ECDC TECHNICAL REPORT

Documentation for use of HIV Modelling Platform

https://shinyapps.ecdc.europa.eu/shiny/hivPlatform/

ecdc
EUROPEAN CENTRE FOR
DISEASE PREVENTION
AND CONTROL

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 Kapodistran University of Athens and Stichting Hiv Monitoring in Amsterdam in 2018 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), Ard van Sighem (Stichting Hiv Monitoring)

Tool developer: Daniel Lewandowski (NextPage Software)

ECDC project manager: Chantal Quinten, Anastasia Pharris

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

Suggested citation: European Centre for Disease Prevention and Control. Documentation for use of the HIV Estimates Accuracy Tool. Stockholm: ECDC; 2019.

Stockholm, January 2023

PDF
ISBN 978-92-9193-900-8
doi: 10.2900/059144
Catalogue number TQ-AP-16-001-EN-N

Contents

Figures

Tables

1 Introduction

The HIV Modelling Platform is an R-based application for both online and stand-alone use that combines features of two previously published ECDC applications: HIV Modelling Tool and HIV Estimates Accuracy Tool. The application aims to facilitate the process of analysing the HIV surveillance data taking into account issues of missing data and reporting delay as well as obtaining modelling estimates of HIV incidence, undiagnosed fraction and time to diagnosis. While it does not replace the knowledge of data analysis with adjustments or modelling, it is intended for routine application in surveillance as no complex programming skills are needed.

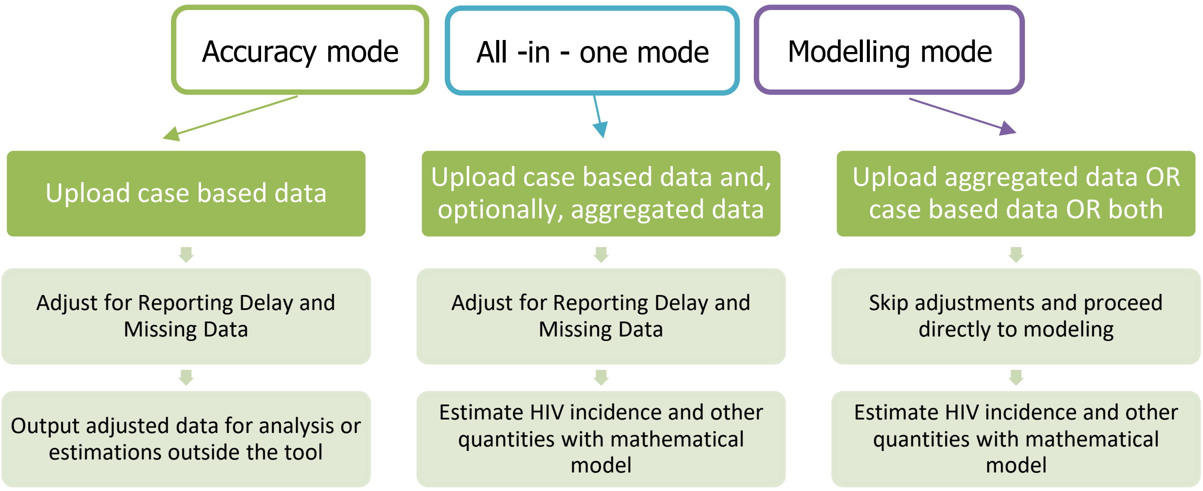
The important change with respect to the HIV Modelling Tool is that the HIV Modelling Platform accepts case-based surveillance data for HIV prepared in the format specified for the European Surveillance System (TESSy) uploads in RecordType=HIV or HIVAIDS. Case-based surveillance data containing the required set of variables and consistent with the TESSy format in coding of variables may also be used. For modelling purposes, the aggregated datasets containing annual numbers of HIV diagnoses by CD4 count and AIDS at diagnosis, AIDS cases and deaths are also accepted.

The tool can be used in three ways:

1) Accuracy, in which the functionalities of HIV Estimates Accuracy Tool are available
2) Modelling, in the functionalities of HIV Modelling Tool for estimation method “Incidence” are available with some modifications
3) All-in-One, in which data quality adjustments may be used to perform modelling

The workflow and the datasets requirements are slightly different depending on the intended application, as depicted in the Figure 1. Please consult `Prerequisites’ section about the details on the input data.

Figure 1. Overview of workflows possible in the HIV Modelling Platform


Accuracy mode
All -in - one mode
Modelling mode
Upload case based data
Upload case based data and, optionally, aggregated data
Upload aggregated data OR case based data OR both
Adjust for Reporting Delay and Missing Data
Adjust for Reporting Delay and Missing Data
Skip adjustments and proceed directly to modeling
Output adjusted data for analysis or estimations outside the tool
Estimate HIV incidence and other quantities with mathematical model
Estimate HIV incidence and other quantities with mathematical model

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 modelling part relies on deterministic progression mathematical model, which describes natural history of the disease from the time of infection to diagnosis or development of AIDS in the absence of antiretroviral treatment.

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

This document guides through the HIV Modelling Platform, 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.

1.4 How this document works

This document 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 documentation. 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 documentation cannot be interpreted and used for any reliable inferences.

2 Prerequisites

2.1 Deployment options

Three options for deployment are possible, including online version (does not require installation on the local computer), stand-alone desktop application (all necessary dependencies are included in the installation package) and R-package that can be run directly from R application on the user's computer.

In all cases the application interface will be displayed in the default internet browser.

For the online version only an active internet connection is required. It is advised to use relatively recent versions of web browsers such as Chrome, Edge, Firefox, Safari with support for JavaScript enabled.

The desktop application and R-package versions are similar in terms of required and optional software. However, the desktop application includes the required software embedded within the deployment package. The only missing piece of software is Latex. R package version requires a working installation of R and pandoc:

Required software:

a) R engine - https://www.r-project.org/ - performs all calculations

b) Pandoc - https://pandoc.org/ - converts Markdown documents (source of reports) to html, latex, Word. It is delivered with RStudio, so if one is using RStudio already, then there is no need to install it separately.

Optional software:

c) Latex - various alternatives exist: TinyTex (https://yihui.name/tinytex/), Miktex (https://miktex.org/), TexLive (https://www.tug.org/texlive/) - generates pdf reports. If it is not installed, then outputting main report to pdf will fail.

2.3.1. Online version

the HIV Platform is freely available as an online tool, that can be accessed through Shinyapps at: https://shinyapps.ecdc.europa.eu/shiny/hivPlatform/

No installation is necessary on the user's computer.

2.3.1. Desktop application

Deployment package includes all required software and R packages. Simply follow the steps:

a) Download the deployment package from here: https://www.nextpagesoft.net/hivPlatform/windows/ (195 MB download size)
b) Unpack the file to an arbitrary folder
c) After unpacking a new folder will appear called "hivPlatform". Browse inside and double-click file "hivPlatform.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).

2.3.1. R packages

The repository of R packages is available here: https://www.nextpagesoft.net/hivPlatform/repo/ . Updated versions will be also posted in this repository. The current version required R 4.1.x or latter. You may need to update your R version prior to installation.

The tool can be installed using standard R commands executed in R console:

1) Type

```
install.packages(
  "hivPlatform",
  repo = "https://www.nextpagesoft.net/hivPlatform/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:

```
hivPlatform::RunApp()
```

3) Periodically, the user can update the tool with the following command:

```
update.packages(repo = “https://www.nextpagesoft.net/hivPlatform/repo")
```

Make sure to start a new R session before running the update.

2.3 Case-based dataset

Case- base dataset is necessary if adjustments are planned (“Adjustment” and “All – in – one” modes).

In the “All-in-one” mode case-based data can be uploaded and used directly in modelling without proceeding through adjustment.

Case-based data can be used alone or in combination with aggregated data for the modelling purposes and the user is able to select which data to use when specifying the model.

- File should contain case-based records of HIV diagnoses.
- There are 19 required attributes/variables by the tool to run the adjustments and modelling. 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.
- Other variables can be also present in the input data, but they will be automatically removed during data manipulation by the tool.

Table 1. Attributes in the case-based dataset used by the tool

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' in the '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.

2.3 Aggregated data

Aggregated data can be used alone or in combination with case-based data for the modelling purposes and the user is able to select which data to use when specifying the model.

Aggregated datasets are prepared in the same way as for the use in the HIV Modelling Tool, for the `Incidence' method, as outlined below.

Preparation of aggregated datasets require surveillance data over multiple years, ideally covering the duration of the HIV epidemic in a country. The modelling method will work with or without CD4 counts at the time of HIV diagnosis, although the first option is preferred. The method will also work if data on CD4 counts are only available for several years, although in such situation the use of case-based data and performing multiple imputations prior to modelling should be considered.

2.3.1. Populations

Before using the tool, the user may define populations, into which the total national or regional HIV population can be divided. Distinguishing one or more populations may be appropriate if the user expects major differences in time between infection and diagnosis between the populations. Differences in time to diagnosis may be indicated by differences in the mean or median CD4 count at the time of diagnosis. Still, the tool will also work when all HIV-positive individuals in a country are considered as one single population. In that case, however, estimates of time to diagnosis will be an average over the total population. Populations can be based, for instance, on route of transmission: men who have sex with men, heterosexual men and women, injecting drug users. Other examples of populations include native HIV-positive individuals and migrants, or people living HIV in a specific city or region. The choice of populations will depend on the nature of the HIV epidemic in each country. When defining populations, the user should also consider the population size. The outcomes of modelling are harder to interpret for smaller populations with only a few new HIV diagnoses per year. A good approach may therefore be to first consider all HIV-positive individuals as a single population and then as a next step disaggregate this single population into smaller populations.

Finally, if the aggregated data are planned to be used in combination with the case-based data, there should be a possibility to generate the same populations in the case-based data as are provided in the aggregated data (see also Section 7.1).

2.3.2. Preparation of the datasets

If only the aggregated data are provided than all datasets marked below are required. If case-based data are provided it is possible to generate all of these from case-based data, but the user may still provide aggregated data to replace some of all datasets generated from case-based data.

It is possible to run the model if there are no data on CD4 counts at the time of diagnosis, but this is not recommended.

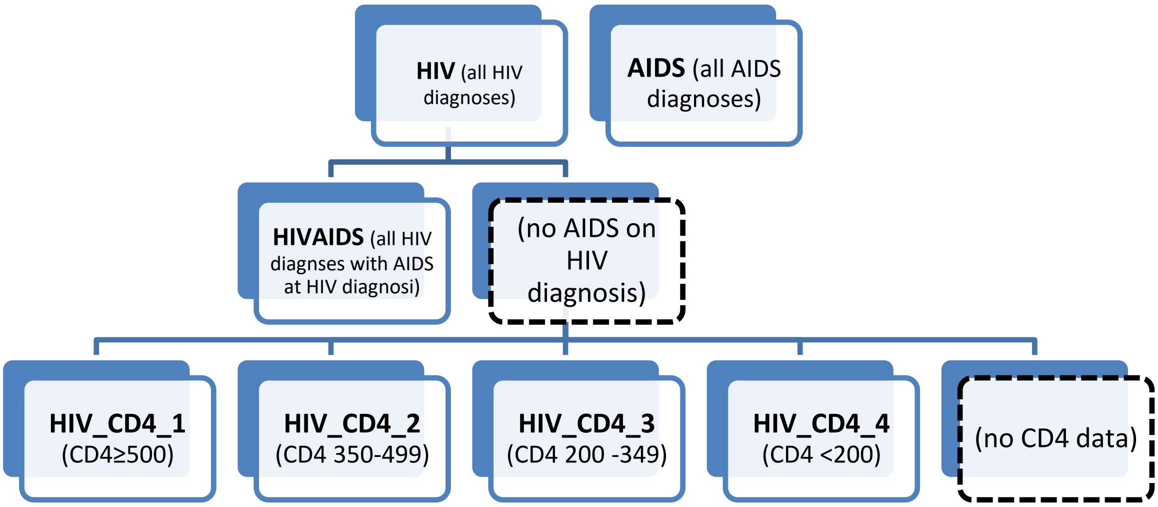
The following datasets can be uploaded:

Table 2. List of data files for aggregated data upload

Required	
Dataset	Description
HIV	total annual number of HIV diagnoses
HIVAIDS	annual number of HIV diagnoses with a concurrent AIDS diagnosis, i.e., an AIDS diagnosis within, for instance, 3 months of HIV diagnosis
AIDS	total annual number of AIDS cases; can be omitted if the calendar years in HIVAIDS cover the duration of the HIV epidemic in a country
Recommended	
Dataset	Description
HIV_CD4_1	annual number of HIV diagnoses with CD4 ≥500 cells/mm3and no concurrent AIDS diagnosis
HIV_CD4_2	annual number of HIV diagnoses with CD4 350-499 cells/mm3and no concurrent AIDS diagnosis
HIV_CD4_3	annual number of HIV diagnoses with CD4 200-349 cells/mm3and no concurrent AIDS diagnosis
HIV_CD4_4	annual number of HIV diagnoses with CD4 < 200 cells/mm3and no concurrent AIDS diagnosis
Optional	
Dataset	Description
Dead	annual number of deaths (of any cause) among diagnosed individuals

These datasets should be prepared according to the scheme below. Only the datasets in the boxes with a solid border need to be provided; data items in the boxes with a dashed border are derived by the tool informed by other data items.

Figure 2. Diagram of data stratification when preparing aggregated datasets



The dataset HIV contains all HIV diagnoses per year, i.e. all contained in HIVAIDS and HIV_CD4_1 - HIV_CD4_4 and possible diagnoses without concurrent AIDS, for which CD4 count is not known (Fig. 2).

AIDS contains all AIDS diagnoses per year, including those that are in HIVAIDS.

HIVAIDS contains all HIV/AIDS diagnoses, i.e., HIV diagnoses with a concurrent AIDS diagnosis, irrespective of the CD4 cell count at the time of diagnosis.

HIV_CD4_1to HIV_CD4_4 contain HIV diagnoses with no concurrent AIDS diagnosis and with CD4 counts at the time of diagnosis in the specified range.

2.3.3. Format of the datasets

The aggregated datasets should be prepared in CSV (comma-separated values) text format and put into a single .zip archive for upload.

The names of the populations have to be the same across all files.

The ready set of the files should be put into zip archive for uploading to the tool.

- CSV files can easily be created from packages like Microsoft Excel or SAS, or in a text editor like Notepad. Each dataset described in the table above should be saved as a separate file, using the names specified in the table. The tool automatically detects both field and decimal separators used in the data sets and can properly interpret values. Output data sets are in CSV format as well and the tool will use the field and decimal separators from the operating systems' regional settings.

- Each data set needs a header row. In a CSV file, the header row is simply a list of variable names and may look like this with comma (",") as list separator:
  - in Notepad:

```
Year, population_1, population_2, population_3
```

- in Excel:

Year	population_1	population_2	population_3

In this case the CSV file contains information on 3 groups. The names of these groups are arbitrary but should be same in all data sets that are necessary for the tool.

- The header row is followed by rows containing epidemiological data, one row for each calendar year. Each data row contains a calendar year followed by one or more numbers, for instance the number of AIDS diagnoses. Numbers do not necessarily have to be integers, which could be the case, for instance, when a correction for reporting delay is made.
  - in Notepad:

```
1982,0,0,0
1983,0,0,0
1984,2,1,3
1985,17,15,20
1987,20,23,20
..,..,..,..
2019,51.5,30,25
```

  - in Excel

1982	0	0	0
1983	0	0	0
1984	2	1	3
1985	17	15	20
1987	20	23	20
..	..	..	..
2019	51.5	30	25

In the example above, for the first population there are zero diagnoses in 1982 and 1983, 2 in 1984, 17 in 1985, and so on. For 2013, there are 50 observed diagnoses, but corrected for a reporting delay of 5% assuming for instance that 5% of the diagnoses has not yet been reported the expected number would be 51.5.

For calendar years in which the surveillance system captured a specific data item but no data (diagnoses) were observed, there should be a corresponding row with 0 diagnoses in the CSV file (as in the example above for 1982 and 1982). In contrast, for calendar years in which the surveillance system did not yet capture a specific data item, there should be no corresponding row (as in the example above for 1981 and earlier years). Putting a 0 in this case may lead to wrong results, because the tool will treat a number of 0 diagnoses the same was a any other number.

Note that the tool will assume 0 observations or diagnoses for intermediate years that are missing in the input datasets (as in the example above for 1986).

3 Using the HIV Modelling Platform

3.1 How to open the tool

When accessing online tool at https://shinyapps.ecdc.europa.eu/shiny/hivPlatform/ the user will be automatically taken to the "Welcome page" of the tool.

If using the Windows x64 desktop package go to the folder "hivPlatform" created when installing the tool. Browse inside and double-click file "hivPlatform.bat". This will open the tool in the default web browser. When done with working with it simply close the browser window.

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

