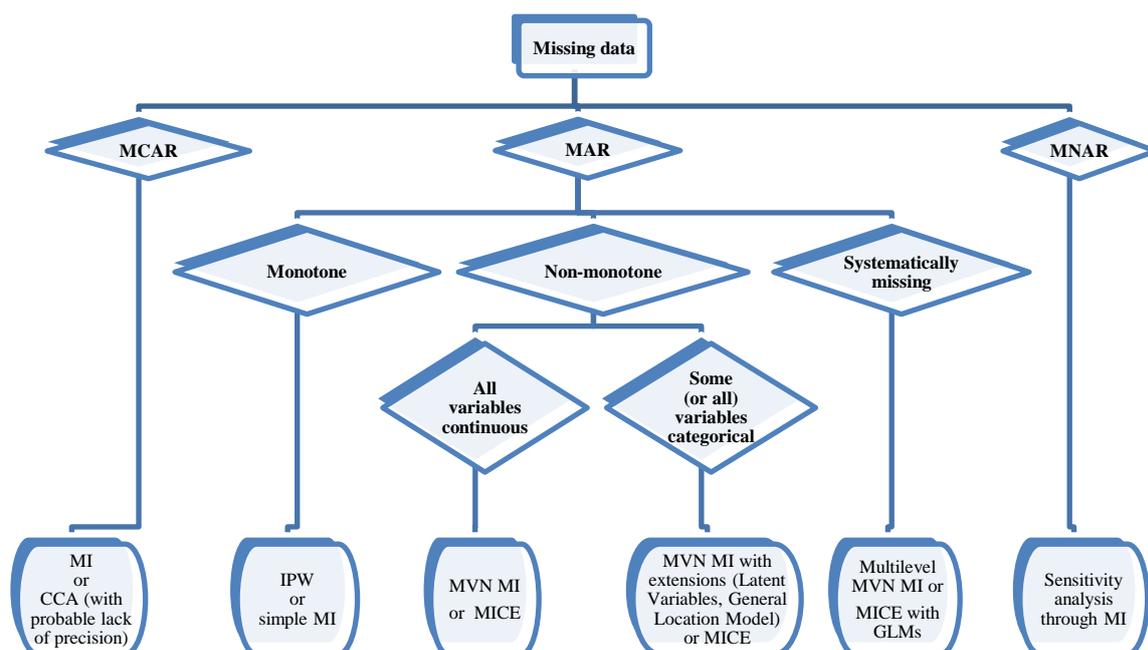
interest (also called “substantive model”) can be fitted to each of the imputed datasets in order to estimate the parameter of interest and its variance M times. These can be combined using Rubin's rules to obtain an overall (average over M) estimator and its associated variance. This variance is enlarged 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 (multiple imputations by chained equations – MICE).

The **multivariate normal imputation** relies on the assumption that the joint distribution of all variables under consideration is multivariate normal. If the 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** (MI by chained equations, MICE). With the MICE method, 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 the full conditional specification approaches can be extended to datasets combining data from different national surveillance systems through multilevel multiple imputation. The suggested approach to the missing data is presented in Figure 1.

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



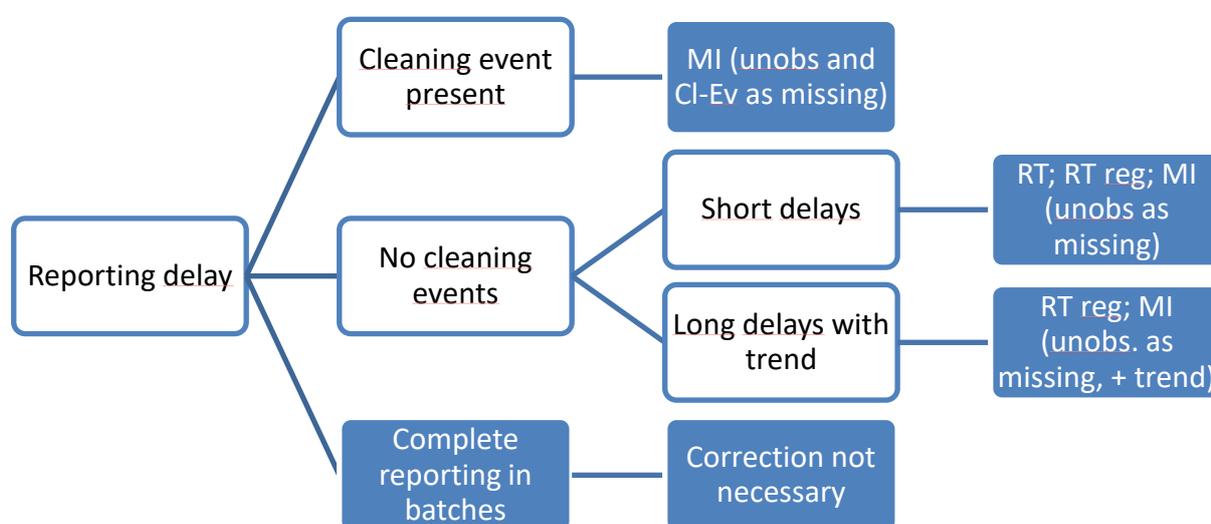
MCAR = missing completely-at-random; MAR = missing-at-random; MNAR = missing not-at-random; MI = multiple imputations; CCA = complete-case analysis; IPW = inverse probability weighting; MVN MI = multivariate normal MI; MICE = MI by chained equations

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 the 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 using the reverse time transform and estimating the survivorship function with left-truncated data. In practice, both the confidence intervals and the point estimates for the 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, the HIV surveillance systems rely on cyclic uploading of complete data on new diagnoses during a pre-defined period of time from laboratory databases. In case of such batch reporting delay may still be calculated, but using the adjustment methods is not necessary. The suggested approach to the reporting delays is presented in Figure 2.

Figure 2. Appropriate methods to deal with reporting delays



MI = multiple imputations of yet unobserved counts and – if applicable – artificially removed counts recorded during cleaning events; CI-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 "jomo" R-package, full conditional specification through "mice" R-package and application of Rubin's rules through "mitools" R-package.

The tool first imputes missing values for **gender (single imputation)**. Since the other variables are imputed separately for males and females and Gender is missing only for a small proportion of case, this simplifies that procedure. Gender "Other" is imputed as either male or female. This is a simplification for the statistical procedures, but for inference it is recommendable 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 be always known (AIDS at diagnosis and, 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 was included in the form of **cubic spline**. The number of knots of the spline may be selected by the user in the range 3 – 5.

Obtaining appropriate imputation requires a procedure that allows estimation of the joint distribution. This is an iterative procedure, which 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 the 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 the estimates obtained with 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 than the actual case counts observed in surveillance.

When adjusting for **reporting delay** in surveillance the time units used vary from 1 day to 1 year. HIV data in Europe were traditionally collected by quarter of an year. In addition the data are usually presented annually, so only longer delays (of several months) can lead to underestimating of the number of diagnoses in the most recent years. Thus 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 the quarter of 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 the reporting delays is imputed along other variables containing missing values, based on other covariates as well as available information about the dates (maximum plausible reporting delay).

The truncation time is assumed to be the latest notification quarter, that occurs in the dataset. However, the **truncation time** may be manually changed by the user in the reporting delay parameters' window, if in e.g. the data do not entirely cover the last quarter. In addition the user may choose to limit the 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 the truncation time and taking the diagnosis quarter as the 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 by default the year of diagnosis 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 the diagnosis and truncation time. The adjusted counts base on application of the adjustment formula to distinct covariates patterns and combining them under the assumption of independence.

The reporting delay estimation models do not account for possible differences in reporting during the year. If in 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 perform **first the imputation and then calculation of reporting delay weight**. The reporting delay distribution estimation is performed separately for each imputed dataset. Weighted (adjusted for reporting delay) estimates are produced for each imputed dataset, which are then combined using the Rubin's rules.

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

1.3 Specific issues

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

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 the 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 minimize the impact of non-normally distributed CD4 count. In case of high percent 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 some 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 (are 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 How this document works

This document takes you step-by-step through the tool 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), you have the following items:

- **Description**
 - o Provides you with a short description on what the corresponding elements of the tool are and what type of output they provide.
- **How-to**
 - o This section tells you what to do with the output provided by Stata
- **Interpretation**
 - o Here the meaning of the output is described
- **Actions to be taken**
 - o This section tells you what to do if there are any issues with the output. This might mean you need to carry out another analysis, modify the data, etc.

Disclaimer:

The dummy dataset based on TESSy HIV dataset has been used as a model for this documentation. This dataset 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 documentation cannot be interpreted and used for any reliable inferences.

2 Prerequisites

2.1 Dataset

1. File should contain case based records of HIV diagnoses.
2. There are 19 required attributes/variables by the tool to run the adjustments. They are presented in the Table 1. with the description of values required for each of the attribute/variable.
3. Upload file should contain all these attributes/variables names, except empty columns and columns containing a single value (e.g. ReportingCountry), which can be created directly in the tool.

4. Different names of variables 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, please note that the variables have to be coded as specified in the below.

No	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	The country reporting the record, according to the ISO list	(see 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	The variable specifies the CD4 cells count at the time of HIV diagnosis. Enter the numeric value of the CD4 (0-6000) or unknown (UNK).	0 - 6000
5	FirstCD4DateYear	Year of first CD4 cell count at time of diagnosis	>1985
6	CountryOfBirth	The country of birth of the patient according to the ISO list. Some additional values used in surveillance are also included (please see the Annex 1). CountryofBirth is the preferred variable for migration status. If Unknown - code as UNK or Blank.	(see Annex 1)
7	CountryOfNationality	Country of nationality of patient, according to the ISO list. Some additional values used in surveillance are also included (please see the Annex 1).	(see Annex 1)
8	RegionOfOrigin	Region of origin of patient	(see Annex 1)
9	DateOfAIDSDiagnosisYear	The 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 of the record.	≥1984
10	DateOfDeathYear	The year of death because of HIV/AIDS	
11	DateOfDiagnosisYear	The year of first HIV diagnosis; clinical or laboratory diagnosis. Missing values are not allowed.	≥1985
12	DateOfDiagnosisQuarter	The quarter of first HIV diagnosis; clinical or laboratory diagnosis.	1,2,3,4
13	DateOfNotificationYear	This is the year on which the HIV case was notified for the first time to the reporting country.	≥1985
14	DateOfNotificationQuarter	This is the quarter on which the HIV case was notified for the first time to the reporting country.	1,2,3,4
15	Gender	Gender of a patient. Transsexual should be coded as O - Other	F = Female M = Male O = Other UNK = Unknown
16	Outcome	Information on whether the case is alive or deceased. The death should be due to the reported disease.	A = Alive D = Died UNK = Unknown
17	PlaceOfNotification	Place of the first notification of the case to a regional authority. Select the most detailed NUTS level possible.	
18	PlaceOfResidence	Place of residence of patient at the time of disease onset. Select the most detailed NUTS level possible.	
19	Transmission	Describes the most probable route of Transmission	Transmission: HAEMO =

	Nosocomial infection includes patients infected in health care settings. Case of occupational exposure should be classified as UNK 'Unknown or undetermined'. Cases which are not fully documented should be coded as UNK.	haemophilic 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
--	--	--

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

If there is one of the variables is not present in the dataset it may be artificially created (see the "Default values" in the Attributes mapping widow description).

