ECDC HIV Modelling Tool User Manual

Version 1.2

European Centre for Disease Prevention and Control

29th of June 2016

Table of Contents

1 Introduction

The ECDC HIV Modelling Tool is a new tool developed by ECDC in collaboration with international partners to provide estimates of the number of people living with HIV, including those not yet diagnosed. The tool can also estimate the annual number of new HIV infections, the average time between infection and diagnosis, and the number of people in need of treatment according to CD4 counts.

To achieve all of this, the tool needs only routinely collected HIV surveillance data. Nearly all countries in the European region report annual HIV and AIDS diagnoses to the TESSy database hosted at ECDC.

The tool produces models based on two different methods. The first method is the so-called Incidence Method, which requires most data but can also provide most detailed estimates[1]. This method first estimates HIV incidence over time and time to diagnosis by CD4 count strata and then estimates the undiagnosed HIV-positive population.

The second method is the London Method[2]. This method requires less data, typically surveillance data for just one year, but, as a result, can only provide less detailed estimates of the undiagnosed population. In particular, the method only makes estimates of the number of undiagnosed individuals in immediate need of antiretroviral treatment, based on CD4 count thresholds of 200 or 350 cells/mm^3, which are also given by the Incidence Method.

1.1 Incidence Method

The Incidence Method uses a mathematical model to estimate the annual number of HIV infections, the probability of being diagnosed with HIV depending on the CD4 count level, and the time between infection and diagnosis. This mathematical model describes the progression of HIV from the time of infection until diagnosis or development of AIDS in the absence of antiretroviral treatment.

The number of HIV infections, the diagnosis probability, and the time to diagnosis are determined by a set of parameters. These parameters are *a priori* unknown and need to be estimated by comparing expected model outcomes with observed data on HIV and AIDS diagnoses. For instance, in a specific calendar year the number of observed HIV diagnoses with CD4 counts above 500 cells/mm^3 can be compared with the model expectation. By repeatedly changing the values of the unknown parameters and comparing expected model outcomes with observed data, a process called fitting, the best-matching set of parameters can be determined. This fitting process takes time and the more parameters need to estimated, the longer the method will take.

Once the number of HIV infections and the time to diagnosis has been estimated, the Incidence Method calculates other outcomes of interest like the number of individuals living with HIV, including those not yet diagnosed.

1.2 London Method

The London Method is based on the assumption that undiagnosed HIV-positive individuals who develop AIDS or other HIV-related symptoms of sufficient severity will present for care and be diagnosed with HIV as a result[2]. The rate at which such symptoms develop depends on the CD4 count level and is known from data about untreated HIV-positive individuals. From the observed number of symptomatic HIV diagnoses in a specific CD4 interval, the tool is able to estimate the total number of HIV-positive individuals with CD4 counts in that particular interval. This method only works for CD4 count levels for which the rate of symptoms is sufficiently large, i.e., for CD4 counts below 350 cells/mm^3.

2 Technical details and installation

The ECDC HIV Modelling Tool is a standalone application executed directly on the user's computer with Windows© operating system. It is a 32-bit application and both 32- and 64 bit versions of Windows Vista SP2, Windows 7 SP1, Windows 8 and later are supported.

2.1 Prerequisites

ECDC HIV Modelling Tool combines an interface written in C# utilizing Microsoft .Net technology with a calculation engine written in C. This imposes certain dependencies on the operating system environment. Two packages must be installed prior to running the tool (both Microsoft products):

1. **Microsoft .Net Framework 4.5.1 or later**
   Download link: http://www.microsoft.com/en-us/download/details.aspx?id=40779
   This version is included in Windows 8.1 by default.

2. **Visual C++ Redistributable Packages for Visual Studio 2013**
   Download link: https://www.microsoft.com/en-us/download/details.aspx?id=40784
   Choose file `vcredist_x86.exe` for download even for 64-bit versions of Windows.

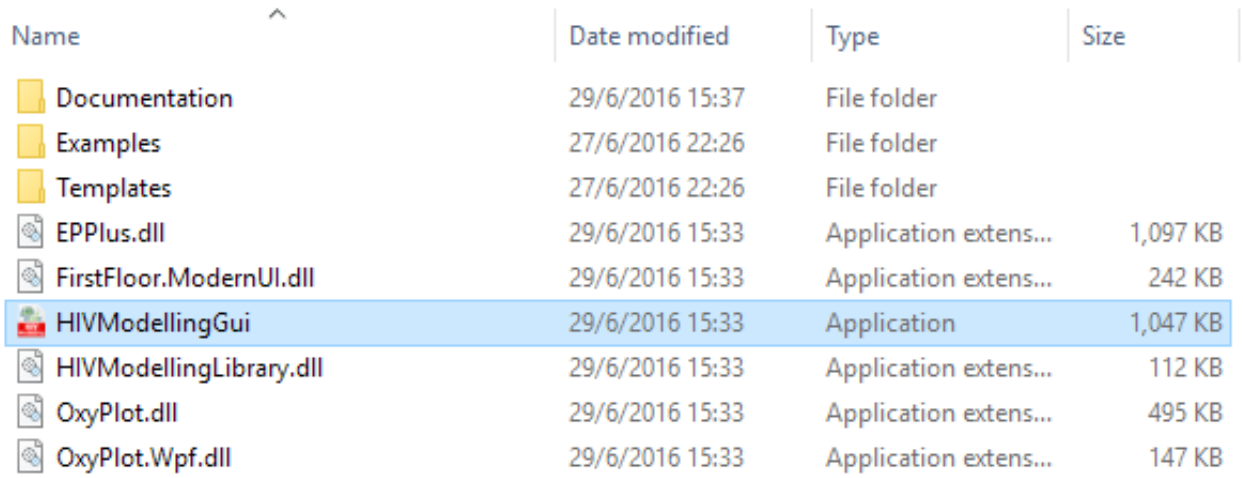
Chances are that these two packages are already installed as prerequisites to other software.

2.2 Installation

The ECDC HIV Modelling Tool can be downloaded from http://ecdc.europa.eu/en/healthtopics/aids/Pages/hiv-modelling-tool.aspx. The latest version (v1.0.2 as of 24 February 2016) is mentioned on the top of that list. The software is distributed as a zip-archive of size around 5MB and requires only a few simple installation steps requiring no computer administration rights:

1. Create a folder in which to install the tool, for example `D:\My Documents\HIV Modelling Tool`.
2. Download the zip-file with the latest version of the tool and save it to the folder created in step 1.
3. Unzip the downloaded zip file by selecting `Extract here` as an option available with a right-click on the file.

The tool zip package contains 6 files in the root folder: one executable `HIVModellingGui.exe` and five auxiliary library files with extension `dll`:



Name	Date modified	Type	Size
Documentation	29/6/2016 15:37	File folder	
Examples	27/6/2016 22:26	File folder	
Templates	27/6/2016 22:26	File folder	
EPPlus.dll	29/6/2016 15:33	Application extens...	1,097 KB
FirstFloor.ModernUI.dll	29/6/2016 15:33	Application extens...	242 KB
HIVModellingGui	29/6/2016 15:33	Application	1,047 KB
HIVModellingLibrary.dll	29/6/2016 15:33	Application extens...	112 KB
OxyPlot.dll	29/6/2016 15:33	Application extens...	495 KB
OxyPlot.Wpf.dll	29/6/2016 15:33	Application extens...	147 KB

The tool can be started by double-clicking the file `HIVModellingGui.exe`. Once the tool is running one

can inspect the about tab. Successful installation can be confirmed by inspecting if the version numbers for the GUI and Library are printed:

A missing Library version number indicates that package 2 of the prerequisites listed above (Visual C++ Redistributable Packages for Visual Studio 2013) is not installed. It should be installed prior to running calculations with the tool.

2.3 Support

For technical support and reporting problems please contact HIVModelling@ecdc.europa.eu.

3 Input data sets

This section describes the data sets that need to be prepared before running the tool. Data are typically retrieved from national or regional HIV surveillance systems. Two examples of input data sets are available in the zip-archive containing the tool.

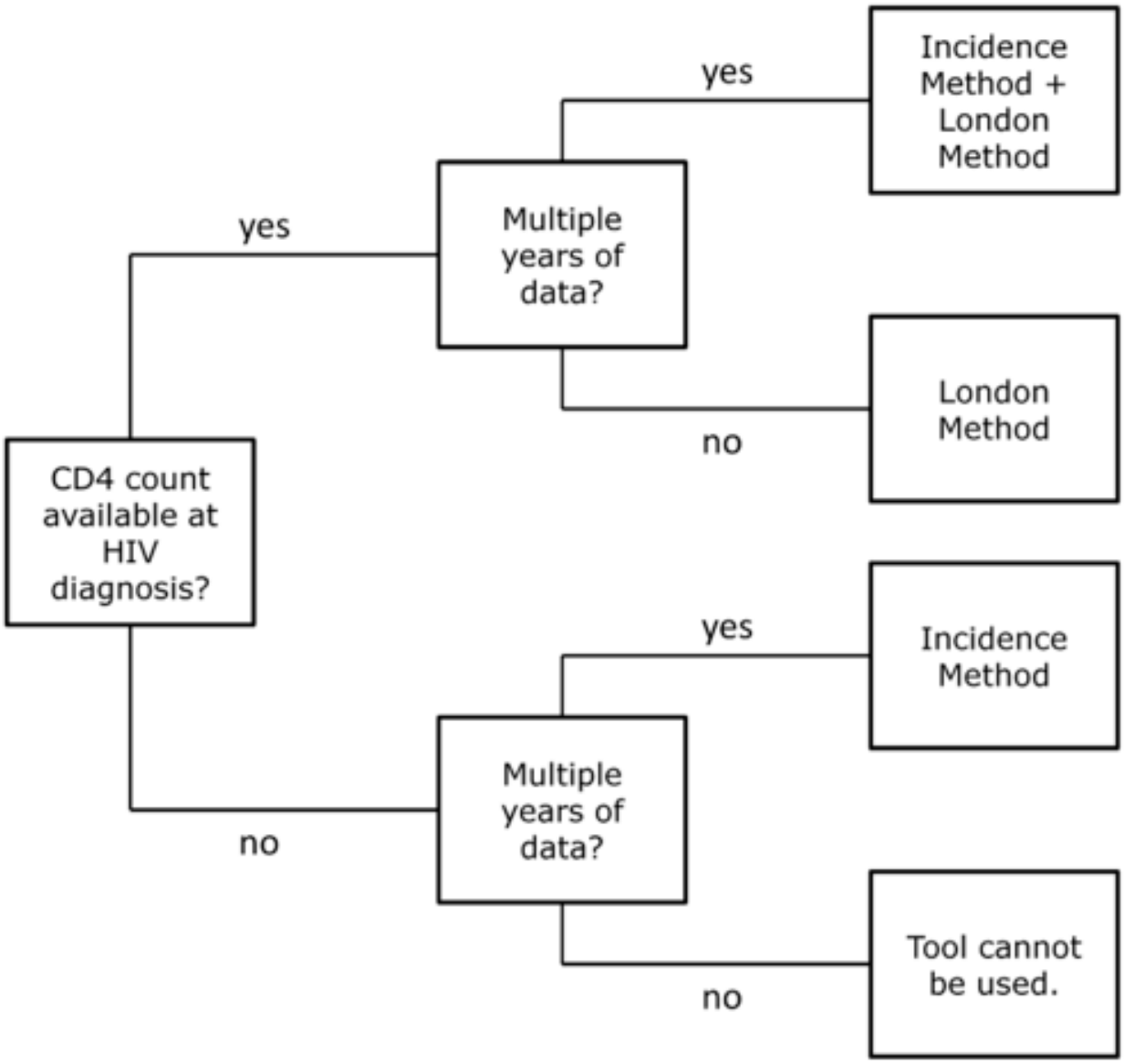
3.1 Datasets

3.1.1 Before creating datasets

Creating the datasets that are required by the two methods may be a considerable amount of work. To avoid unnecessary preparation of datasets there are two main issues to consider:

- Are there data available on CD4 counts at the time of HIV diagnosis?
- Are surveillance data available for multiple calendar years?

Depending on the answers to these questions, one or both methods may be applicable according to the scheme below.



The Incidence Method requires surveillance data over multiple years, ideally covering the duration of the HIV epidemic in a country. This method will work both with or without CD4 counts at the time of diagnosis, although the first option is the preferred one. The Incidence Method will also work if data on CD4 counts are only available for several years.