```
> hivPlatform::RunApp()
```

In both cases, the tool will open as a new window in the default browser. The tool supports most of the commonly used browsers in their current versions.

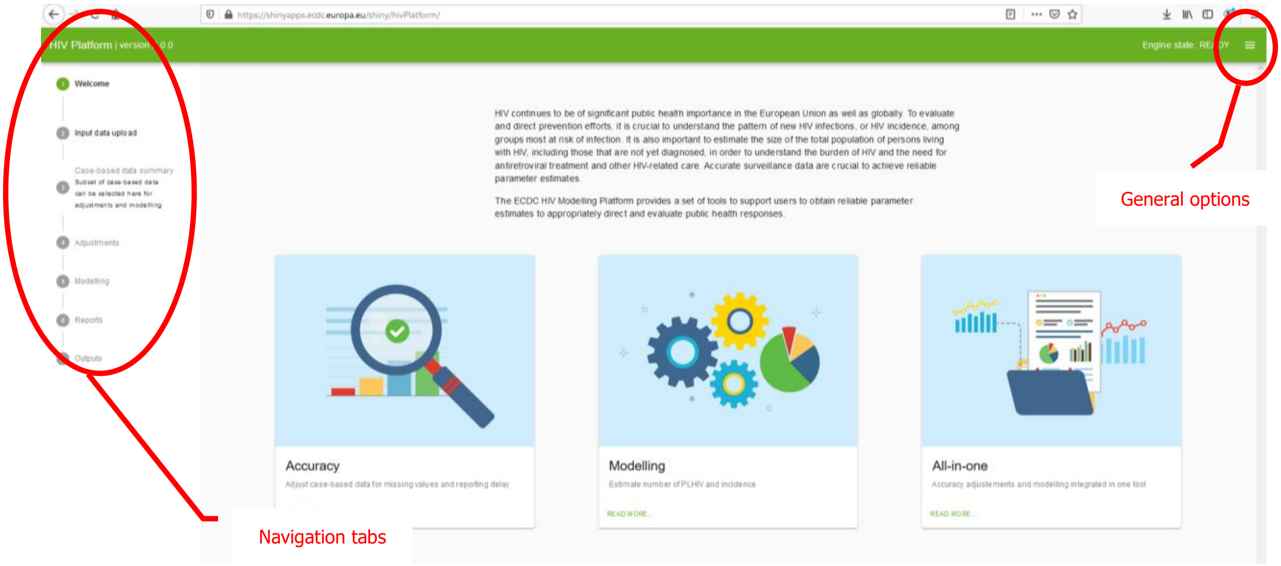
3.2 Construction of the tool

Both `General options' and `Navigation tabs' are visible at all times, although some tabs may be inactive depending on what actions were already performed.

When one of the main `Navigation tabs' is selected it expands and detailed options or steps are displayed under the main tab heading to enable navigating within this tab.

The tool is organised into tabs displayed on the left-hand panel. The tool automatically opens at the `Welcome page' that provides general information about the functions of the tool.

To use the tool the user navigates through the `Navigation tabs' on the left hand side. The order of the tabs represents the general order of the workflow and, as some steps require prior action, not all tabs are active all the time. Depending on the type of data that are uploaded some tool options may be not available and these tabs will be inactive. For example, if only aggregated data are uploaded the `Adjustments' tab will be inactive.



HIV continues to be of significant public health importance in the European Union as well as globally. To evaluate and direct prevention efforts, it is crucial to understand the pattern of new HIV infections, or HIV incidence, among groups most at risk of infection. It is also important to estimate the size of the total population of persons living with HIV, including those that are not yet diagnosed, in order to understand the burden of HIV and the need for antiretroviral treatment and other HIV-related care. Accurate surveillance data are crucial to achieve reliable parameter estimates.

The ECDC HIV Modelling Platform provides a set of tools to support users to obtain reliable parameter estimates to appropriately direct and evaluate public health responses.

https://shinyapps.ecdc.europa.eu/shiny/hivPlatform/
HIV Platform | version 0.0
Engine state: READY
Welcome
Input data upload
Case-based data summary
Subset of case-based data can be selected here for adjustments and modelling
Adjustments
Modelling
Reports
Outputs
Accuracy
Adjust case-based data for missing values and reporting delay
Modelling
Estimate number of PLHIV and incidence
READ MORE...
All-in-one
Accuracy adjustements and modelling integrated in one tool
READ MORE...
General options
Navigation tabs

Navigation Tabs:

- Input data upload. Further tabs are not active unless the data are uploaded and validated in this tab
- Case-based data summary. Allows exploration of data and selection of filters for adjustments – summarizes only the case-based data.

- Adjustments. Tab to specify adjustments and parameters for adjustments as well as to examine diagnostic charts.
- Modelling. Tab to define modelling options and parameters as well as to examine goodness of fit and estimations.
- Reports. Allows creating and exporting a predefined report.
- Outputs. Contains output datasets that can be used for further analysis, and exportable results of modelling

General options:

These options are available clicking on the `≡' symbol and include:

- Save state. Allows to save all steps that were performed up to the moment of saving in a file that can be uploaded latter.
- Load state. Allows to load the data and the steps that were performed up to the moment when the tool's state was saved
- Open new instance in separate tab. Opens the tool again in a new tab in the same browser window. The tool in the second tabs runs completely independently.
- Access Manual. Opens Manual in pdf format.
- Set seed. Advanced option, which allows to obtain the same results each time the tool is running. Otherwise, since the tool relies on procedures that involve random sampling the results could be slightly different each time the tool is run. If empty field is selected the prior seed specification will be cleared.

3.3 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, previously defined modelling parameters (including diagnosis matrices) and modelling results.

The previously saved application state can be uploaded at any time. The current application state will be overwritten by the one uploaded. If this is not intended the user should first open a new instance of the tool in the second tab and then upload the saved file there.

Process

Select the 'Load state' option form the General options menu on the top right-hand corner - `≡' symbol 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.

3.4 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 or saved application states must 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 selecting the 'Load state' option form the General options menu on the top right-hand corner - `≡' symbol.

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 and/or modelling.

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

Description

The tool uses a random number generator when pre-processing data, imputing missing values, as well as running main fit and creating bootstrap confidence intervals while modelling. This means that each time the adjustments/modelling 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). In addition, the process should be run in exactly the same steps as the seed is not reset after each of the steps.

Process

To set up the seed, fill in an integer in the "Random seed" box form the General options menu on the top right-hand corner - `≡' symbol and click "Apply". Enter an empty value to remove the fixed seed.

Interpretation

The seed is set that will be used in the further analysis. It will be applied to the first step requiring sampling, after the seed is set. The subsequent steps will be applied according to the current status of the random number generator. This implies that in order to receive the same results the user must set seed as well as perform the tasks in exactly the same order.

Further actions

Proceed with adjustments and/or modelling.

4 Input data upload

This tab is always active and allows for uploading and preprocessing data. Clicking on `Input data upload' opens automatically on the tab for uploading the case-based data. In order of switch to aggregated data format select `Aggregated data' from the `Navigation tabs' on the left-hand side.

4.1. Uploading case-based data

Description

The tool allows case-based data sets corresponding to TESSy format. Supported file 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. The maximum file size allowed is 100MB.

Process

Select the 'Upload' button in the Input data section and navigate to the location of the data file. The progress bar under the `Upload data button' shows the proportion of task done.

Interpretation

The tool provides data summary (i.e. file name and path, size and type of file, number of records, variable names) and opens the section: 'Attributes mapping'. Migrant variables regrouping is not available until the data are successfully mapped.

Further actions

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

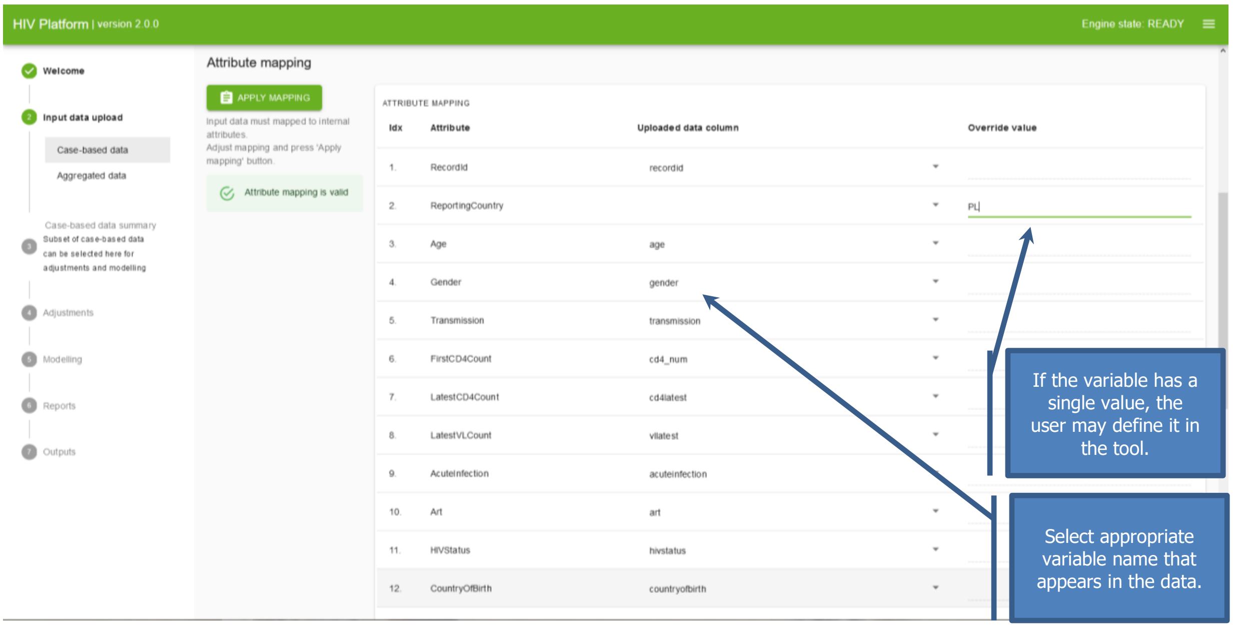
4.2 Mapping and validating case-based data

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 the '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 dataset) from the dropdown menu.



Attribute mapping

Input data must mapped to internal attributes.
Adjust mapping and press 'Apply mapping' button.

Attribute mapping is valid

ATTRIBUTE MAPPING

Idx	Attribute	Uploaded data column	Override value
1.	RecordId	recordid	
2.	ReportingCountry		PL
3.	Age	age	
4.	Gender	gender	
5.	Transmission	transmission	
6.	FirstCD4Count	cd4_num	
7.	LatestCD4Count	cd4latest	
8.	LatestVLCount	vllatest	
9.	AcuteInfection	acuteinfection	
10.	Art	art	
11.	HIVStatus	hivstatus	
12.	CountryOfBirth	countryofbirth	

If the variable has a single value, the user may define it in the tool.

Select appropriate variable name that appears in the data.

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 each covariate. A successful mapping process is indicated by the statement ‘Attributed applied correctly’.

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.

Further actions

Once the validity of mapping and values of the variables are confirmed, proceed to regrouping of the migrant variables. You may chose the default by immediately pressing “Apply regrouping” to activate adjustment and modelling tabs or create the desired categorisation of the migrant variable.

4.4 Defining the migrant variable categorisation

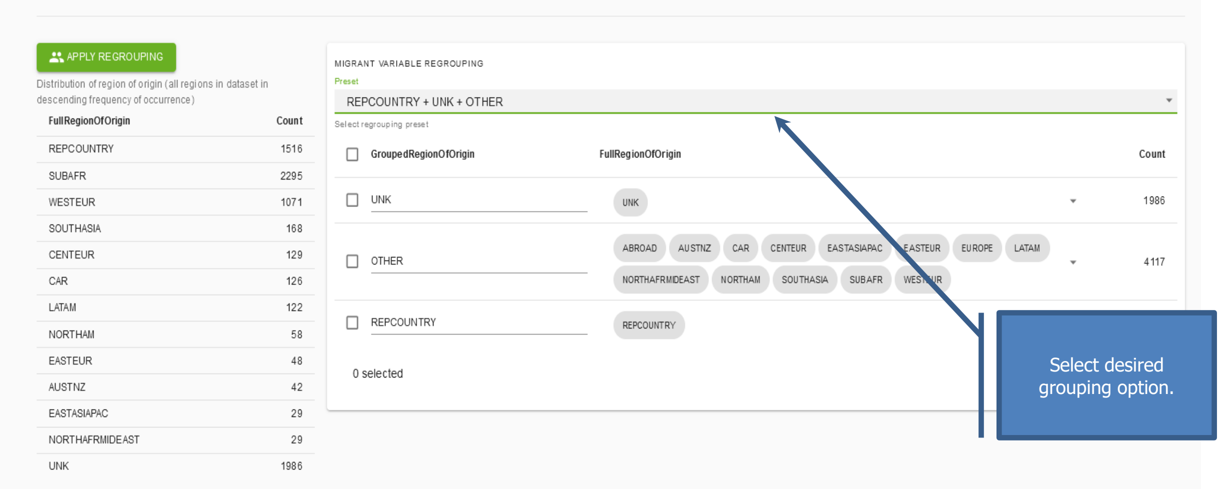
Description

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 country.

Process

The following options are available:

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



Distribution of region of origin (all regions in dataset in descending frequency of occurrence)

FullRegionOfOrigin	Count
REPCOUNTRY	1516
SUBAFR	2295
WESTEUR	1071
SOUTHASIA	168
CENTEUR	129
CAR	126
LATAM	122
NORTHAM	58
EASTEUR	48
AUSTNZ	42
EASTASIAPAC	29
NORTHAFRMIDEAST	29
UNK	1986

MIGRANT VARIABLE REGROUPING

Preset: REPCOUNTRY + UNK + OTHER

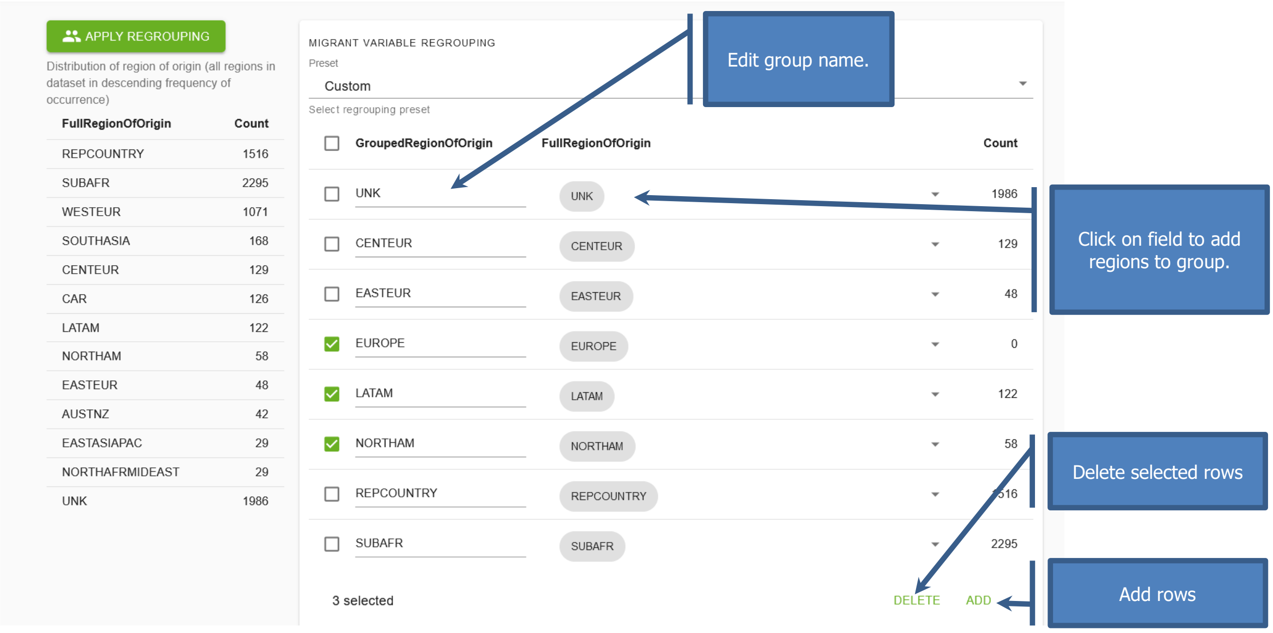
Select regrouping preset

□ GroupedRegionOfOrigin	FullRegionOfOrigin	Count
□ UNK	UNK	1986
□ OTHER	ABROAD AUSTNZ CAR CENTEUR EASTASIAPAC EASTEUR EUROPE LATAM NORTHAFRMIDEAST NORTHAM SOUTHASIA SUBAFR WESTEUR	4117
□ REPCOUNTRY	REPCOUNTRY	

0 selected

Select desired grouping option.

Custom regrouping may be created by selecting ‘Custom option’. This option is initialized to all available categories. In order to create custom group, remove the categories to be grouped with others and add them to an appropriate place. The categories can be removed by selecting them with a tick an clicking deleted at the bottom.



Distribution of region of origin (all regions in dataset in descending frequency of occurrence)

FullRegionOfOrigin	Count
REPCOUNTRY	1516
SUBAFR	2295
WESTEUR	1071
SOUTHASIA	168
CENTEUR	129
CAR	126
LATAM	122
NORTHAM	58
EASTEUR	48
AUSTNZ	42
EASTASIAPAC	29
NORTHAFRMIDEAST	29
UNK	1986

MIGRANT VARIABLE REGROUPING

Preset: Custom

Select regrouping preset

□	GroupedRegionOfOrigin	FullRegionOfOrigin	Count
□	UNK	UNK	1986
□	CENTEUR	CENTEUR	129
□	EASTEUR	EASTEUR	48
☑	EUROPE	EUROPE	0
☑	LATAM	LATAM	122
☑	NORTHAM	NORTHAM	58
□	REPCOUNTRY	REPCOUNTRY	1516
□	SUBAFR	SUBAFR	2295

3 selected

Edit group name.

Click on field to add regions to group.

Delete selected rows

Add rows

Regions may be added to the group by clicking on the 'FullRegionOfOrigin' field and adding regions from the drop-down menu. Regions are added/removed 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, click `Apply grouping' button and proceed to further tabs.

4.4 Aggregated data upload

Description

Aggregated data can be uploaded in the format described in the Section 2.3. Only CSV files are supported in ZIP archive. The set of files to be uploaded should be placed in a single ZIP archive prior to uploading. Please note that RAR archives are not supported.

Process

Select the 'Upload' button in the Input data section and navigate to the location of the data file. The progress bar under the `Upload data button' shows the proportion of task done.

Interpretation

Aggregated data can be used for the process of modelling. Combining aggregated and case-based data in modelling is allowed and this combination is specified in the modelling tab. Not all uploaded aggregated data need to be used in modelling, but they will be available.

Further actions

Check if all intended datasets are uploaded. Check if the population names are correct. Proceed to further tabs as intended.

5 Input data summary tab

This tab allows for inspecting data quality issues present in the input case-based data. Filters with an effect on adjustments can be defined. The tab is organized in three sections: filters ("Select case-based data for summary"); missing data summary and reporting delay summary.

5.1 Inspecting missing data and reporting delay patterns

Description

Section `Missing data summary: key variables' provides a summary of the missing values for: 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 `Reporting delay summary' displays the observed (unadjusted) distribution of reporting delay.

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.

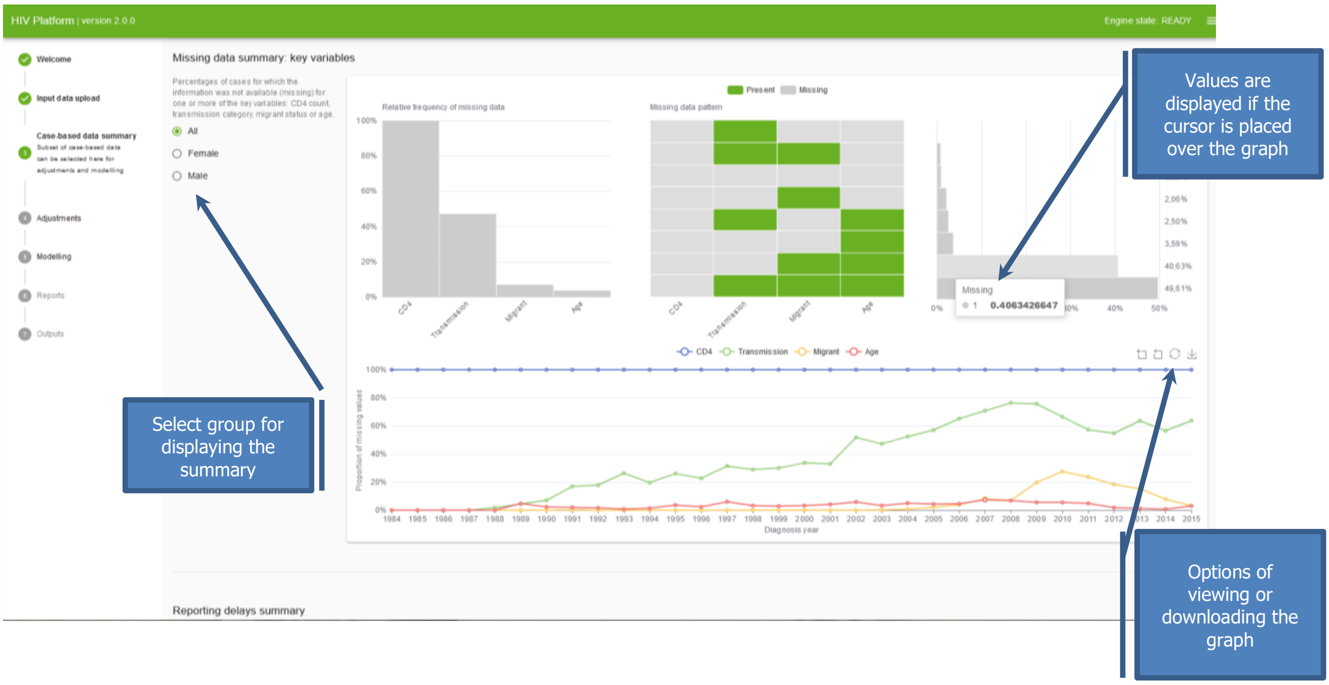
Interpretation

For all cases, as well as separately for males and females, two panels are presented in the section relating to missing values.

The top panel specifies:

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

Bottom panel – proportion missing for each of the key variables over time.



Missing data summary: key variables

Percentages of cases for which the information was not available (missing) for one or more of the key variables: CD4 count, transmission category, migrant status or age.

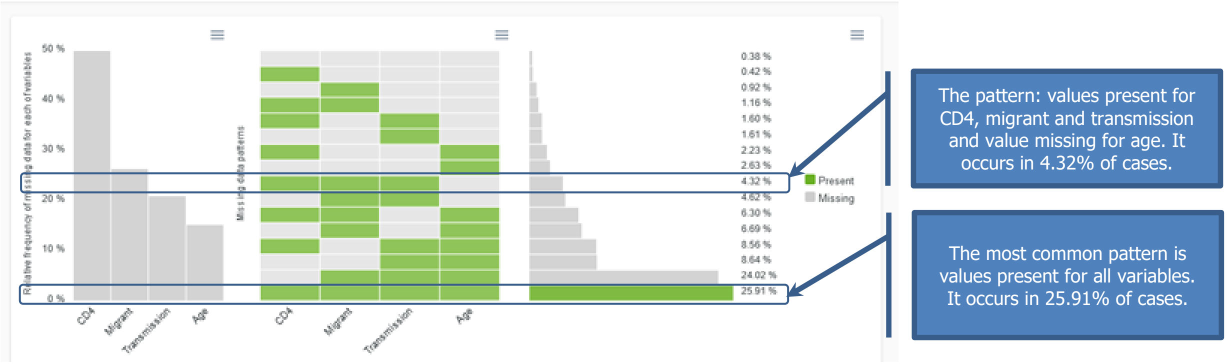
● All
○ Female
○ Male

Values are displayed if the cursor is placed over the graph

Select group for displaying the summary

Options of viewing or downloading the graph

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 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 pattern: values present for CD4, migrant and transmission and value missing for age. It occurs in 4.32% of cases.

The most common pattern is values present for all variables. It occurs in 25.91% of cases.

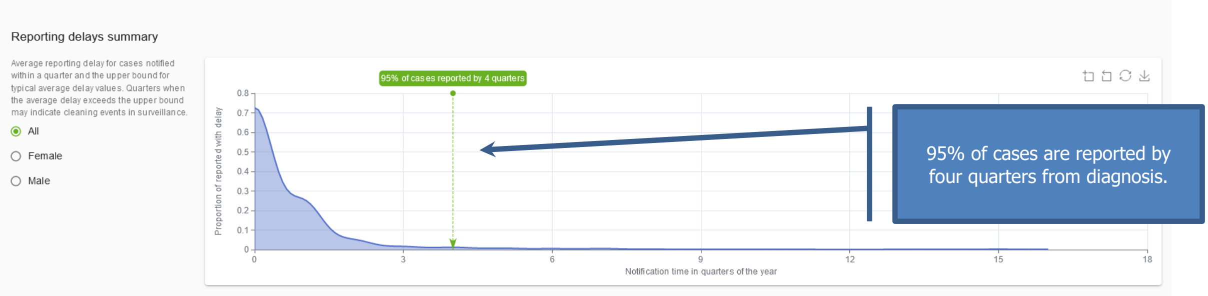
The graphs allow for checking if the reported levels of missing values are correct for the input data and may help in deciding whether the data should be restricted for further analysis.

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. The adjustment methods implemented in the tool assume non-monotone missing patterns. They are also valid (although less efficient) for the monotone missing pattern.

With respect to the bottom panel, 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.

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 four quarters, indicating moderate delays.


Reporting delays summary

Average reporting delay for cases notified within a quarter and the upper bound for typical average delay values. Quarters when the average delay exceeds the upper bound may indicate cleaning events in surveillance.

● All

○ Female

○ Male

95% of cases reported by 4 quarters
Proportion of reported with delay
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
3
6
9
12
15
18
Notification time in quarters of the year

95% of cases are reported by four quarters from diagnosis.

All graphs can be inspected in detail by using "Zoom" option at the right top corner of graphs. To zoom in click the "Zoom" option and then click and drag over the area to be enlarged. Use "Restore" button to go back to the initial graph. Graphs can be downloaded in .svg format.

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

5.2 Applying filters

Description

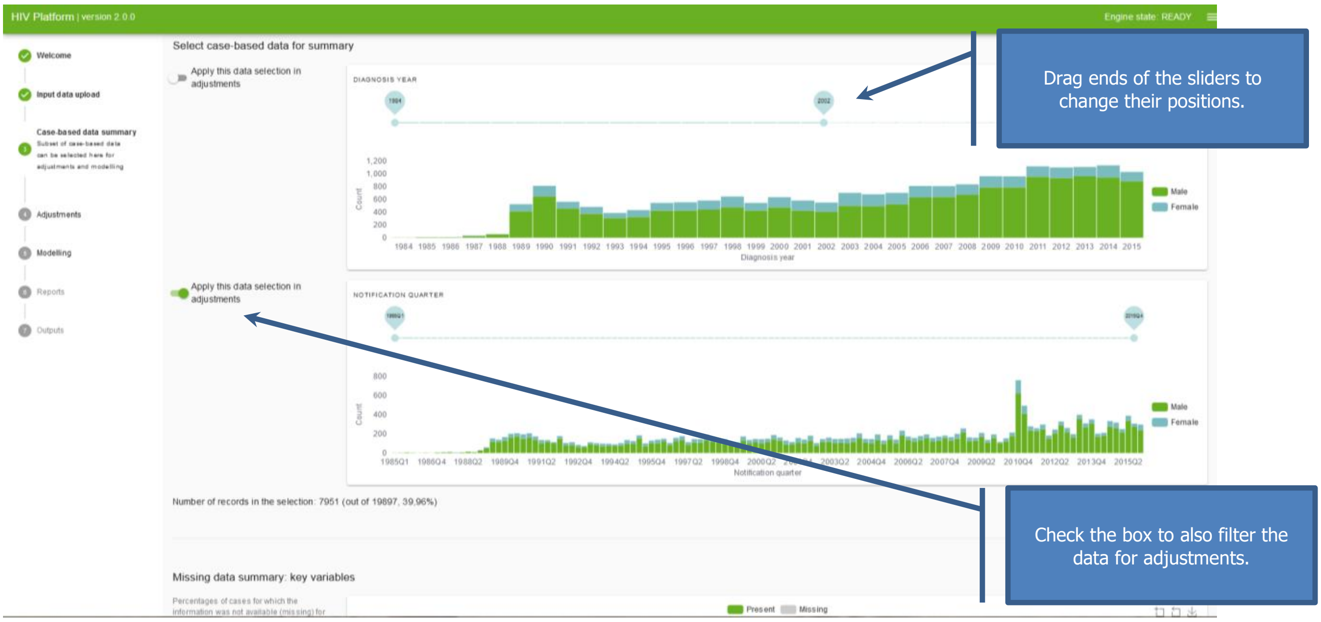
The Input data summary tab allows for applying filters on the year of diagnosis and the year and quarter of notification. These filters may be applied to inspect the data in the Input data summary tab, but can be also passed onto the adjustments. When passed onto the adjustments, the filtering will also have an effect on the output datasets.

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.



Select case-based data for summary

Apply this data selection in adjustments

Number of records in the selection: 7951 (out of 19897, 39.96%)

Missing data summary: key variables

Percentages of cases for which the information was not available (missing) for

Drag ends of the sliders to change their positions.

Check the box to also filter the data for adjustments.

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.

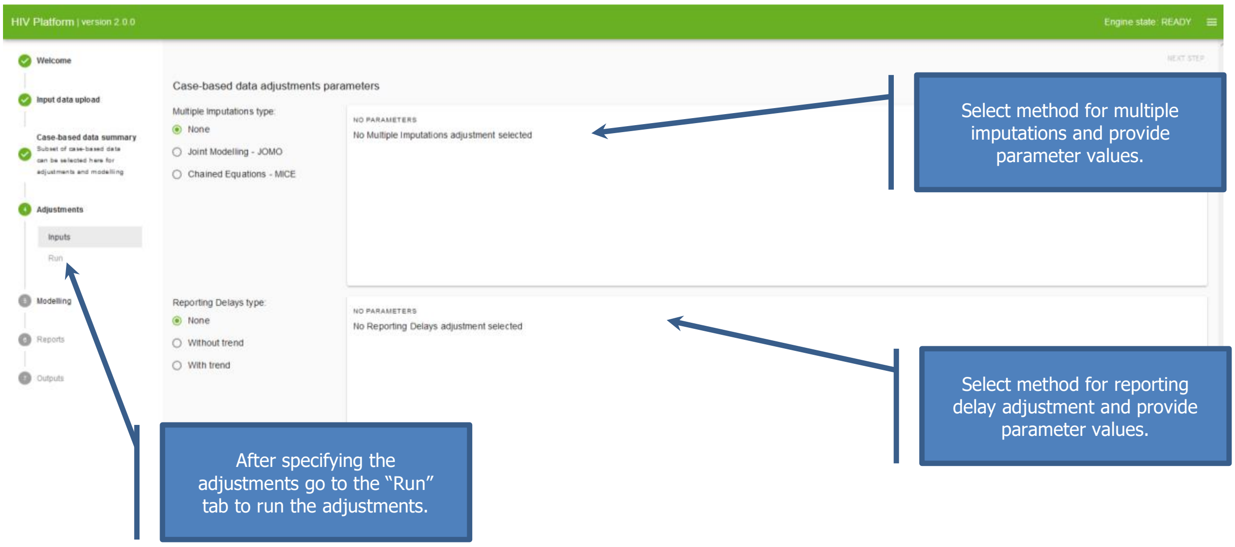
Further actions

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.



Case-based data adjustments parameters

Multiple imputations type:
● None