6. If the file which is to be upload to the tool was previously uploaded to TESSy database and successfully passed TESSy validation, there should be no problem with using it with the tool unless all the 19 required by the too 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 <https://ecdc.shinyapps.io/hivestimatesaccuracyui/>:

To access shinyapp.io, the user needs to create an account on [Shinyapps.io](https://shinyapps.io). Authentication is possible through one of the following three methods:

- Shinyapps.io authentication. A new Shinyapps.io account can be created during the authentication process and requires only e-mail address and a new password for the account.
- Google Authorization
- GitHub authorization

To access the link, please send an e-mail to HIV.Modelling@ecdc.europa.eu with the subject "Registration for HIV estimates accuracy tool + Full Name". An invitation e-mail will be send back.

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

2.3 Offline version

There are two possible ways to download the offline version.

1. For experienced R users a CRAN-like repository is set up for installing the tool as 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:

```
1)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.
2) Once R is done with installation the tool can be run with command:
> hivEstimatesAccuracy::RunApp()
3) Periodically, the user can update the tool with the following command:
> update.packages(repo = "http://www.nextpagesoft.net/hiv-estimates-accuracy/repo/")
```

2. An offline Windows x64 deployment package with R environment embedded

Last option for deployment of the tool locally is to use a deployment package, which includes all required software and R packages. Simply follow the steps:

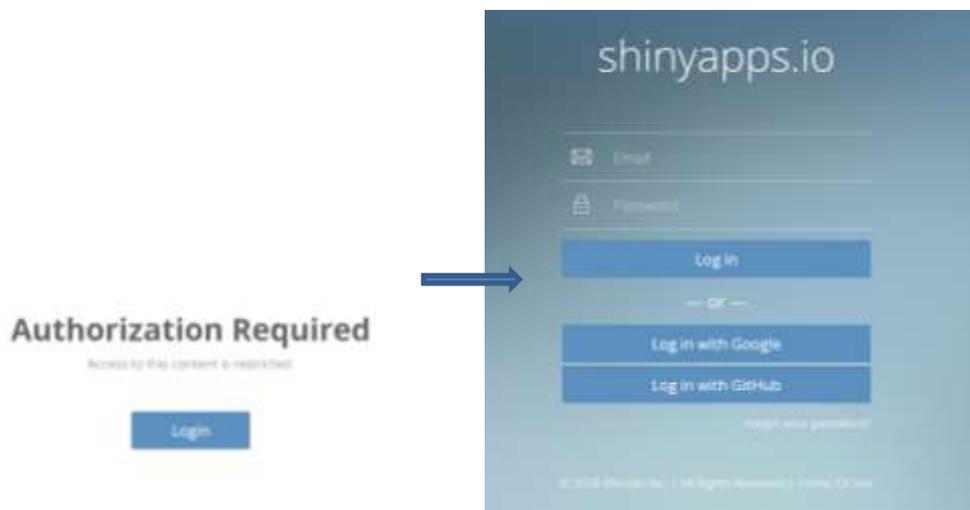
- 1) Download the deployment package by clicking the following link:
<http://www.nextpagesoft.net/hiv-estimates-accuracy/win/x64/hivEstimatesAccuracy-1.0.0.x64.zip>
(201 MB download size)
- 2) Unpack the file to an arbitrary folder.
- 3) After unpacking a new folder will appear called "hivEstimatesAccuracy". Browse inside and double-click file "hivEstimatesAccuracy.bat". This will open the tool in the default web browser. When done with working with it simply close the browser window.

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

3 Using the HIV Estimates Accuracy tool

3.1 How to open the tool

When following the link to the online tool <https://ecdc.shinyapps.io/hivestimatesaccuracyui/> you will be taken to Authorisation screen.



To log in, the user needs to have an account on [Shinyapps.io](https://shinyapps.io). Authentication is possible through one of the following three methods:

- Shinyapps.io authentication
- Google Authorization
- GitHub authorization

ECDC must register the user prior to the first use. Please, send the 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:

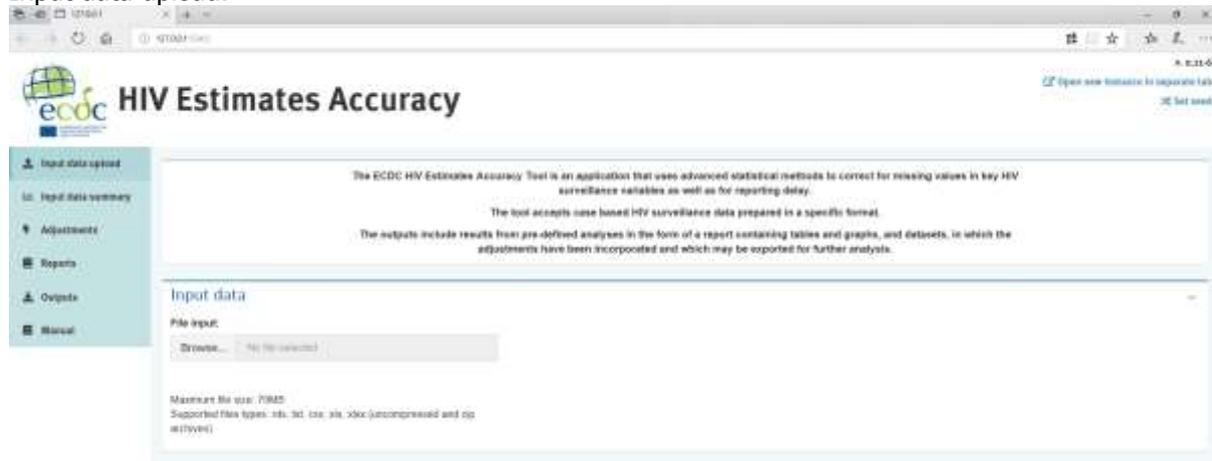
```
> hivEstimatesAccuracy::RunApp()
```

The offline version installed using the Windows x64 deployment package can be opened by double-clicking file "hivEstimatesAccuracy.bat".

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

3.2 Construction of the tool

The tool is organised into tabs displayed on the left-hand panel. It automatically opens at the first tab – Input data upload.



Tabs:

1. **Input data upload.** Further tabs are not active unless the data are uploaded and validated in this tab.
2. **Input data summary.** Allows exploration of data and selection of filters for adjustments
3. **Adjustments.** Tab to specify adjustments and parameters for adjustments as well as to examine diagnostic charts.
4. **Reports.** Allows to create and export a pre-defined report
5. **Outputs.** Contains output datasets that can be used for further analysis
6. **Manual**

You may navigate freely between the tabs once the dataset is uploaded.

4. Input data upload tab

This tab is always active and allows to upload and pre-process data.

4.1. Uploading data

Description

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

How-to

Select "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 the file, number of records, variables names) and opens new sections: "Attributes mapping" and "Migrant variables regrouping".

Actions to be taken

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

4.2. Uploading a saved application state

Description

The tool allows uploading of a previously saved application state that contains uploaded and pre-processed data as well as adjustments, which were previously applied.

How-to

Select "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. In particular mapped and pre-processed data are available for further analysis.

Actions to be taken

Proceed to further tabs to continue the analysis.

4.3. Mapping and validating data

Description

The "Attributes mapping" section provides a possibility to match between variables' names used internally by the tool (column "Attribute") and the variables present in the input data (column "Input Data"). The names of the variables used by the tool correspond to the names of the variables used in the TESSy metadataset. If the variables in the input data have the same or similar names they will be automatically identified by the Tool and suggested in the "Input Data" column. If the tool cannot identify the mapping the field will be left blank.

How-to

The mapping automatically proposes assigning the variables with names similar or the same as the ones used by the tool. Other variables are mapped manually by selecting the appropriate variable (from the Input dataset) from the drop-down menu.

Please, provide mapping between attributes used internally by the tool (column "Attribute") and the input data dimensions (column "Input data column"). If "Input data column" is not specified, their value in column "Default value" is used.

Attribute	Input data column	Default value
Region	region	
RegionCountry		NA
Age	age	
FirstCD4Count	cd4_num	
FirstCD4DateYear	firstcd4year	
CountryOfBirth	countryofbirth	
CountryOfNationality	countryofnationality	
RegionOfOrigin	regionoforigin	
DateOfFirstDiagnosisYear	dateofdiagnosisyear	
DateOfDeathYear	dateofdeathyear	
DateOfDiagnosisYear	dateofdiagnosisyear	
DateOfDiagnosisQuarter	dateofdiagnosisquarter	
DateOfInfectionYear	dateofinfectionyear	
DateOfInfectionQuarter	dateofinfectionquarter	

Input data has to be mapped to internal attributes and validated. Adjust mapping and press "Apply mapping" button to the left.

If the variable has a single value you may define it in the tool

Select appropriate variable name that appears in your data

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

In particular 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 "Apply mapping" button on the top of this section.

Interpretation

Clicking on the "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. Successful mapping process is indicated by the statement "mapping is valid" and "Values are valid".

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

Valid mapping automatically triggers pre-processing of data. During the pre-processing a migrant status variable is created based on the following variables: ContryOfBirth, CountryOfNationality, RegionOfOrigin;

and AIDS at diagnosis based on: DateOfAIDSDiagnosisYear, DateOfDiagnosis. Moreover, a single imputation of Gender is performed.

The pre-processed data may be inspected at the bottom of the page, in the "Input data records pre-processed".

Actions to be taken

Once the validity of mapping and the 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: ContryOfBirth, CountryOfNationality, RegionOfOrigin in combination with the variable ReportingCountry. Based on this regrouping a variable FullRegionOfOrigin is created, based on categorisation used in TESSy (see Annex 1). The FullRegionOfOrigin variable may be regrouped into categories that are the most relevant to the particular country.