The London Method requires at least one year of data, including CD4 counts at the time of diagnosis. This method will also work with multiple years of data.

Both the Incidence Method and the London Method are based on CD4 cell decline in adult HIV-1-positive individuals. Therefore, *HIV-2-positive individuals and HIV-positive children below 16 years of age should be excluded* from the datasets.

3.1.2 Populations

Before using the tool, the user may define *populations* in 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. An indication for

differences in time to diagnosis may be 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 in the Incidence Method will be an average over the total population.

Populations can be based, for instance, on transmission route: men who have sex with men, heterosexual men and women, injecting drug users. Other examples of populations are native HIV-positive individuals and migrants, or people living HIV in a specific city. The choice of populations will depend on the nature of the HIV epidemic in the country.

When defining populations, the user should also consider the size of the population. The outcomes of the ECDC HIV Modelling Tool 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 in a next step consider smaller populations.

3.1.3 Preparation of datasets

The Incidence Method and the London Method each require their own datasets. There is no overlap between these datasets, although they will most likely be extracted from the same surveillance system.

Incidence Method

The following datasets are necessary for the Incidence Method:

Required

Data set	Description
HIV	total annual number of HIV diagnoses
AIDS	total annual number of AIDS cases
HIVAIDS	annual number of HIV diagnoses with a concurrent AIDS diagnosis, i.e., an AIDS diagnosis within, for instance, 3 months after HIV diagnosis

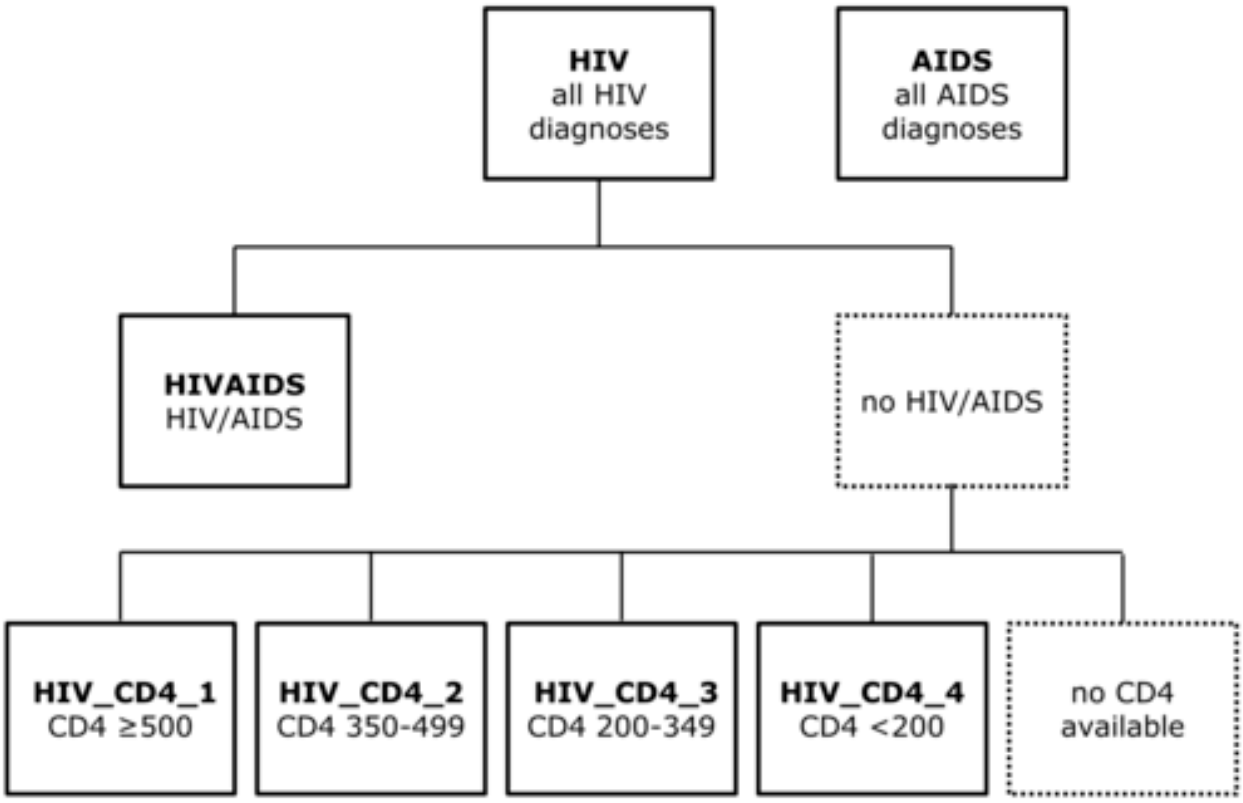
Recommended

Data set	Description
HIV_CD4_1	annual number of HIV diagnoses with CD4 $\geq$ 500 cells/mm^3 and no concurrent AIDS diagnosis
HIV_CD4_2	annual number of HIV diagnoses with CD4 350-499 cells/mm^3 and no concurrent AIDS diagnosis
HIV_CD4_3	annual number of HIV diagnoses with CD4 200-349 cells/mm^3 and no concurrent AIDS diagnosis
HIV_CD4_4	annual number of HIV diagnoses with CD4 < 200 cells/mm^3 and no concurrent AIDS diagnosis

Optional

Data set	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 determined by the tool from the other data items.



The dataset **HIV** contains all HIV diagnoses per year.

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_1 to **HIV_CD4_4** contain HIV diagnoses with *no* concurrent AIDS diagnosis and with CD4 counts at the time of diagnosis in the specified range.

Notes

- It is possible to run the Incidence Method if there are no data on CD4 counts at the time of diagnosis, but this is not recommended.
- The Incidence Method calculates the number of HIV-positive individuals who are still alive by subtracting the number who died, as given in dataset **Dead**, from the estimated number of individuals ever infected. Therefore, if the annual number of deaths among diagnosed individuals is (partly) missing, the Incidence Method cannot correctly determine the number of HIV-positive individuals who are still alive. This does, however, not affect estimation of the annual number of new infections.

London Method

The London Method is based on the assumption that undiagnosed HIV-positive individuals who develop AIDS or other HIV-related symptoms of sufficient severity, or which are sufficiently specific to HIV, will present for care and be diagnosed with HIV as a result[2]. Such HIV-related symptoms would typically refer to those listed as category B and C conditions (CDC-B and -C events) in the 1993 revised CDC classification system[3]. Only symptoms which are assumed to be caused by HIV are of interest, so symptoms related to a bacterial sexually transmitted infection, for example, should not count.

Ideally, the London Method would require data on the number of symptomatic HIV diagnoses. However, quite often the only data that are available in surveillance systems, are data on HIV/AIDS diagnoses, which may underestimate the true number of HIV diagnoses as a result of HIV-related symptoms.

The datasets for the London Method need to be specified as given below. Datasets starting with **HIV_CD4_LM_1** contain data on HIV/AIDS diagnoses, while datasets starting with **HIV_CD4_LM_2** contain data on symptomatic HIV diagnoses, i.e., HIV/AIDS diagnoses plus HIV diagnoses with other HIV-related symptoms. The tool will run the London Method when either or both are provided. The datasets should contain data for at least one calendar year.

Required

Data set	Description
HIV_CD4_LM_1_0	number of HIV/AIDS diagnoses without CD4 count
HIV_CD4_LM_1_1	number of HIV/AIDS diagnoses with CD4 < 20 cells/mm^3
HIV_CD4_LM_1_2	number of HIV/AIDS diagnoses with CD4 20-49 cells/mm^3
HIV_CD4_LM_1_3	number of HIV/AIDS diagnoses with CD4 50-99 cells/mm^3
HIV_CD4_LM_1_4	number of HIV/AIDS diagnoses with CD4 100-149 cells/mm^3
HIV_CD4_LM_1_5	number of HIV/AIDS diagnoses with CD4 150-199 cells/mm^3
HIV_CD4_LM_1_6	number of HIV/AIDS diagnoses with CD4 200-249 cells/mm^3
HIV_CD4_LM_1_7	number of HIV/AIDS diagnoses with CD4 250-299 cells/mm^3
HIV_CD4_LM_1_8	number of HIV/AIDS diagnoses with CD4 300-350 cells/mm^3
HIV_CD4_LM_1_9	number of HIV/AIDS diagnoses with CD4 > 350 cells/mm^3

Required

Data set	Description
HIV_CD4_LM_2_0	number of HIV/AIDS diagnoses without CD4 count
HIV_CD4_LM_2_1	number of HIV/AIDS diagnoses with CD4 < 20 cells/mm^3
HIV_CD4_LM_2_2	number of HIV/AIDS diagnoses with CD4 20-49 cells/mm^3
HIV_CD4_LM_2_3	number of HIV/AIDS diagnoses with CD4 50-99 cells/mm^3
HIV_CD4_LM_2_4	number of HIV/AIDS diagnoses with CD4 100-149 cells/mm^3
HIV_CD4_LM_2_5	number of HIV/AIDS diagnoses with CD4 150-199 cells/mm^3
HIV_CD4_LM_2_6	number of HIV/AIDS diagnoses with CD4 200-249 cells/mm^3
HIV_CD4_LM_2_7	number of HIV/AIDS diagnoses with CD4 250-299 cells/mm^3
HIV_CD4_LM_2_8	number of HIV/AIDS diagnoses with CD4 300-350 cells/mm^3
HIV_CD4_LM_2_9	number of HIV/AIDS diagnoses with CD4 > 350 cells/mm^3

3.2 General format of data sets

Data sets need to be prepared in CSV (comma-separated values) format. CSV files can easily be created from packages like Microsoft Excel or SAS, or in a text editor like Notepad.

In a CSV file, the data are arranged in records, or rows in a spreadsheet, and the records are made up of a set of fields that represent epidemiological variables. Each field is separated from the next by a comma (","), except for the last field in the record, which is followed by the `[RETURN]` character.

Each data set contains epidemiological data for a certain number of mutually exclusive populations that together form the total HIV population in a country or region.

3.2.1 Header row

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:

- 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.

3.2.2 Data row

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
..,..,..,..
2013,51.5,30,25
```

- in Excel

```
1982     0   0   0
1983     0   0   0
1984     2   1   3
1985    17  15  20
1987    20  23  20
..      ..  ..  ..
2013  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).

3.2.3 Excel and regional settings

Please be aware of the regional settings of your computer when preparing the data sets in Excel. As mentioned already the tool requires comma (",") as the field and dot (".") as the decimal separator. Excel follows this convention for many Western European countries, including United Kingdom. However, for other countries, for instance Germany, it applies a different convention - semi-colon (";") as the field separator and comma (",") as the decimal separator. Therefore the same data set will be saved by Excel in different formats depending on the regional settings applied:

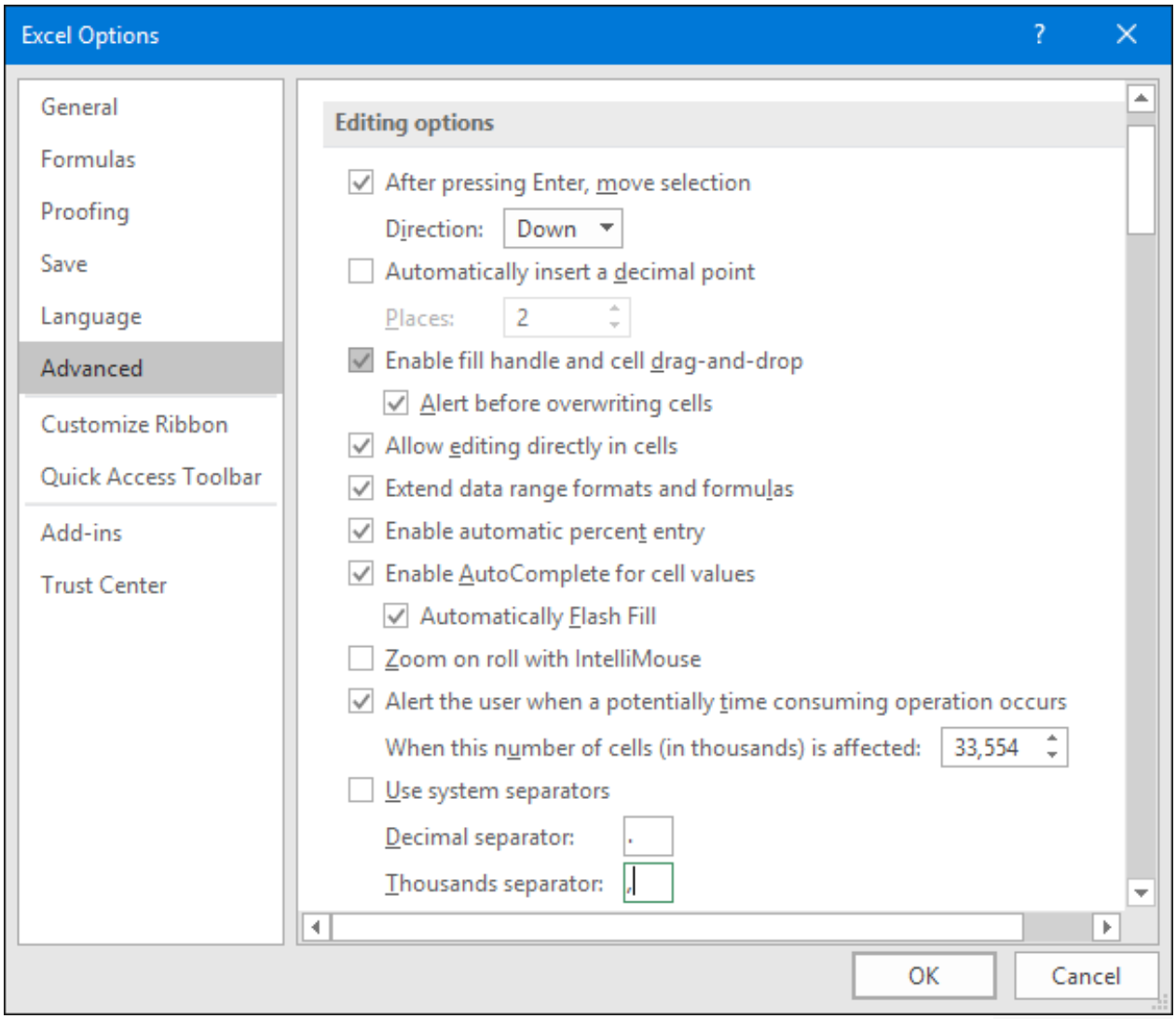
- in UK

```
1982,0,0,0
1983,0,0,0
1984,2,1,3
1985,17,15,20
1987,20,23,20
..,..,..,..
2013,51.5,30,25
```

- in Germany

```
1982;0;0;0
1983;0;0;0
1984;2;1;3
1985;17;15;20
1987;20;23;20
..;..;..,..
2013;51,5;30;25
```

It can be easily check which setting is currently applied by opening one of the csv files distributed with the tool in Excel. If Excel does not put each value in to a separate cell, then it means that it did not recognize comma (",") as the field separator. Csv files edited and saved in Excel with such settings will not be usable for the tool. In order to be able to prepare input csv files for the tool in Excel, please go to Excel options (menu `File -> Options`). In the newly open `Excel Options` dialog select tab `Advanced`. Go to section `Editing options` and deselect option `Use system separators`. Type dot (".") in the newly enabled `Decimal separator` box and comma (",") as `Thousands separator` (the latter is not used by Excel when saving csv file so comma will be properly applied as field separator):



Once all input data files editing is finalized you can revert the original settings by opening `Excel Options` dialog and selecting option `Use system separators`.

4 Starting the tool

To start the ECDC HIV Modelling Tool:

1. go to the folder where the tool was installed.
2. double-click HIVModellingGui.

The tool will open as shown in the figure below and is ready for use.

4.1 SETTINGS

In the `SETTINGS` menu, the user can specify several general settings for the tool.

4.1.1 Interface

Show RunLog tab

If ticked, the tool will show the `run log` tab in the `models` window. The `run log` can be used for monitoring the models' progression, but this is mainly for development purposes.

Charts as png files

- **Export after run**
  If ticked, charts generated by the tool will be exported as png files.
- **Width**
  Specify the width of the output charts (default: 800 pixels).
- **Height**
  Specify the height of the output charts (default: 400 pixels).

Charts as Excel template

- **Export after run**
  If ticked, charts generated by the tool will be exported as Excel files.

4.1.2 Model

This section specifies several default settings for the Incidence Method and London Method.

Range of calculation

'Range of calculations' is the range visible in the slider bars in the `advanced` tab in the `models` window. There is generally no need to change the default settings.

Minimum year

Specify the minimum year of the sliders bars (default: 1975).

Maximum year

Specify the maximum year of the sliders bars (default: 2020).

4.2 HELP

In the `HELP` menu, the user can browse through the manual or download the complete manual as a PDF file.

Browse manual

Here the user can browse through the manual. By clicking items in the table of contents on the right hand side of the window, it is possible to navigate through the manual.

Open PDF version

Open a PDF version of the complete manual.

4.3 ABOUT

The `ABOUT` tab shows information on the following items:

About

Current version of the Graphical User Interface (GUI) and the library implementing the Incidence Method and the London Method and the date of release of the current version of the tool.

Team in charge

Team of international experts and experts at ECDC who are responsible for developing the tool.

Acknowledgement

Member States who participated in developing and piloting the tool.

Funding

Information on how the development of the tool and the Graphical User Interface was funded.

Further reading

References to publicly accessible literature relating to the tool.

Contact us

Email address for suggestions and questions regarding the tool.

Keeping informed

You can choose to

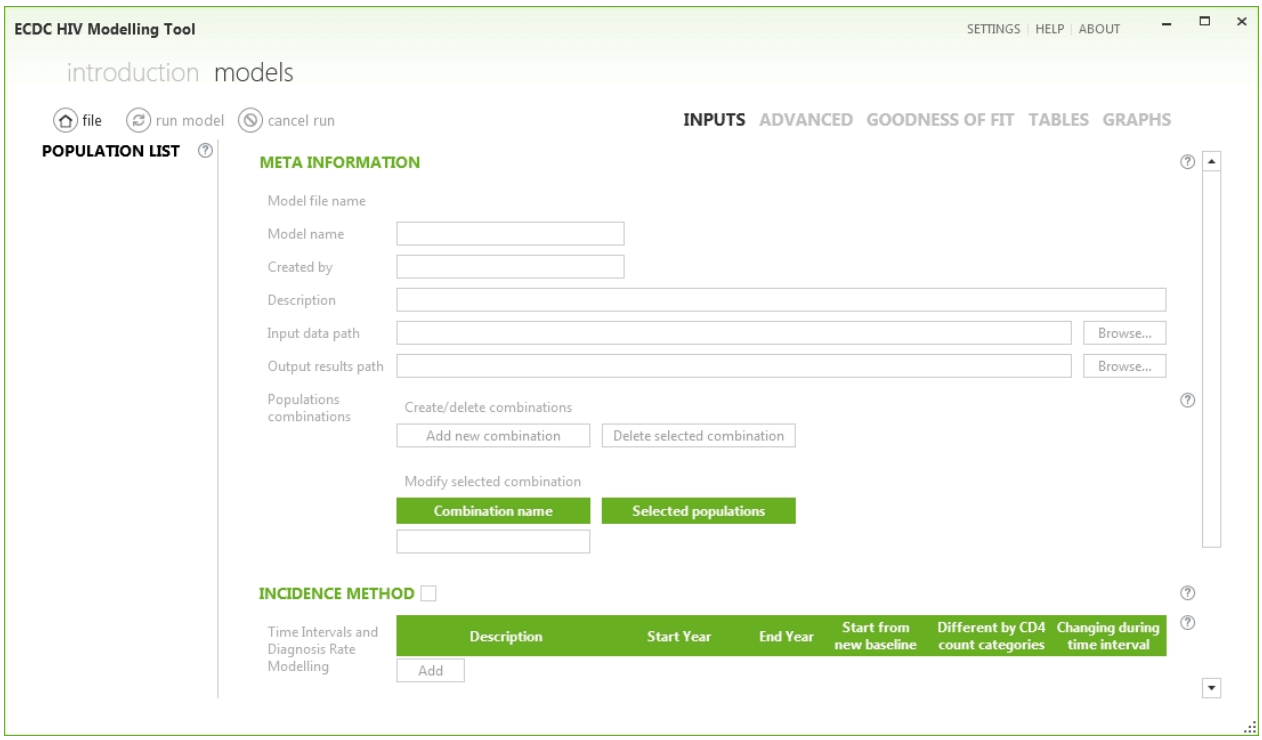
- Receive news about updates of the tool by clicking `Subscribe for updates`.
- Stop receiving news about the tool by clicking `Unsubscribe from updates`.

4.4 introduction

The `introduction` window gives a brief introduction of the ECDC HIV Modelling Tool.

4.5 models

Select `models` to go to the main window of the ECDC HIV Modelling Tool. This window will look like this:



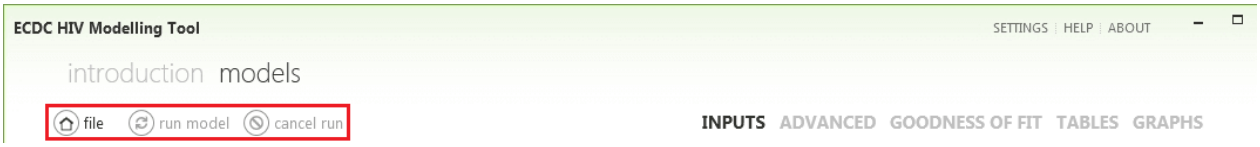
4.5.1 Population list

On the left-hand side of the window, `population list` lists all models that were defined by the user and loaded into the tool. Within each model there may be different *populations*.

The currently selected population is shown in green and further information is shown in the main window.

4.5.2 Managing populations

On the top left are three buttons that are used to manage models and to run the Incidence Method and the London Method.



file

This menu is used to create, save, and open models in the `population list`.

- **New**:
  Create a new model.
- **Open**:
  Open an existing model and add it to the `population list`.
- **Save**:
  Save the currently selected model and its parameter settings. This option will be faded out as long as there were no changes in the settings for the selected model.
- **Save as**:
  Save and rename the current selected model and its parameter settings.
- **Close**:
  Remove the currently selected model from the tool. Note that this option will only remove the model from the population list. It will still exist in the Models folder.

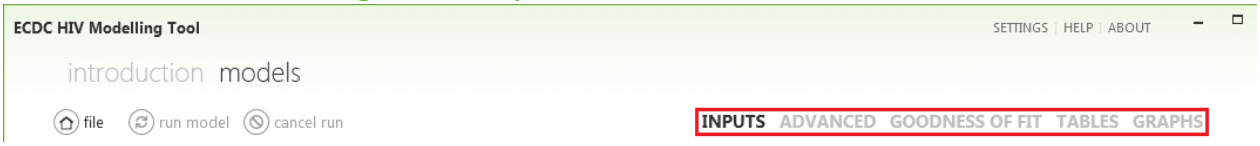
run model

Run the Incidence Method and/or London Method for the selected population in the `population list`.

cancel run

Stop running the Incidence Method and/or London Method for the selected population in the list.

4.5.3 Parameter settings and outputs



Using the buttons on the right the user can switch between different frames in the `models` window. The selected frame will be indicated in bold font, the other frames will be faded out.

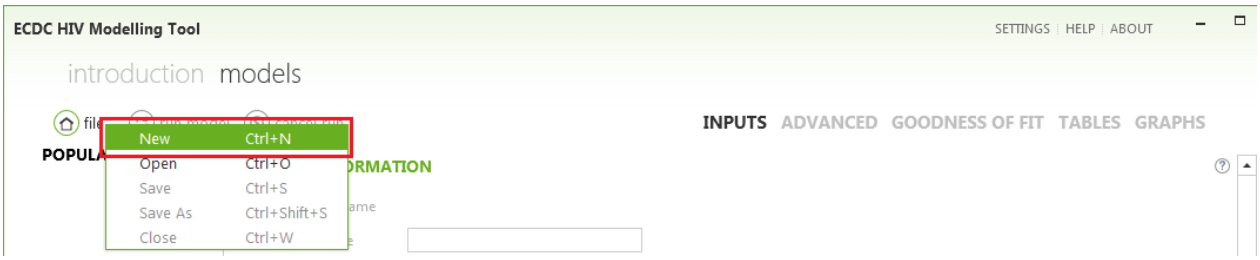
- **Inputs**:
  Basic input parameters and model settings.
- **Advanced**:
  Advanced input parameters and model settings.
- **Run log**:
  (for development purposes only) Show progression of the tool and interim results. By default, this button is hidden unless `INTERFACE->Show RunLog` is checked.
- **Goodness of fit**:
  Tables and figures showing input data with model fits.