○ Joint Modelling - JOMO
○ Chained Equations - MICE

NO PARAMETERS
No Multiple Imputations adjustment selected

Reporting Delays type:
● None
○ Without trend
○ With trend

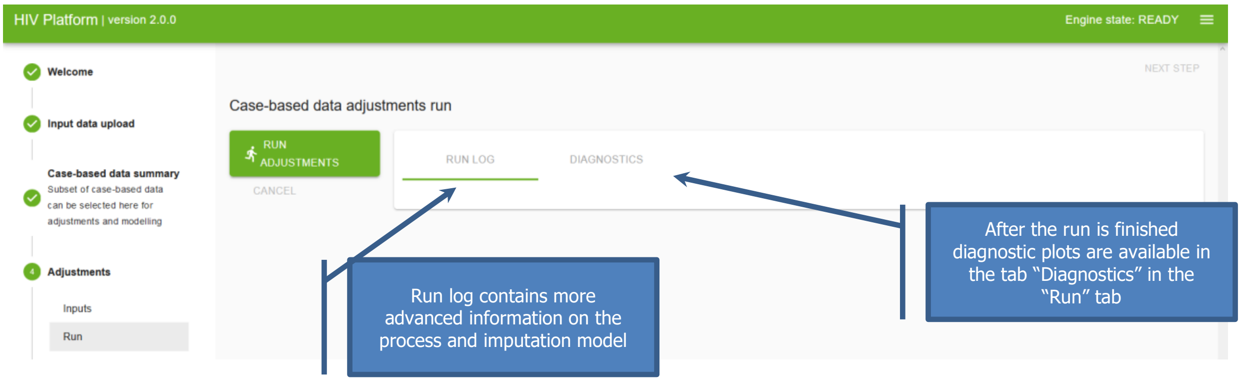
NO PARAMETERS
No Reporting Delays adjustment selected

Select method for multiple imputations and provide parameter values.

Select method for reporting delay adjustment and provide parameter values.

After specifying the adjustments go to the "Run" tab to run the adjustments.

"Run" tab is only active once at least one adjustment has been specified.



Case-based data adjustments run

Run log contains more advanced information on the process and imputation model

After the run is finished diagnostic plots are available in the tab "Diagnostics" in the "Run" tab

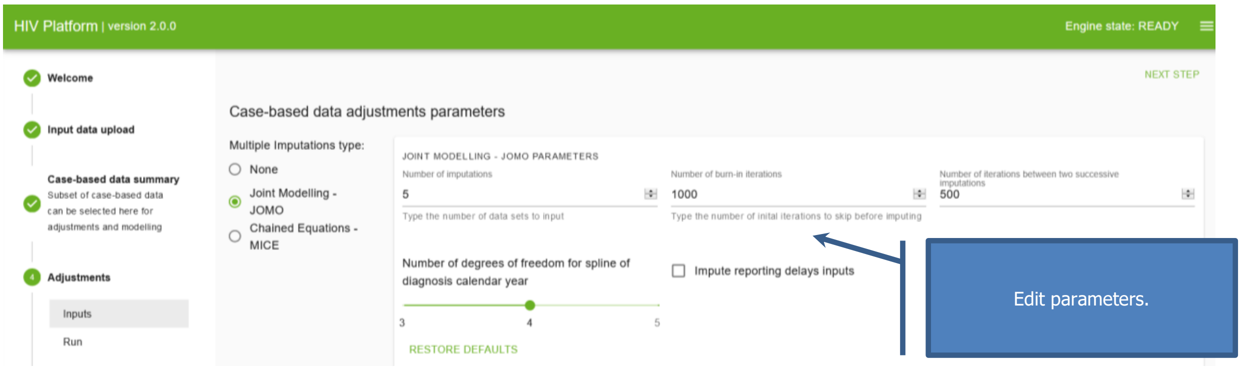
6.1 Joint Modelling Multiple Imputation

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 of imputation datasets, burn-in iterations and iterations between two consecutive datasets. Depending on the size of the datasets allow up to several hours for the final runs. Imputations are run separately by gender.

How-to

Select 'Joint Modelling Multiple Imputation' in the 'Multiple Imputations adjustment' field. The options of editable parameters will appear on the right.



Case-based data adjustments parameters

Multiple Imputations type:

○ None
● Joint Modelling - JOMO
○ Chained Equations - MICE

JOINT MODELLING - JOMO PARAMETERS

Number of imputations: 5
Type the number of data sets to input

Number of burn-in iterations: 1000
Type the number of inital iterations to skip before imputing

Number of iterations between two successive imputations: 500

Number of degrees of freedom for spline of diagnosis calendar year: 3 – 4 – 5

□ Impute reporting delays inputs

Edit parameters.

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 "Impute reporting delays inputs".

Interpretation

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

Table 3. Parameters for Jomo imputations

Parameter	Description	What to select
Number of imputations	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	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 dataset, which should limit autocorrelation of imputed datasets	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.

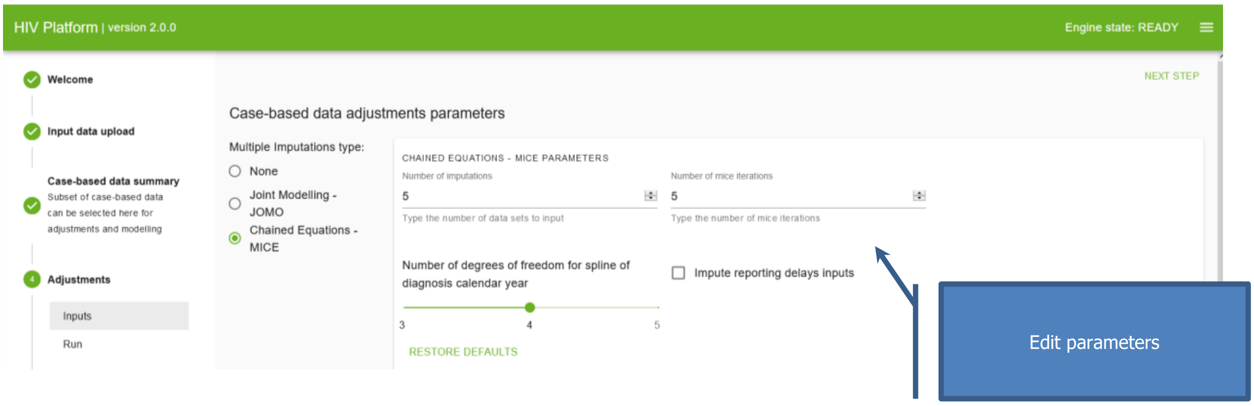
6.2 Multiple Imputation using Chained Equations (MICE)

Description

This option performs multiple imputation using chained equations. 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. Imputations are run separately by gender.

Process

Select ‘Multiple Imputation using Chained Equations’ in the ‘Multiple Imputations adjustment’ field. It is possible to edit parameters on the right.



Case-based data adjustments parameters

Multiple Imputations type:

○ None
○ Joint Modelling - JOMO
● Chained Equations - MICE

CHAINED EQUATIONS - MICE PARAMETERS

Number of imputations: 5 — Type the number of data sets to input

Number of mice iterations: 5 — Type the number of mice iterations

Number of degrees of freedom for spline of diagnosis calendar year: 4 (3–5)

□ Impute reporting delays inputs

Edit parameters

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 “Impute reporting delays inputs”.

Interpretation

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

Table 4. Parameters for MICE imputations

Parameter	Description	What to select
Number of imputations	Number of imputed datasets 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.

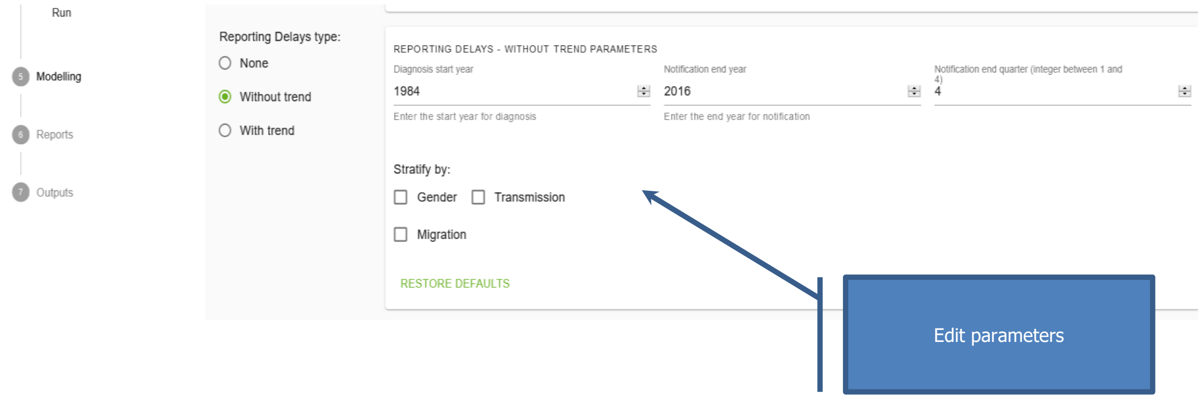
6.3 Simple reporting delay

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 without trend’ from the ‘Reporting delay type’ field. The option to edit parameters will appear on the right.



Reporting Delays type:

○ None

● Without trend

○ With trend

REPORTING DELAYS - WITHOUT TREND PARAMETERS

Diagnosis start year: 1984 (Enter the start year for diagnosis)

Notification end year: 2016 (Enter the end year for notification)

Notification end quarter (integer between 1 and 4): 4

Stratify by:

□ Gender □ Transmission

□ Migration

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.

Table 5. Parameters for reporting delay adjustment

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 using only the later data. CAUTION: if delays are long this may cause underestimation of the number of cases.
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. Migration refers to the regrouped migration status as in Section 4.4.	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.

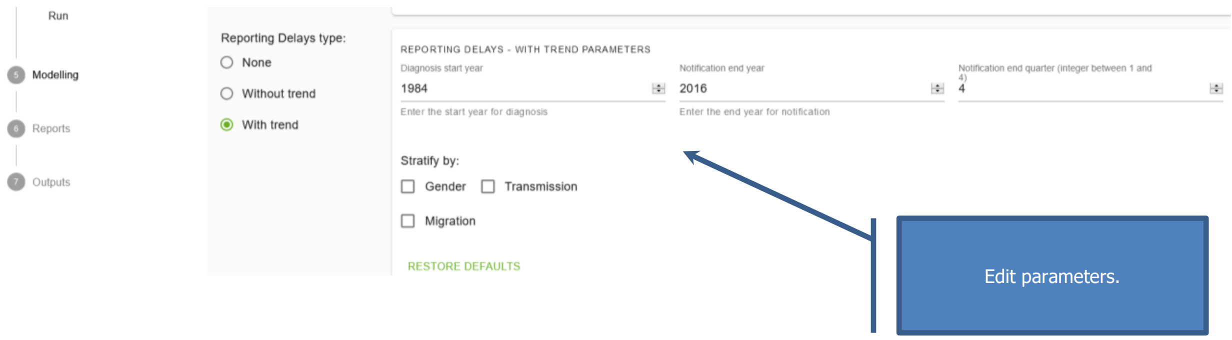
6.4 Reporting delay with trend

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 ‘Reporting delay type’ field. The option to edit parameters will appear on the right.



Reporting Delays type:
○ None
○ Without trend
● With trend

REPORTING DELAYS - WITH TREND PARAMETERS

Diagnosis start year: 1984 (Enter the start year for diagnosis)

Notification end year: 2016 (Enter the end year for notification)

Notification end quarter (integer between 1 and 4): 4

Stratify by:
□ Gender □ Transmission
□ Migration

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 using only the later data. CAUTION: if delays are long this may cause underestimation of the number of cases.
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 “Run” tab will be populated with the results of the imputation model (“Log” tab) and the imputation diagnostics (“Diagnostics” tab).

The “Log” tab provides information on what models were used, with what parameters and the estimated parameters of the imputation models. Should the estimations fail, indication why it could have happened will be displayed. These results are intended for more advanced users.

The “Diagnostics” tab is organised in 3 section related to convergence of and autocorrelation in the imputation procedure, as well as comparison of distribution of imputed vs observed values. Examples and interpretation are provided below.

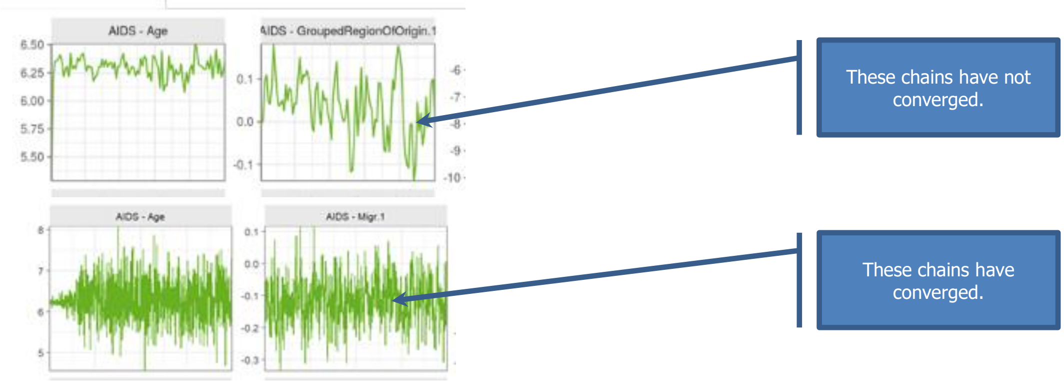
Process

Intermediate outputs are generated automatically when running adjustments.

Interpretation

The output related to the **convergence** contains trace plots for all covariates.

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’.

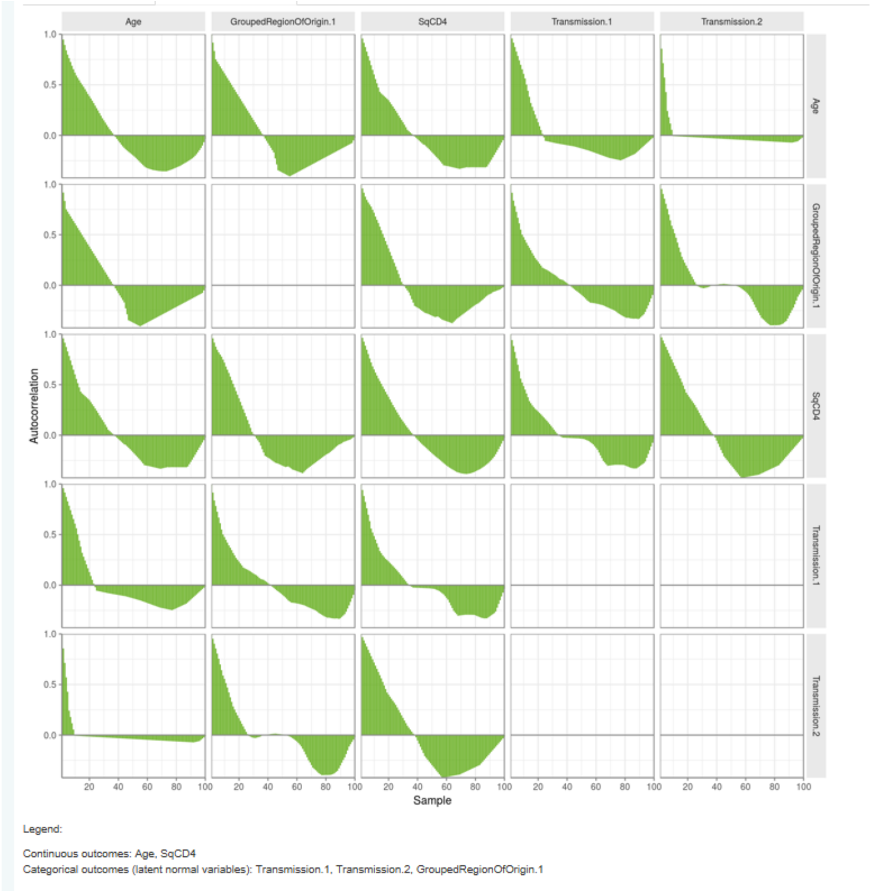


These chains have not converged.

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 autocorrelation varies from -1 to 1. The aim is that the autocorrelation should be insignificant (near the 0 line).

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.

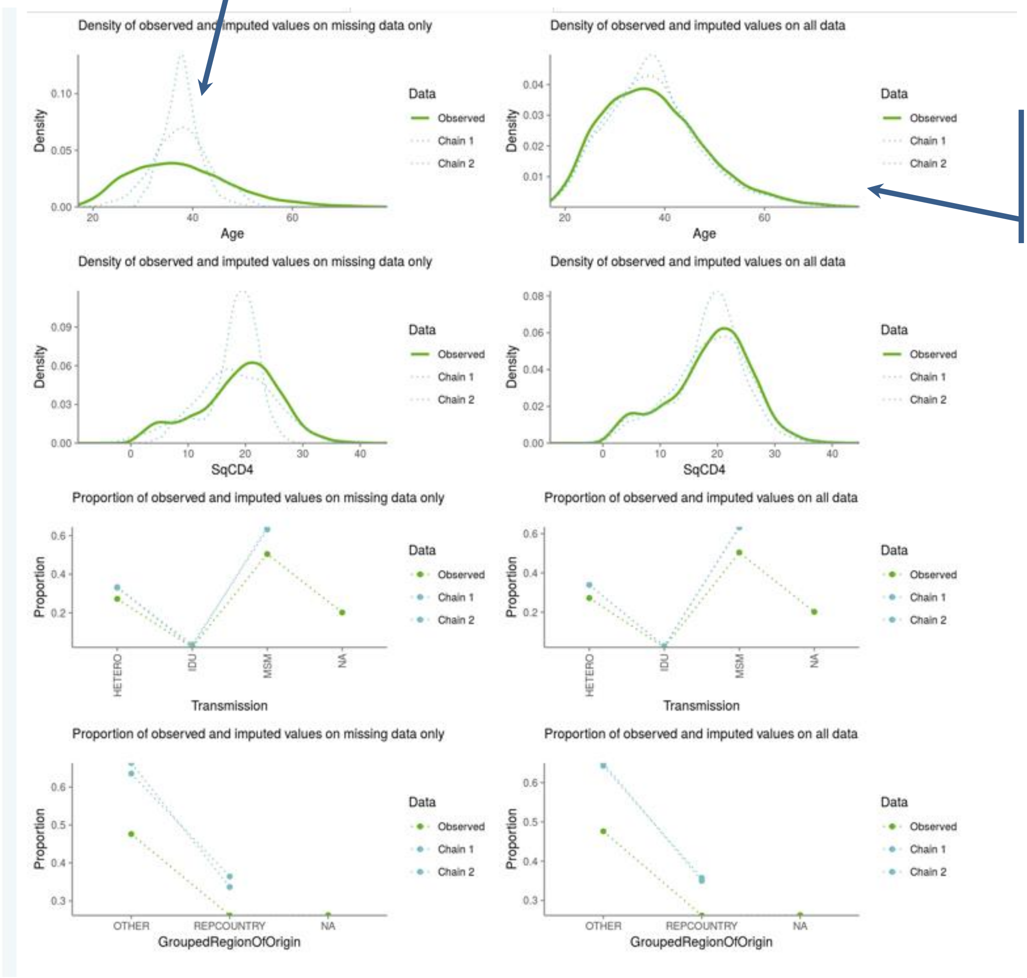


Legend:

Continuous outcomes: Age, SqCD4
Categorical outcomes (latent normal variables): Transmission.1, Transmission.2, GroupedRegionOfOrigin.1

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.

Imputed values of age are more likely to focus around 40


Density of observed and imputed values on missing data only

Density of observed and imputed values on all data

Density of observed and imputed values on missing data only

Density of observed and imputed values on all data

Proportion of observed and imputed values on missing data only

Proportion of observed and imputed values on all data

Proportion of observed and imputed values on missing data only

Proportion of observed and imputed values on all data

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 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 significant autocorrelation is detected after achieving convergence.

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 – MICE

Description

the "Run" tab will be populated with the results of the imputation model ("Log" tab) and the imputation diagnostics ("Diagnostics" tab).

The "Log" tab provides information on what models were used, with what parameters and the estimated parameters of the imputation models. Should the estimations fail, indication why it could have happened will be displayed. These results are intended for more advanced users.

The "Diagnostics" tab is organised in 3 section related to convergence of and autocorrelation in the imputation procedure, as well as comparison of distribution of imputed vs observed values. Examples and interpretation are provided below.

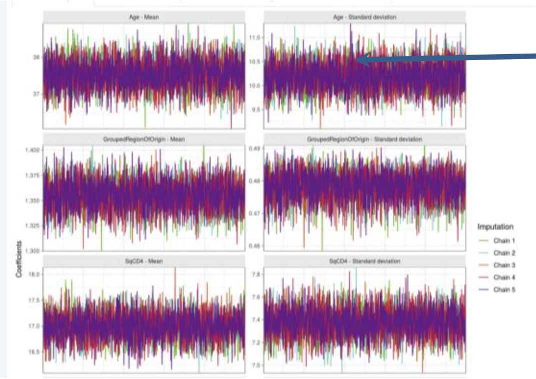
Process

Intermediate outputs are generated automatically when running the adjustments.

Interpretation

The output related to the convergence contains trace plots.

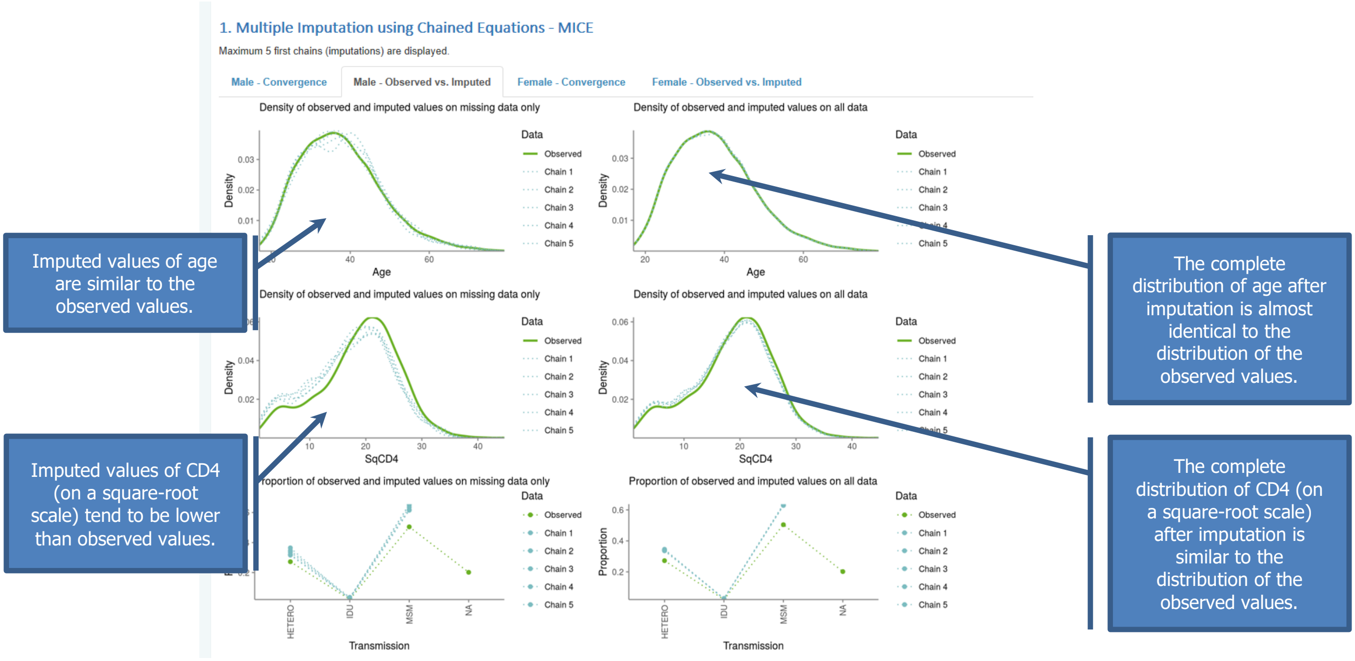
The use of the trace plots determines whether the procedure converged, assuring that the missing values are imputed from the correct distribution. In case of convergence, the trace plot for every parameter does not display any pattern. More iterations are needed in case certain parameters display ertain 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.



The plots represent a converged procedure:

- No evident trends; and
- 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 and imputed values and the other comparing the distribution of observed and all values observed and imputed.



1. Multiple Imputation using Chained Equations - MICE

Maximum 5 first chains (imputations) are displayed.

Male - Convergence
Male - Observed vs. Imputed
Female - Convergence
Female - Observed vs. Imputed
Density of observed and imputed values on missing data only
Density of observed and imputed values on all data
Density
Age
SqCD4
Proportion of observed and imputed values on missing data only
Proportion of observed and imputed values on all data
Proportion
Transmission
HETERO
IDU
MSM
NA
Data
Observed
Chain 1
Chain 2
Chain 3
Chain 4
Chain 5

Imputed values of age are similar to the observed values.

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

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

The complete distribution of CD4 (on a 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 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.

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

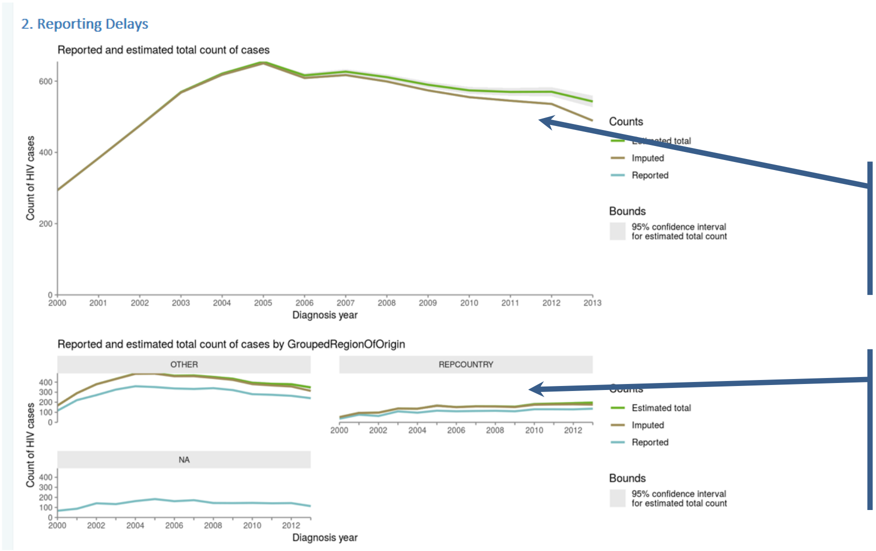
Process

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.

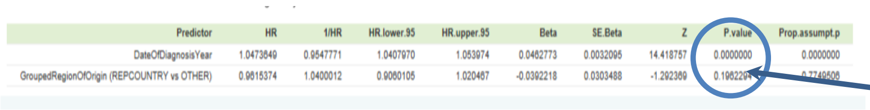
If stratification was introduced, trends by stratification variables are also displayed. In case the reporting delay adjustment was run together with the imputations, the graphs will display the observed trend, trends after imputations and trend after both imputations and the reporting delay adjustment.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



Overall plot comparing observed vs adjusted values

Stratified plots are generated for stratification variables specified in the reporting delay parameters.

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.



Predictor	HR	1/HR	HR.lower.95	HR.upper.95	Beta	SE.Beta	Z	P.value	Prop.assumpt.p
DateOfDiagnosisYear	1.0473649	0.9547771	1.0407970	1.053974	0.0462773	0.0032095	14.418757	0.0000000	0.0000000
GroupedRegionOfOrigin (REPCOUNTRY vs OTHER)	0.9615374	1.0400012	0.9060105	1.020467	-0.0392218	0.0303488	-1.292369	0.1962294	0.7749506

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.

Further actions

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

7 Modelling

This tab takes the user through the process of creating the model, including defining the populations, specifying the diagnosis matrix and model parameters the, as well as running the main fit and boostrap (to create confidence intervals).

7.1 Populations: creating populations from case-based data

Description

The tool can use the case-based data to automatically create input for modelling. The models can be defined separately for different subgroups of cases, defined by variables present in the dataset.

The process of defining populations from case-base data is similar to the process of preparing the aggregated data by populations. The advantage is that is can be done directly in the tool.

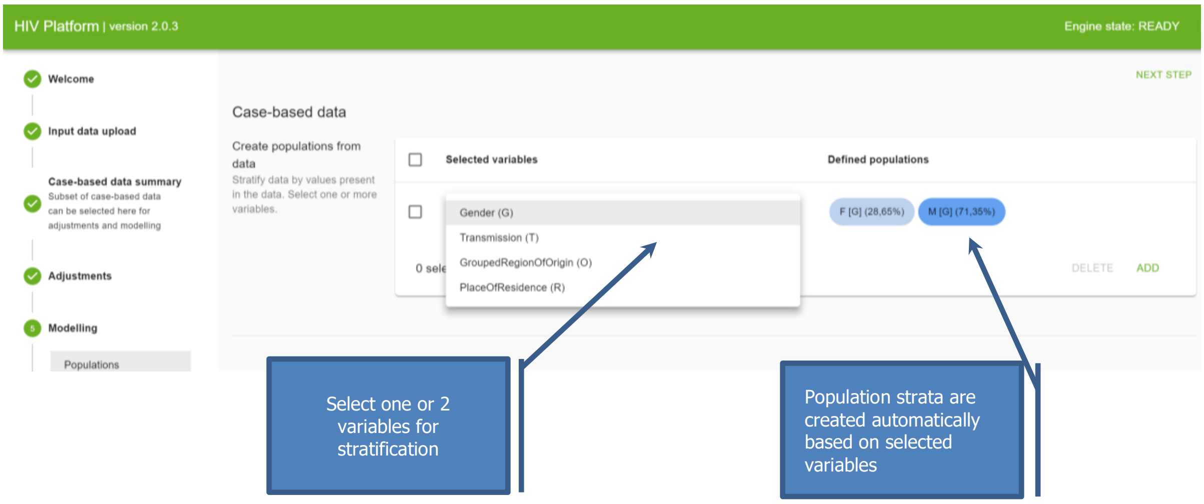
As in case of aggregated data the user should reflect whether major differences in terms of time-to-diagnosis or incidence trends by population are expected. Distinguishing one or more populations may be than appropriate. As in case of aggregate data, the tool will also work when all HIV-positive individuals in a country are considered as one single population. In that case, however, estimates of time to diagnosis will be an average over the total population and may not fully reflect important epidemiological differences.

Populations that can be created using case-based data can be based on the following variables, present in the dataset: 1) gender; 2) route of transmission: men who have sex with men, heterosexual men and women, injecting drug users; 3) migration status; 4) place of residence.

Process

The population tab is automatically adjusted to the dataset uploaded. Defining populations for case-based data is available if these data were uploaded.

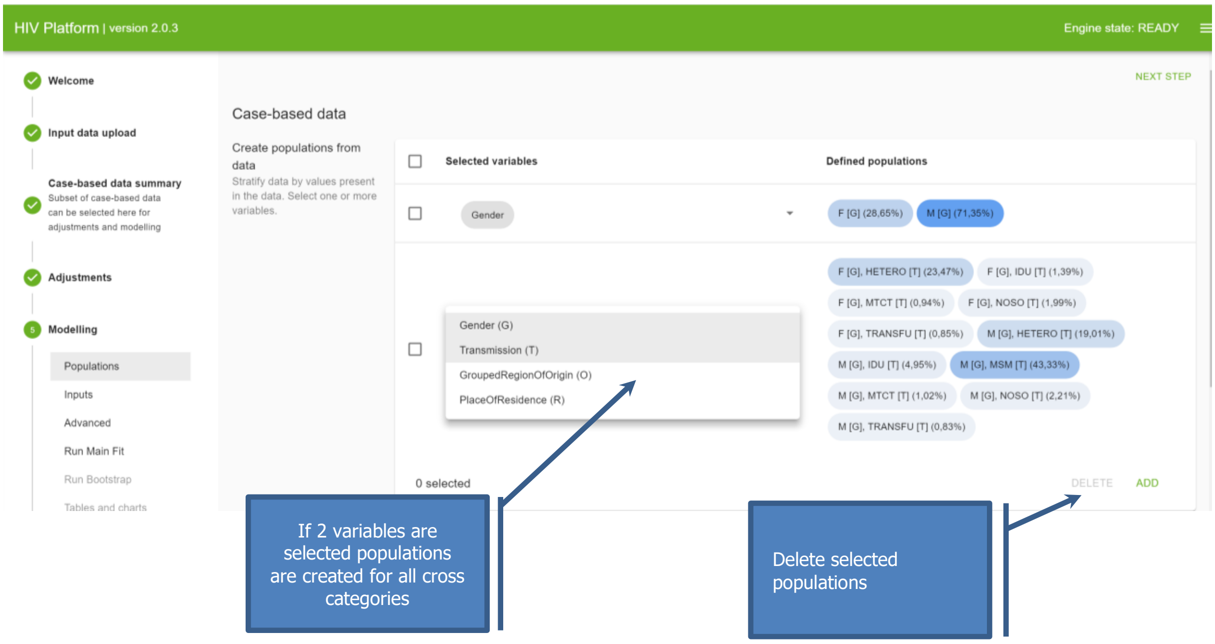
In order to define populations for case-based data select variables for stratification. Click `Add' in the `Case-based data: Create populations' panel.



Select one or 2 variables for stratification

Population strata are created automatically based on selected variables

Several different stratifications may be defined using different variables. If additional stratification is required, click `ADD' in the right lower corner of the panel. The unnecessary stratifications may be deleted.



Case-based data

Create populations from data

Stratify data by values present in the data. Select one or more variables.

□ Selected variables	Defined populations
□ Gender	F [G] (28,65%) M [G] (71,35%)
□ Gender (G) Transmission (T) GroupedRegionOfOrigin (O) PlaceOfResidence (R)	F [G], HETERO [T] (23,47%) F [G], IDU [T] (1,39%) F [G], MTCT [T] (0,94%) F [G], NOSO [T] (1,99%) F [G], TRANSFU [T] (0,85%) M [G], HETERO [T] (19,01%) M [G], IDU [T] (4,95%) M [G], MSM [T] (43,33%) M [G], MTCT [T] (1,02%) M [G], NOSO [T] (2,21%) M [G], TRANSFU [T] (0,83%)

0 selected

DELETE ADD

If 2 variables are selected populations are created for all cross categories

Delete selected populations

Interpretation

Populations are defined as strata with respect to single variable or 2 variables. For e.g. selecting gender the user can create populations: males and females, while selecting transmission category: MSM, PWID, Hetero, Other. When both gender and transmission category are selected the following populations will be created: MSM, PWID – males; PWID – females, Hetero – males, Hetero – females, MtC – males, MtC – females, Other – males, Other – females.

If a variable contains missing values (e.g. unknown transmission) – this will become a separate population. If missing values adjustment is run beforehand the unknown category will not appear even if present in the original data. The populations for modelling will be created based on the imputed dataset.

Not all groups created by this stratification process have to be used in modelling- see `Creating combinations' section.

The proportion of all cases in the specific population defined are displayed next to the created subpopulation. Keep in mind that modelling is often impossible and /or inaccurate for populations with small numbers.

Further actions

If all desired populations are created the user can pass to further sections. The user can come back to this section and update the definitions at a later stage if needed. Review the aggregated data section to determine if appropriate aggregated datasets are switched on. The case-based populations and populations available in the switched-on aggregated datasets must be aligned.

7.2 Populations: selecting aggregate data

Description

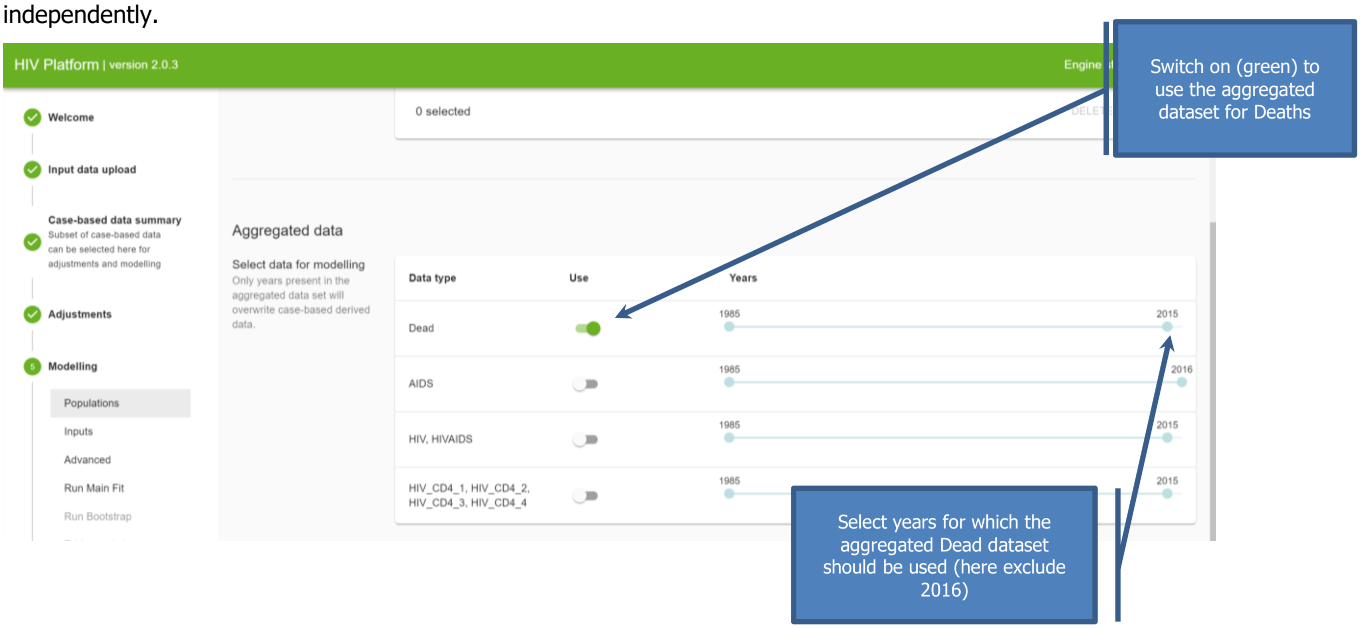
The tool allows to use aggregated data for modelling either on their own or along with the case-based data. If aggregated data are uploaded, the user needs specify exactly the datasets and time periods for which the aggregated data should be used.