How-to

The following options are available:

- REPCOUNTRY+UNK+OTHER
- REPCOUNTRY+UNK+SUBAFR+OTHER
- REPCOUNTRY+UNK+3 most prevalent regions+OTHER
- Custom

Migrant variable regrouping

Distribution of region of origin:
All regions in detail in descending frequency

FullRegionOfOrigin	Count
REPCOUNTRY	1023
SUBAFR	2295
WESTEUR	1071
SOUTHASA	168
LATAI	138
CAS	128
CENTEUR	100
NORTHAM	58
EASTEUR	48
AUSTNC	42
NORTHAFRMEAS	28
EASTASAPAC	28
UNK	1088

Grouping options:
REPCOUNTRY + UNK + SUBAFR + OTHER

GroupedRegionOfOrigin	FullRegionOfOrigin	Count
REPCOUNTRY	REPCOUNTRY	1023
SUBAFR	SUBAFR	2295
OTHER	AUSTNC, CAS, CENTEUR, EASTASAPAC, EASTEUR, LATAI, NORTHAM, SOUTHASA, WESTEUR	1018
UNK	UNK	1088

Select desired grouping option

Custom regrouping may be created by selecting "Custom option" and creating group using option "Add group".

Migrant variable regrouping

Distribution of region of origin:
All regions in detail in descending frequency

FullRegionOfOrigin	Count
REPCOUNTRY	1023
SUBAFR	2295
WESTEUR	1071
SOUTHASA	168
LATAI	138
CAS	128
CENTEUR	100
NORTHAM	58
EASTEUR	48
AUSTNC	42
NORTHAFRMEAS	28
EASTASAPAC	28
UNK	1088

Grouping options:
Custom

GroupedRegionOfOrigin	FullRegionOfOrigin	Count
GROUP 1		

Edit group name

Click on this field to add regions to the group.

Regions may be added to the group by clicking on the "FullRegionOfOrigin" field and then adding regions from drop down menu. Regions are added by clicking on appropriate names. Unselected regions are automatically grouped into group "OTH".

Interpretation

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

Actions to be taken

Select appropriate grouping and then proceed to further tabs.

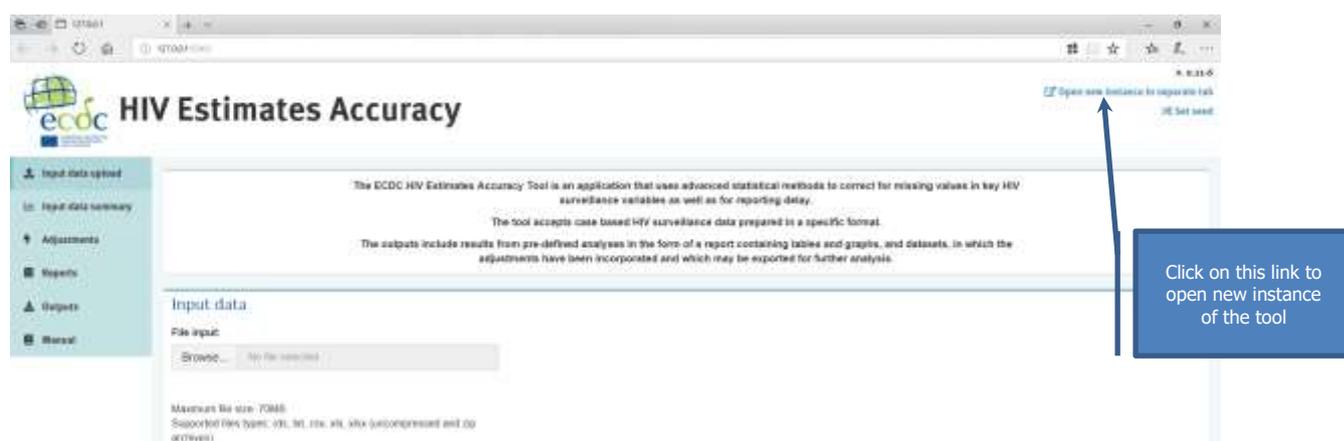
4.5. Opening a new instance of the tool

Description

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

How-to

In order to open a new instance of the tool select the button "Open new instance in separate tab" in the right-hand top 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 new data upload. You may avoid duplicating the mapping process by first saving the workspace with pre-processed data for further upload in the new instance of the tool.

Actions to be taken

Proceed with adjustments.

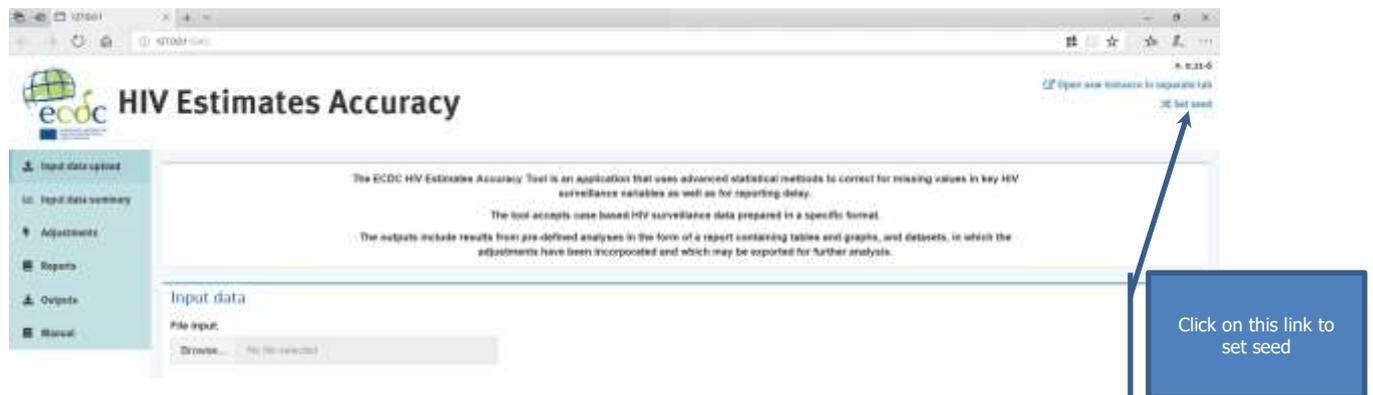
4.6. Setting the seed for the random processes used by the tool

Description

The tool uses a random number generator when imputing the 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 initialized with the same number (seed).

How-to

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

**Interpretation**

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

Actions to be taken

Proceed with adjustments.

5. Input data summary tab

This tab allows to inspect data quality issues present in the inputted data.

5.1. Inspecting missing data patterns

Description

Section 1 provides 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 pre-processing step.

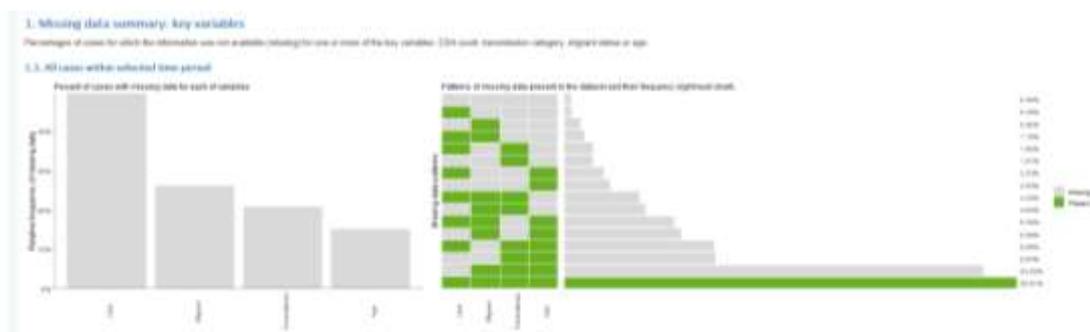
Section 2 provides summary of the trends of the proportion of the missing values in key variables by diagnosis year.

How-to

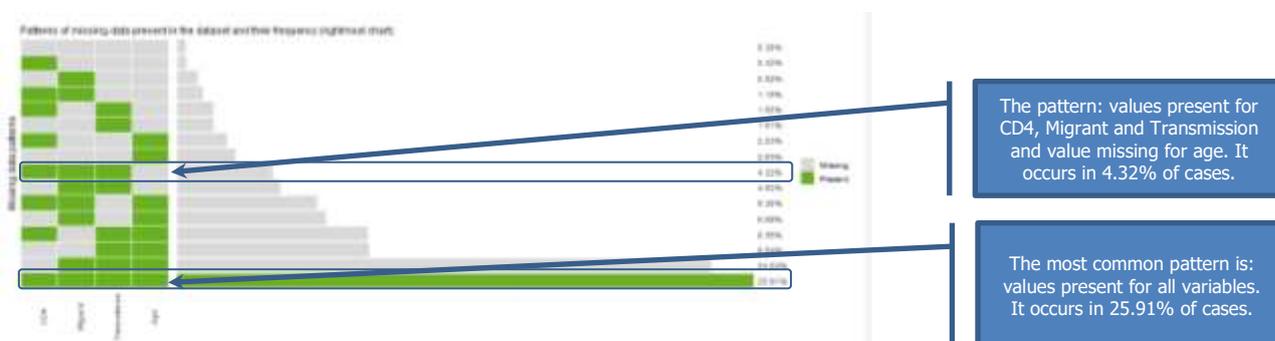
The output is generated automatically when moving to the Input data summary tab. You may select time periods for which the data are summarised by selecting filters.

Interpretation

In section 1, for all cases as well as separately for males and for females two graphs are presented: a) Percent of cases with missing data for each of variables and b) Patterns of missing data present in the dataset and their frequency (rightmost 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 which are 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 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 that they occur in data.

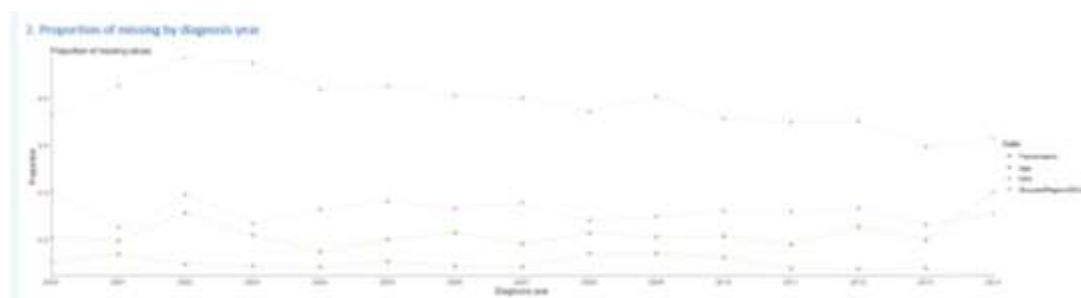


The graphs allow to check if the reported levels of missing values are correct for the input data and may help in deciding whether the data should be restricted for further analysis.