- **Tables**:
  Tables with data from the models and model fits.

- **Graphs**:
  Figures with model outcomes.

5 Creating a new model

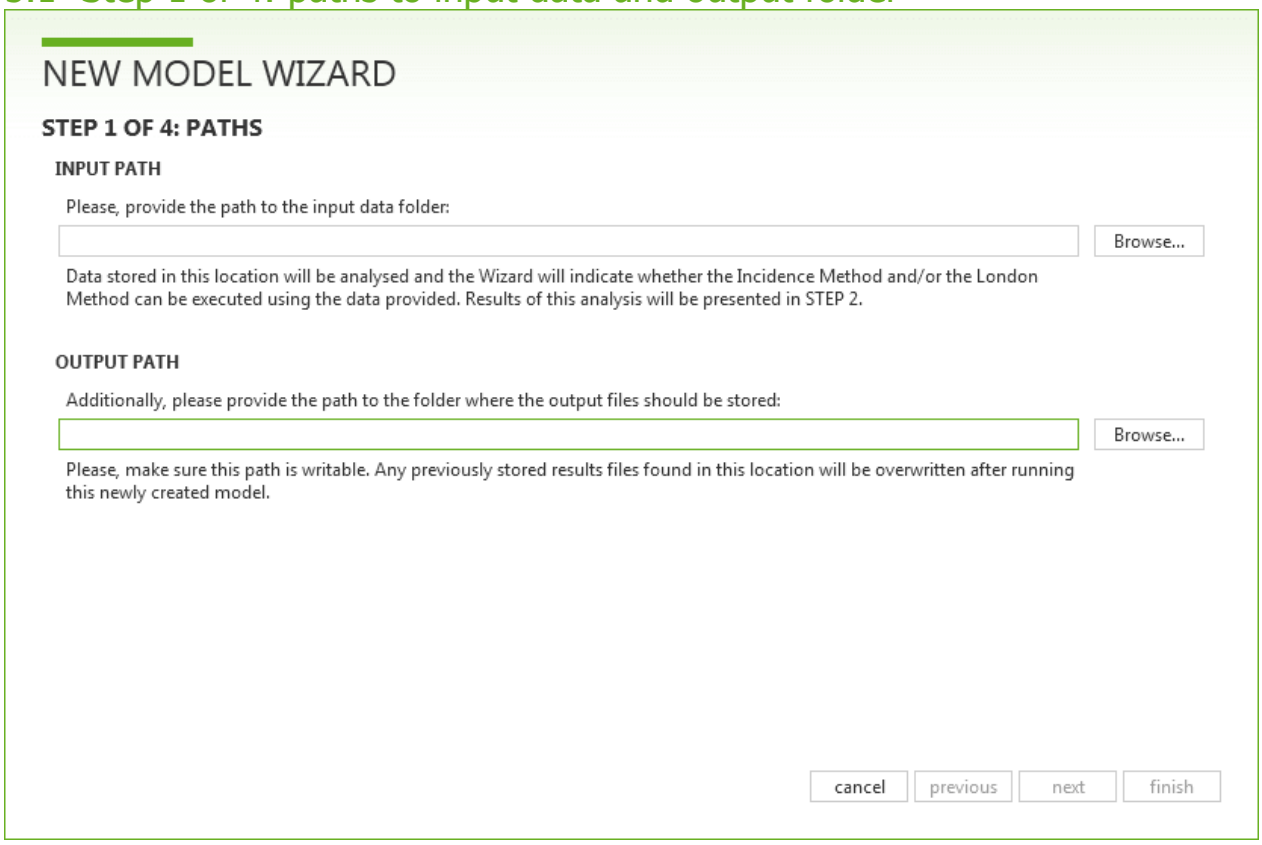
To create a new population in the list select `file->New` on the top left hand side of the `models` window.



A wizard will guide you through the process of creating a new model.

You can the next and previous buttons to move to the next or previous step in the wizard. Use cancel if you want to stop the wizard.

5.1 Step 1 of 4: paths to input data and output folder

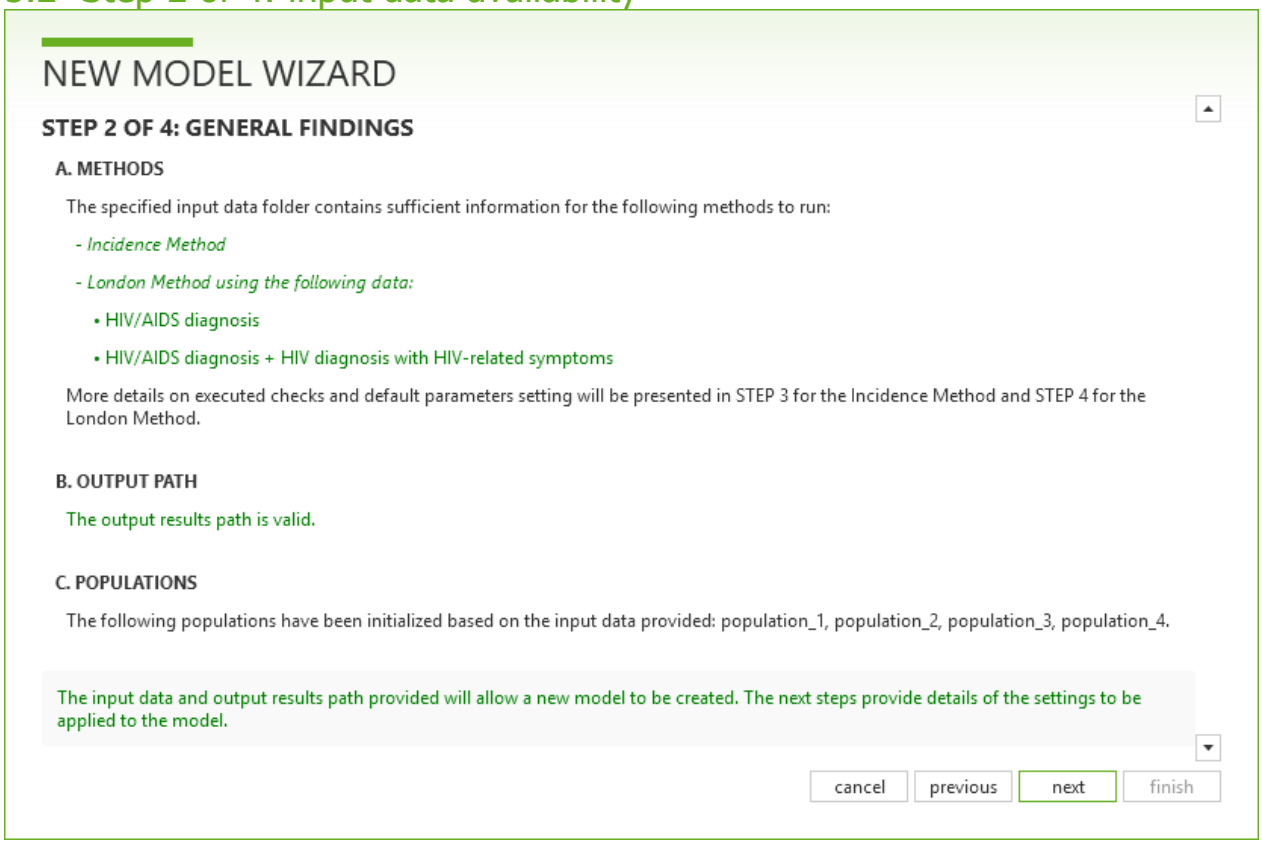


In the first step of the wizard you are asked to specify

- the location of the input data folder.
- the location of the folder where the outcomes of the tool should be stored.

For each model, which can contain one or more populations, you need a separate input data folder.

5.2 Step 2 of 4: input data availability



A. Methods

The tool will check whether the required and recommended datasets for each of the two methods are available and correctly specified. For each of the methods, the wizard will show whether the input data files were correctly specified (green) or not (red). More specific information on problems in the input data is given in Step 3 and 4.

Note that the names of the input datasets should be exactly as specified before in .

After amending the input data files, you need to go back to Step 1 and continue from there or restart the wizard.

B. Output path

The tool will check whether the specified output folder exists and if not, it will be created.

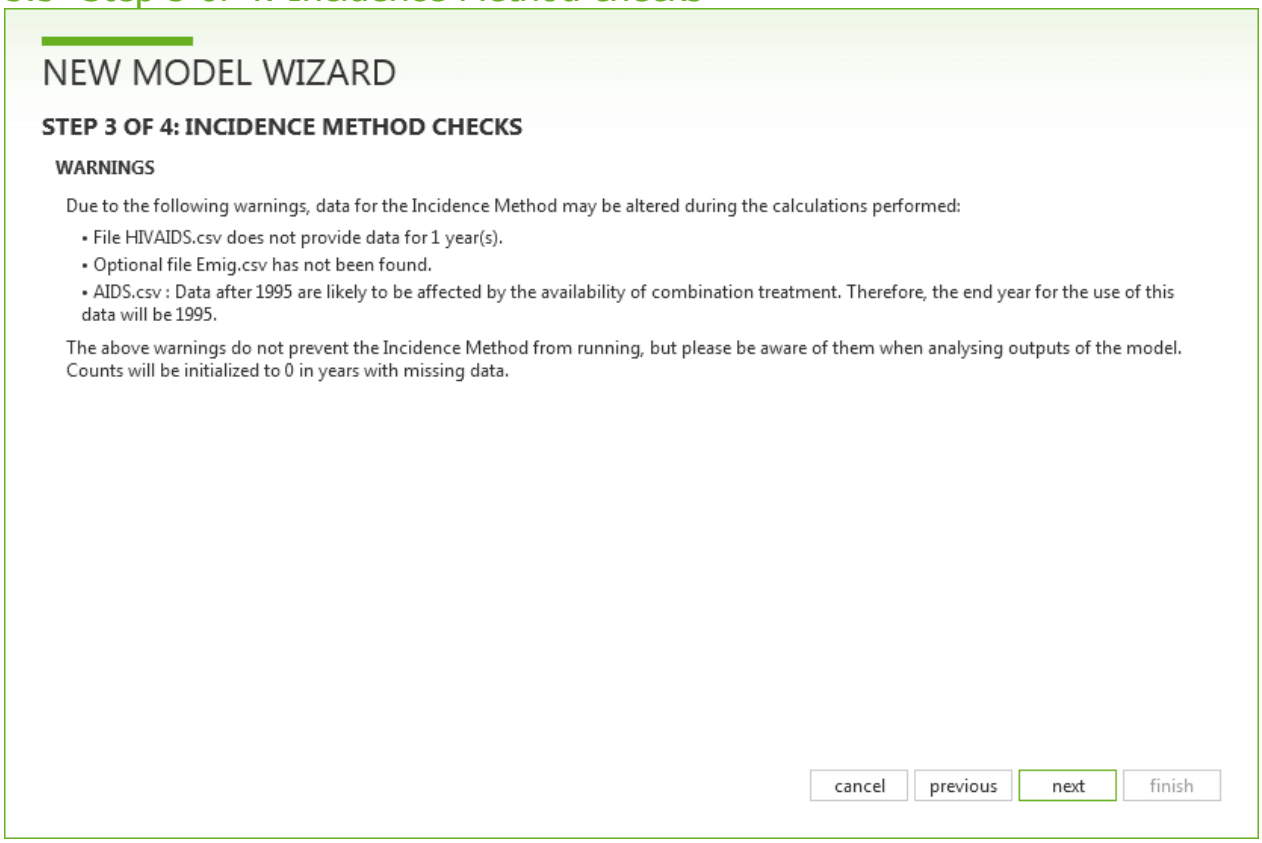
C. Populations

The wizard will show the populations that were specified in the header rows of the input data files. A warning will be given if the number of groups or the names of the groups differ between any of the input files.

Example

The tool has sufficient input data files in order to be able to run both the Incidence Method and the London Method. The output folder for the results of the tool is correct. The wizard found four different risk groups or populations in the input data files.

5.3 Step 3 of 4: Incidence Method checks



In Step 3, the wizard may give some Errors or Warnings about the input files for the Incidence Method.