If both case-based and aggregated data are uploaded then the user can control the time period, for which to use the aggregated data. For the remaining time period the case-based data will be used by default.

Process

The "Aggregate data: Select data" section allows the use to control which aggregated data to use. If only the aggregated data are uploaded than all aggregated datasets are switched on (green) by default. In case case-based data are uploaded they take precedence by default. Switching on the aggregated dataset, this dataset will replace the case-based data for modelling. This can be further refined by selecting years for which the aggregated dataset should overwrite the case-based dataset. For years outside of the selected period case-based data will be used.

For consistency reasons datasets based on new HIV diagnoses (i.e. HIV, HIV_CD4_1, HIV_CD4_2, HIV_CD4_3, HIV_CD4_4, HIVAIDS) should come for a given year from the same source, either case-based or aggregated and may be only replaced all together. Aggregated AIDS and Dead files and specific years can be selected independently.



Aggregated data

Select data for modelling

Only years present in the aggregated data set will overwrite case-based derived data.

Data type	Use	Years
Dead	●	1985 – 2015
AIDS	○	1985 – 2016
HIV, HIVAIDS	○	1985 – 2015
HIV_CD4_1, HIV_CD4_2, HIV_CD4_3, HIV_CD4_4	○	1985 – 2015

Switch on (green) to use the aggregated dataset for Deaths

Select years for which the aggregated Dead dataset should be used (here exclude 2016)

Interpretation

Do not confuse this selection with the selection of time periods for modelling in the tab "Advanced". This section creates a single dataset based on the different types of data uploaded. This dataset may be than subset for modelling in the "Advanced" section.
The aggregated data will overwrite the case-based data for the selected time periods. It is up to the user to make sure that the populations from the aggregated data and from the case data describe the same populations and can be used together for modelling.
In case only aggregated data are used, all aggregated datasets must be switched on for modelling.

Further actions

Proceed to defining the actual population combinations to use in modelling.

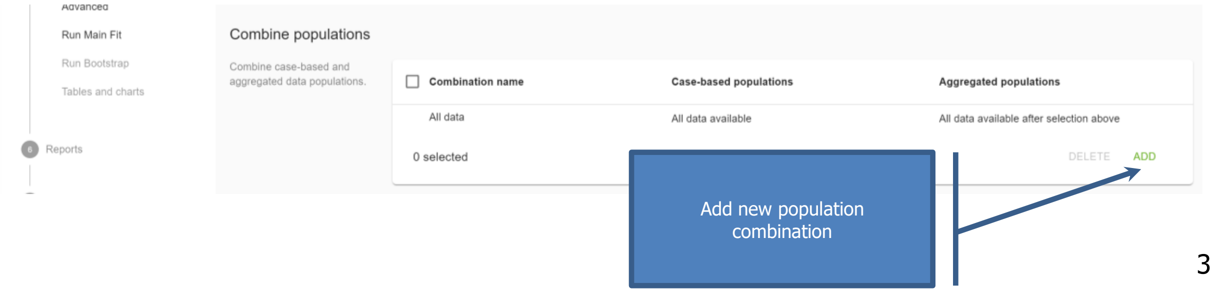
7.3 Populations: combining populations

Description

Populations defined from the case-based data (Section 7.1) and the populations present in the aggregated data, i.e. the groups that were specified in the header row of the aggregated input data files, may be used to create new populations. The user can easily add combinations of these risk groups as new populations for modelling.

Process

By default "All data" population is included, which includes all cases from case-based data and the sum of all populations columns for aggregated data. New combinations may be added with "Add" button.



Combine populations

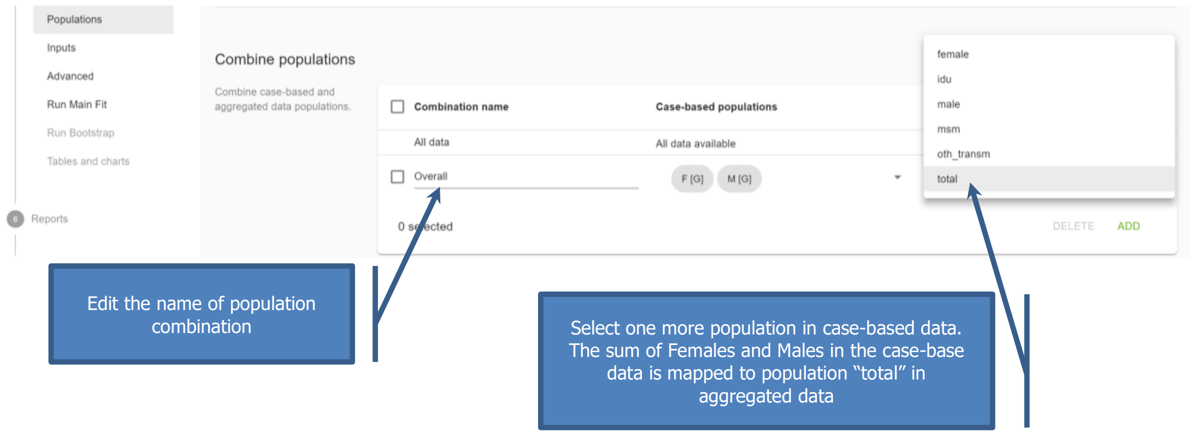
Combine case-based and aggregated data populations.

□ Combination name	Case-based populations	Aggregated populations
All data	All data available	All data available after selection above

0 selected

Add new population combination

“All data” may not be useful in case the aggregated dataset does not contain exclusive categories. For instance, the aggregated dataset may already have a column describing total population. It can also contain columns with the same data adjusted using multiple methods with intended use to compare the final modelling outcomes. The user should therefore pay attention to define appropriate mapping of populations between case-based and the aggregated datasets.



Edit the name of population combination

Select one more population in case-based data. The sum of Females and Males in the case-base data is mapped to population “total” in aggregated data

Interpretation

The user is free to combine aggregated and case-based data from modelling. Different datasets may be used for different time periods.

It is up to the user to make sure that the aggregated populations selected for a new combination are mutually exclusive. Combining of populations defined for case-based data can be interpreted as including all cases that belong to either of populations, so no overlap is produced. In case of aggregated data – the sum of the counts are calculated.
It is up to the used to make sure that the combination defined by case-based populations and the one defined by aggregated populations refer to the same population group.

Selection of aggregated and/or case-base data has impact on availability of bootstrap method to be applied. The tool determines the appropriate method automatically. In general, if aggregated data for HIV (i.e. HIV, HIVAIDS and/or HIV cases by CD4 count category) data are selected the tool will use parametric bootstrap. For case-based HIV data, including adjusted case-based data – non-parametric bootstrap is preferable.

Some stratifications may lead to creating very small subsets of cases. This can cause troubles for convergence of the model and should be avoided. A good approach may therefore be to first consider all HIV-positive individuals as a single population and then as a next step disaggregate this single population into smaller populations.

Further actions

Once the populations are defined the user can proceed to define diagnosis matrices. It is possible to go back to defining additional populations and / or combinations of populations at later time if necessary.

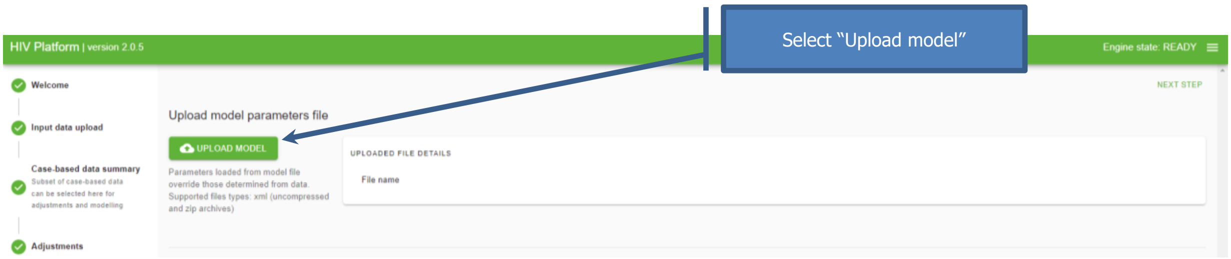
7.4 Inputs: uploading the saved model from ECDC Modelling Tool

Description

Existing model files created by the ECDC Modelling tool (https://www.ecdc.europa.eu/en/publications-data/hiv-modelling-tool), can be uploaded. Only files from ECDC Modelling tool created with the method “Incidence” are accepted by the ECDC HIV Platform.

Process

Select "Inputs" tab under the "Modelling section". Click on "Upload model" and select appropriate file. Supported files types include xml (uncompressed and zip archives).



Select "Upload model"

HIV Platform | version 2.0.5
Engine state: READY
Welcome
Input data upload
Case-based data summary
Subset of case-based data can be selected here for adjustments and modelling
Adjustments
NEXT STEP
Upload model parameters file
UPLOAD MODEL
UPLOADED FILE DETAILS
File name

Parameters loaded from model file override those determined from data. Supported files types: xml (uncompressed and zip archives)

Interpretation

Importing prior model is a utility to allow the user to transition the parameters set up in the ECDC Modelling Tool to the HIV Platform tool. In particular diagnosis matrices as well as advanced parameters for the incidence curves will be imported.

Parameters loaded from model file override those determined from data. It is important to review the parameters, especially the calculation ranges in the "Advanced settings" section. For example if the current dataset contains additional years of data, beyond the time periods used for the saved file, the user needs to manually extend the ranges.

Further actions

Review the imported parameters including the diagnosis matrices and the Advanced parameters, and add changes if necessary. Proceed to modelling.

7.5 Inputs: defining diagnosis matrices

Description

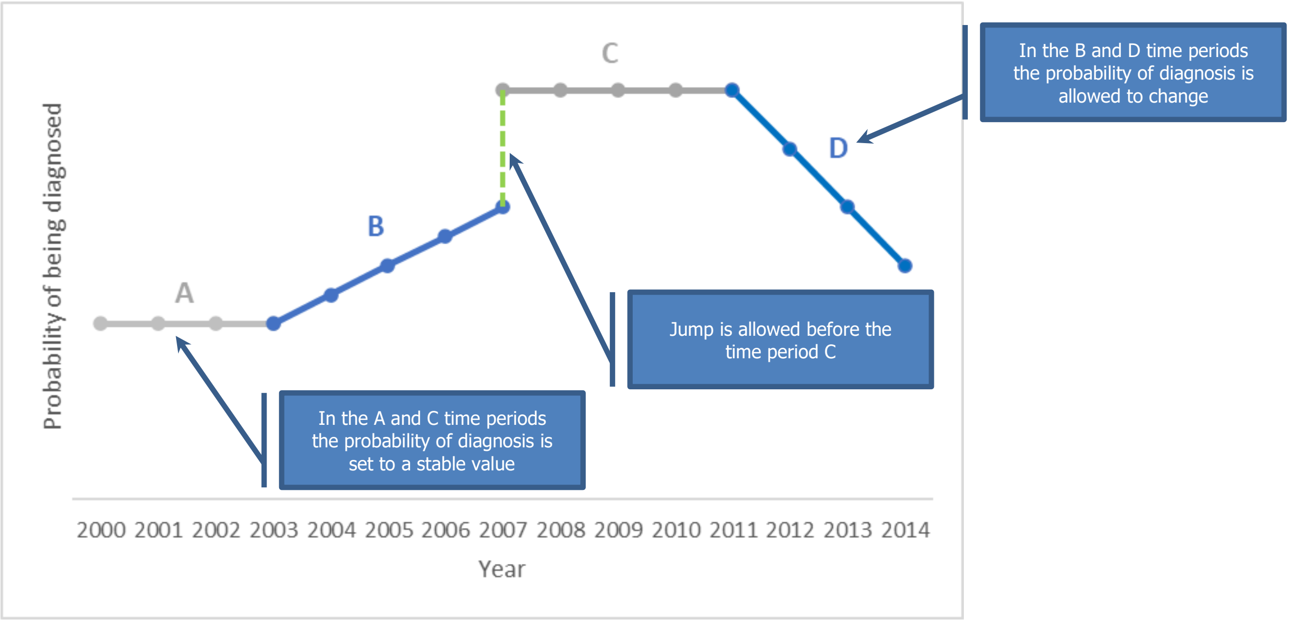
The diagnosis matrix defines the shape of the diagnosis probability. The tool automatically displays the default matrix pre-specified by the wizard, but the user is strongly encouraged to change this specification. In case model file from ECDC Modelling Tool is uploaded the tool will display the uploaded matrix as default. Currently, the tool allows to specify several matrices, which can be applied to either of the populations.

The diagnosis matrix provides expert input to how the probability of being diagnosed might have changed over the time. The user is not required to specify the exact probability, rather he or she should provide the following information:

- Time intervals: indicate when the probability of being diagnosed may change.
- Presumed shape: indicate how the probability of being diagnosed may change.

To specify the presumed shape the user has to decide whether within a given time period the probability of being diagnosed remains the same or is likely to change, for example due to testing campaigns or increasing availability of testing sites. In addition, in some particular situations jumps in the diagnostic probability may be expected – e.g. legal change that allows self-testing, which previously was not in place.

In the following example there are two time periods for which the diagnosis probability is allowed to change: B and D. Note, that the user does not need to define whether is it likely to increase or decrease. This will be determined by the tool. In turn in the time periods A and C the diagnosis probability should stay stable, but the probability of being diagnosed in the time period C may differ from what is expected based on the time period B (i.e. the jump is allowed).



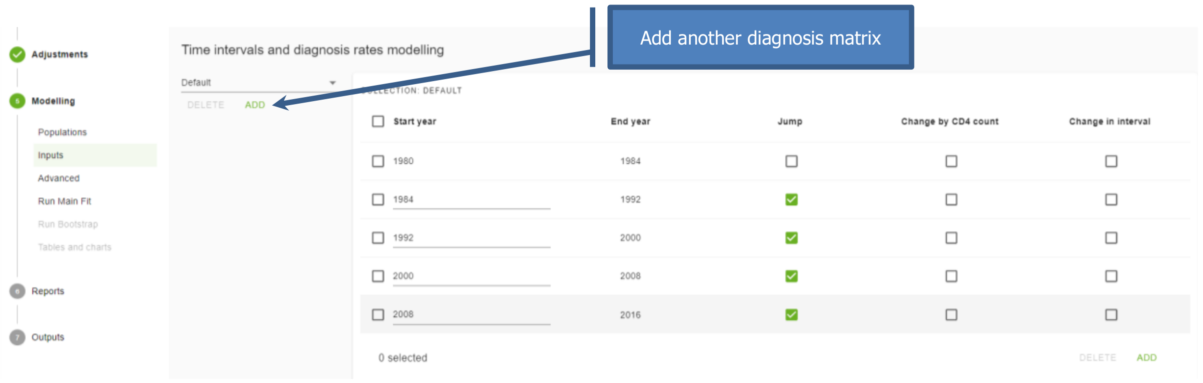
In the B and D time periods the probability of diagnosis is allowed to change

Jump is allowed before the time period C

In the A and C time periods the probability of diagnosis is set to a stable value

Process

Default matrix is provided by the tool wizard to start with. The user can edit this matrix or add another one to match their specific knowledge.



Add another diagnosis matrix

Time intervals and diagnosis rates modelling

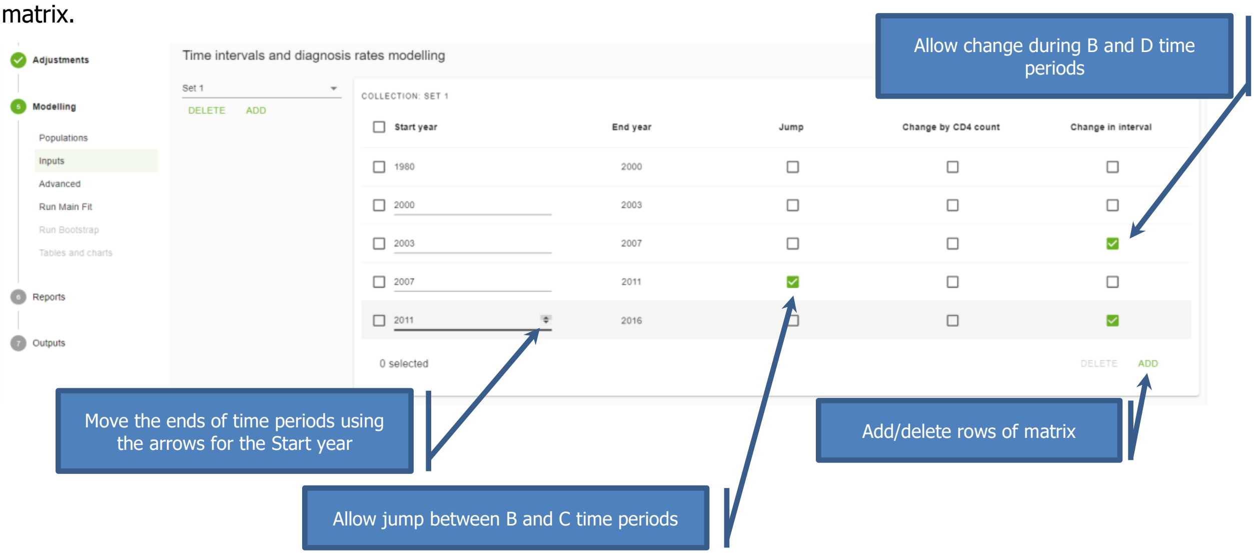
Default

COLLECTION: DEFAULT

☐ Start year	End year	Jump	Change by CD4 count	Change in interval
☐ 1980	1984	☐	☐	☐
☐ 1984	1992	☑	☐	☐
☐ 1992	2000	☑	☐	☐
☐ 2000	2008	☑	☐	☐
☐ 2008	2016	☑	☐	☐

0 selected

The number of time periods with different probability of diagnosis is controlled by adding/ deleting rows of the matrix.



Allow change during B and D time periods

Time intervals and diagnosis rates modelling

Set 1

COLLECTION: SET 1

☐ Start year	End year	Jump	Change by CD4 count	Change in interval
☐ 1980	2000	☐	☐	☐
☐ 2000	2003	☐	☐	☐
☐ 2003	2007	☐	☐	☑
☐ 2007	2011	☑	☐	☐
☐ 2011	2016	☐	☐	☑

0 selected

Move the ends of time periods using the arrows for the Start year

Add/delete rows of matrix

Allow jump between B and C time periods

The start year in the first row and the end year of the last row correspond to the range of the data that should be used for calculations. By defaults this range is the range of data available. If the user instead wishes to run modelling only on a subset of data, this can be changed by adjusting the calculation range in the Advances parameters section.

The following options are available to adjust the diagnosis matrix:

A. Start from new baseline

If this box is not ticked, the diagnosis probability at the start of the interval will be the same as a the end of the previous time interval. If the tick box is checked, the diagnosis probability will start at a new value. An example of when this box should be ticked is when diagnosis of HIV by means of serologic tests became possible in 1984.

B. Different by CD4 count categories

Tick this box if you want to diagnosis probabilities to be different for each of the four CD4 count strata. This could for example reflect the fact that people with more advanced infection may have higher probability of diagnosis. In order to limit the number of parameters the user can leave out this option and add "Extra diagnosis rate due to non-AIDS symptoms" in the Advanced option section.

Caution: This option should only be used if there are data on HIV diagnoses by CD4 count.

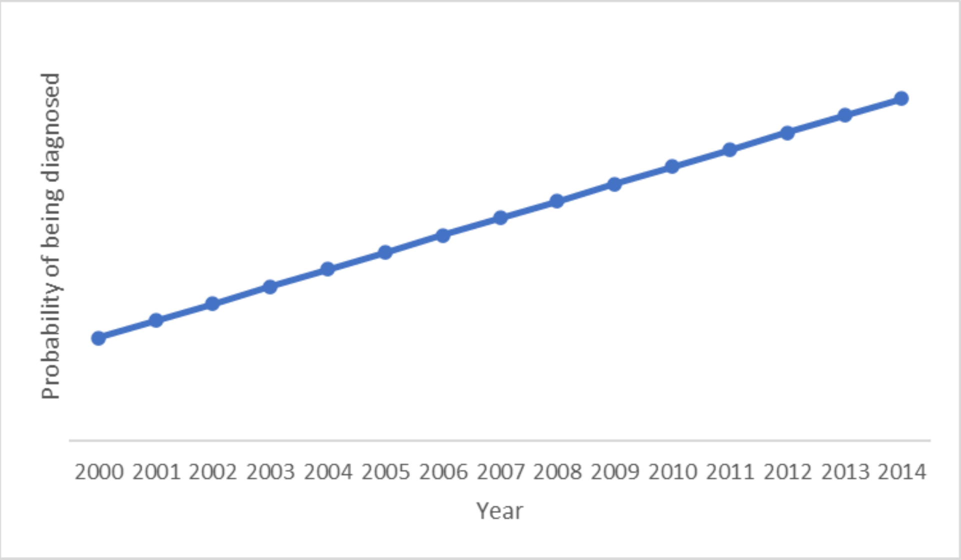
C. Changing during time interval

If the tick box is unchecked, the tool will assume that the probability of being diagnosed will not change during the time interval.

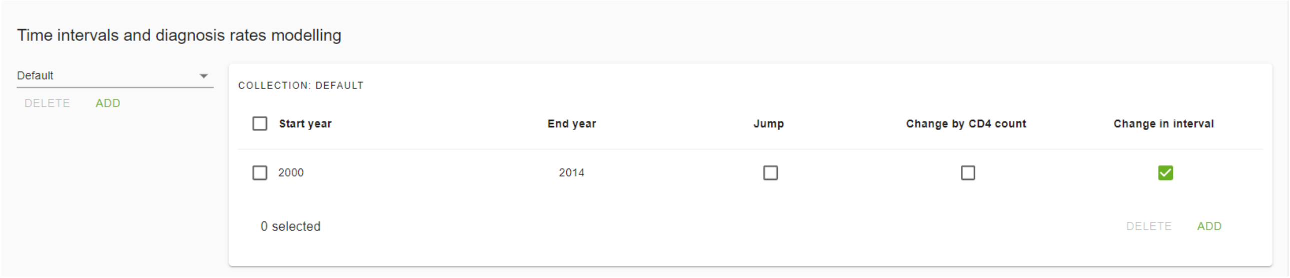
If the tick box is checked, the diagnosis probability can increase or, less likely, decrease during the time interval. An increase would be expected when, for instance, people are more frequently tested for HIV due to increasing awareness. The pace at which the diagnosis probability increases (or decreases) is determined from the input data.

Simplified examples of use:

1. The probability of diagnosis gradually increasing in a stable way during the study period



The corresponding matrix in the tool would be as follows.

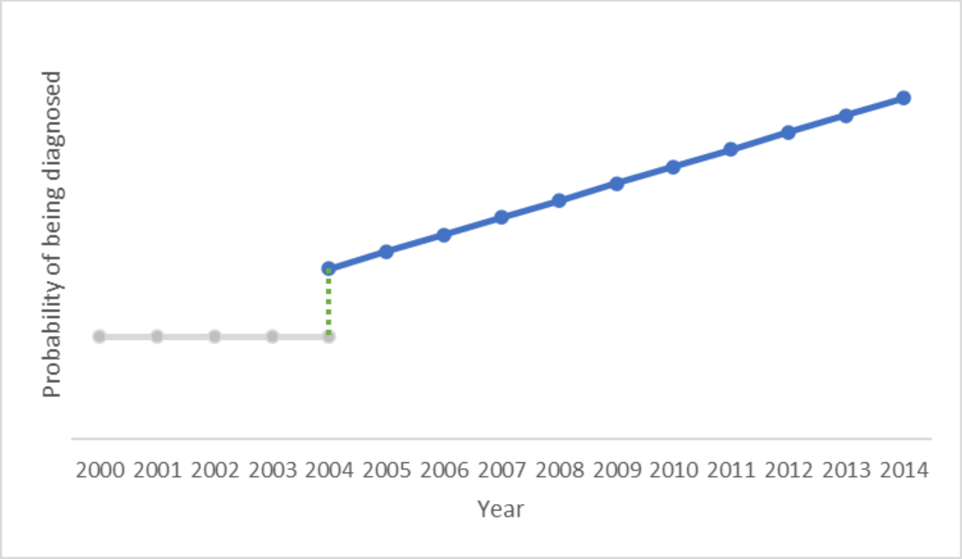

Time intervals and diagnosis rates modelling

COLLECTION: DEFAULT

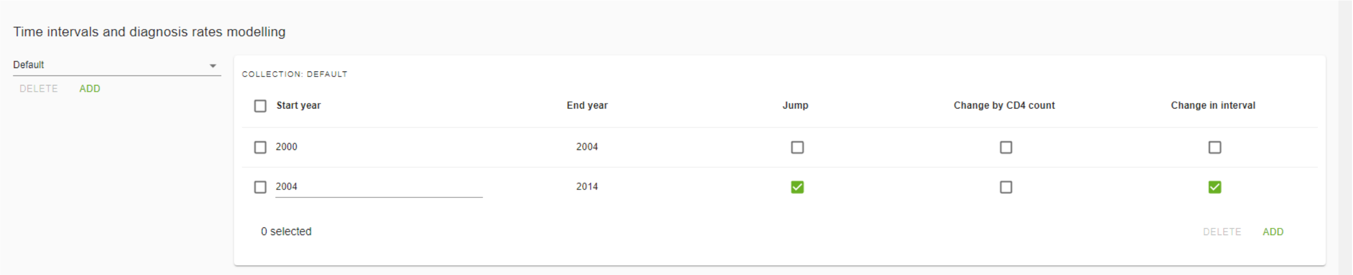
□ Start year	End year	Jump	Change by CD4 count	Change in interval
□ 2000	2014	□	□	☑

0 selected

2. Introduction of new testing system and gradual increase in the probability of diagnosis



The corresponding matrix in the tool would be as follows.



Time intervals and diagnosis rates modelling

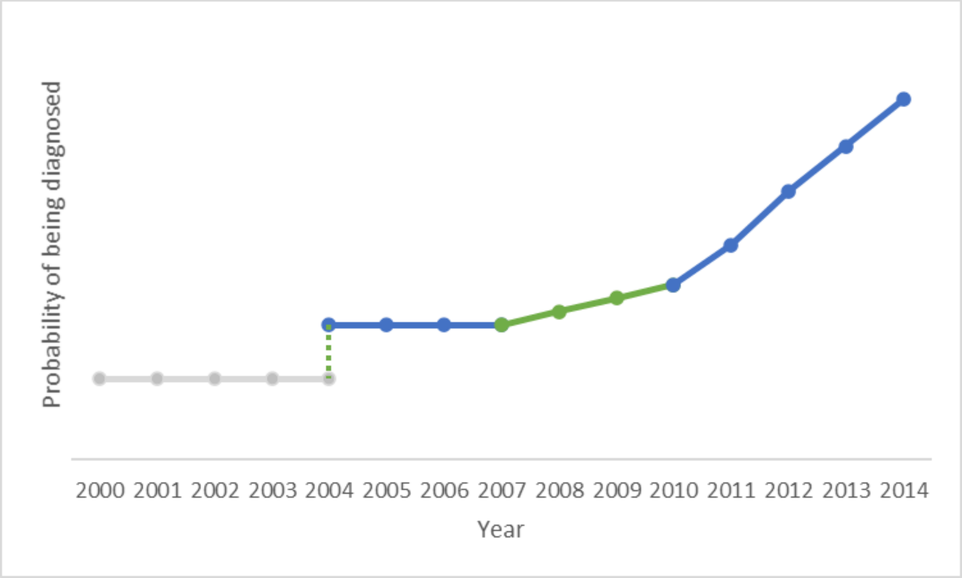
COLLECTION: DEFAULT

□ Start year	End year	Jump	Change by CD4 count	Change in interval
□ 2000	2004	□	□	□
□ 2004	2014	☑	□	☑

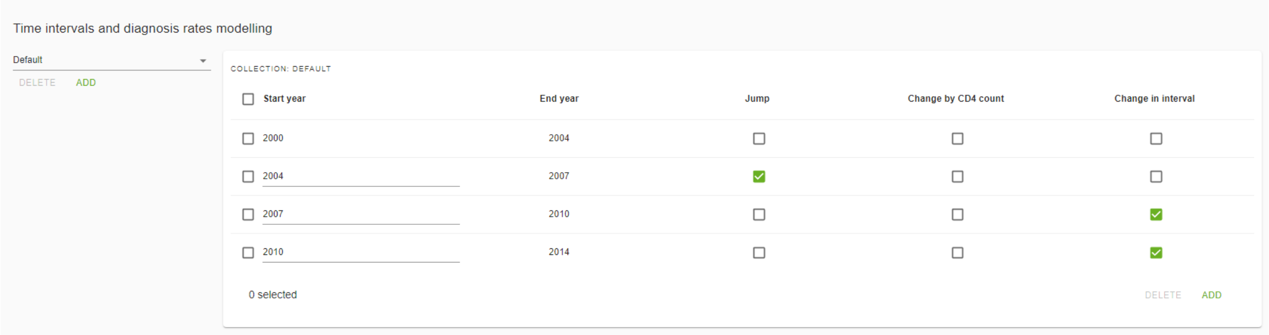
0 selected

3. Introduction and accelerating the roll-out of testing programme

Changes is testing that are not linear can still be modelled using more than one interval, which will allow the tool to match the changing slope to subsequent time periods.



The corresponding matrix in the tool would be as follows.



Time intervals and diagnosis rates modelling

Default

COLLECTION: DEFAULT

□ Start year	End year	Jump	Change by CD4 count	Change in interval
□ 2000	2004	□	□	□
□ 2004	2007	☑	□	□
□ 2007	2010	□	□	☑
□ 2010	2014	□	□	☑

0 selected

Interpretation

- Diagnosis matrix is used in the tool to determine the best-matching diagnosis probability over calendar time.

Apart from estimating the annual number of new HIV infections, the tool also estimates the probability that HIV-positive individuals are diagnosed with HIV when their (unobserved) CD4 count is in one of the four CD4 intervals. This probability is usually unknown and needs to be estimated from the observed input data. The tool uses the diagnosis probabilities to calculate the expected time between infection and diagnosis by year of infection.

- The more time periods are specified, the closer the fit could get. However, many time periods also imply many parameters to be estimated by the tool. This can cause problems with convergence especially with small size populations and short time series, in which cases it is advisable to reduce the size of the matrix.

- In the time interval 1980 to 1984, there was no testing for HIV and all three boxes should be unchecked. HIV could only be 'diagnosed' when AIDS symptoms appeared. However, it is not necessary to specify the probability of being diagnosed with HIV when AIDS symptoms appear, because this is taken into account by the tool.

- Testing data for a given population, if available, can be used to guide the initial selection of the time intervals. Especially introduction of new testing modalities or wider roll-out of testing campaigns may indicate changing probability of diagnosis. However, jumps are a rarely used option, there should be a good reason behind using it.

Further actions

The users can specify as many matrices as applicable given the specificities of the populations for whom they wish to model HIV incidence. The user can go back to this section and specify additional matrices if necessary at any later stage of modelling. Adjustment of the matrix based on the modelling results can improve the fit. For example if the estimated number of new diagnoses deviates significantly from the observed data in a specific time period, this can indicate that the diagnosis probability is not optimally specified.

7.6 Advanced parameters specification

Description

Advanced options refer to the ranges of data to be used by modelling as well as additional flexibility around the modelling. The default values determined by the tool should be sufficient in most applications.