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 monotone or heterogeneous (non-monotone). Monotone missing pattern would be represented as "grey triangle" without green cells within. Patchy pattern indicates heterogeneity. At the moment the adjustment methods implemented in the tool assume non-monotone missing pattern. 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 this 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 some 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.

Actions to be taken

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

5.2. Inspecting reporting delay patterns

Description

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

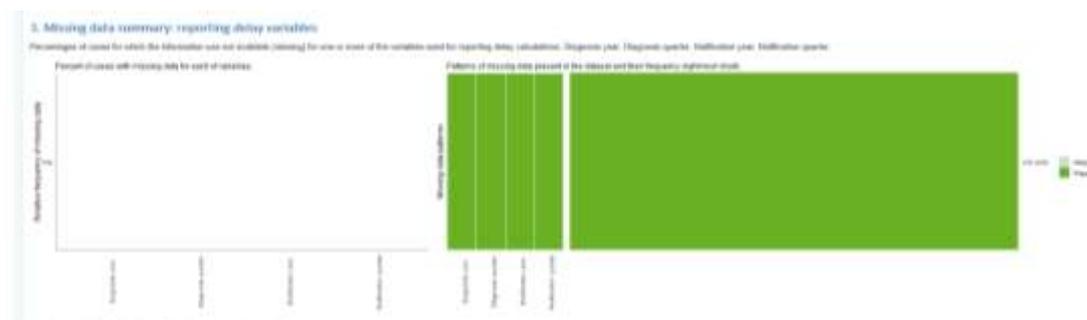
Section 5 displays the average delay, by notification quarter. It also provides a line for trend in the reporting delay by notification quarter and the upper limit of typical delays given the variability of the average reporting delay by quarter.

How-to

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

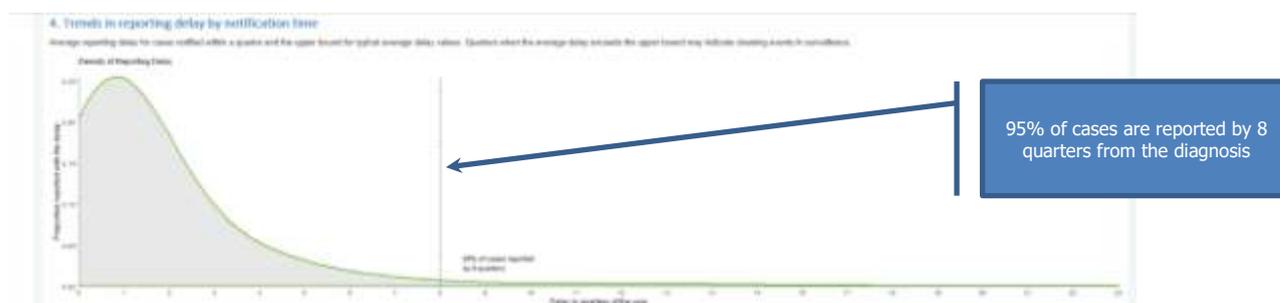
Interpretation

The two graphs displaying the patterns of missing values are similar to graphs for the other variables: a) Percent of cases with missing data for each of variables and b) Patterns of missing data present in the dataset and their frequency. In the example below the required variables are 100% complete, so only one pattern of missing/available values is present.

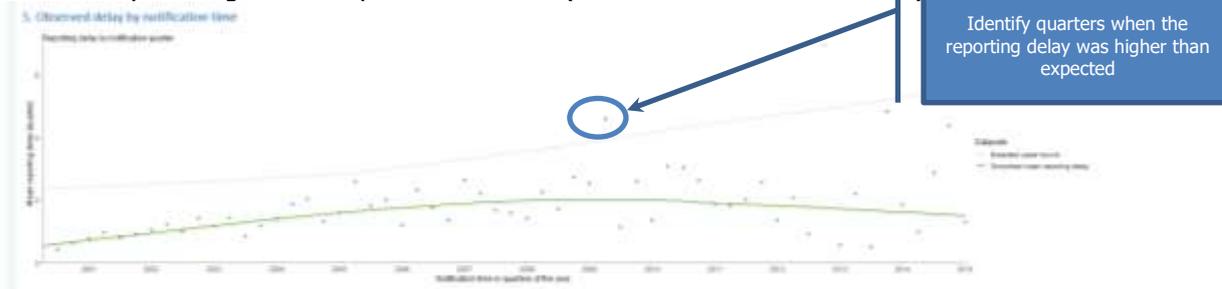


The reporting delay weight is estimated based on cases for which all the four variables are available (100% complete pattern). If missing data are present the results may be less accurate, especially in case of substantial missingness for the reporting delay variables.

Next section provides observed distribution of the reporting delay. It provides an overview how important is the reporting delay in the Input dataset. Note that 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 represent the quarter by which 95% of cases were reported. Since usually the data are also analysed with some delay, if 95% of cases are reported within 2 quarters than the delay adjusting for reporting delay will not make much difference. In case of the example data – this is 8 quarters indicating important delays.



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



Actions to be taken

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

5.3. Applying filters

Description

The Input data summary tab allows to apply 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 on to the adjustments. When passed on to the adjustments, the filtering will also have an effect on the output datasets.

How-to

The filters may be applied by using sliders. Both the start time and the end time may be changed for both the year of diagnosis and the time of notification. The chart below each slider shows the distribution of cases by gender among the included and excluded cases. The application of filter 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 checks an appropriate box.

Drag the ends of the sliders to change their position

Check the box to filter the data also for adjustments.

Interpretation

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

Both the slider for the diagnosis year and the 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 filter is included the tool issues a warning.



Actions to be taken

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

6. Adjustments tab

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

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



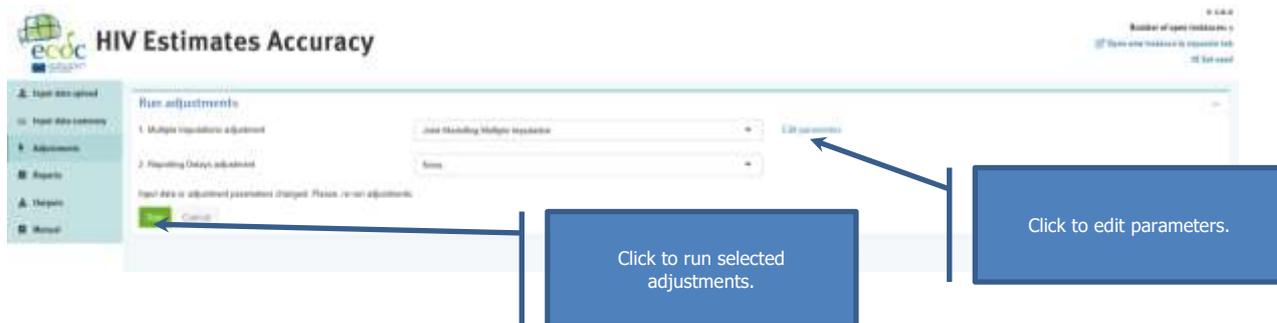
6.1. Joint Modelling Multiple Imputation

Description

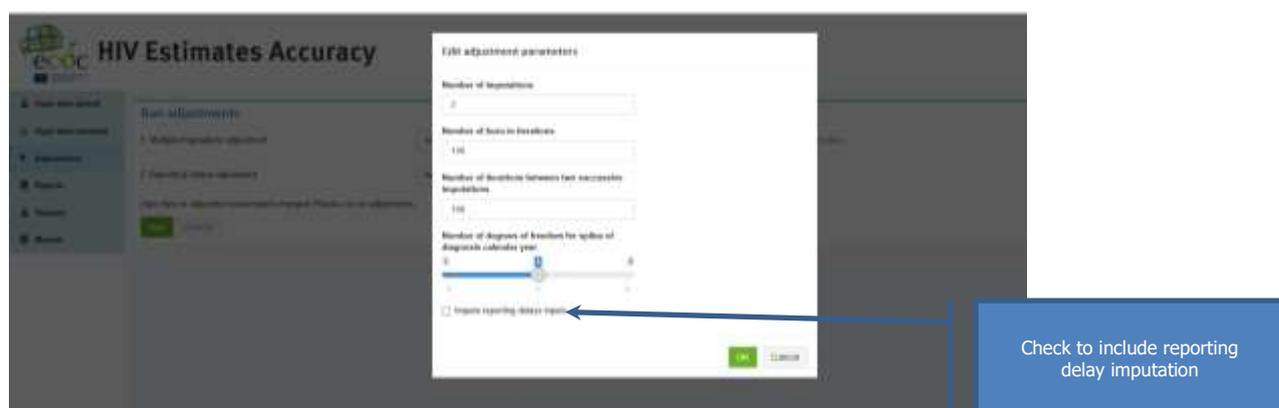
This option performs multiple imputation with joint multivariate normal modelling. This is an iterative procedure that can be very time consuming. The amount of the time needed depends on the parameters set. It is recommended to start with lower numbers to look at the outputs and allow up to several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

How-to

Select the "Joint Modelling Multiple imputation" from the drop-down menu in the "Multiple Imputations adjustment" field. A possibility to edit parameters will appear at the right-hand side.



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



Interpretation

The parameters relating to the imputation procedure and the imputation model are displayed below. Please refer to the diagnostics section in order to select proper values.

Parameter	description	What to select
Number of imputations	The number of imputed datasets that will be produced	For test runs select 2. For the final adjustments at least 5 – 10 imputations
Number of burn-in iterations	The number of iterations after which the 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	The number of iterations between outputting the successive imputed dataset, which should limit autocorrelation of imputed datasets.	For the test runs select 100. Usually this is sufficient or too high. Please refer to the adjustment diagnostics.
Number of degrees of freedom for splines of diagnosis calendar year	The 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 of: quarter of diagnosis, notification year, quarter of notification are missing	Should be applied in case of substantial proportion of missing values in reporting delay variables.