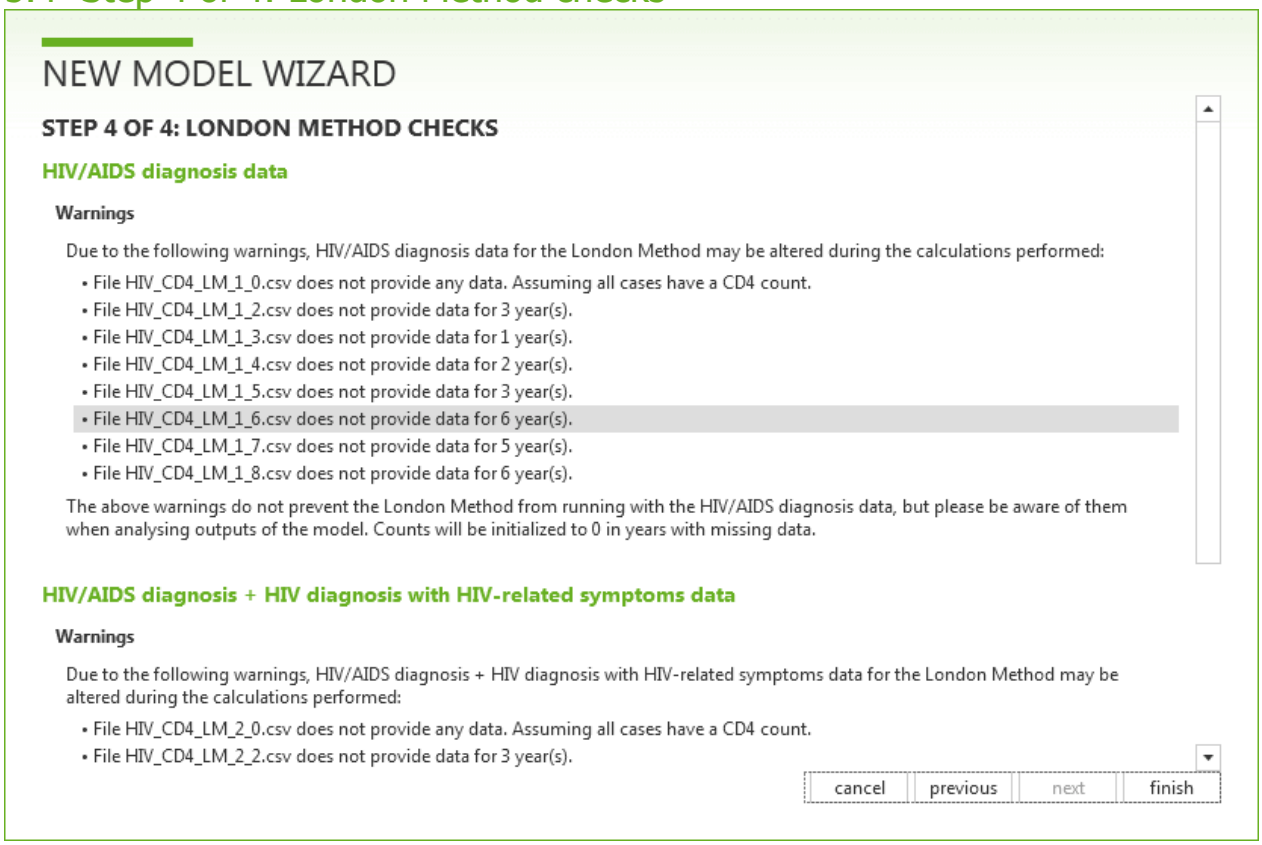
Errors will prevent the Incidence Method from running, while Warnings can be ignored although we encourage the user to check the decisions taken by the wizard.

Example

In the example, the file **HIVAIDS.csv** has no record for one calendar year and the tool will assume there were 0 HIV/AIDS diagnoses in this year.

The input file with total number of AIDS cases contains data after 1995. These data will be ignored by the tool, because these data are likely to be affected by combination treatment.

5.4 Step 4 of 4: London Method checks



In Step 4, the wizard may give some Errors or Warnings about the input files for the London Method.

Errors will prevent the London Method from running, while Warnings can be ignored although we encourage the user to check the decisions taken by the wizard.

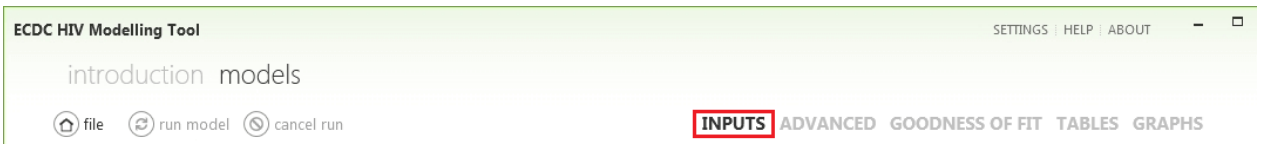
Example

In the example, there is no data in input file **HIV_CD4_LM_1_0.csv**, which contains number of cases without a CD4 count measurement, and the tool will assume that all cases have a CD4 count. In most of the other input data files there appear to be missing records for some calendar years. For instance, in **HIV_CD4_LM_1_2.csv**, which contains numbers of patients with an HIV/AIDS diagnosis and a CD4 count 20-49 cells/mm^3, data are missing for three years. The tool assumes that there were zero patients in these years.

6 Parameter specification

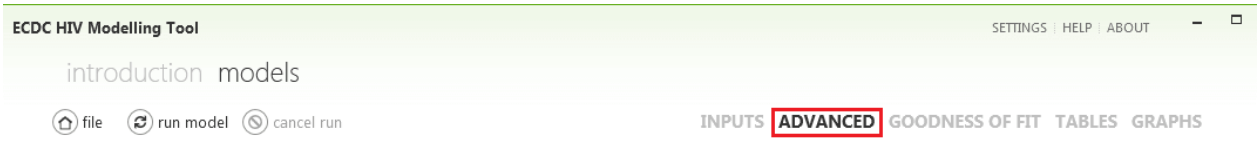
The ECDC HIV Modelling Tool has two windows where parameters can be specified:

- `INPUTS`: basic input parameters and model settings.



  - **Meta Information**: general information on the selected model with location of the data and the model output.
  - **Incidence Method**: basic parameters for the Incidence Method.
  - **London Method**: basic parameters for the London Method.

- `ADVANCED`: advanced input parameters and model settings.



  - **Incidence Method**: advanced parameters for the Incidence Method.
  - **London Method**: advanced parameters for the London Method.

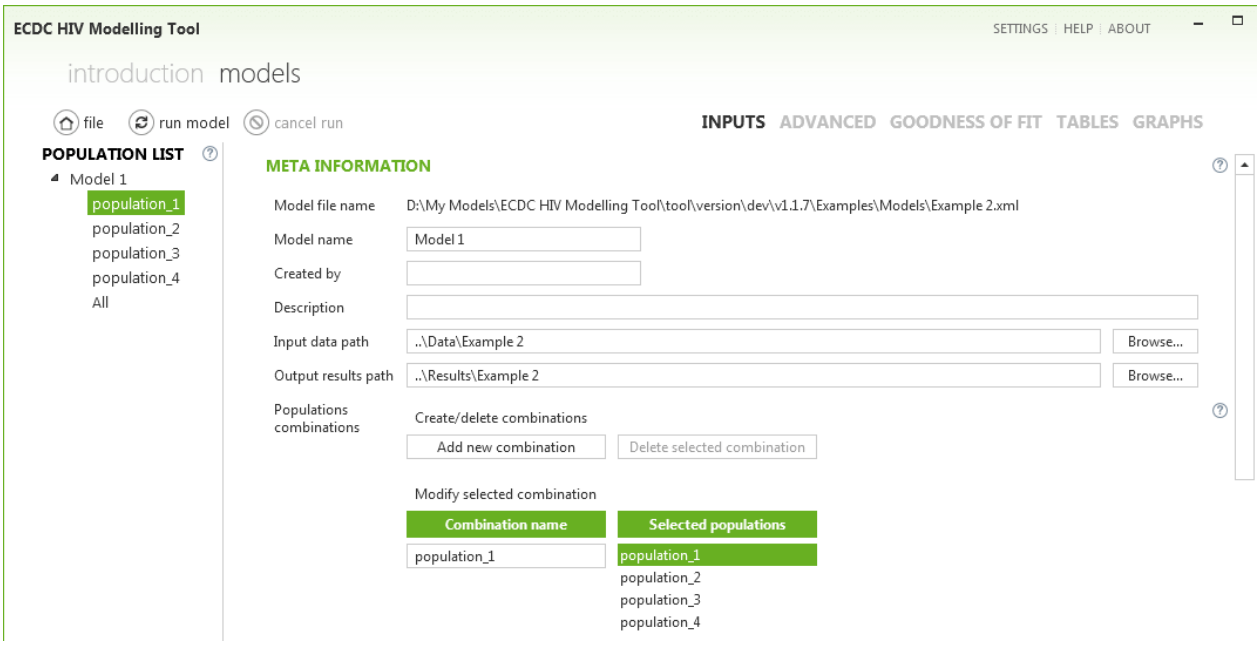
All parameters are pre-specified when initialising a population in the `POPULATION LIST` using the wizard.

Notes

- The parameters specified in the `inputs` and `advanced` window will be the same for all populations within a model.

6.1 Meta Information

`Meta information` shows general information for the selected model.



Model file name

Information on each population in the `population list` is stored in a XML file in the folder Models that was created when installing the tool.

Model name

This is the name of the model that appears in the `population list`. The name can be changed by the user.

Created by

Here the user can enter information on who created this particular model.

Description

The user can provide an additional description of the model and the population.

Input data path

This shows the location of the folder containing the datasets as specified in the wizard.

Output results path

This shows the location of the folder for the results of the tool as specified in the wizard.

Populations combinations

Here the user can manage combinations of populations within a population.

6.1.1 Combinations of populations

By default, populations within a model in the `population list` are the risk groups that were specified in the header row of the input data files. The user can easily add combinations of these risk groups as new populations.

The combination `All` is automatically created by the tool and includes all populations in the model. Hence, `All` will typically contain all diagnosed HIV infections in a country.

Combination name

This is the name of the population or combination of populations that is currently highlighted in green in the `population list`.

Selected populations

This shows which populations specified in the input data, highlighted in green, form the currently selected combination in the `population list`.

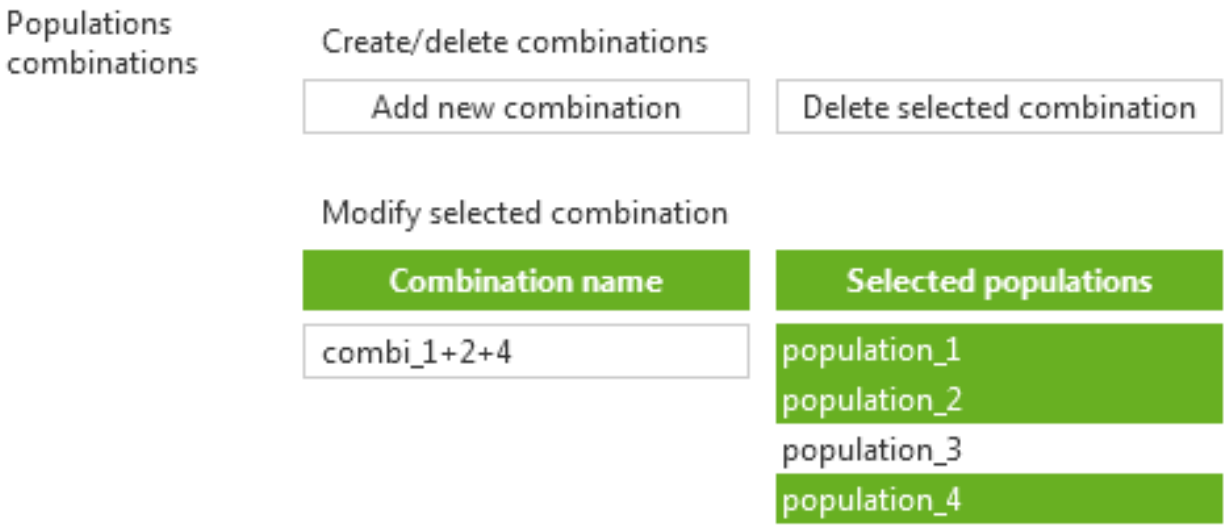
The followings options are available for managing combinations of populations:

- **Add new combination**: Adds a new combination to the `population list`. By default, the new combination includes all populations specified in the input data.
- **Delete selected combination**: Deletes a combination from the `population list`. The combination `All` and the populations specified in the input data cannot be deleted.