Nonetheless, the user should always review these options at least in the following 2 situations:

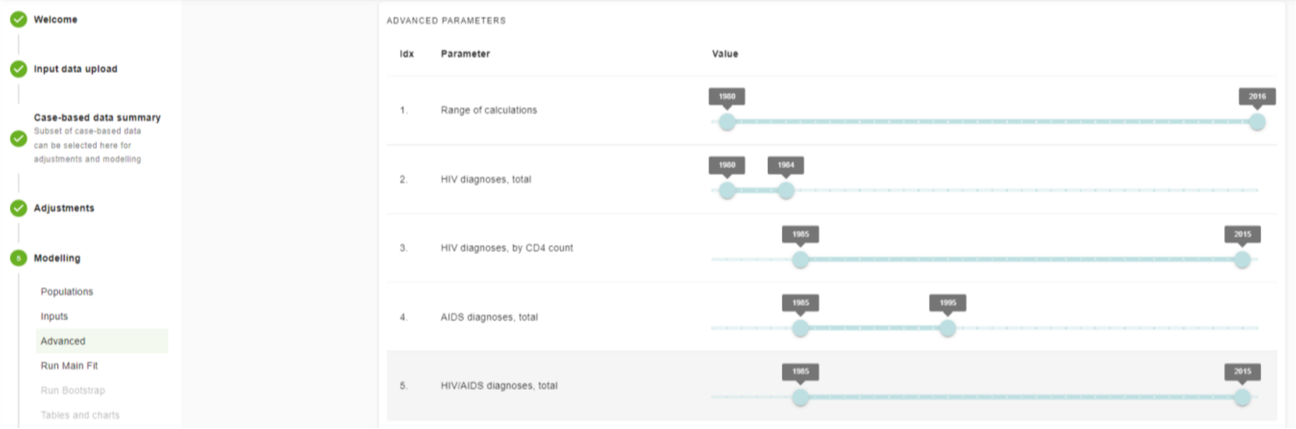
- in case they do not have data from the beginning of the epidemic
- in case they uploaded model parameters from the existing file

In case the parameters are uploaded from an existing file the ranges of calculation may be inappropriate for the new data.

Process

The first part refers to the data ranges used for the calculation. The ranges are determined by tool based on available data based on minimum and maximum years in the input dataset. Note that the input dataset for modelling can be a composite of aggregated and case-based data. The description how to specify which data are use for which time periods is provided in Section 7.2.

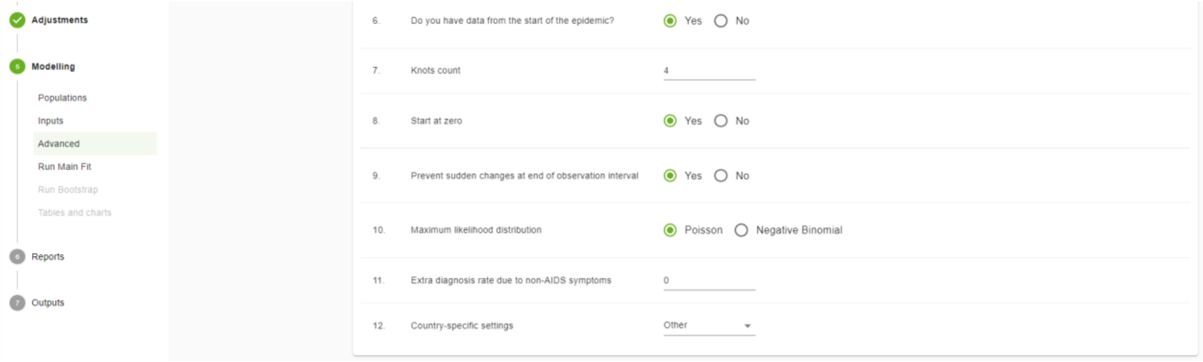
The default ranges of calculation are overridden, in case Model Parameter File is uploaded in the Input tab (see Section 7.4). In this situation the tools takes the range of available data within the range provided by the Model Parameter File.



ADVANCED PARAMETERS

Idx	Parameter	Value
1.	Range of calculations	1980 – 2016
2.	HIV diagnoses, total	1980 – 1984
3.	HIV diagnoses, by CD4 count	1985 – 2015
4.	AIDS diagnoses, total	1985 – 1995
5.	HIV/AIDS diagnoses, total	1985 – 2015

The second part allows to better define the incidence curve and confidence intervals.



6.	Do you have data from the start of the epidemic?	● Yes ○ No
7.	Knots count	4
8.	Start at zero	● Yes ○ No
9.	Prevent sudden changes at end of observation interval	● Yes ○ No
10.	Maximum likelihood distribution	● Poisson ○ Negative Binomial
11.	Extra diagnosis rate due to non-AIDS symptoms	0
12.	Country-specific settings	Other

Full or partial data. If (complete) HIV surveillance data are not available from the start of the HIV epidemic but only from a certain year Y onwards, "no" should be selected in the "**Do you have data from the start of the epidemic?**". This may also apply to a situation when only partial data, e.g. only the data for the latest time period are judged reliable enough.

"Knots count". Select an integer number controlling the flexibility curve. Higher values (>6) are not recommended.

"Start at zero". This parameter allows to control how the epidemic is modelled. If the aim is to model the whole epidemic (irrespective of available data range) then it is assumed that there were no cases before the start year of the calculation range. If this box is ticked the tool assumes that the HIV incidence curve is zero on January 1st of the year in which the model calculations start (Start year of Range of calculations). This also means that one of the parameters necessary to specify the HIV incidence curve is fixed at zero and does not need to be estimated. The user could also chose to restrict the model and then assume that already cases exist before the start year (i.e. not starting from zero cases).

"Prevent sudden changes at the end of observation interval" - in most of applications this option should be selected. It prevents the undue instability at the end of the calculation range (the most recent years), that comes from the way the incidence is modelled (for more details see methodological annex A1.2). At the end of the calendar year range, the incidence curve is only constrained by individuals who have been diagnosed relatively shortly after becoming infected. The majority of these individuals will have been diagnosed with a CD4 count ≥ 500 cells/mm3 and fluctuations in this number will have a large impact on the behaviour of the incidence curve. The impact of such fluctuations can be attenuated by requiring that sudden increases or decreases in the incidence curve should be prevented. Generally, this does not give a worse fit to the data. However, confidence intervals will become narrower and the estimated incidence curve may appear more precise than it really is.

"Maximum likelihood distribution" could be changed from Poisson to Negative Binomial in case the confidence bounds do not cover the observed data sufficiently well. Maximum likelihood methods are used to find the set of parameters that best fit the observed data. To define the likelihood, it is assumed that all data items are distributed according to a certain probability density function around a mean defined by the model. For convenience, instead of maximising the likelihood, the tool minimises the equivalent deviance measure.

The default distribution is a Poisson distribution in which the mean is equal to the variance. In practice this distribution works well enough. The other option is a negative binomial distribution, which is a generalisation of the Poisson distribution such that the variance can be larger than in a Poisson distribution. The increase in variance is determined by a so-called dispersion parameter. If the negative binomial distribution is selected, the main model fit will be identical to the one obtained with the Poisson distribution. In addition, the tool will estimate two dispersion parameters, one for AIDS cases and one for HIV diagnoses

"Extra diagnosis rate due to non-AIDS symptoms" could be added if "Change by CD4 count" is not specified in the diagnosis matrix, but generally the non-AIDS symptoms are believed to trigger diagnosis. When this value is larger than zero, the tool will add an extra contribution to the probability of being diagnosed when (unobserved) CD4 counts are below 200 cells/mm3. The parameter should be adjusted manually.

"Country-specific settings" parameter generally should not be changed, as is are relevant for specific situations only.

Interpretation

Extra parameters for better adjusting modelling. To be used as a second step after default modelling.

Table 6. Advanced selection of ranges for modelling

Parameter	Description	What to select
Range of calculations: Start year	Only diagnoses made during this year or later will be included in estimation. Calculations start on 1 January of Start year and end on 31 December of End year.	Typically, the Start year should be the approximate year in which the HIV epidemic started in a country (default value 1980). It may be earlier than availability of data.
Range of calculations: End year	Only diagnoses made during this year or before will be included in estimation. Calculations start on 1 January of Start year and end on 31 December of End year.	End year should not be larger than the most recent calendar year for which data are available. It is not possible to make future projections with the tool.
Data items range: HIV diagnoses, total	Range of calendar years for which data on the total annual number of HIV diagnoses are used in the model fit.	HIV by CD4 category data are by default preferred to overall HIV counts (HIV diagnoses, total) and this is reflected by the default settings. For years with no or insufficient data on CD4 counts at the time of diagnosis, total number of diagnoses should be used
Data items range: HIV diagnoses, by CD4	Range of calendar years for which data on the annual number of HIV diagnoses by CD4 count interval are	HIV by CD4 category data are by default preferred to overall HIV counts (HIV diagnoses, total) and this is reflected by the default settings. If the CD4 count data are available for

	Used rather than HIV diagnoses, total.	a subset of cases it is a good idea to run the multiple imputations first and run modelling on adjusted data.
Data items range: AIDS diagnoses, total	Range of calendar years for which data on the annual number of AIDS diagnoses are used.	The upper boundary of this range should be earlier than the year in which combination antiretroviral treatment (cART) became widely available. After the introduction of cART, the annual total number of AIDS diagnoses will strongly depend on how many individuals are treated. The effect of treatment on the time to developing AIDS is difficult to quantify. Therefore, and also because treatment is not taken into account in the tool, total number of AIDS diagnoses should not be used in the era of cART.
Data items range: HIV/AIDS diagnoses, total	Range of calendar years for which data are available on the number of HIV diagnoses with a concurrent AIDS diagnosis.	Since HIV diagnosis generally precedes start of treatment, HIV/AIDS diagnoses can be used during the entire course of the epidemic. There is no limitation on the upper boundary as there is for total number of AIDS diagnoses.

Recommended settings if data available only for the latest time period

The following settings are recommended if data are only available from year Y onwards:

- Start year of Range of calculations is the approximate year in which the HIV epidemic started in a country (default value 1980). Stop year is the most recent year for which data are available.
- The slider bars for the data items (HIV diagnoses, total; HIV diagnoses, by CD4 count; AIDS diagnoses, total; HIV/AIDS diagnoses, total) should start in the year from which surveillance data are available. In the figure below, data on HIV diagnoses, stratified by the presence of a concurrent AIDS event or CD4 cell count, are available from 2003 onwards.
- In the diagnosis matrix, make sure Start Year for the first time interval equals the start year of the range of calculations and End Year equals Y. For this first time interval, select Start from new baseline and, if data on CD4 counts are available in year Y, also Different by CD4 count categories, but do not select Changing during time interval. Since there are no data before year Y, it is not possible to estimate diagnosis probabilities in more than one time interval.
- Do not specify too many knots for the incidence curve. The default value of 4 is likely to be enough.

Note: The estimates basing on partial data will only be reliable for the most recent calendar years, i.e., for years where the undiagnosed proportion of people who acquired their HIV infection before year Y is small. Since data before year Y are not available, the tool cannot estimate the total population living with HIV.

Note: Model outcomes for recent calendar years may be similar if “yes” is selected despite data not being available from the start of the HIV epidemic. However, the estimation process will be less efficient, especially if a large number of parameters needs to be estimated or if confidence intervals are calculated.

Table 7. Advanced selection of modelling settings

Parameter	Description	What to select
Do you have data from the start of the epidemic?	If yes, the tool assumes that the were no cases diagnosed before what is supplied	The question refers to the data defined by the data items range. Even if data are available from the beginning of the epidemic, the user may chose to restrict the data range for calculations, to drop the earliest years. In this case “no” should be selected.
Knots count	The value controlling the flexibility of incidence curve.	The higher number of knots allows to model unusual incidence curve shapes. The more knots the closer the fit of the curve, but the more parameters to estimate, so models with less knots should be used if the fit is comparable. The recommended values are 4 to 6 knots. The default value of 4 is likely to be enough.
Start at zero	If yes, the tool assumes that the epidemic started at the Start year of the calculation range.	The question refers to the epidemic period defined by the calculation range. Typically, we chose to model the whole epidemic, in which case “yes” is selected.
Prevent sudden changes at end of observation interval	Alters the way the incidence is modelled, to keep incidence at similar level at the end of the interval	Choosing “yes” is recommended

Maximum likelihood distribution	Controls the assumed variability of data points. The Poisson distribution assumes the data are less variable and the Negative Binomial allows extra variability.	If data points (new diagnoses counts) lie outside of confidence bands (confidence bands seem too narrow) it is recommended to use the Negative Binomial distribution.
Extra diagnosis rate due to non-AIDS symptoms	The non-AIDS symptoms are believed to trigger diagnosis. When this value is larger than zero, the tool will add an extra contribution to the probability of being diagnosed when (unobserved) CD4 counts are below 200 cells/mm3	If you specify estimating diagnosis rate by CD4 category in the diagnosis matrix, there is no need to chose the "Extra diagnosis … " option. The actual value of this parameter depends on the testing system if it is focused on risk group screening or indicator condition testing. Recommended values are between 0.4 – 0.8, but in case a lot of cases are diagnosed late, with symptoms, even a bit higher values may be applicable.

Further actions

The user can go back to this section and specify additional matrices if necessary at any later stage of modelling. Correct specification of the parameters may help to achieve better fit of the model and more accurate estimation of the incidence and other outcomes of modelling.

7.7 Modelling with adjustments

Description

The tool allows to run modelling on adjusted data. This can include reporting delay adjustments and/or missing values imputations depending on what was selected in the adjustment tab. The methods are discussed in the methodological annex A1.3.

Process

In case adjustments are used the modelling section will automatically use the adjusted data. No additional user action is required.

In case of reporting delay adjustment the adjusted counts are used to fit the model. In case of imputed data the model is run separately on each of the imputed datasets, and then combined to obtain final modelling estimates. The imputations will have an effect on the final modelling output both for overall modelling, as the CD4 counts are imputed, and for modelling in subpopulations formed based on variables that originally contained missing values.

Interpretation

Reporting delay adjustment and imputations will have an effect on the counts of new diagnoses provided to the model so the counts displayed in the "Goodness-of-fit" section of the "Tables and charts" tab in modelling will be the corrected counts and may differ from the counts in the original data.

Further actions

If the user wants to run modelling without adjustments he or she should upload the desired data and skip immediately to modelling. The user can also save the tool state after mapping and regrouping steps to be able to use it for both adjusted and unadjusted analyses latter on.

7.8 Running main fit

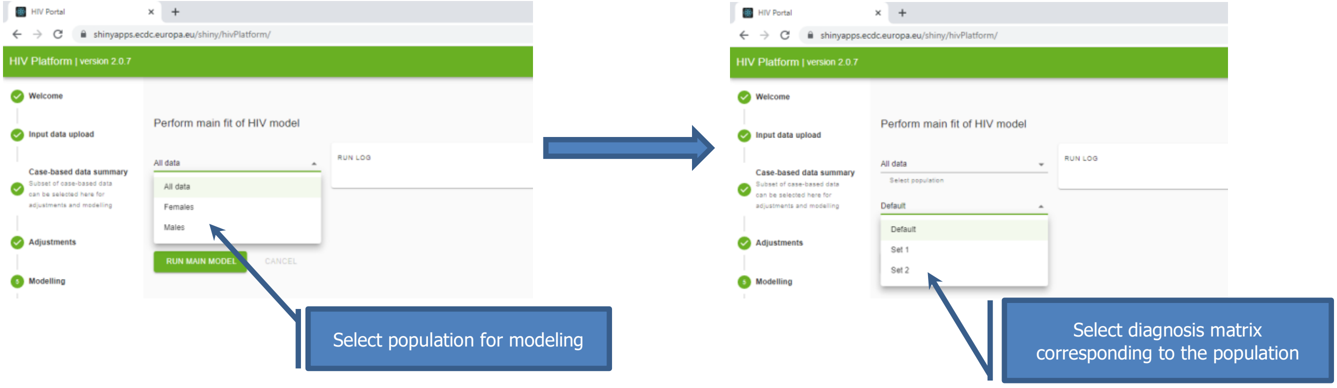
Description

The main fit finds the incidence curve best fitting to the data given the parameters defined by the user (mainly the number of knots and the diagnosis matrix). Please see the Methodology annex for details on the estimation procedure.

All populations defined in the Populations tab and all diagnosis matrices defined in the Inputs tab are automatically available in this tab. The main fit can be only run one at a time – for one selected population and one selected diagnosis matrix.

Process

To perform the main fit select population and select diagnosis matrix. "Run main model" button will activate the estimation procedure.

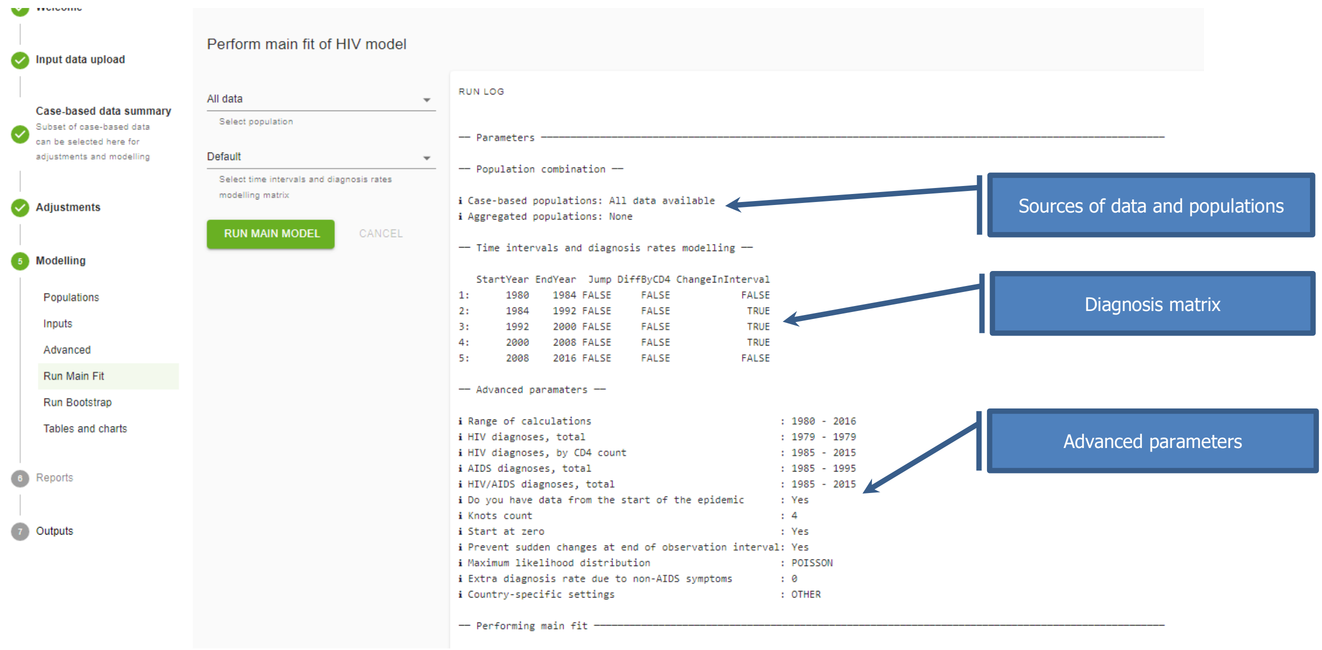

HIV Portal
shinyapps.ecdc.europa.eu/shiny/hivPlatform/
HIV Platform | version 2.0.7
Welcome
Input data upload
Case-based data summary
Subset of case-based data can be selected here for adjustments and modelling
Adjustments
Modelling
Perform main fit of HIV model
All data
Females
Males
RUN MAIN MODEL
CANCEL
RUN LOG
Select population
Default
Set 1
Set 2

Select population for modeling

Select diagnosis matrix corresponding to the population

Interpretation

The tool will signal once the run is finished, which usually takes several seconds. Inspect he run log before proceeding.
The first part summarises the parameters and data selected for the estimation.



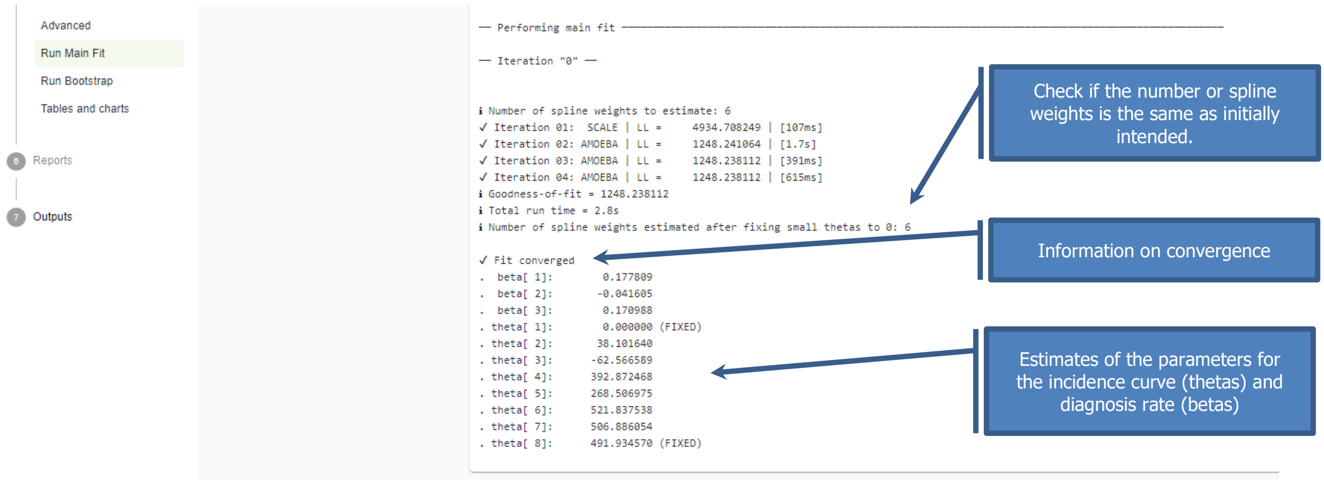
```
RUN LOG