Actions to be taken

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

6.2. Multiple Imputation using Chained Equations - MICE

Description

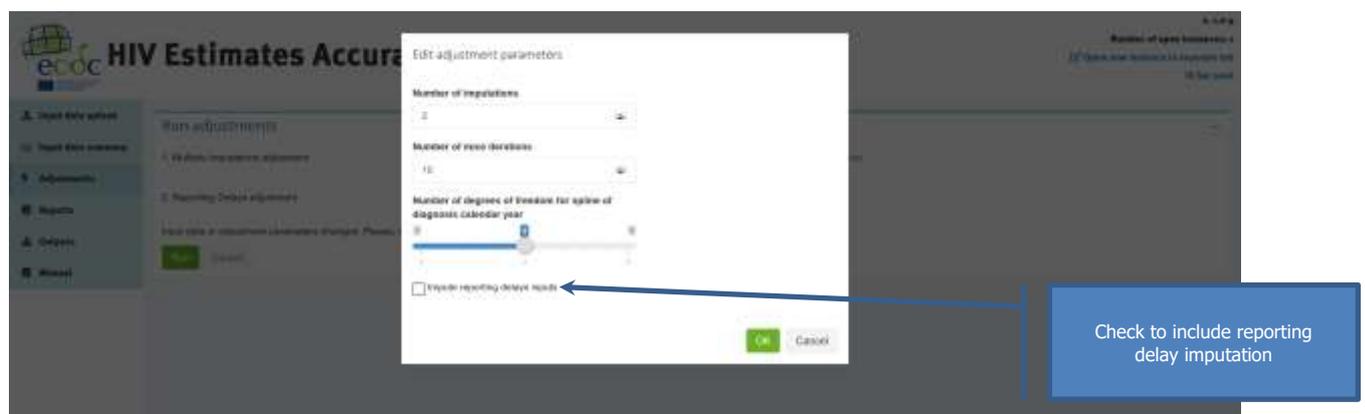
This option performs multiple imputation using chained equations. This is an iterative procedure that can be very time consuming. The amount of the time needed depends on the parameters set. It is recommended to start with lower numbers to look at the outputs and allow up to several hours for the final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

How-to

Select the “Multiple Imputation using Chained Equations” from the drop-down menu in the “Multiple Imputations adjustment” field. A possibility to edit parameters will appear at the right-hand side.



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



Interpretation

The parameters relating to the imputation procedure and the imputation model are displayed below. Please refer to the diagnostics section in order to select proper values.

Parameter	description	What to select
Number of imputations	The number of imputed datasets that will be produced	For test runs select 2. For the final adjustments at least 5 – 10 imputations
Number of mice iterations	The number of iterations after which the 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	The 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. Chose 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 of: quarter of diagnosis, notification year, quarter of notification are missing	Should be applied in case of substantial proportion of missing values in reporting delay variables.

Actions to be taken

Test run the selected adjustments with smaller values for the “Number of imputations”, “Number of mice iterations”. Inspect the diagnostics section. Re-run 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.

How-to

Select the “Reporting delay” from the drop-down menu in the “Reporting delay adjustment” field. A possibility to edit parameters will appear at the right-hand side.

A pop-up window is displayed with pre-set values of the parameters, which can be edited. The interpretation and selection of proper values is provided in the section below.



Interpretation

The parameters relating to the reporting delay estimation are displayed. Note that filtering on diagnosis year 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 the 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 the estimation could be performed on the later data.
Notification end year and quarter	Only the 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 instable if the stratification results in small numbers of cases in some strata.

Actions to be taken

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

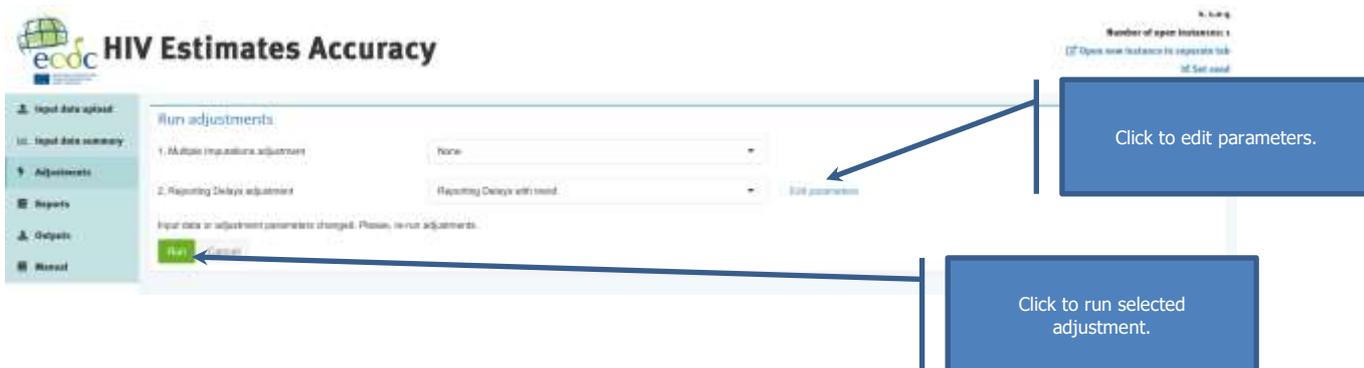
6.4. Reporting delay with trend

Description

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

How-to

Select the "Reporting delay with trend" from the drop-down menu in the "Reporting delay adjustment" field. A possibility to edit parameters will appear at the right-hand side.



A pop-up window is displayed with pre-set values of the parameters, which can be edited. The interpretation and selection of proper values is provided in the section below.



Interpretation

The parameters relating to the reporting delay estimation are displayed. Note that filtering on diagnosis year 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 the 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 the estimation could be performed on the later data.
Notification end year and quarter	Only the 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 instable if the stratification results in small numbers of cases in some strata.

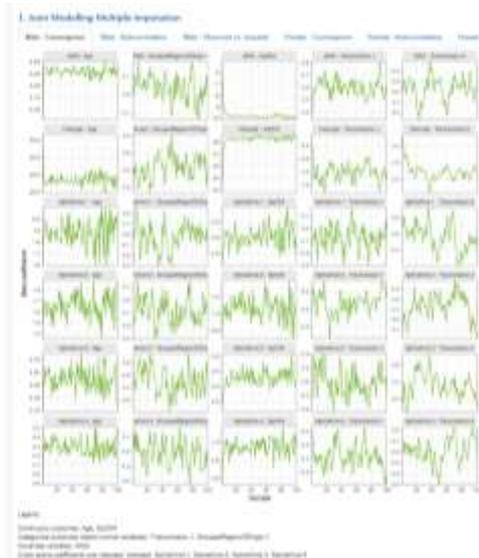
Actions to be taken

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

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

Description

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



Tabs contain outputs related to convergence of and autocorrelation in the imputation procedure as well as comparison of distribution of imputed vs observed values

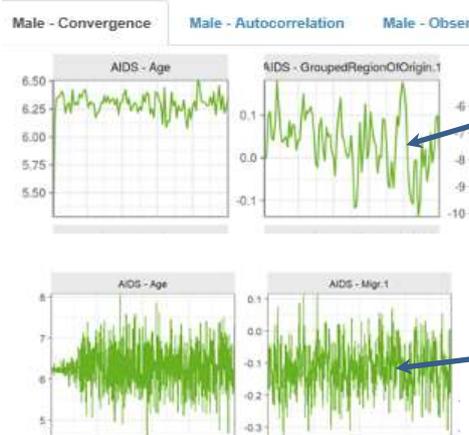
How-to

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 is to determine 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 some of the parameters display some trends, which do not level off at the right hand side of the graph. In case more iterations are needed this can be controlled with the "Number of burn-in iterations".

Joint Modelling Multiple Imputation

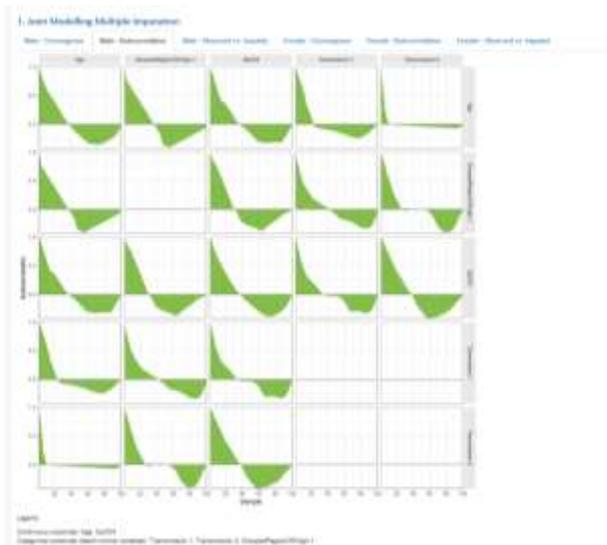


These chains have not converged

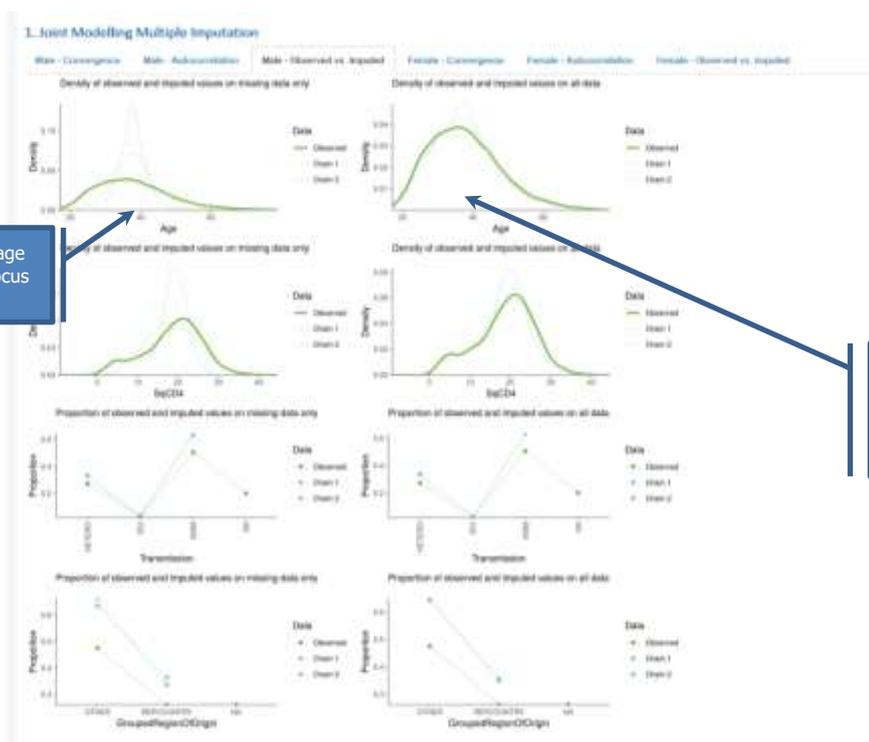
These chains have converged

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 plot below 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 values and the distribution of imputed values and the second one – comparing the distribution of observed values and the distribution of all values observed and imputed.