Example

Creating a new combination of populations `population_1`, `population_2`, and `population_4`:

- Click **Add new combination**. This will create a new combination `new name 1` that consists of all risk groups.
- Change **Combination name** to `combi_1+2+4`, for instance.
- Deselect `category_3` under **Selected populations**.



The combination `combi_1+2+4` has now been added to the `population list`.

6.2 Incidence Method

`Incidence Method` contains basic parameter settings for the Incidence Method, which currently only includes the specification of the presumed shape of the diagnosis probability. This diagnosis probability is pre-specified by the wizard, but the user is *strongly encouraged* to change this specification.

If the tick box is checked the tool will run the Incidence Method for the selected population. Unchecking the tick box will prevent the tool from running the Incidence Method and the table specifying the diagnosis probability will be faded out.

6.2.1 Diagnosis probability

Apart from estimating the annual number of new HIV infections, the Incidence Method 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.

Before using the Incidence Method, the user needs to specify:

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

Notes

- It is only necessary to specify the *presumed shape* of the diagnosis probability, as shown in the examples. The tool will determine the best-matching diagnosis probability over calendar time.
- The specification of the diagnosis probability applies to *all* populations within a model.

Intermezzo: diagnosis probability

The tool assumes that the probability that an HIV-positive individual is diagnosed with HIV within t years after infection or entering a certain CD4 count interval is given by the expression $1 - e^{-\delta t}$. The parameter δ is a rate parameter, or *diagnosis rate*, per time unit. For instance, when δ is 0.2 per year, the probability of being diagnosed within 0.5 years is $1 - e^{-0.1 \times 0.5}$, which is approximately 0.05 or 5%. The diagnosis probability can thus be seen in terms of proportions: suppose there are 100 undiagnosed HIV-positive individuals at time 0, then 5% of them will be diagnosed after 0.5 years.

An increase in the rate parameter δ will result in an increase in diagnosis probability. Within the tool, the user can specify the presumed shape of δ and how it may change over calendar time.

6.2.2 Specifying time intervals

Time intervals in which the diagnosis probabilities may change are given in the left part of the table.

Deleting and adding time intervals

The following options are available for managing time intervals:

- **Add**: Add a new time interval. Time intervals will be automatically sorted according to Start Year.
- **Delete**: Delete a time interval.

Description

Here you can give a short description of the time interval or why this specific time interval was chosen. For instance, the wizard specifies a time interval 1980 to 1984 in which the probability of being diagnosed with HIV is zero due to the absence of serological testing. Other examples of a time interval could be the

time interval corresponding to the pre-cART era, or a new time interval could start in the year when there was a change in HIV testing policy, triggered, for instance, by an outbreak of HIV in a specific population.

Start and End Year

By clicking on `Start Year` you can change the start year of an interval. This can be done by using the up and down arrows.

Notes

- It is not necessary to specify the end year of a time interval, because it will be automatically set to the start year of the next time interval.
- The start year of the first time interval will always be 1980.
- The end year of the last time interval will be the last year for which data are available.
- A higher number of time intervals will result in a more flexible shape of the diagnosis probability curve. However, this also means that a higher number of unknown parameters need to be estimated from the data.

6.2.3 Specifying presumed shape

In the right part of the table you can specify how the diagnosis probability may change in a time interval.

Start from new baseline

If this box is not ticked, the diagnosis probability at the start of the interval will be the same as at 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.

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 should only be done if there are data on HIV diagnoses by CD4 count.

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.

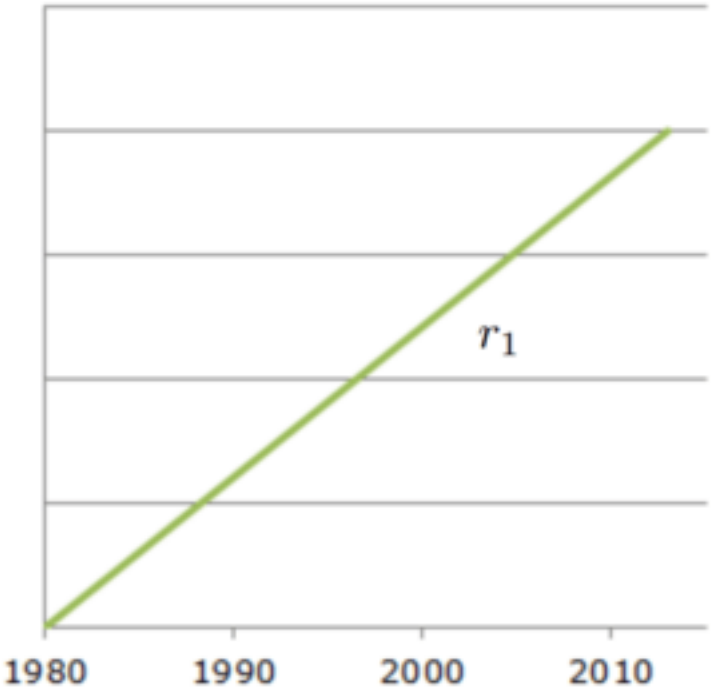
Notes

- The more boxes are ticked, the higher the number of parameters that need to be estimated from the data and the longer the Incidence Method will run.
- Be careful with ticking `Start from new baseline` and `Changing during time interval` at the same time as this may lead to negative diagnosis probabilities.
- Parameters describing the diagnosis probability curve are largely determined by the observed number of HIV diagnoses by CD4 count. Therefore, if data on CD4 counts are missing or sparse, it may not be possible to accurately determine the diagnosis probability curve.
- 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.

6.2.4 Examples

The examples below show a few possible shapes of the diagnosis probability curve and how these shapes are specified in the tool.

Example I



Presumed shape

The diagnosis probability is zero in 1980 and then gradually increases with calendar time at a rate r_1.

Time intervals and diagnosis probabilities

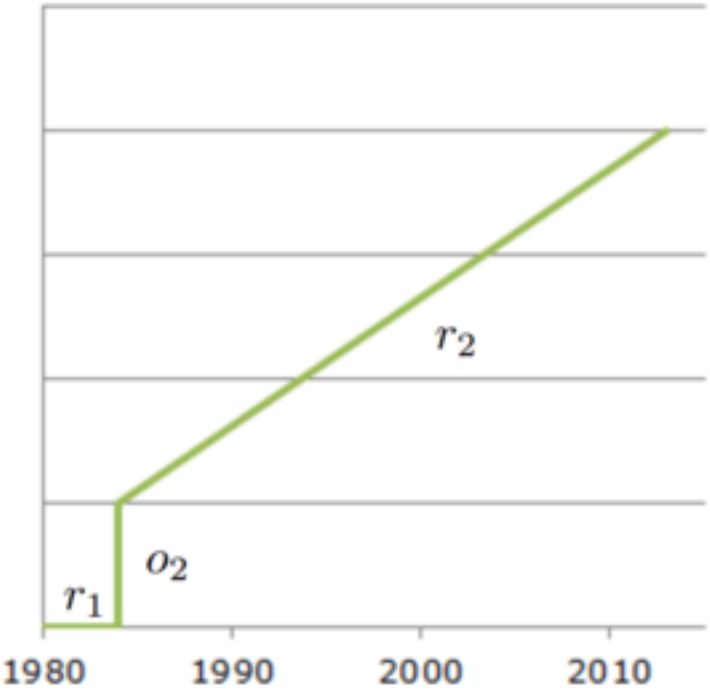
INCIDENCE METHOD ☑

Time Intervals and Diagnosis Rate Modelling

	Description	Start Year	End Year	Start from new baseline	Different by CD4 count categories	Changing during time interval
Delete	Time interval 1	1980	2012	☐	☐	☑

Add

Example II



Presumed shape

The diagnosis probability is zero from 1980 until 1984. In 1984, the diagnosis probability is o_2 and afterwards increases linearly at a rate r_2.

Time intervals and diagnosis probabilities

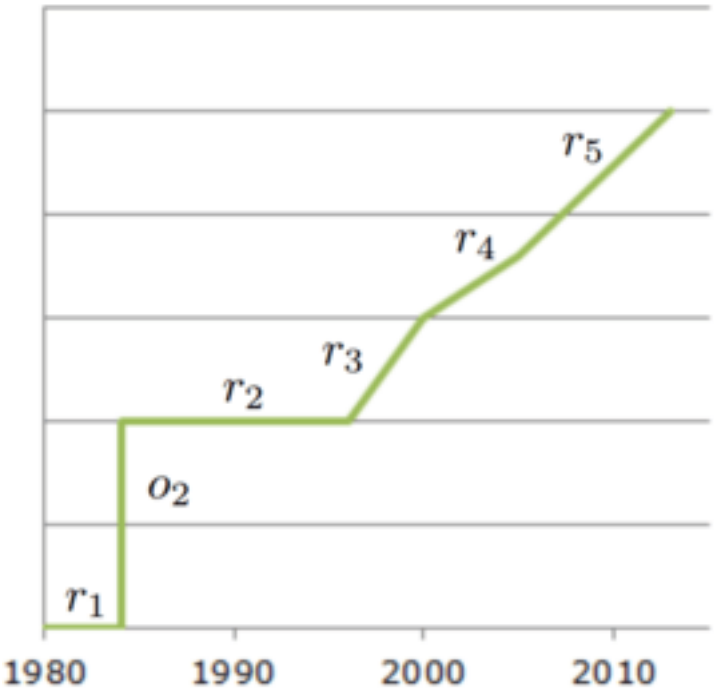
INCIDENCE METHOD ☑

Time Intervals and Diagnosis Rate Modelling

	Description	Start Year	End Year	Start from new baseline	Different by CD4 count categories	Changing during time interval
Delete	Time interval 1	1980	1984	☐	☐	☐
Delete	Time interval 2	1984	2012	☑	☐	☑

Add

Example III



Presumed shape

This is the presumed shape used in[1]. The diagnosis probability is zero from 1980 until 1984. Between 1984 and 1996, the diagnosis probability is at a constant value o_2. From 1996 onwards, the probability increases linearly, but the rate at which the probability increases with calendar time changes in 2000 from r_3 to r_4 and in 2005 from r_4 to r_5. Diagnosis rates can be different for each of the four CD4 count categories.

Time intervals and diagnosis probabilities

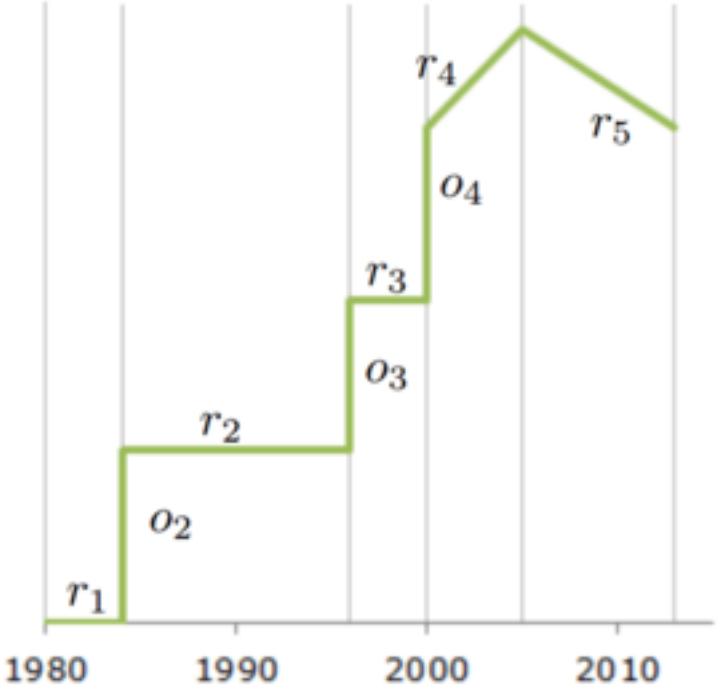
INCIDENCE METHOD ☑

Time Intervals and Diagnosis Rate Modelling

	Description	Start Year	End Year	Start from new baseline	Different by CD4 count categories	Changing during time interval
Delete	Time interval 1	1980	1984	☐	☐	☐
Delete	Time interval 2	1984	1996	☑	☑	☐
Delete	Time interval 3	1996	2000	☐	☑	☑
Delete	Time interval 4	2000	2005	☐	☑	☑
Delete	Time interval 5	2005	2012	☐	☑	☑

Add

Example IV



Presumed shape

The diagnosis probability is zero from 1980 until 1984, o_2 from 1984 until 1996, o_3 from 1996 until 2000. In 2000, the diagnosis probability is o_4 and then changes linearly. The rate at which the probability increases with calendar changes in 2005 from r_4 to r_5.