── Parameters ──────────────────────────────────────────────────────────────────────────────

── Population combination ──

ℹ Case-based populations: All data available
ℹ Aggregated populations: None

── Time intervals and diagnosis rates modelling ──

   StartYear EndYear  Jump DiffByCD4 ChangeInInterval
1:      1980    1984 FALSE     FALSE            FALSE
2:      1984    1992 FALSE     FALSE             TRUE
3:      1992    2000 FALSE     FALSE             TRUE
4:      2000    2008 FALSE     FALSE             TRUE
5:      2008    2016 FALSE     FALSE            FALSE

── Advanced paramaters ──

ℹ Range of calculations                            : 1980 - 2016
ℹ HIV diagnoses, total                             : 1979 - 1979
ℹ HIV diagnoses, by CD4 count                      : 1985 - 2015
ℹ AIDS diagnoses, total                            : 1985 - 1995
ℹ HIV/AIDS diagnoses, total                        : 1985 - 2015
ℹ Do you have data from the start of the epidemic  : Yes
ℹ Knots count                                      : 4
ℹ Start at zero                                    : Yes
ℹ Prevent sudden changes at end of observation interval: Yes
ℹ Maximum likelihood distribution                  : POISSON
ℹ Extra diagnosis rate due to non-AIDS symptoms    : 0
ℹ Country-specific settings                        : OTHER

── Performing main fit ─────────────────────────────────────────────────────────────────────
```

Sources of data and populations

Diagnosis matrix

Advanced parameters

Further details on the outcome of the fit are provided. If there are issues with convergence they will be listed here.



```
— Performing main fit ———————————————————————————————————————————————————

— Iteration "0" —

ℹ Number of spline weights to estimate: 6
✓ Iteration 01:  SCALE | LL =     4934.708249 | [107ms]
✓ Iteration 02: AMOEBA | LL =     1248.241064 | [1.7s]
✓ Iteration 03: AMOEBA | LL =     1248.238112 | [391ms]
✓ Iteration 04: AMOEBA | LL =     1248.238112 | [615ms]
ℹ Goodness-of-fit = 1248.238112
ℹ Total run time = 2.8s
ℹ Number of spline weights estimated after fixing small thetas to 0: 6

✓ Fit converged
.  beta[ 1]:       0.177809
.  beta[ 2]:      -0.041605
.  beta[ 3]:       0.170988
. theta[ 1]:       0.000000 (FIXED)
. theta[ 2]:      38.101640
. theta[ 3]:     -62.566589
. theta[ 4]:     392.872468
. theta[ 5]:     268.506975
. theta[ 6]:     521.837538
. theta[ 7]:     506.886054
. theta[ 8]:     491.934570 (FIXED)
```

Check if the number or spline weights is the same as initially intended.

Information on convergence

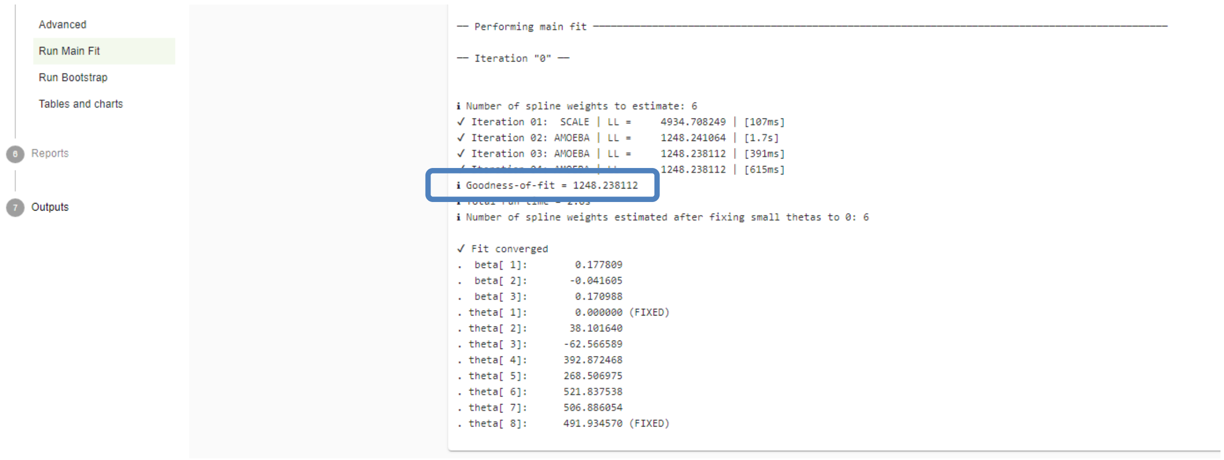
Estimates of the parameters for the incidence curve (thetas) and diagnosis rate (betas)

The tool has built in functionalities that allows the fit to converge most of the times. In case convergence is not reached it usually indicates that there it too many parameters specified. This could be solved by simplifying the diagnosis matrix (reducing the number of rows, unselecting estimation by CD4 count categories) and/or the number of knots. The tool usually attempts as first step to reduce the number of parameters by setting the small ones to zero. These are then displayed as "(FIXED)". First and last θs are also affected by the selected options. In the example above $\theta_1$is fixed to zero which corresponds to selecting the option "Start at zero" and the last one θ_8 is also fixed due to selecting the option "Prevent sudden changes at the end of the observation interval".

Issues with convergence may also occur when data are inconsistent, i.e. contradict each other. This may for example happen when specifying populations on both case based and aggregated data and incorrectly matching the populations. Most likely such situation would be visible "Goodness of fit" section of the "Tables and charts" tab.

Goodness of fit

The tool also outputs the goodness-of-fit statistic in terms of the deviance, which is an overall measure of how far are the model predictions from the data that they try to estimate. The best-fitting model is the model that minimises the sum of the deviances for all data items. As a rule of thumb a model gives an adequate fit to the data if the deviance is approximately equal to the number of observations. This statistic may also help to identify the best fitting model.
Notice, that the Goodness-of-fit statistic displayed is a sum of deviances for all data items, so larger number of data items (e.g. longer time series of data for estimation) the overall deviance statistics may increase. It therefore only makes sense to compare it for models fitted to the same dataset.



```
— Performing main fit ———————————————————————————————————————————————————

— Iteration "0" —

ℹ Number of spline weights to estimate: 6
✓ Iteration 01:  SCALE | LL =     4934.708249 | [107ms]
✓ Iteration 02: AMOEBA | LL =     1248.241064 | [1.7s]
✓ Iteration 03: AMOEBA | LL =     1248.238112 | [391ms]
✓ Iteration 04: AMOEBA | LL =     1248.238112 | [615ms]
ℹ Goodness-of-fit = 1248.238112
ℹ Total run time = 2.8s
ℹ Number of spline weights estimated after fixing small thetas to 0: 6