Imputed values of age are more likely to focus around 40

The complete distribution of age after imputation is similar to the distribution of the observed values

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 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, normally the main trends are preserved. In any case graphs should be judged only if convergence is suggested by the previous type of graphs.

Actions to be taken

In case of lack of convergence increase the number of iteration through increasing the “Number of burn-in iterations”.

Increase the number of “Iterations between subsequent imputations” if needed.

Re-run the analysis.

In case the distributions of the imputed values are very different from the observed values – re-run the analysis with MICE.

6.6. Intermediate outputs of adjustments and diagnostics – multiple imputation with chained equations, MICE

Description

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

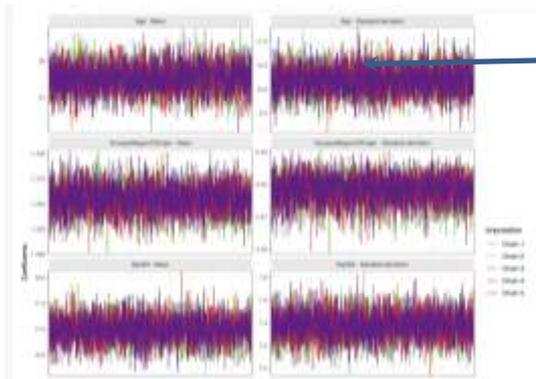
**How-to**

Intermediate outputs are generated automatically when running the adjustments.

Interpretation

The output related to the convergence contains trace plots.

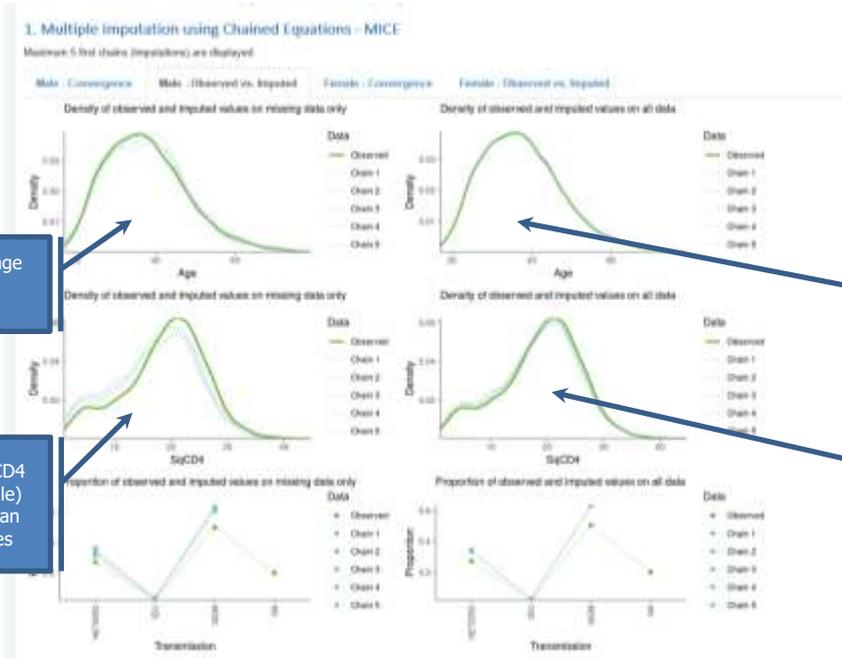
The use of the trace plots is to determine whether the procedure converged, assuring that the missing values are imputed from the correct distribution. In case of convergence the trace plot for every parameter does not display any pattern. More iterations are needed in case some of the parameters display some trends, which do not level off at the right hand side of the graph. In case more iterations are needed this can be controlled with the “Number of mice iterations”. The picture below represent a converged procedure.



The plots represent a converged procedure:

- No evident trends
- Particular chains are mixed

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 values and the distribution of imputed values and the second one – comparing the distribution of observed values and the distribution of all values observed and imputed.



Imputed values of age are similar to the observed values

Imputed values of CD4 (on square-root scale) tend to be lower than the observed values

The complete distribution of age after imputation is almost identical to the distribution of the

The complete distribution of CD4 (on square-root scale) after imputation is similar to the distribution of the observed values

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 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, normally the main trends are preserved. In any case graphs should be judged only if convergence is suggested by the previous type of graphs.

Actions to be taken

In case of lack of convergence increase the number of iteration through increasing the “Number of mice iterations”. Re-run the analysis.
 In case the distributions of the imputed values are very different from the observed values – re-run the analysis with joint modelling.

6.6. 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 univariable analysis of the selected predictors of the reporting delay adjustments.

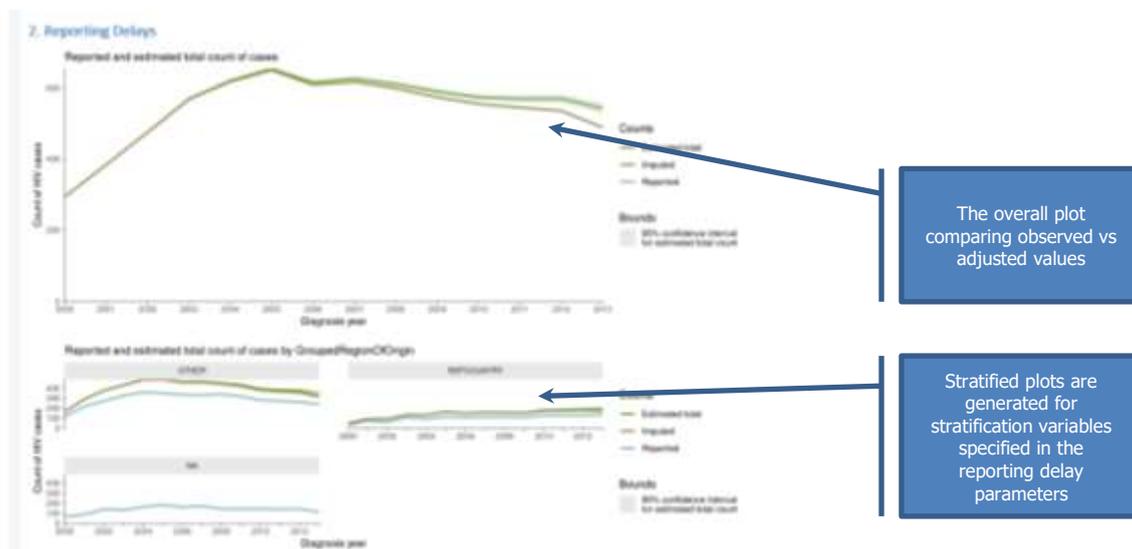
How-to

Intermediate outputs are generated automatically when running the adjustments.

Interpretation

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

If stratification was introduced, also trends by stratification variables are displayed. In case the reporting delay adjustment was run together with the imputations the graphs will display the observed trend, the trends after imputations and the 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.

	Factor	HR	HR lower CI	HR upper CI	Sex	CI Sex	p	P-value	Probability
	Sex(Diagnosis)	1.07394	0.84771	1.34767	1.00074	0.940775	0.02280	4.41e-017	0.00000
	Sex(Diagnosis) (REF: MALE) vs (FEMALE)	0.810374	1.04002	0.62910	1.02487	-0.250215	0.00146	1.30e-019	0.174699

Based on these p-values the migration status is not an important predictor of reporting delay and could be dropped, but the year of diagnosis is.

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.

Actions to be taken

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

7. Reports

7.1. Creating report

Description

In this tab a pre-defined report with main findings is provided. In the version 1 of the tool only the "Main report" is available. However, some parameters may be set for the report.



How-to

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

- Adjust counts of cases for reporting delay. This option is selected by default if the reporting delay adjustment was applied. It can be un-checked to produce a report on imputed data, but excluding the reporting delay correction.
- Apply plot curves smoothing. This option refers to the way the imputations are dealt with when producing the plots. If no smoothing is selected, that is by 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 the CD4 counts. If the 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-quartile range in CD4 count plot. This option affect 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 additionally provides comparison of the overall counts observed and adjusted for the reporting delay. The column "Weight not estimated" provides information on the number of cases for which it was not possible to estimate the reporting delay weight. Estimated number of yet unreported cases is also provided.

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

Reported year	Reported	Weight estimated	Weight not estimated	Estimated unreported (2010-2016)	Estimated year (2010-2016)
2000	200	100	100	100	200,000
2001	200	100	100	100	200,000
2002	200	100	100	100	200,000
2003	200	100	100	100	200,000
2004	200	100	100	100	200,000
2005	200	100	100	100	200,000
2006	200	100	100	100	200,000
2007	200	100	100	100	200,000
2008	200	100	100	100	200,000
2009	200	100	100	100	200,000
2010	200	100	100	100	200,000
2011	200	100	100	100	200,000
2012	200	100	100	100	200,000
2013	200	100	100	100	200,000
2014	200	100	100	100	200,000
2015	200	100	100	100	200,000
2016	200	100	100	100	200,000
Total	2000	1000	1000	1000	2000,000

Actions to be taken

The report may be exported

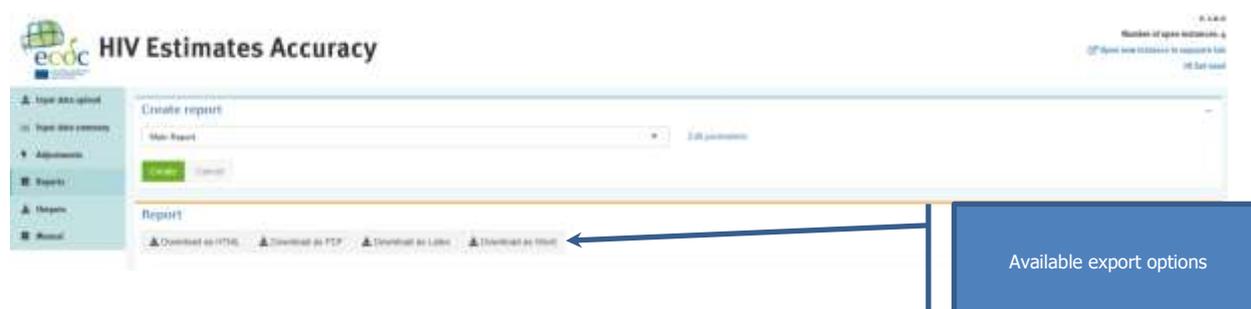
7.2. Exporting report

Description

The report may be exported to different formats: html, PDF, Word or Latex. If using the offline version, you need to have Latex installed in order to generate the PDF version of the report.