Time intervals and diagnosis probabilities

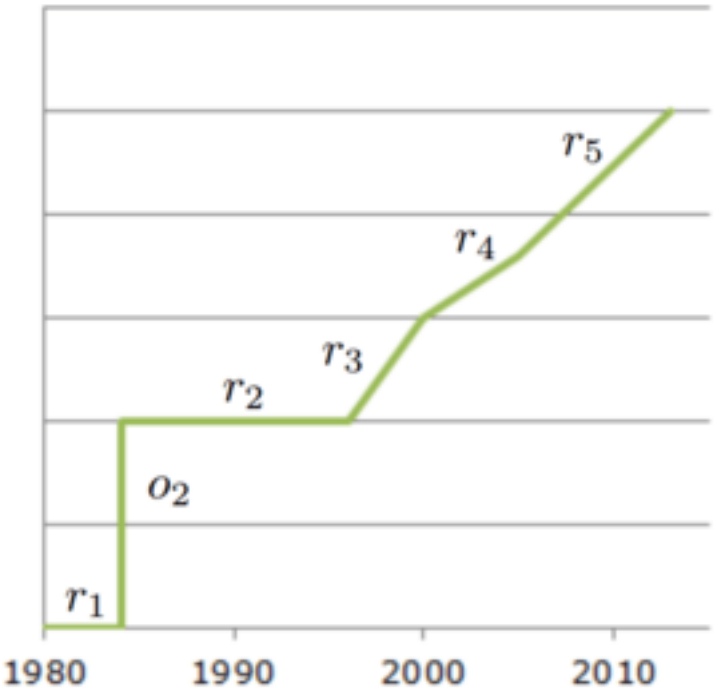
INCIDENCE METHOD ☑

Time Intervals and Diagnosis Rate Modelling

	Description	Start Year	End Year	Start from new baseline	Different by CD4 count categories	Changing during time interval
Delete	Time interval 1	1980	1984	☐	☐	☐
Delete	Time interval 2	1984	1996	☑	☐	☐
Delete	Time interval 3	1996	2000	☑	☐	☐
Delete	Time interval 4	2000	2005	☑	☐	☑
Delete	Time interval 5	2005	2012	☐	☐	☑

Add

Example V



Presumed shape

Same as in Example III. The diagnosis probability is zero from 1980 until 1984. Between 1984 and 1995, the diagnosis probability is a constant value o_2. After 1996, the probability increases linearly, but the rate at which the probability increase with calendar changes in 2000 from r_3 to r_4 and in 2005 from r_4 to r_5. Diagnosis rates are the same for all CD4 count intervals.

Time intervals and diagnosis probabilities

INCIDENCE METHOD ☑

Time Intervals and Diagnosis Rate Modelling

	Description	Start Year	End Year	Start from new baseline	Different by CD4 count categories	Changing during time interval
Delete	Time interval 1	1980	1984	☐	☐	☐
Delete	Time interval 2	1984	1996	☑	☐	☐
Delete	Time interval 3	1996	2000	☐	☐	☑
Delete	Time interval 4	2000	2005	☐	☐	☑
Delete	Time interval 5	2005	2012	☐	☐	☑

Add

6.3 London Method

`London Method` contains basic parameter settings for the London Method.

LONDON METHOD ☑

Calculation based on

- ✓ HIV/AIDS diagnoses
- ✓ HIV/AIDS diagnoses + HIV diagnoses with HIV-related symptoms

Multiplication factor for rate of HIV symptoms from rate of AIDS: 2.000

If the tick box is checked the tool will run the London Method for the selected population. Unchecking the tick box will prevent the tool from running the London Method.

Intermezzo: London Method

The calculations by the London Method are based on data about HIV/AIDS diagnoses or on data about symptomatic HIV diagnoses, i.e., HIV/AIDS diagnoses plus HIV diagnoses with HIV-related symptoms. For both types of input data, the tool will run the London Method using data from the selected population, highlighted in green in the `population list`. The tool will also do calculations using data from combination `All` and then multiply the result with the proportion of HIV/AIDS diagnoses observed in the highlighted population. This latter method is less sensitive to fluctuations as a result of small number of observations in CD4 count intervals.

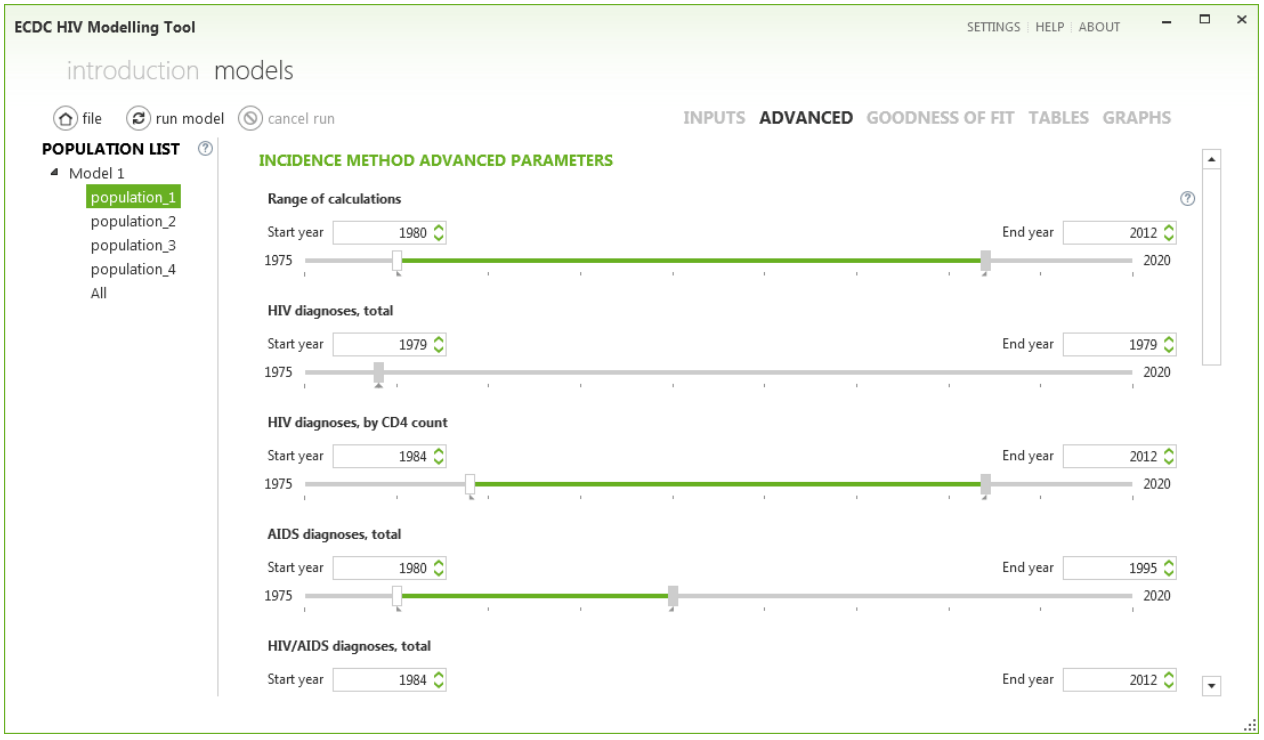
6.3.1 Calculation based on

A green tick mark indicates whether the London Method will be based on HIV/AIDS diagnoses only and/or on HIV/AIDS diagnoses plus HIV diagnoses with HIV-related symptoms.

When calculations are based on HIV/AIDS diagnoses plus HIV diagnoses with HIV-related symptoms, the London Method requires knowledge of the CD4 count-specific rate of HIV-related symptoms. While the CD4 count-specific rate of occurrence of AIDS is known, the CD4 count-specific rate of occurrence of such HIV-related symptoms is less well described. There is evidence to suggest that the rate of developing HIV-related symptoms is approximately two- to four-fold higher compared to the rate of AIDS[2]. By default, the CD4 count-specific rates of HIV-related symptoms are assumed to be two-fold the CD4 count-specific rates of AIDS, but this can be changed by the user.

6.4 Incidence Method advanced parameters

The `advanced` tab displays shows advanced parameters settings for the selected population in the `population list`. Parameters in this window are pre-specified by the wizard.



6.4.1 Calendar year ranges

The five slider bars show the range of calendar years used in the calculations of the Incidence Method. These ranges are specified by the wizard and are based on the minimum and maximum calendar year in the input datasets.

Ranges can be made wider or narrower by dragging the rectangles at either end of the green bar. Alternatively, the lower boundary (`Start year`) and upper boundary (`End year`) of each calendar year range can be selected by clicking the up or down button in the white boxes above the slider bars.

Range of calculations

Specifies the range of the model calculations. Calculations start on 1 January of `Start year` and end on 31 December of `End year`.

HIV diagnoses, total

Range of calendar years for which data on the total annual number of HIV diagnoses are used in the model fit. This range should correspond to years for which there is no or insufficient information on CD4 counts at the time of diagnosis.

HIV diagnoses, by CD4 count

Range of calendar years for which data on the annual number of HIV diagnoses by CD4 count interval are used by the Incidence Method. For years with no or insufficient data on CD4 counts at the time of diagnosis (see below), total number of diagnoses should be used (see HIV diagnoses, total)

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.

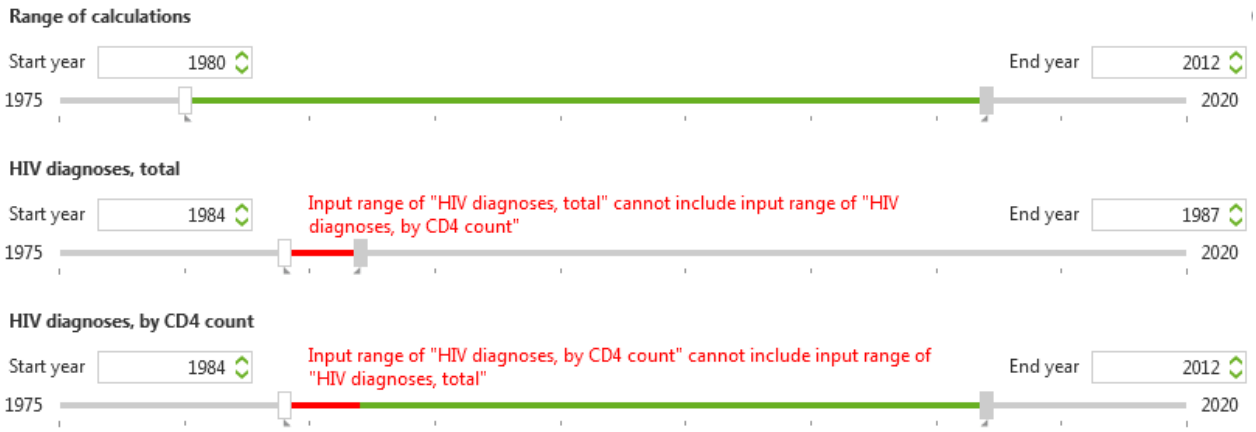
HIV/AIDS diagnoses, total

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 (see AIDS diagnoses, total).

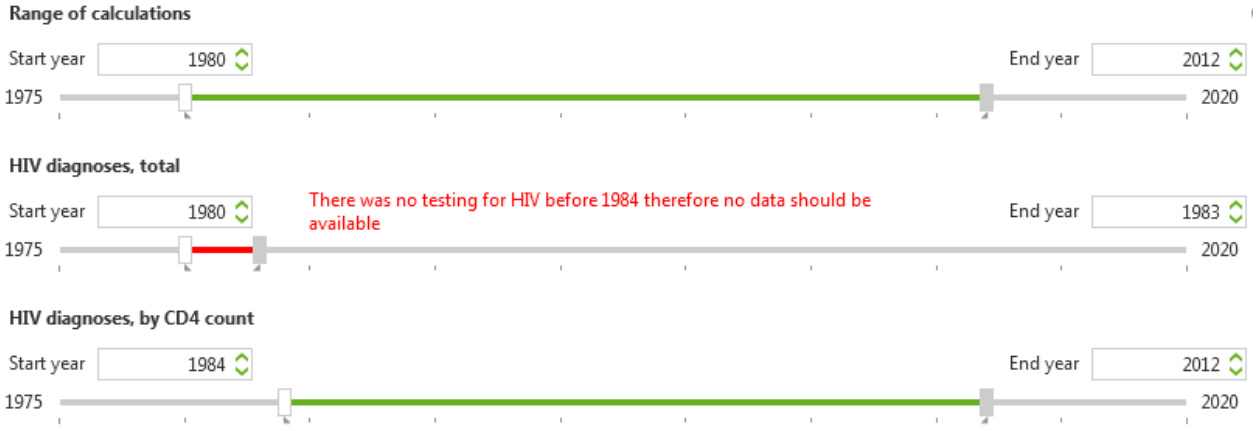
Notes

- The calendar year ranges for `HIV diagnoses, total` and `HIV diagnoses, by CD4 counts` *should not overlap*. The tool will issue a warning when the ranges overlap.