✓ Fit converged
.  beta[ 1]:       0.177809
.  beta[ 2]:      -0.041605
.  beta[ 3]:       0.170988
. theta[ 1]:       0.000000 (FIXED)
. theta[ 2]:      38.101640
. theta[ 3]:     -62.566589
. theta[ 4]:     392.872468
. theta[ 5]:     268.506975
. theta[ 6]:     521.837538
. theta[ 7]:     506.886054
. theta[ 8]:     491.934570 (FIXED)
```

Further actions

If the model converged, inspect the Goodness-of-fit statistic and the "Goodness-of-fit" section in the "Tables and charts" tab (see Section 7.10)

7.9 Running bootstraps to generate confidence intervals

Description

The confidence intervals around the estimated quantities are produced using the bootstrap method. The bootstrap in general is a statistical technique that mimics taking of random sample from the population, except that sampling with replacement is performed based on the available data. By sampling we obtain similar data set as the original one. The procedure is repeated for a number of times, which creates a series of replicated datasets. The analysis is then repeated on each of the datasets and the same parameter is estimated from all of them. We thus obtain a series of estimates for the parameter, which is considered to approximate the distribution of the parameter. The confidence interval is derived from this distribution.

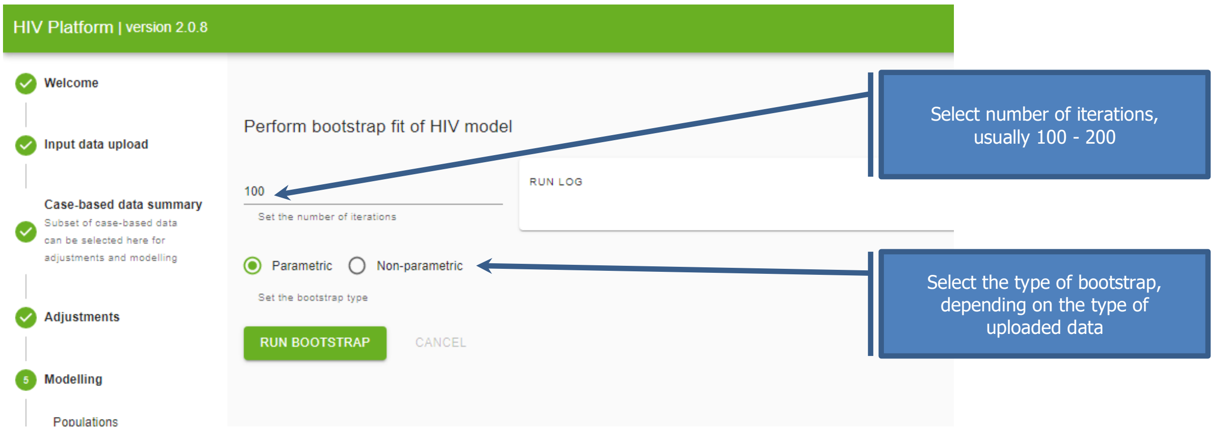
Two options are available: parametric and non-parametric bootstrap.

Parametric bootstrap is available both in case of using case-based data and in case of using aggregated data or in mixture of these. The parametric bootstrap analysis works as follows. Assuming that the data are distributed according to a certain probability distribution, in this case either a Poisson or a negative binomial distribution, with a mean defined by the best-fitting model, the tool generates a new dataset by sampling from this distribution for every year for each of the relevant data items. The model is then refitted to this new dataset starting from the parameter values found in the main fit. This procedure of sampling and refitting is repeated many times. From these many fits, 95% confidence intervals around the estimated model parameters and model outcomes can then be determined as the 2.5-th and 97.5-th percentile.

Non-parametric bootstrap can be performed only on case based data. If aggregated data are used, even for selected time periods parametric bootstrap has to be used. In non-parametric bootstrap replicated datasets are constructed by sampling individual cases from the uploaded dataset, with replacement. Similarly to the parametric bootstrap the model is refitted to the replicated datasets to obtain the estimated model parameters and outcomes and the 95% confidence intervals around the estimated model parameters and model outcomes can then be determined as the 2.5-th and 97.5-th percentile.

Process

The bootstrap option is available after the main fit. To activate the bootstrap the used needs to select the type of bootstrap and the number of iterations, defining how many replicate datasets will be created.



Perform bootstrap fit of HIV model

100
Set the number of iterations

● Parametric ○ Non-parametric
Set the bootstrap type

Select number of iterations, usually 100 - 200

Select the type of bootstrap, depending on the type of uploaded data

Interpretation

To run the bootstrap the tool starts with the model determined in the main fit. If in the main fit some parameters are fixed to 0 (see section 7.8) then these will be fixed to zero in the bootstrap runs. The remaining parameters are required to be estimated for the bootstrap sample, which may not be possible for all replicated datasets. The tool has a safeguard to drop replicated datasets, for which convergence is not reached in 3 times duration needed for convergence of the main fit. On rare occasions if more than 10% of replicated datasets fail the bootstrap procedure will fail. In this case it may be advisable to inspect the main model specification and "Goodness-of-fit"

charts to look for possible data inconsistencies. Reduction of the input parameters may also be helpful (see section 7.8).
In case the bootstrap is run on imputed data, the tools first performs the multiple imputations and then bootstraps each imputed dataset (see methodological Annex A1.3).

To get a feeling for the variation in the estimated model fits a value of 20 would suffice. For a full calculation of confidence intervals at least 100 to 200 iterations are recommended.

Warning: A bootstrap analysis can be time-consuming because it involves running the model multiple times on bootstrap replicates of the data. Confidence intervals should, therefore, only be determined when the main model gives a satisfactory description of the observed data.

Further actions

The bootstrap analysis provides confidence intervals around the estimated parameters. The number of new diagnoses is also estimated together with the confidence interval and the observed diagnoses should fall inside the confidence intervals for correctly specified models. If goodness-of-fit are judged acceptable (see section 7.10), the output charts and datasets are available in the 'Outputs' tab.

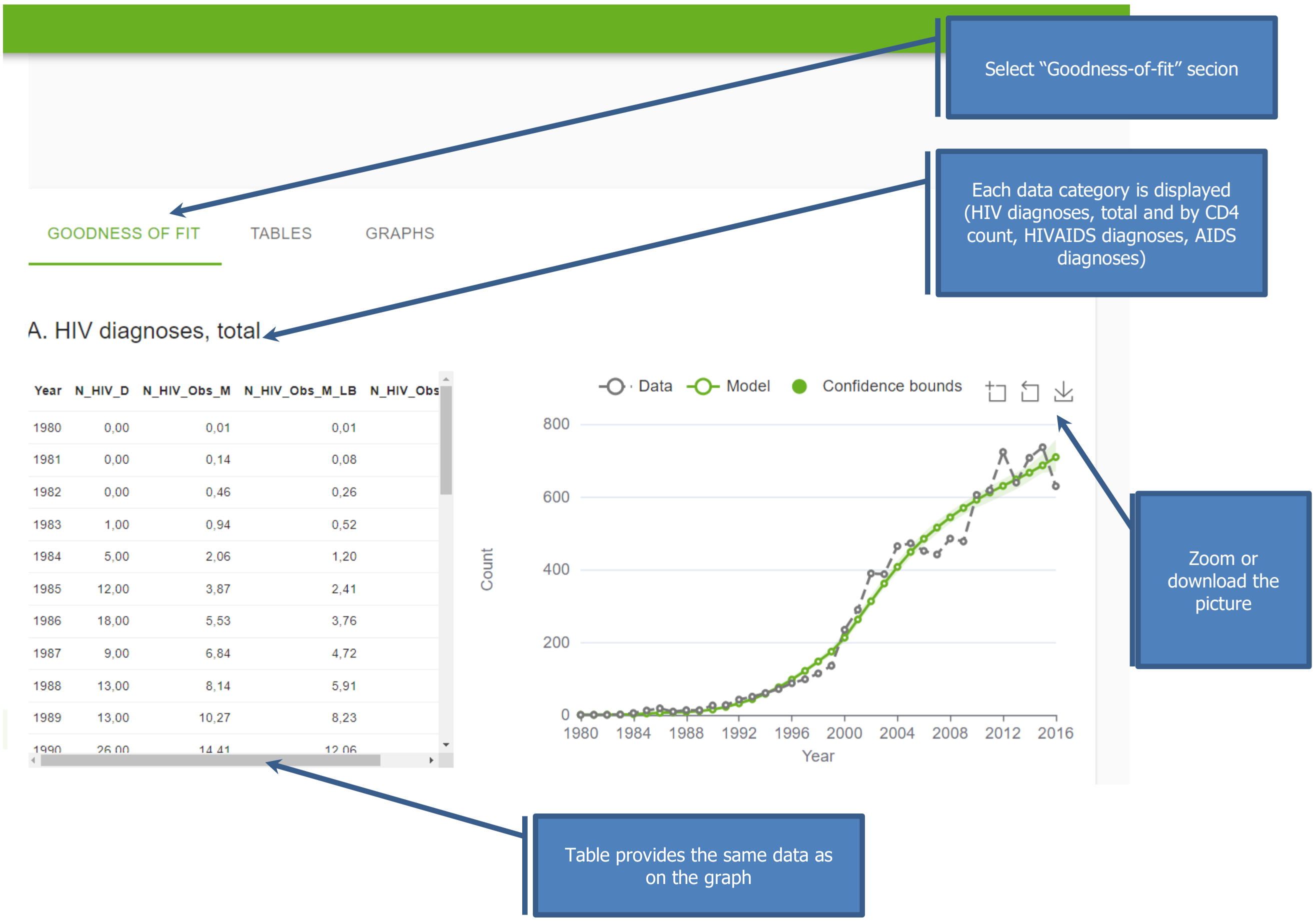
7.10 Tables and charts: inspecting goodness of fit

Description

The "Tables and charts" tab is active both after running the main fit and after running the bootstrap. The difference is that if bootstrap is run, the confidence intervals are also displayed, as in the examples below.
This tab has 3 sections: "Goodness-of-fit" , "Tables" and "Graphs". The first allows to visually inspect how well the model is representing the data. See also section 7.8 for the goodness-of-fit statistic.

Process

"Goodness of fit" tab contains tables and charts comparing the observed data to the fitted values.



A. HIV diagnoses, total

Year	N_HIV_D	N_HIV_Obs_M	N_HIV_Obs_M_LB	N_HIV_Obs
1980	0,00	0,01	0,01	
1981	0,00	0,14	0,08	
1982	0,00	0,46	0,26	
1983	1,00	0,94	0,52	
1984	5,00	2,06	1,20	
1985	12,00	3,87	2,41	
1986	18,00	5,53	3,76	
1987	9,00	6,84	4,72	
1988	13,00	8,14	5,91	
1989	13,00	10,27	8,23	
1990	26,00	14,41	12,06	

Select "Goodness-of-fit" secion

Each data category is displayed (HIV diagnoses, total and by CD4 count, HIVAIDS diagnoses, AIDS diagnoses)

Zoom or download the picture

Table provides the same data as on the graph

Interpretation

The data point should lie in proximity of the model fitted curves, although some variation due to chance is expected, larger if the case counts are small. If the confidence interval is available then most of the observed datapoint should lie within the interval.
Inspecting the charts may provide clues how to improve the model. In case the overall fit is good, but not the fits by CD4 count, it may be reasonable to choose "Change by CD4" option in the diagnosis rate matrix (see section 7.5). If the model fits poorly to a specific time period data, then the diagnosis matrix time periods may be changed.

Total annual number of AIDS diagnoses after 1996 are not used in the model fit because they are likely to be effected antiretroviral treatment and (lack of) adherence to it. Since the treatment process is not modelled, the model prediction is usually not fitting the number of AIDS diagnoses.

Further actions

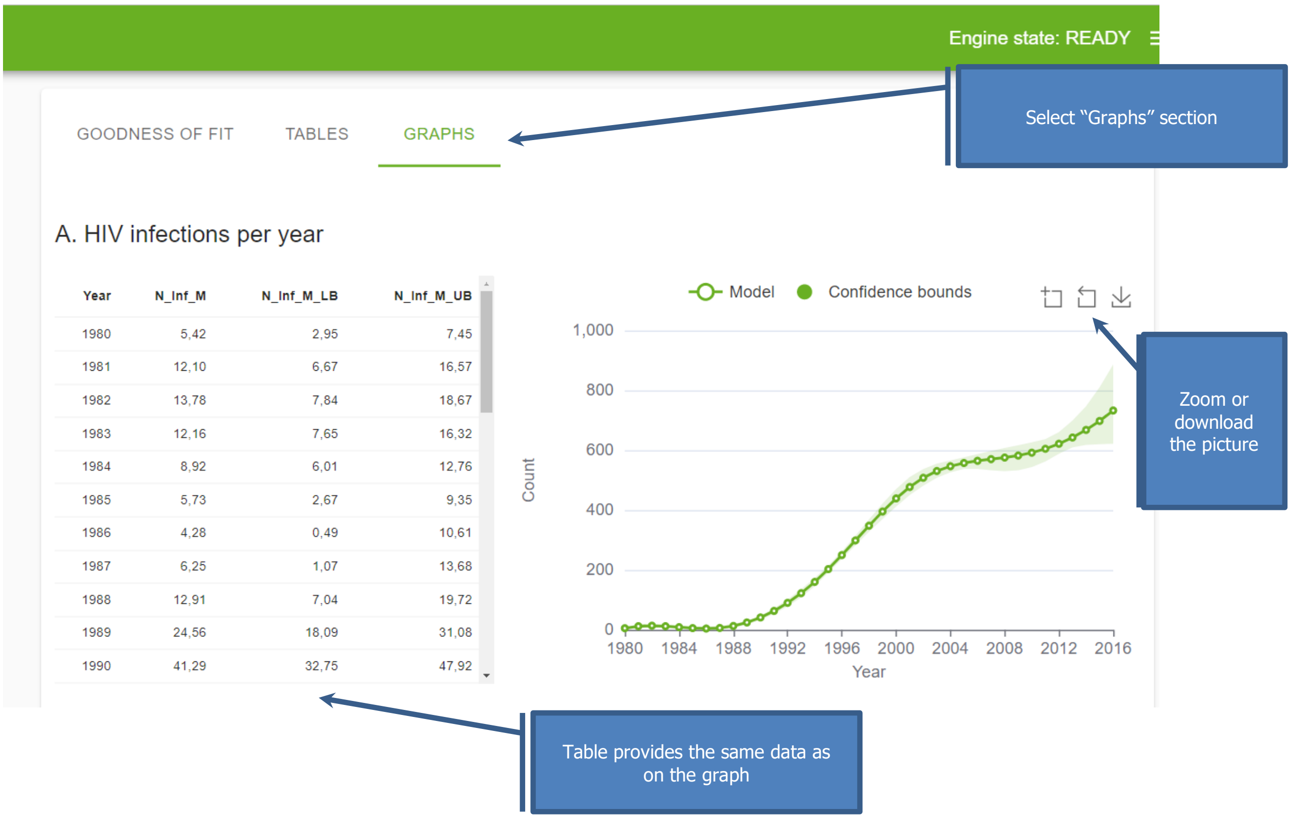
If the model fit is satisfactory, the output charts and tables can be viewed in the "Tables" and "Charts" sections and the output charts and datasets are available for exporting to different formats in the 'Outputs' tab.

7.11 Tables and charts: modelling estimates

Description

The "Tables" and "Charts" sections are active both after running the main fit and after running the bootstrap. The difference is that if bootstrap is run, the confidence intervals are displayed. These section contain the modelling results.

Process



A. HIV infections per year

Year	N_Inf_M	N_Inf_M_LB	N_Inf_M_UB
1980	5,42	2,95	7,45
1981	12,10	6,67	16,57
1982	13,78	7,84	18,67
1983	12,16	7,65	16,32
1984	8,92	6,01	12,76
1985	5,73	2,67	9,35
1986	4,28	0,49	10,61
1987	6,25	1,07	13,68
1988	12,91	7,04	19,72
1989	24,56	18,09	31,08
1990	41,29	32,75	47,92

Select "Graphs" section

Zoom or download the picture

Table provides the same data as on the graph

The following parameters (estimated for each calendar year are displayed:
A. HIV infections per year
Estimated number of HIV infections in each calendar year.
B. Time to diagnosis, by year of infection
Estimated time between infection and diagnosis by year of infection if diagnosis probabilities would remain the same as in the year of infection. By default this graph shows the average time to diagnosis.
C. Total number living with HIV
Estimated number of individuals living with HIV by the end of each calendar year. The three lines include the total number living with HIV (green), the number of diagnosed individuals living with HIV (grey), and the number living with undiagnosed HIV (blue).
Please note that if Full/partial data is set to no, the graph will only show the number of individuals living with undiagnosed HIV.
D. Proportion undiagnosed of all those alive
Percentage of individuals with undiagnosed HIV among those living with HIV. This percentage is equal to the ratio of the blue and the green line in graph.
Please note that if data are not available from the start of the HIV epidemic onwards, it is not possible to determine the total number of people living with HIV. Therefore, this graph is not shown if Full/partial data set to no.

Interpretation

This section serves to review the modelling results. Critical interpretation of the modelling results should take into account external expertise on the developments of the epidemic in the country or studied region. The estimates for the latest years are usually subject to substantial uncertainty and special attention should be paid to the confidence interval limits.

Further actions

Proceed to "Outputs" tab to save the required outputs and return to the modelling inputs tabs to continue modelling additional populations if needed.

8 Reports

8.1 Creating report

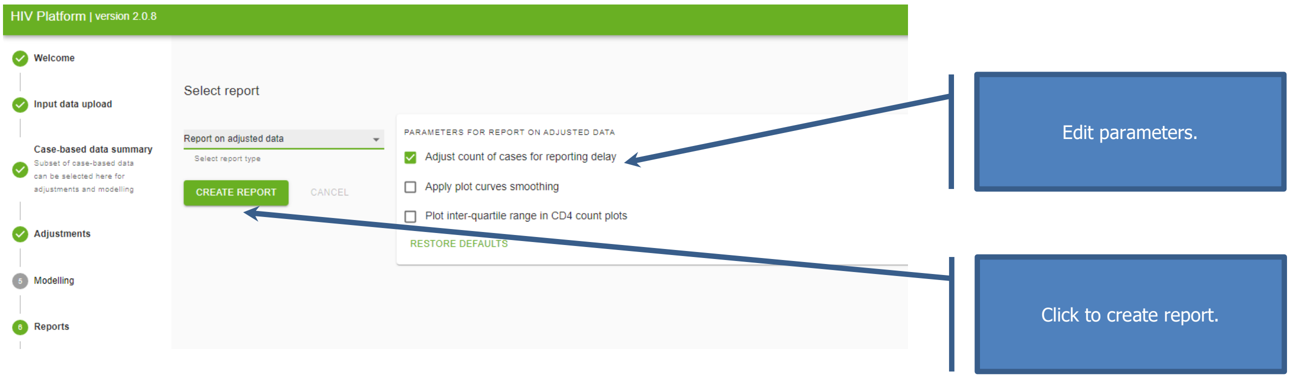
Description

In this tab, a predefined report with main findings of the reporting delay and missing values adjustment is provided. 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.



Select report

Report on adjusted data
Select report type

PARAMETERS FOR REPORT ON ADJUSTED DATA

☑ Adjust count of cases for reporting delay

□ Apply plot curves smoothing

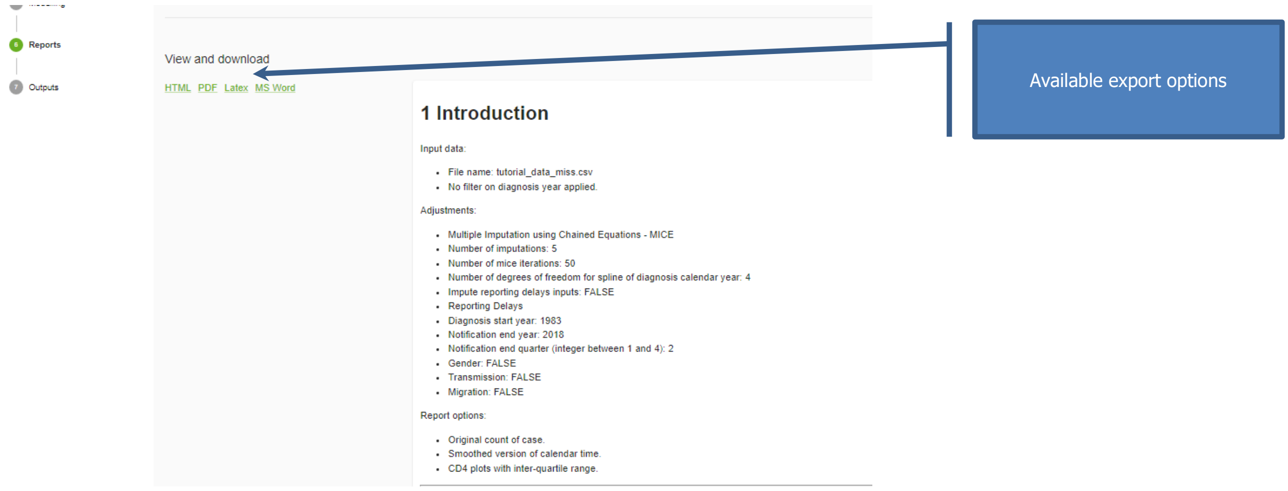
□ Plot inter-quartile range in CD4 count plots

Edit parameters.

Click to create report.

Interpretation

In the first section of the report, all selected options are summarised for convenience. Note that the export options appear on the top.



Available export options

1 Introduction

Input data:

- File name: tutorial_data_miss.csv
- No filter on diagnosis year applied.

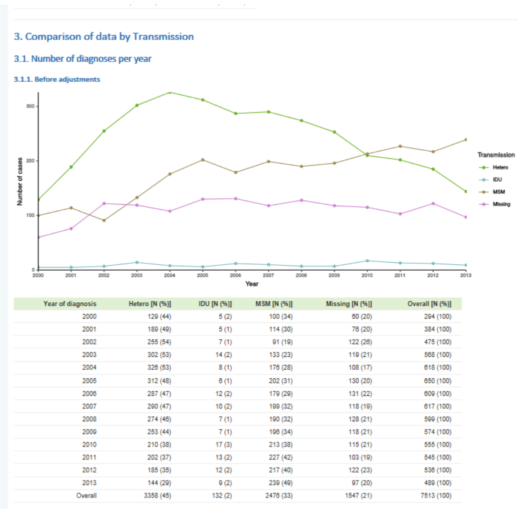
Adjustments:

- Multiple Imputation using Chained Equations - MICE
- Number of imputations: 5
- Number of mice iterations: 50
- Number of degrees of freedom for spline of diagnosis calendar year: 4
- Impute reporting delays inputs: FALSE
- Reporting Delays
- Diagnosis start year: 1983
- Notification end year: 2018
- Notification end quarter (integer between 1 and 4): 2
- Gender: FALSE
- Transmission: FALSE
- Migration: FALSE

Report options:

- Original count of case.
- Smoothed version of calendar time.
- CD4 plots with inter-quartile range.

The following section contain comparisons of trends by covariates (gender, transmission category and migration status as defined by grouping, Section 4.4.) for unadjusted and adjusted data.



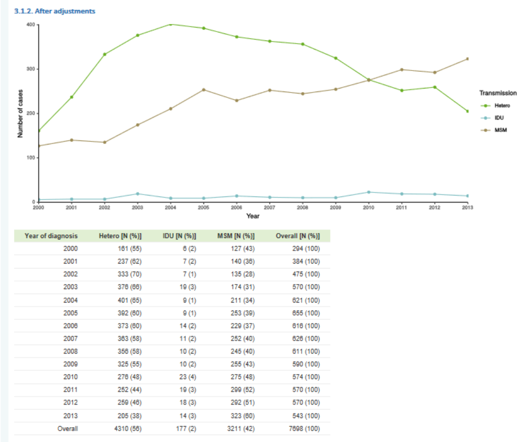
3. Comparison of data by Transmission

3.1. Number of diagnoses per year

3.1.1. Before adjustments

Year of diagnosis	Hetero [N (%)]	IDU [N (%)]	MSM [N (%)]	Missing [N (%)]	Overall [N (%)]
2000	129 (44)	5 (2)	100 (34)	60 (20)	294 (100)
2001	189 (49)	5 (1)	114 (30)	76 (20)	384 (100)
2002	255 (54)	7 (1)	91 (19)	122 (26)	475 (100)
2003	302 (53)	14 (2)	133 (23)	119 (21)	568 (100)
2004	326 (53)	8 (1)	176 (28)	108 (17)	618 (100)
2005	312 (48)	6 (1)	202 (31)	130 (20)	650 (100)
2006	287 (47)	12 (2)	179 (29)	131 (22)	609 (100)
2007	290 (47)	10 (2)	199 (32)	118 (19)	617 (100)
2008	274 (46)	7 (1)	190 (32)	128 (21)	599 (100)
2009	253 (44)	7 (1)	196 (34)	118 (21)	574 (100)
2010	210 (38)	17 (3)	213 (38)	115 (21)	555 (100)
2011	202 (37)	13 (2)	227 (42)	103 (19)	545 (100)
2012	185 (35)	12 (2)	217 (40)	122 (23)	536 (100)
2013	144 (29)	9 (2)	239 (49)	97 (20)	489 (100)
Overall	3358 (45)	132 (2)	2476 (33)	1547 (21)	7513 (100)

Unadjusted data. Note the presence of 'UNK'"transmission category (pink)

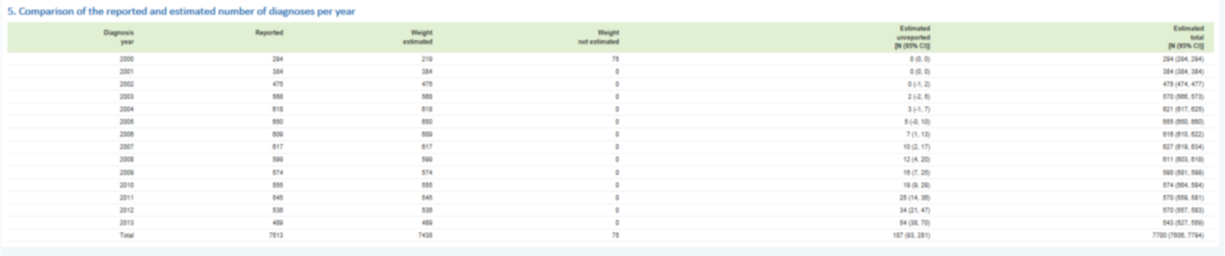


3.1.2. After adjustments

Year of diagnosis	Hetero [N (%)]	IDU [N (%)]	MSM [N (%)]	Overall [N (%)]
2000	161 (55)	6 (2)	127 (43)	294 (100)
2001	237 (62)	7 (2)	140 (36)	384 (100)
2002	333 (70)	7 (1)	135 (28)	475 (100)
2003	376 (66)	19 (3)	174 (31)	570 (100)
2004	401 (65)	9 (1)	211 (34)	621 (100)
2005	392 (60)	9 (1)	253 (39)	655 (100)
2006	373 (60)	14 (2)	229 (37)	616 (100)
2007	363 (58)	11 (2)	252 (40)	626 (100)
2008	356 (58)	10 (2)	245 (40)	611 (100)
2009	325 (55)	10 (2)	255 (43)	590 (100)
2010	276 (48)	23 (4)	275 (48)	574 (100)
2011	252 (44)	19 (3)	299 (52)	570 (100)
2012	259 (46)	18 (3)	292 (51)	570 (100)
2013	205 (38)	14 (3)	323 (60)	543 (100)
Overall	4310 (56)	177 (2)	3211 (42)	7698 (100)

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

9 Outputs

9.1 Adjusted dataset and reporting delay weights

Description

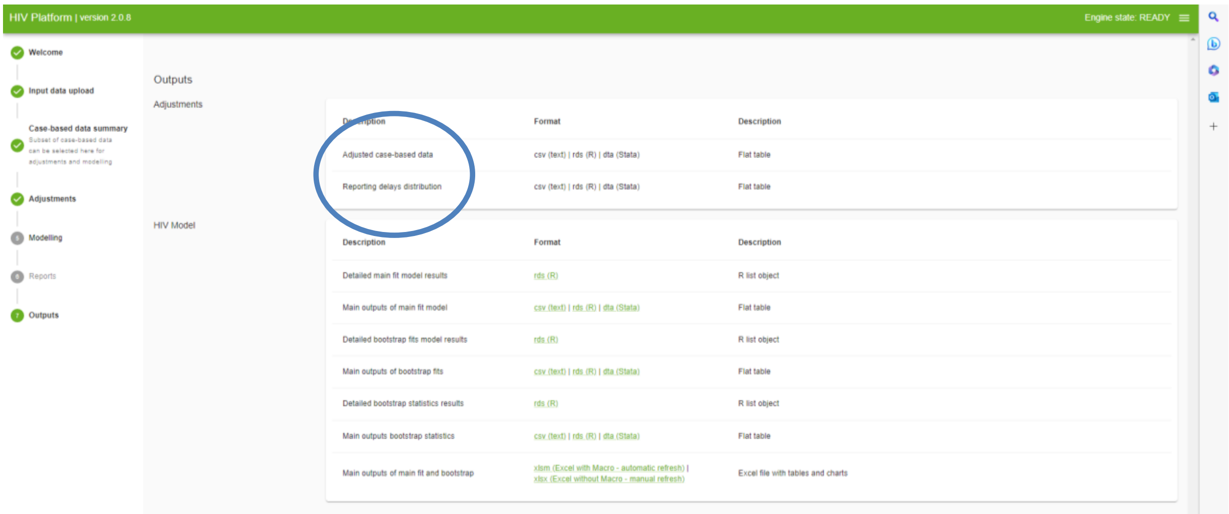
Adjusted case-based dataset. The full dataset with adjustments may be exported. If both imputation and reporting delay adjustments are run, the output data will be a multiply imputed dataset with reporting delay weight. This data set contains the original data as uploaded to the tool, variables created 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 datasets, imputation=1,2).

Reporting delays distribution. 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.

Both datasets 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 dataset and export format from the 'Outputs section'.


HIV Platform | version 2.0.8
Engine state: READY
Welcome
Input data upload
Case-based data summary
Subset of case-based data can be selected here for adjustments and modelling
Adjustments
Modelling
Reports
Outputs

Outputs

Adjustments

Description	Format	Description
Adjusted case-based data	csv (text) | rds (R) | dta (Stata)	Flat table
Reporting delays distribution	csv (text) | rds (R) | dta (Stata)	Flat table

HIV Model

Description	Format	Description
Detailed main fit model results	rds (R)	R list object
Main outputs of main fit model	csv (text) | rds (R) | dta (Stata)	Flat table
Detailed bootstrap fits model results	rds (R)	R list object
Main outputs of bootstrap fits	csv (text) | rds (R) | dta (Stata)	Flat table
Detailed bootstrap statistics results	rds (R)	R list object
Main outputs bootstrap statistics	csv (text) | rds (R) | dta (Stata)	Flat table
Main outputs of main fit and bootstrap	xlsm (Excel with Macro - automatic refresh) | xlsx (Excel without Macro - manual refresh)	Excel file with tables and charts

9.2 Modelling output files

Description

Detailed main fit model results (R list object)

Main outputs of main fit model (Flat table)

Detailed bootstrap fits model results (R list object)

Main outputs of bootstrap fits (Flat table)

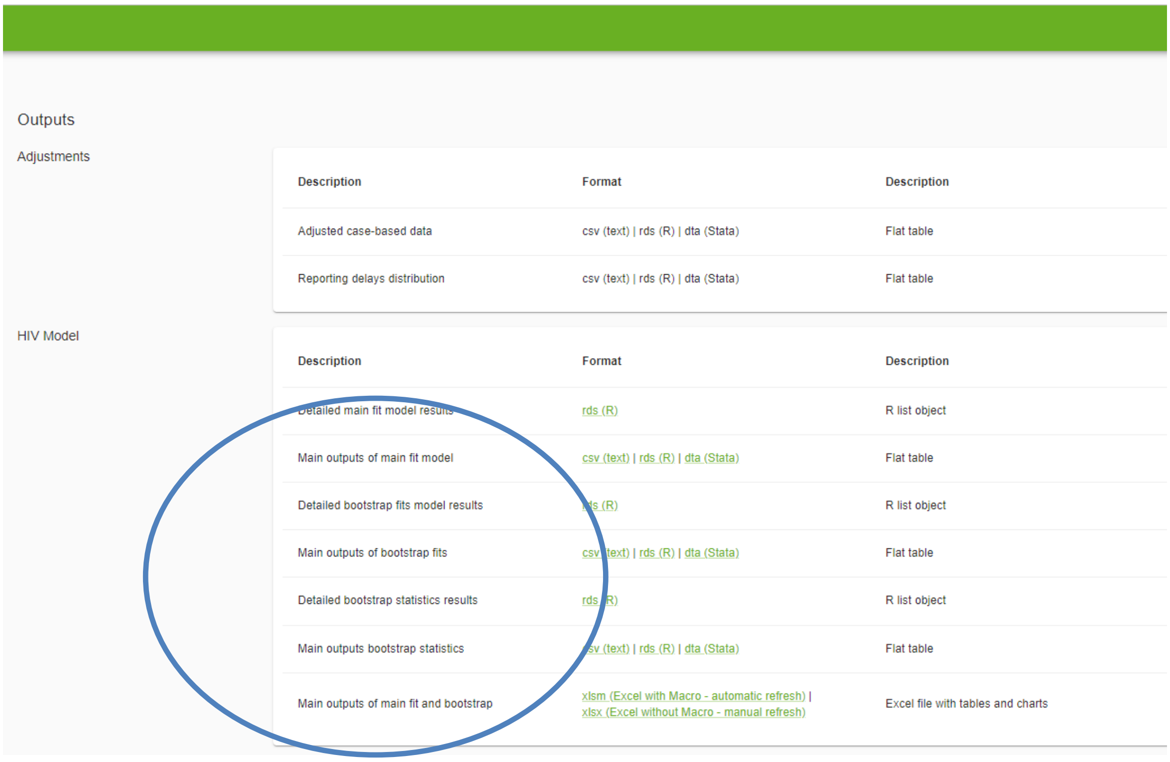
Detailed bootstrap statistics results (R list object)

Main outputs bootstrap statistics (Flat table)

Main outputs of main fit and bootstrap (Excel file with tables and charts). This file contains tables and charts as in the modelling tab, editable in excel.

Process

Select the desired dataset and export format from the 'Outputs section'.



Outputs

Adjustments

Description	Format	Description
Adjusted case-based data	csv (text) | rds (R) | dta (Stata)	Flat table
Reporting delays distribution	csv (text) | rds (R) | dta (Stata)	Flat table

HIV Model

Description	Format	Description
Detailed main fit model results	rds (R)	R list object
Main outputs of main fit model	csv (text) | rds (R) | dta (Stata)	Flat table
Detailed bootstrap fits model results	rds (R)	R list object
Main outputs of bootstrap fits	csv (text) | rds (R) | dta (Stata)	Flat table
Detailed bootstrap statistics results	rds (R)	R list object
Main outputs bootstrap statistics	csv (text) | rds (R) | dta (Stata)	Flat table
Main outputs of main fit and bootstrap	xlsm (Excel with Macro - automatic refresh) | xlsx (Excel without Macro - manual refresh)	Excel file with tables and charts

References