How-to

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



8. Outputs

8.1. Adjusted dataset

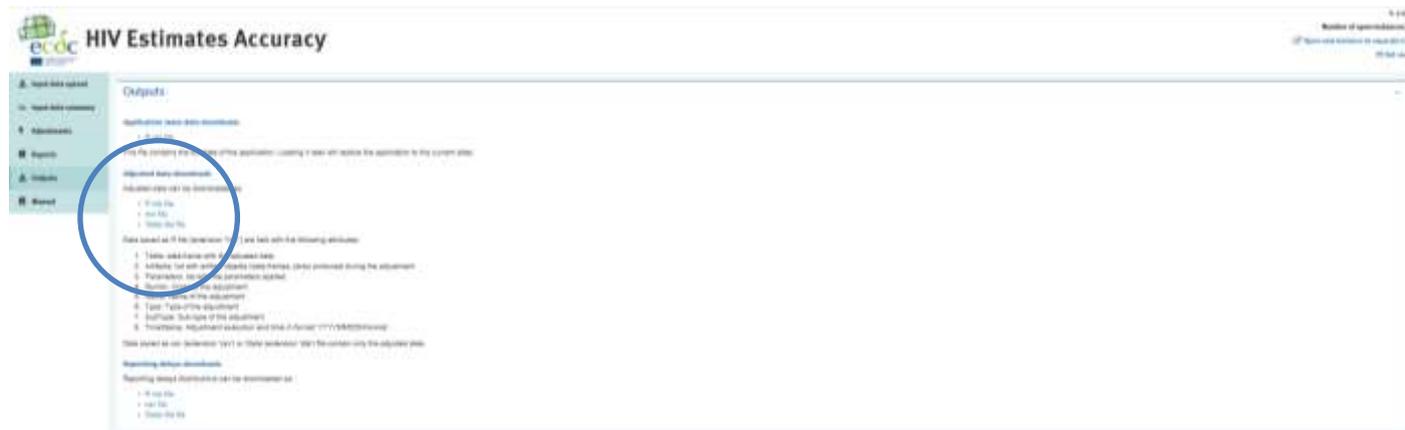
Description

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

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.

How-to

Select the desired format from the “Adjusted data downloads section”



8.2. Reporting delay weights

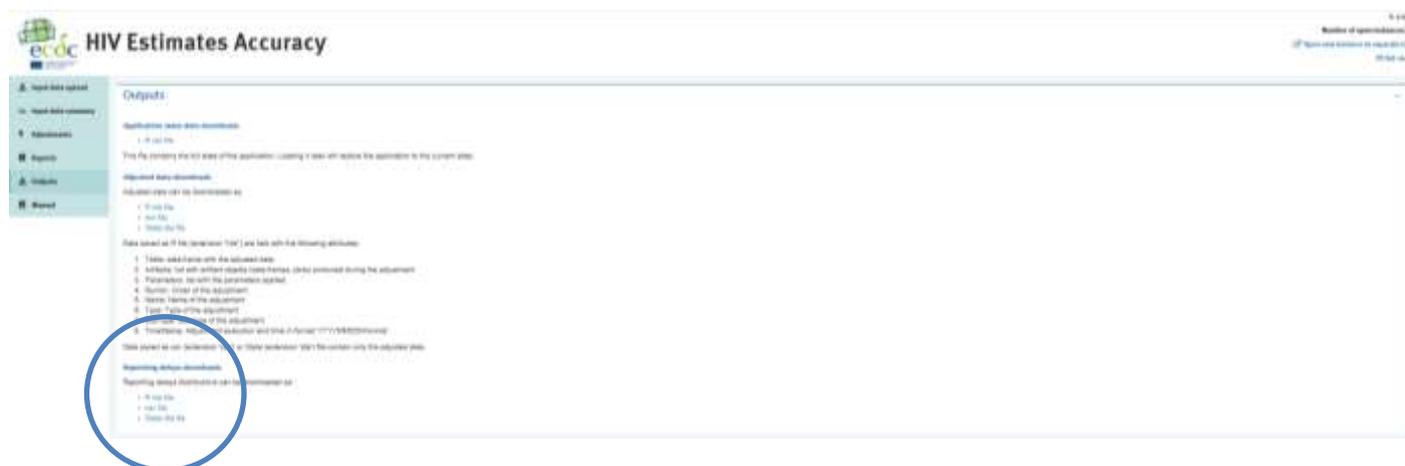
Description

The dataset 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.

How-to

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 pre-processed data and adjusted date if needed. This file can be uploaded the next time the tool is used or the a new instance of the tool in order to continue or modify the adjustments. 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.

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, N.J: Wiley; 2002. 381 p. (Wiley series in probability and statistics).

Carpenter JR, Kenward MG. *Missing data in randomised controlled trials — a practical guide* [Internet]. Birmingham: National Institute for Health Research; (Publication RM03/JH17/MK). Available from: <http://www.missingdata.org.uk>

Schafer JL. *Analysis of incomplete multivariate data*. 1. ed., 1. CRC Press reprint. Boca Raton: Chapman & Hall/CRC; 2000. 430 p. (Monographs on statistics and applied probability).

Quartagno M, Carpenter JR. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Stat Med.* 2016;35(17):2938-54. doi: 10.1002/sim.6837.

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: S. JOLANI ET AL. *Stat Med.* 2015;34(11):1841–63.

Quartagno M, Carpenter J. (2018). *Jomo: A package for Multilevel Joint Modelling Multiple Imputation*. <https://CRAN.R-project.org/package=jomo>

van Buuren S., Groothuis-Oudshoorn K. (2011). *mice: Multivariate Imputation by Chained Equations in R*. *Journal of Statistical Software*, 45(3), 1-67. URL <http://www.jstatsoft.org/v45/i03/>.

Lumley T. (2014). *Mitools: Tools to perform analyses and combine results from multiple-imputation datasets*. <https://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 Rev Can 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;75(3):515–23.

Kalbfleisch JD, Lawless JF. Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Stat Sin.* 1991;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

	<i>Name</i>	<i>Code</i>	<i>FormalName</i>	<i>RegionOrigin</i>
1	Taiwan	TW	Republic of China	EASTASIAPAC
2	Afghanistan	AF	the Islamic Republic of Afghanistan	SOUTHASIA
3	Albania	AL	the Republic of Albania	CENTEUR
4	Algeria	DZ	the People's Democratic Republic of Algeria	NORTHAFRMIDEAST
5	American Samoa	AS	NA	EASTASIAPAC
6	Andorra	AD	the Principality of Andorra	WESTEUR
7	Angola	AO	the 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	the Argentine Republic	LATAM
12	Armenia	AM	the Republic of Armenia	EASTEUR
13	Aruba	AW	NA	CAR
14	Australia	AU	Australia	AUSTNZ
15	Austria	AT	the Republic of Austria	WESTEUR
16	Azerbaijan	AZ	the Republic of Azerbaijan	EASTEUR
17	Bahamas	BS	the Commonwealth of the Bahamas	CAR
18	Bahrain	BH	the Kingdom of Bahrain	NORTHAFRMIDEAST
19	Bangladesh	BD	the People's Republic of Bangladesh	SOUTHASIA
20	Barbados	BB	Barbados	CAR
21	Belarus	BY	the Republic of Belarus	EASTEUR
22	Belgium	BE	the Kingdom of Belgium	WESTEUR
23	Belize	BZ	Belize	LATAM
24	Benin	BJ	the Republic of Benin	SUBAFR
25	Bermuda	BM	NA	CAR
26	Bhutan	BT	the Kingdom of Bhutan	SOUTHASIA
27	Bolivia (Plurinational State of)	BO	the 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	the Republic of Botswana	SUBAFR
31	Bouvet Island	BV	NA	CAR
32	Brazil	BR	the 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	the Republic of Bulgaria	CENTEUR
37	Burkina Faso	BF	Burkina Faso	SUBAFR
38	Burundi	BI	the Republic of Burundi	SUBAFR
39	Cabo Verde	CV	the Republic of Cabo Verde	SUBAFR
40	Cambodia	KH	the Kingdom of Cambodia	SOUTHASIA
41	Cameroon	CM	the Republic of Cameroon	SUBAFR
42	Canada	CA	Canada	NORTHAM

43	Cayman Islands	KY	NA	CAR
44	Central African Republic	CF	the Central African Republic	SUBAFR
45	Chad	TD	the Republic of Chad	SUBAFR
46	Chile	CL	the Republic of Chile	LATAM
47	China	CN	the 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	the Republic of Colombia	LATAM
53	Comoros	KM	the Union of the Comoros	SUBAFR
54	Congo	CG	the Republic of the Congo	SUBAFR
55	Cook Islands	CK	the Cook Islands	EASTASIAPAC
56	Costa Rica	CR	the Republic of Costa Rica	LATAM
57	Croatia	HR	the Republic of Croatia	CENTEUR
58	Cuba	CU	the Republic of Cuba	CAR
59	Curaçao	CW	NA	CAR
60	Cyprus	CY	the Republic of Cyprus	CENTEUR
61	Czechia	CZ	the Czech Republic	CENTEUR
62	Côte d'Ivoire	CI	the Republic of Côte d'Ivoire	SUBAFR
63	Democratic People's Republic of Korea	KP	the Democratic People's Republic of Korea	EASTASIAPAC
64	Democratic Republic of the Congo	CD	the Democratic Republic of the Congo	SUBAFR
65	Denmark	DK	the Kingdom of Denmark	WESTEUR
66	Djibouti	DJ	the Republic of Djibouti	SUBAFR
67	Dominica	DM	the Commonwealth of Dominica	CAR
68	Dominican Republic	DO	the Dominican Republic	CAR
69	Ecuador	EC	the Republic of Ecuador	LATAM
70	Egypt	EG	the Arab Republic of Egypt	NORTHAFRMIDEAST
71	El Salvador	SV	the Republic of El Salvador	LATAM
72	Equatorial Guinea	GQ	the Republic of Equatorial Guinea	SUBAFR
73	Eritrea	ER	the State of Eritrea	SUBAFR
74	Estonia	EE	the Republic of Estonia	EASTEUR
75	Ethiopia	ET	the Federal Democratic Republic of Ethiopia	SUBAFR
76	Falkland Islands (Malvinas)	FK	NA	LATAM
77	Faroe Islands	FO	NA	WESTEUR
78	Fiji	FJ	the Republic of Fiji	EASTASIAPAC
79	Finland	FI	the Republic of Finland	WESTEUR
80	France	FR	the 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	the Gabonese Republic	SUBAFR
85	Gambia	GM	the Republic of the Gambia	SUBAFR
86	Georgia	GE	Georgia	EASTEUR
87	Germany	DE	the Federal Republic of Germany	WESTEUR

88	Ghana	GH	the Republic of Ghana	SUBAFR
89	Gibraltar	GI	NA	WESTEUR
90	Greece	EL	the 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	the Republic of Guatemala	LATAM
96	Guernsey	GG	NA	WESTEUR
97	Guinea	GN	the Republic of Guinea	SUBAFR
98	Guinea-Bissau	GW	the Republic of Guinea-Bissau	SUBAFR
99	Guyana	GY	the Republic of Guyana	LATAM
100	Haiti	HT	the Republic of Haiti	CAR
101	Heard Island and McDonald Islands	HM	NA	AUSTNZ
102	Holy See	VA	the Holy See	WESTEUR
103	Honduras	HN	the Republic of Honduras	LATAM
104	Hungary	HU	Hungary	CENTEUR
105	Iceland	IS	the Republic of Iceland	WESTEUR
106	India	IN	the Republic of India	SOUTHASIA
107	Indonesia	ID	the Republic of Indonesia	SOUTHASIA
108	Iran (Islamic Republic of)	IR	the Islamic Republic of Iran	SOUTHASIA
109	Iraq	IQ	the Republic of Iraq	NORTHAFRMIDEAST
110	Ireland	IE	Ireland	WESTEUR
111	Isle of Man	IM	NA	WESTEUR
112	Israel	IL	the State of Israel	WESTEUR
113	Italy	IT	the Republic of Italy	WESTEUR
114	Jamaica	JM	Jamaica	CAR
115	Japan	JP	Japan	EASTASIAPAC
116	Jersey	JE	NA	WESTEUR
117	Jordan	JO	the Hashemite Kingdom of Jordan	NORTHAFRMIDEAST
118	Kazakhstan	KZ	the Republic of Kazakhstan	EASTEUR
119	Kenya	KE	the Republic of Kenya	SUBAFR
120	Kiribati	KI	the Republic of Kiribati	SOUTHASIA
121	Kuwait	KW	the State of Kuwait	NORTHAFRMIDEAST
122	Kyrgyzstan	KG	the Kyrgyz Republic	EASTEUR
123	Lao People's Democratic Republic	LA	the Lao People's Democratic Republic	SOUTHASIA
124	Latvia	LV	the Republic of Latvia	EASTEUR
125	Lebanon	LB	the Lebanese Republic	NORTHAFRMIDEAST
126	Lesotho	LS	the Kingdom of Lesotho	SUBAFR
127	Liberia	LR	the Republic of Liberia	SUBAFR
128	Libya	LY	Libya	NORTHAFRMIDEAST
129	Liechtenstein	LI	the Principality of Liechtenstein	WESTEUR
130	Lithuania	LT	the Republic of Lithuania	EASTEUR
131	Luxembourg	LU	the Grand Duchy of Luxembourg	WESTEUR
132	Madagascar	MG	the Republic of Madagascar	SUBAFR
133	Malawi	MW	the Republic of Malawi	SUBAFR

134	Malaysia	MY	Malaysia	SOUTHASIA
135	Maldives	MV	the Republic of Maldives	SOUTHASIA
136	Mali	ML	the Republic of Mali	SUBAFR
137	Malta	MT	the Republic of Malta	WESTEUR
138	Marshall Islands	MH	the Republic of the Marshall Islands	SOUTHASIA
139	Martinique	MQ	NA	CAR
140	Mauritania	MR	the Islamic Republic of Mauritania	SUBAFR
141	Mauritius	MU	the Republic of Mauritius	SUBAFR
142	Mayotte	YT	NA	SUBAFR
143	Mexico	MX	the United Mexican States	LATAM
144	Micronesia (Federated States of)	FM	the Federated States of Micronesia	SOUTHASIA
145	Monaco	MC	the Principality of Monaco	WESTEUR
146	Mongolia	MN	Mongolia	EASTASIAPAC
147	Montenegro	ME	Montenegro	CENTEUR
148	Montserrat	MS	NA	CAR
149	Morocco	MA	the Kingdom of Morocco	NORTHAFRMIDEAST
150	Mozambique	MZ	the Republic of Mozambique	SUBAFR
151	Myanmar	MM	the Republic of the Union of Myanmar	SOUTHASIA
152	Namibia	NA	the Republic of Namibia	SUBAFR
153	Nauru	NR	the Republic of Nauru	SOUTHASIA
154	Nepal	NP	the Federal Democratic Republic of Nepal	SOUTHASIA
155	Netherlands	NL	the Kingdom of the Netherlands	WESTEUR
156	New Caledonia	NC	NA	SOUTHASIA
157	New Zealand	NZ	New Zealand	AUSTNZ
158	Nicaragua	NI	the Republic of Nicaragua	LATAM
159	Niger	NE	the Republic of the Niger	SUBAFR
160	Nigeria	NG	the 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	the Kingdom of Norway	WESTEUR
165	Oman	OM	the Sultanate of Oman	NORTHAFRMIDEAST
166	Pakistan	PK	the Islamic Republic of Pakistan	SOUTHASIA
167	Palau	PW	the Republic of Palau	SOUTHASIA
168	Panama	PA	the Republic of Panama	LATAM
169	Papua New Guinea	PG	Independent State of Papua New Guinea	EASTASIAPAC
170	Paraguay	PY	the Republic of Paraguay	LATAM
171	Peru	PE	the Republic of Peru	LATAM
172	Philippines	PH	the Republic of the Philippines	SOUTHASIA
173	Pitcairn	PN	NA	SOUTHASIA
174	Poland	PL	the Republic of Poland	CENTEUR
175	Portugal	PT	the Portuguese Republic	WESTEUR
176	Puerto Rico	PR	NA	CAR
177	Qatar	QA	the State of Qatar	NORTHAFRMIDEAST
178	Republic of Korea	KR	the Republic of Korea	EASTASIAPAC

179	Republic of Moldova	MD	the Republic of Moldova	EASTEUR
180	Romania	RO	Romania	CENTEUR
181	Russian Federation	RU	the Russian Federation	EASTEUR
182	Rwanda	RW	the 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	the Independent State of Samoa	EASTASIAPAC
192	San Marino	SM	the Republic of San Marino	WESTEUR
193	Sao Tome and Principe	ST	the Democratic Republic of Sao Tome and Principe	SUBAFR
194	Saudi Arabia	SA	the Kingdom of Saudi Arabia	NORTHAFRMIDEAST
195	Senegal	SN	the Republic of Senegal	SUBAFR
196	Serbia	RS	the Republic of Serbia	CENTEUR
197	Seychelles	SC	the Republic of Seychelles	SUBAFR
198	Sierra Leone	SL	the Republic of Sierra Leone	SUBAFR
199	Singapore	SG	the Republic of Singapore	SOUTHASIA
200	Sint Maarten (Dutch part)	SX	NA	CAR
201	Slovakia	SK	the Slovak Republic	CENTEUR
202	Slovenia	SI	the Republic of Slovenia	CENTEUR
203	Solomon Islands	SB	Solomon Islands	EASTASIAPAC
204	Somalia	SO	the Federal Republic of Somalia	SUBAFR
205	South Africa	ZA	the Republic of South Africa	SUBAFR
206	South Georgia and the South Sandwich Islands	GS	NA	LATAM
207	South Sudan	SS	the Republic of South Sudan	NORTHAFRMIDEAST
208	Spain	ES	the Kingdom of Spain	WESTEUR
209	Sri Lanka	LK	the Democratic Socialist Republic of Sri Lanka	SOUTHASIA
210	State of Palestine	PS	State of Palestine	NORTHAFRMIDEAST
211	Sudan	SD	the Republic of the Sudan	NORTHAFRMIDEAST
212	Suriname	SR	the Republic of Suriname	LATAM
213	Svalbard and Jan Mayen Islands	SJ	NA	WESTEUR
214	Swaziland	SZ	the Kingdom of Swaziland	SUBAFR
215	Sweden	SE	the Kingdom of Sweden	WESTEUR
216	Switzerland	CH	the Swiss Confederation	WESTEUR
217	Syrian Arab Republic	SY	the Syrian Arab Republic	NORTHAFRMIDEAST
218	Tajikistan	TJ	the Republic of Tajikistan	EASTEUR
219	Thailand	TH	the Kingdom of Thailand	SOUTHASIA
220	The former Yugoslav Republic of Macedonia	MK	the former Yugoslav Republic of Macedonia	CENTEUR
221	Timor-Leste	TL	the Democratic Republic of Timor-Leste	SOUTHASIA
222	Togo	TG	the Togolese Republic	SUBAFR
223	Tokelau	TK	NA	EASTASIAPAC

224	Tonga	TO	the Kingdom of Tonga	EASTASIAPAC
225	Trinidad and Tobago	TT	the Republic of Trinidad and Tobago	CAR
226	Tunisia	TN	the Republic of Tunisia	NORTHAFRMIDEAST
227	Turkey	TR	the 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	the Republic of Uganda	SUBAFR
232	Ukraine	UA	Ukraine	EASTEUR
233	United Arab Emirates	AE	the United Arab Emirates	NORTHAFRMIDEAST
234	United Kingdom of Great Britain and Northern Ireland	UK	the United Kingdom of Great Britain and Northern Ireland	WESTEUR
235	United Republic of Tanzania	TZ	the 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	the United States of America	NORTHAM
239	Uruguay	UY	the Eastern Republic of Uruguay	LATAM
240	Uzbekistan	UZ	the Republic of Uzbekistan	EASTEUR
241	Vanuatu	VU	the Republic of Vanuatu	EASTASIAPAC
242	Venezuela (Bolivarian Republic of)	VE	the Bolivarian Republic of Venezuela	LATAM
243	Viet Nam	VN	the Socialist Republic of Viet Nam	SOUTHASIA
244	Wallis and Futuna Islands	WF	NA	EASTASIAPAC
245	Western Sahara	EH	NA	NORTHAFRMIDEAST
246	Yemen	YE	the Republic of Yemen	NORTHAFRMIDEAST
247	Zambia	ZM	the Republic of Zambia	SUBAFR
248	Zimbabwe	ZW	the Republic of Zimbabwe	SUBAFR
249	Åland Islands	AX	NA	WESTEUR