van Sighem A, Nakagawa F, De Angelis D, et al. Estimating HIV incidence, time to diagnosis, and the undiagnosed HIV epidemic using routine surveillance data. Epidemiology. 2015;26(5):653-660. doi:10.1097/EDE.0000000000000324.

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.

Bootstrap and imputations

Little, Roderick J. A., Donald B. Rubin. 2002. Statistical Analysis with Missing Data, Second Edition. Wiley-Interscience: Hoboken, New Jersey.

Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med. 2018 Jun 30;37(14):2252-2266. doi: 10.1002/sim.7654.

Brand, J, van Buuren, S, le Cessie, S, van den Hout, W. Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. Statistics in Medicine. 2019; 38: 210– 220. https://doi.org/10.1002/sim.7956

Bartlett JW, Hughes RA. Bootstrap inference for multiple imputation under uncongeniality and misspecification. Statistical Methods in Medical Research. 2020;29(12):3533-3546. doi:10.1177/0962280220932189

Annex 1. Methodological considerations

This Annex brings together the details about the methods used by the HIV Modelling Platform. Some of these methods were also described in respective documents related to HIV Modelling Tool and HIV Estimates Accuracy Tools.

A1.1 Methodological background for statistical adjustments

Missing values

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) – A value is missing independently of the value itself and of any other factors including observable covariates.
- 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.
- 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.

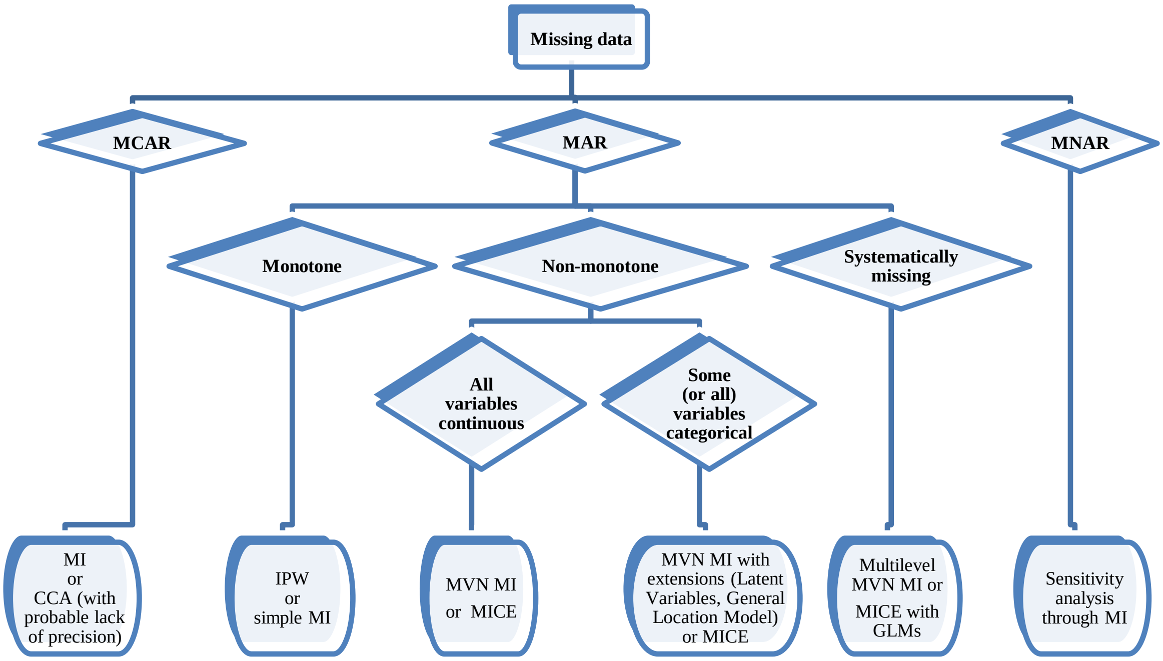
The most popular and flexible method of dealing with missing data (MCAR or MAR) involves multiple imputations (MI), firstly introduced by Rubin in 1987. The MI method involves filling each of the missing values with values randomly sampled from an appropriate distribution. The imputation is performed M times (typically 5–10) and in effect M so called pseudo-complete datasets are obtained. 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 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 3. Appropriate methods to deal with missing data depending on missing data characteristics


Missing data
MCAR
MAR
MNAR
Monotone
Non-monotone
Systematically missing
All variables continuous
Some (or all) variables categorical
MI or CCA (with probable lack of precision)
IPW or simple MI
MVN MI or MICE
MVN MI with extensions (Latent Variables, General Location Model) or MICE
Multilevel MVN MI or MICE with GLMs
Sensitivity analysis through MI

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.

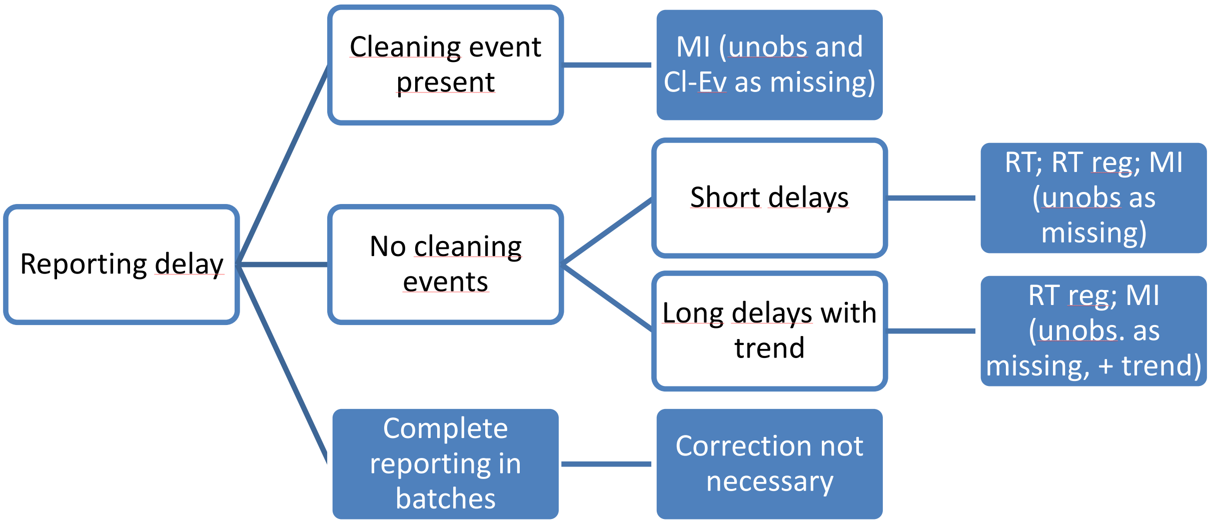
Reporting delay

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 4. Appropriate methods to deal with reporting delays


Reporting delay
Cleaning event present
MI (unobs and Cl-Ev as missing)
No cleaning events
Short delays
RT; RT reg; MI (unobs as missing)
Long delays with trend
RT reg; MI (unobs. as missing, + trend)
Complete reporting in batches
Correction not necessary

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.

Methods used in the 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 dataset. 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.

Some specific issues that may be encountered in the analysis are summarized in the table below.

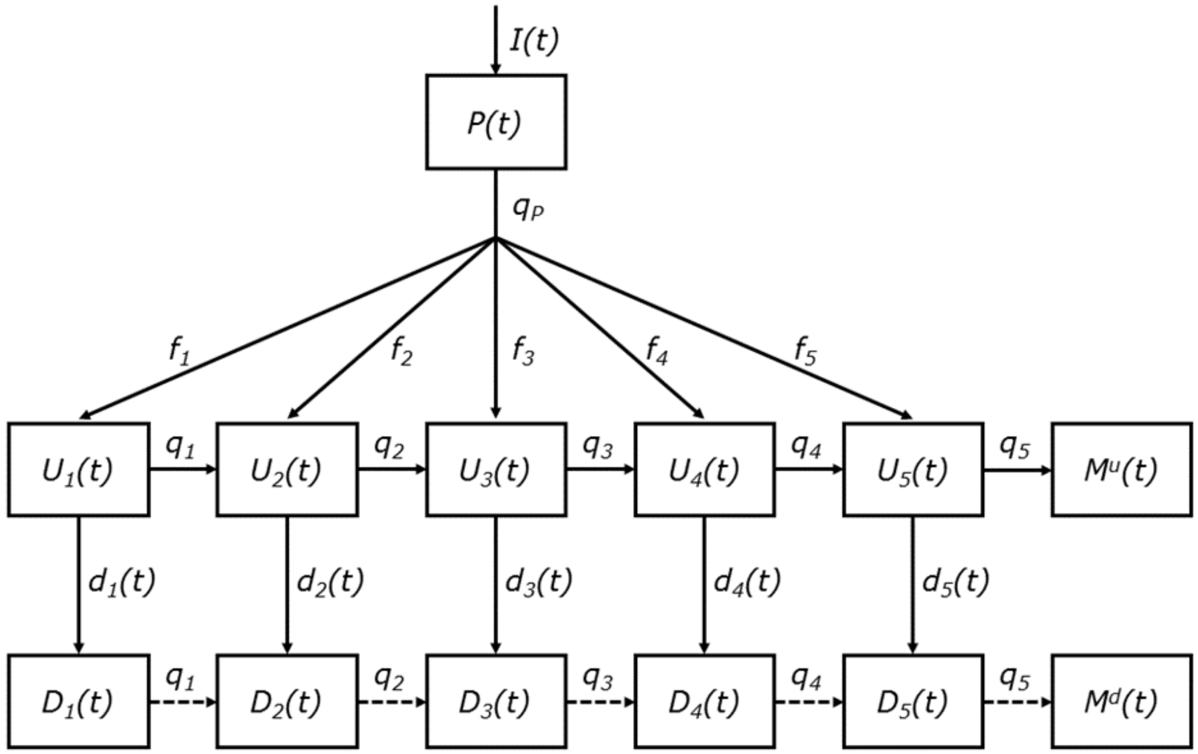
Table 8. Specific issues around adjustments of missing values and reporting delays

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.

A1.2 Modelling method

The tool is based on deterministic compartmental model defined by a system of ordinary differential equations to describe the process of CD4 count depletion in the course of HIV disease. The schematic representation of the model is provided below.

Figure 5. Schematic representation of the model

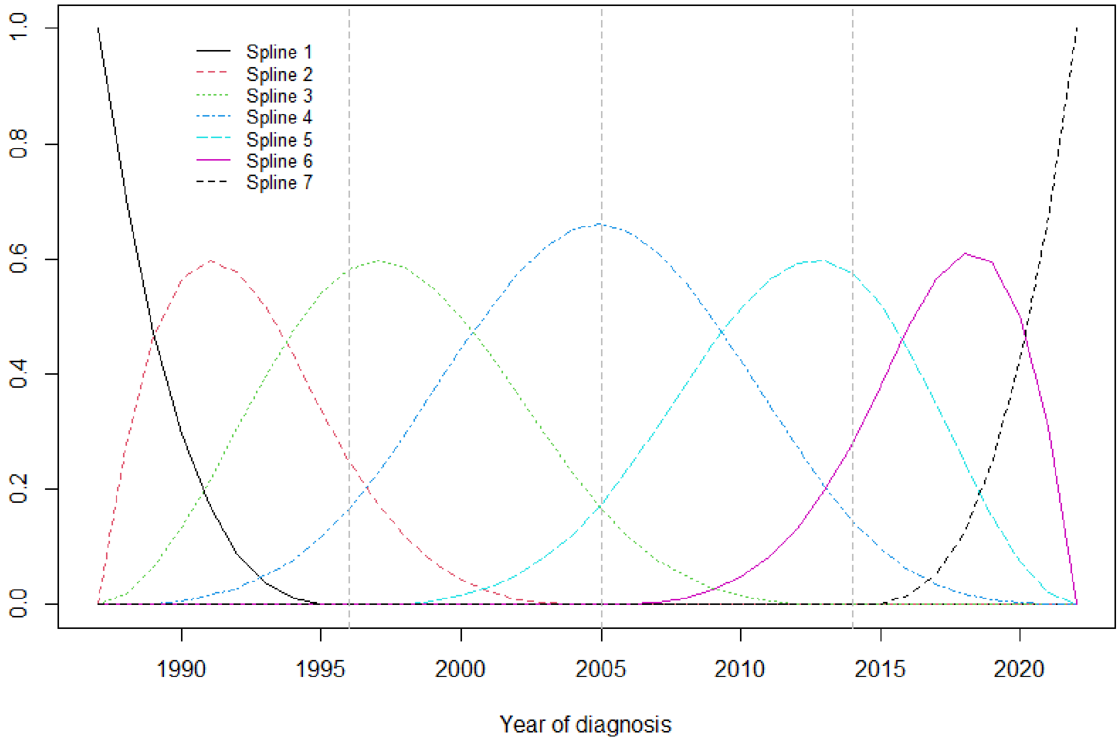


Source: van Sighem A, et al. Estimating HIV incidence, time to diagnosis, and the undiagnosed HIV epidemic using routine surveillance data. Epidemiology. 2015;26(5):653-660. doi:10.1097/EDE.0000000000000324. eAppendix. I(t) – number of incident cases by calendar year, P(t) – primary infection phase, $U_i(t)$ – number of undiagnosed cases whose CD4 count is in respective ranges i: ≥500, 350-499, 200-349, and <200 cells/mm3, $U_5(t)$ – the number of people living with AIDS (undiagnosed). $D_i(t)$ – number of undiagnosed cases whose CD4 count is in respective ranges i: ≥500, 350-499, 200-349, and <200 cells/mm3, $D_5(t)$ – the number of people living with AIDS (diagnosed), $M^u(t)$, $M^d(t)$ – number of deaths among undiagnosed and diagnosed cases respectively. Further f_i, q_i – progression parameters derived from literature.

The model describes progression of new infection (in state I and P), which is undiagnosed, through stages of deteriorating CD4 counts (U_i) to AIDS (U_5) and death (M^U). The progression parameters f_i, q_i are taken from literature. In each state of undiagnosed infection, the case can be diagnosed at the rate d_i. States D_i represent diagnosed cases in one of the CD4 count categories.

The unknowns in this model are estimated by fitting the model to the observed data. They include: incidence over the calendar years (I(t)) and the diagnoses rates by CD4 count $d_i(t)$, which are allowed to change in time.
The I(t) is defined as a superposition of k+4 cubic B-splines, where k is the number of internal knots. The number of knots, k, is a user defined parameter.
Cubic splines are polynomial functions of the form $Spline = a * t^3 + b * t^2 + c * t + d$, where t is calendar time and a, b, c and d are constant numbers. By definition, the spline has a non-zero value within a certain time interval that is specified by knots. Outside this time interval the function is zero.

The figure below shows the splines that are created by the tool in case the range of calculations is between 1987 and 2022 and the numbers of knots selected is 4. The knot placement is indicated by the dashed grey lines.



Note that the number of knots defined by the user refers to the so called internal knots. Thus if the user defines 3 knots, they divide the time range into 4 time periods.

The tool places the knots equally spaced throughout the time range used for calculation.

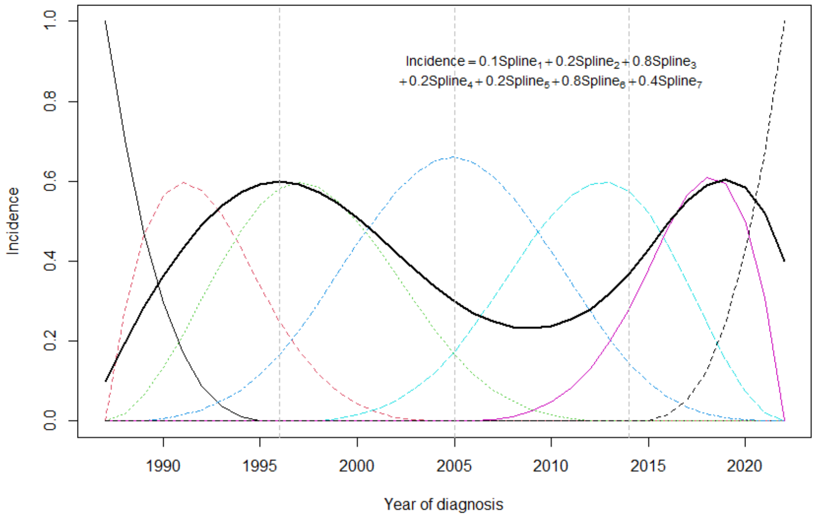
For the 3 knots the incidence curve is modelled as:

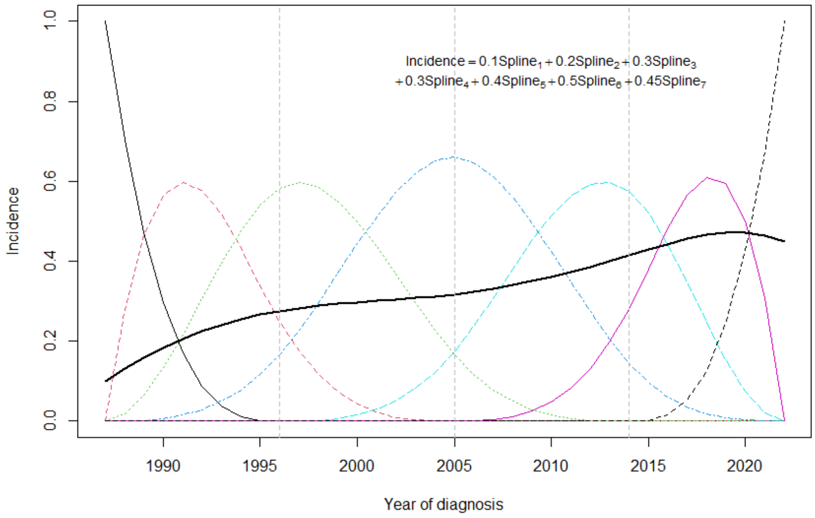
$$Incidence = \sum_{i=1..7} \theta_i * Spline_i$$

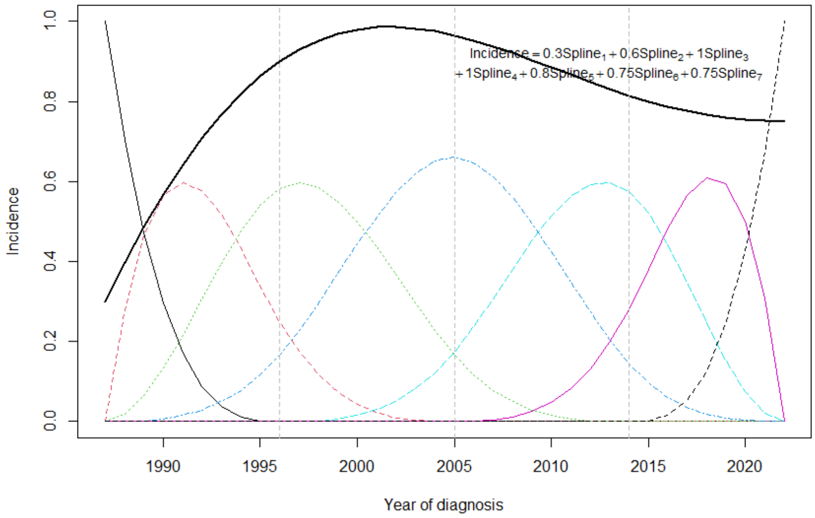
Adding up all the different splines, each with a different weight, θ_i, determines the shape of the incidence Curve. The parameters θ_i are estimated during the model fitting procedure (see below) and the final number of infections in a given year is provided based on this curve.
A property of cubic B-splines is that if all weight parameters, θ_i, are equal, the incidence curve will be a horizontal line. Another property, which is used by the tool to control the incidence curve at the end of the calendar year range, is that straight lines can be obtained by setting $\theta_i = 2 * \theta_{i-1} - \theta_{i-2}$

The estimation precision depends on the amount of data available and generally is lesser at the end of the curve. Given the shape of the Spline 7 (with very high values at the very end of the interval) this lack of precision could cause extreme values to appear at the end of the calculation range. To avoid this undesirable feature a constraint is put on the last parameter fixing it as linear combination necessary to obtain flat line, i.e. in our example $\theta_7 = 2 * \theta_6 - \theta_5$.

The spline approach gives significant flexibility to the shape of incidence curve, which may be especially useful in case of epidemic which has had several peaks. The example curves are shown on the pictures below, with the incidence equation generating these curves specified.


$\text{Incidence} = 0.1\text{Spline}_1 + 0.2\text{Spline}_2 + 0.8\text{Spline}_3 + 0.2\text{Spline}_4 + 0.2\text{Spline}_5 + 0.8\text{Spline}_6 + 0.4\text{Spline}_7$
Incidence
0.0
0.2
0.4
0.6
0.8
1.0
1990
1995
2000
2005
2010
2015
2020
Year of diagnosis


$\text{Incidence} = 0.1\text{Spline}_1 + 0.2\text{Spline}_2 + 0.3\text{Spline}_3 + 0.3\text{Spline}_4 + 0.4\text{Spline}_5 + 0.5\text{Spline}_6 + 0.45\text{Spline}_7$
Incidence
0.0
0.2
0.4
0.6
0.8
1.0
1990
1995
2000
2005
2010
2015
2020
Year of diagnosis


$\text{Incidence} = 0.3\text{Spline}_1 + 0.6\text{Spline}_2 + 1\text{Spline}_3 + 1\text{Spline}_4 + 0.8\text{Spline}_5 + 0.75\text{Spline}_6 + 0.75\text{Spline}_7$
Incidence
0.0
0.2
0.4
0.6
0.8
1.0
1990
1995
2000
2005
2010
2015
2020
Year of diagnosis

On the other hand $d_i(t)$ are defined as step functions or linear functions on time intervals specified by the user. The presumed shape of $d_i(t)$ is defined by the diagnosis matrix, see Section 7.5.

The equations are solved numerically, and the maximum likelihood methods are used to find the values of parameters. The likelihood is constructed assuming that the number of new diagnoses has either Poisson or negative binomial distribution with a mean value corresponding to an integral of d_i*U_i over one year period. The maximum of the likelihood function is found with downhill simplex optimisation algorithm. The algorithm is started from various starting values to ensure that the optimisation is robust and that local optima are avoided.
The tool will do an initial run to determine if there are any splines with a very small weight parameter.
Since such splines will not contribute significantly to the incidence curve their weight parameter is then set to zero. If complete surveillance data are not available from the start of the HIV epidemic (see Full/partial data), the tool will do additional runs to determine the number of weight parameters that can be set to zero before the model fit to the data gets worse.

In case the procedure does not converge the number of knots is reduced by one and the procedure is re-started. In case a parameter for a specific spline is very close to 0 it is set to 0 and the other parameters are re-estimated.

A1.3 Bootstrap confidence intervals for modelling with imputed data

Original HIV Modelling Tool (https://www.ecdc.europa.eu/en/publications-data/hiv-modelling-tool) relied on parametric bootstrap to estimate the confidence intervals and this method is used if the HIV Modelling Platform is used with aggregated input datasets. On the other hand if the case-based data are provided the both multiple imputation and bootstrap can be performed on the case based data. In this case non-parametric bootstrap can be applied, i.e. cases can be sampled with replacement prior to modelling. There is a number of options that could be used to approach bootstrap (BS) in the context of multiple imputations (MI). Accordingly, the validity in terms of coverage of the resulting confidence intervals of combining bootstrap and multiple imputations was reviewed by several authors for various scenarios.

According to the classical theory of Little – Rubin the combining of the imputations with the bootstrap should run as follows:

1. Generate bootstrap samples from the unimputed data;
2. Impute missing values in each bootstrap sample;
3. Run MI analyses in each of the bootstrap samples.

In a study of Schomaker the following approaches are considered

I. B bootstrap samples of the original data set (including missing values) are drawn and in each of these samples the data are multiply imputed (M sets). Latter to achieve the final estimates there are 2 options:
   a. as proposed by Little-Rubin – create a single estimate using imputed data for each of the bootstrapped datasets (**BOOT MI**) OR
   b. B × M estimates of the pooled data are used for interval estimation (**BOOT MI POOLED**)

II. M imputed datasets are created and bootstrap estimation is applied to each of them
   a. Estimate standard error based on bootstrap (t-method) in each imputed data set and apply the standard MI combining rules (**MI BOOT**) OR
   b. the B × M estimates could be pooled and 95% confidence intervals could be calculated based on the 2.5th and 97.5th percentiles of the respective empirical distribution (**MI BOOT POOLED**)

Simulation results (MAR assumption) confirm that:
- the time needed for estimation is always longer form BOOT MI than for MI BOOT
- point estimates are unbiased for all methods
- confidence intervals have coverage probabilities close to the nominal level except for higher levels of missingness (here 38% and 45% for two covariates with missing values). In these settings Boot MI, Boot MI POOLED performed better then MI Boot and MI Boot POOLED. MI Boot to reach the nominal coverage M's of 20 or higher were required. For **MI Boot and MI Boot POOLED it is recommended to use larger M.-**
- Boot MI may perform well even for M < 5, but the Boot MI POOLED has a tendency towards coverage probabilities > 95%. Boot MI POOLED is an inefficient estimate which follows from MI theory. In this case each of

the estimated parameter values are treated as an estimate of the parameter, i.e. we say that we take more bootstrap samples, while only 1 imputation (M=1).
The MI theory said that the relative efficiency of the MI estimator compared to true variance is:

$$(1+\frac{\gamma}{M})^{-1}$$

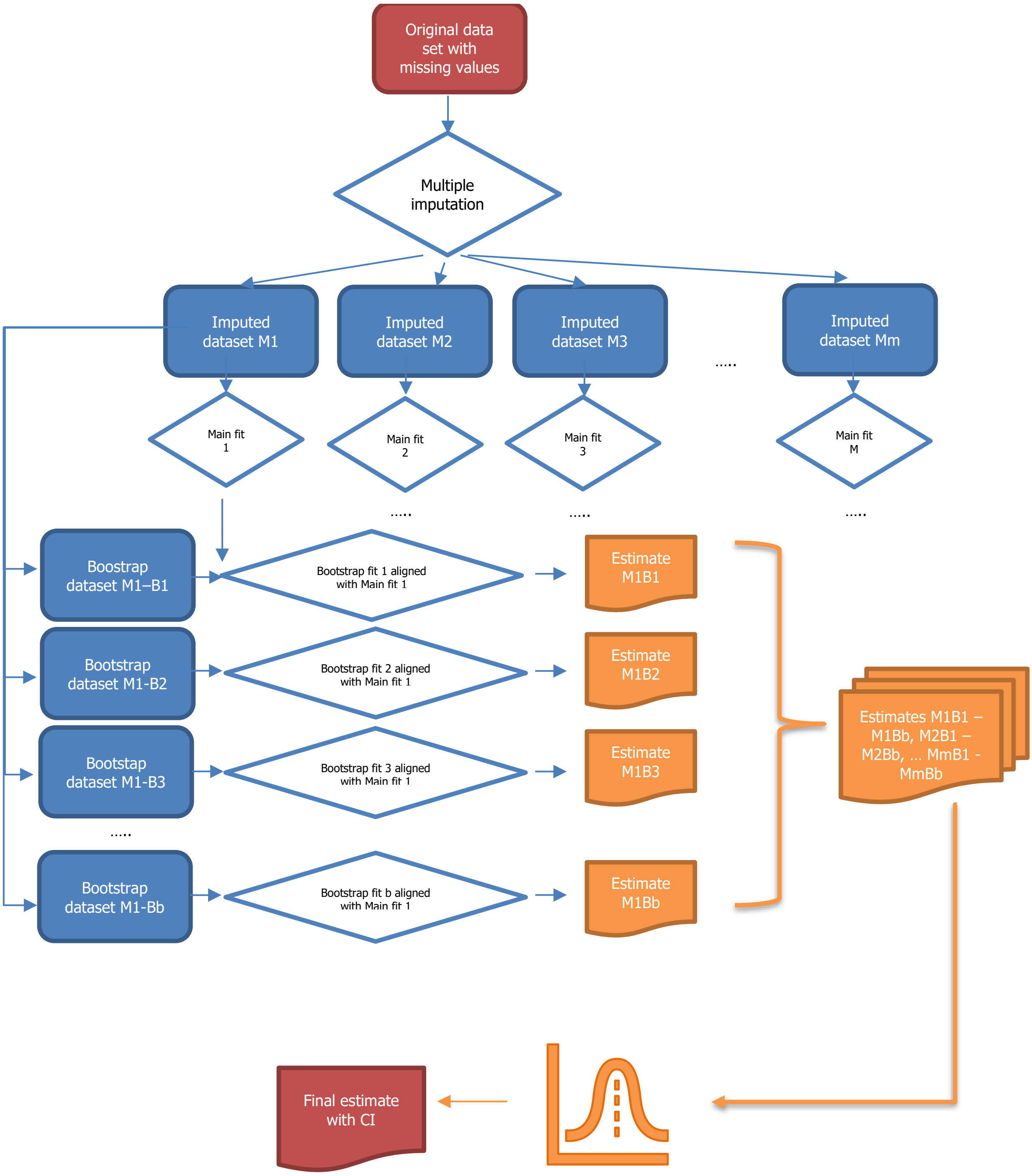
Where γ is the fraction missing.

The paper of Brand also mentions a simplification of **Boot MI (Boot MI S)** in which only a single imputation is made in each of the bootstrapped datasets. They argue that this single imputation accounts for the uncertainty both of the imputed missing values and of underlying distribution being bootstrapped.
Of note they also find all methods unbiased, with a different data structure then considered in the simulation studies of Schomaker. Lower coverage of confidence intervals was found for **MI Boot** when the percentile confidence intervals were used in each imputed set and then averaged; and **Boot MI S** with t-method for confidence interval but not with the percentile method. The previous study indicated that MI Boot POOLED coverage could be improved with increasing number of imputed datasets (here only 5 were used).

The paper of Bartlett also considered these methods under the violation of the model congeniality assumption (i.e. when the imputation model differs from the analysis model). For all the methods the point estimates were unbiased. However, only the Boot MI percentile (with moderate M) and von Hippel approaches give intervals with nominal coverage (provided the point estimator is consistent). All of the other methods either under- or over-cover, depending on the particular situation. Nonetheless the divergence from the nominal coverage was not very high, with 95% confidence interval coverage being at least 89%.
The approaches for which the bootstrap is run before the multiple imputations are associated with the very high computational costs, in our case mainly associated with the time needed to produce imputed dataset. The modelling would be then associated with prohibitively long time.

Therefore, in the tool we used the MI Boot POOLED procedure.
Briefly for each imputed dataset there is the main fit performed. Next the replicated datasets are created either through parametric or non-parametric bootstrap procedure. For each replicated dataset the tool starts with the main fit model for the respective imputed dataset. The number of model parameters is not allowed to change in within the bootstrap procedure, i.e. simplification of the model to obtain convergence is not allowed as for the main fit. A replicated dataset, for which the convergence is not obtained is discarded. If more than 10% of the replicated datasets fail with the main fit model specification, the tool will issue an error message.
Once the fits are obtained for the required number of replicated datasets for all imputed datasets, each is used to derive a set of estimates of interest including annual number of incident cases, annual number of PLWH, annual number of undiagnosed PLWH, etc. This means that for each such parameters the procedure outputs MxB values (M – number of imputations, B – number of bootstrap iterations). As per the the MI Boot POOLED procedure 95% confidence interval is derived by taking 2.5% and 97.5% percentiles of all MxB estimates.

Figure 6. Visual representation of MI BOOT procedure implemented to obtain confidence intervals around modelling estimates as described above.



Annex 2. 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

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
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
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