The tool automatically calculates the yearly proportion of HIV diagnoses with and without a CD4 count and uses this proportion in the model fit. Therefore, the two data items are not independent and cannot be used simultaneously in the model fit.

- The tool will also issue a warning when specifying a calendar year range that includes years before 1984 for `HIV diagnoses, total` or `HIV diagnoses, by CD4 count`.



- Data items in calendar years outside `Range of calculations` will not be taken into account in the model fit.
- For calendar years for which no data are provided for one or more data items the tool assumes that the number of observations is 0. This assumption may not be correct if the data item only started to be collected from a specific calendar year onwards. Users are advised to be cautious when extending calendar ranges beyond the pre-specified setting.

6.4.2 Country-specific settings

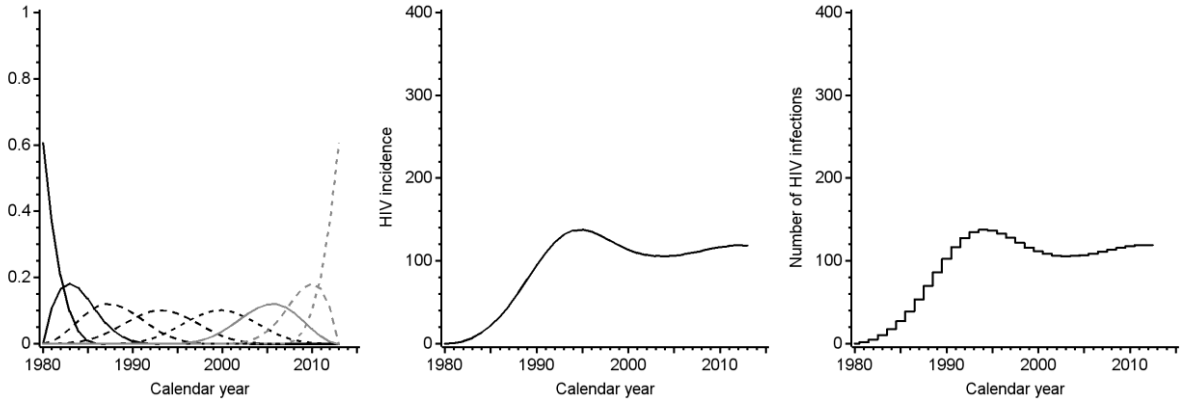
The tool has the possibility to incorporate county-specific settings.

Currently, this option is only implemented for the Netherlands where a correction needs to be done for truncation of HIV diagnoses before 1996. National surveillance of HIV diagnoses only started in 1996 and patients diagnosed before that time were recorded if they survived up to 1996.

6.4.3 Incidence curve

For each calendar year, the Incidence Method determines the annual number of HIV infections from the estimated HIV incidence curve. In the tool, this incidence curve is fully described by a limited number of parameters that is determined by the number of so-called *knots*. A higher number of knots will result in more flexibility for the incidence curve, but also involves estimating a higher number of parameters.

Intermezzo: HIV incidence



The annual number of HIV infections (right figure) is determined by the tool from the input data on diagnosed HIV cases. Usually it is not possible to estimate this number separately for each year, because this would involve estimating a too large number of unknown parameters, one for each calendar year. Therefore, the number of infections is derived from an HIV incidence curve (middle figure) that is approximated using so-called *cubic splines*, which can be *M-splines* or *B-splines*.

Cubic splines are polynomial functions of the form $at^3 + bt^2 + ct + d$ 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. In the figure on the left, there are 4 knots equally spaced between 1980 and 2013 and each curve is a different M-spline.

Adding up all the different splines, each with a different weight, determines the shape of the incidence curve (middle figure). The number of cubic splines needed is always the number of knots plus 4, so 8 splines in this case. For each of the splines the tool needs to estimate a weight parameter, which determines its contribution to the incidence curve. These weight parameters are estimated from the input data. In case of 4 knots, 8 parameters are necessary, one for each spline. Even though only 8 splines are used, the resulting incidence curve is very flexible. These 8 parameters are much easier to estimate than one parameter for each calendar year.

The tool will do a few initials runs to determine which splines contribute significantly to the incidence curve. Splines with a very small contribution compared to their neighbouring splines will not contribute to the final incidence curve and their associated parameter is set to zero.

There is a small difference between M-splines and B-splines. M-splines are normalised to 1, i.e., integrating a M-spline over the time interval where the M-spline has a non-zero value gives 1. B-splines are not normalised but have the property that if all weight parameters 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 $\theta_i = 2\,\theta_{i-1} - \theta_{i-2}$, where θ_{i-2}, θ_{i-1}, and θ_i are the weight parameters associated with the $i-2^{\text{th}}$, $i-1^{\text{th}}$, and i^{th} spline function.

Note that in earlier versions of the tool (version 1.0.2 and before), the weight parameters associated with each M-spline were required to be non-negative, which is a sufficient condition for having a non-negative

HIV incidence curve. The current version of the tool allows for negative weight parameters as long as the incidence curve remains non-negative.

Type

Here the user can specify whether the HIV incidence curve is modelled with cubic M-splines or cubic B-splines (default).

Both types of splines should give very similar results and M-splines are only retained for backward compatibility with earlier versions of the tool (version 1.0.2 and before). Note that full backward compatibility may not always be achieved because of changes in the requirements for the weight parameters associated with each M-spline (see Intermezzo: HIV incidence).

Knots count

This is the number of knots used in the incidence curve (default value 4).

We recommend using 4 to 6 knots. Models with less internal knots are preferred if the fit to the data does not become worse.

Start at zero

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 means that one of the parameters necessary to specify the HIV incidence curve is fixed at zero and does not need to be estimated.

Prevent sudden changes at end of observation interval

If this box is ticked the tool will prevent sudden increases or decreases in the estimated HIV incidence curve at the end of the calendar year range.

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 $\geq$ 500 cells/mm^3 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.

6.4.4 Maximum likelihood

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, we minimise the equivalent deviance measure.

Distribution

The default distribution is a Poisson distribution in which the mean is equal to the variance. In practice this distribution will work well enough.

It is also possible to select a negative binomial distribution which is a generalisation of the Poisson distribution such that the variance can be larger than in a Poisson distribution. This will involve estimating a *dispersion parameter* r which is initially set at a value of 1000.

6.4.5 Diagnosis rate

Extra rate due to non-AIDS symptoms

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/mm^3. This contribution takes into account that HIV-positive individuals may be diagnosed because of HIV-related non-AIDS symptoms. The publication on the Incidence Method uses a value of 0.4 per year[1]. Note that the extra rate will be added to the parameter δ (see Intermezzo: diagnosis probability) for CD4 counts below 200 cells/mm^3.

Setting this contribution to a non-zero value is particularly useful if there are no CD4 counts available and/or if for one or more time interval the box `Different by CD4 count categories` in the diagnosis probability matrix in the `inputs` tab is unchecked. In this case, diagnosis probabilities will be the same for the three highest CD4 categories and higher for CD4 counts below 200 cells/mm^3.

If CD4 counts are used in the fit and all boxes `Different by CD4 count categories` are checked (except for the interval 1980-1984), the specified value of the extra rate will not have a large impact. This is because the extra contribution to the diagnosis probability is almost entirely be offset by the estimated probabilities.

6.4.6 Confidence intervals

The tool has the option to determine confidence intervals on estimated parameters and model outcomes via a so-called bootstrap analysis (see below Intermezzo: bootstrap analysis).

A bootstrap analysis can be time-consuming because it involves running the Incidence Method 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.

Intermezzo: bootstrap analysis

The Incidence Method calculates 95% confidence intervals on estimated parameters and model outcomes by doing a bootstrap analysis. In brief, a 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 2.5th and 97.5th percentile.

Number of iterations

The user can specify the number of iterations in the bootstrap analysis, ranging from 0 to 1000. A value of 0 means that no bootstrap analysis is done. 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.

6.5 London Method advanced parameters

This section contains advanced parameter settings for the London Method.

6.5.1 Calendar year ranges

The slider bar shows the range of calendar years used in the calculations of the London Method.

The range can be made wider or narrower by dragging the rectangles at either end of the green bar. Alternatively, the lower boundary (`Start year`) and upper boundary (`End year`) of each calendar year range can be selected by clicking the up or down button in the white boxes above the slider bars.

Range of calculations

Specifies the range of the model calculations. Calculations are done for all the years from `Start year` to `End year`.

6.5.2 95% confidence intervals

The tool has the option to determine 95% confidence intervals using a simple simulation method[2]. These confidence intervals include two sources of uncertainty for the estimated number of people living with undiagnosed HIV: the stochastic uncertainty concerning the CD4 count-specific rate of symptoms and the stochastic uncertainty associated with the possibility that the observed number of HIV diagnoses with HIV-related symptoms may not correspond to the expected number based on the CD4 count-specific rate of symptoms.

Number of iterations

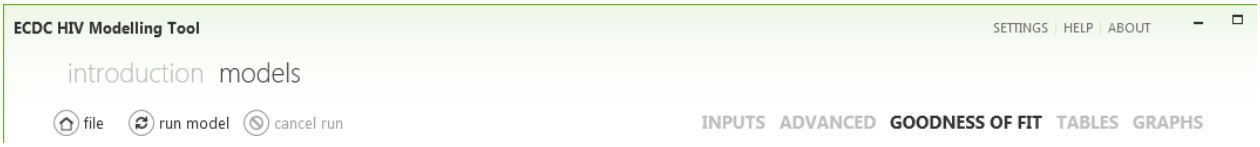
The user can specify the number of iterations in the simulation method, ranging from 0 to 100,000 (default 50,000). A value of 0 means that confidence intervals are not determined. Note that calculation of 95% confidence intervals for the London Method is much faster than for the Incidence Method and can, therefore, always be done.

7 Outputs

7.1 Goodness-of-fit

In the `goodness of fit` window you can find:

- Tables and graphs showing input data and model fits for the Incidence Method (panels to A to G).
- Summary statistics of goodness-of-fit for the Incidence Method (section H).



Each of the panels labelled to shows a table with the following columns

- `year`: calendar year.
- `N_xxx_D`: observed number of patients in the input data.
- `N_xxx_Obs_M`: 'observed' number of patients according to the best-fitting model.

The names of the columns are the same as those used in the output CSV files.

On the right of each panel is a graph showing

- observed number of patients in the input data (green line).
- best-fitting model (dashed grey line).
- 95% confidence intervals (grey band, labelled `Min-max`), only available when `Number of iterations` under `Bootstrap` in the `advanced` window is set to a value larger than zero during running of the Incidence Method.

Results are shown for the selected population in the `population list`.

A. HIV diagnoses, total

Total annual number of HIV diagnoses. This number will be the total of the four CD4 count categories (B, C, D, and E), the number of HIV/AIDS cases (G), and the number of HIV diagnoses without AIDS and no CD4 count available.

B. HIV diagnoses, CD4 ≥ 500

Annual number of HIV diagnoses with CD4 counts ≥ 500 cells/mm^3.

C. HIV diagnoses, CD4 350-499

Annual number of HIV diagnoses with CD4 counts 350-499 cells/mm^3.

D. HIV diagnoses, CD4 200-349

Annual number of HIV diagnoses with CD4 counts 200-349 cells/mm^3.

E. HIV diagnoses, CD4 < 200

Annual number of HIV diagnoses with CD4 counts <200 cells/mm^3.

F. HIV/AIDS diagnoses

Annual number of HIV/AIDS diagnoses (HIV diagnoses with a concurrent AIDS diagnosis).

G. AIDS diagnoses, total

Total annual number of AIDS diagnoses. The model prediction is likely to be larger than the observed data for calendar years after the introduction of antiretroviral treatment.

H. Goodness-of-fit statistics

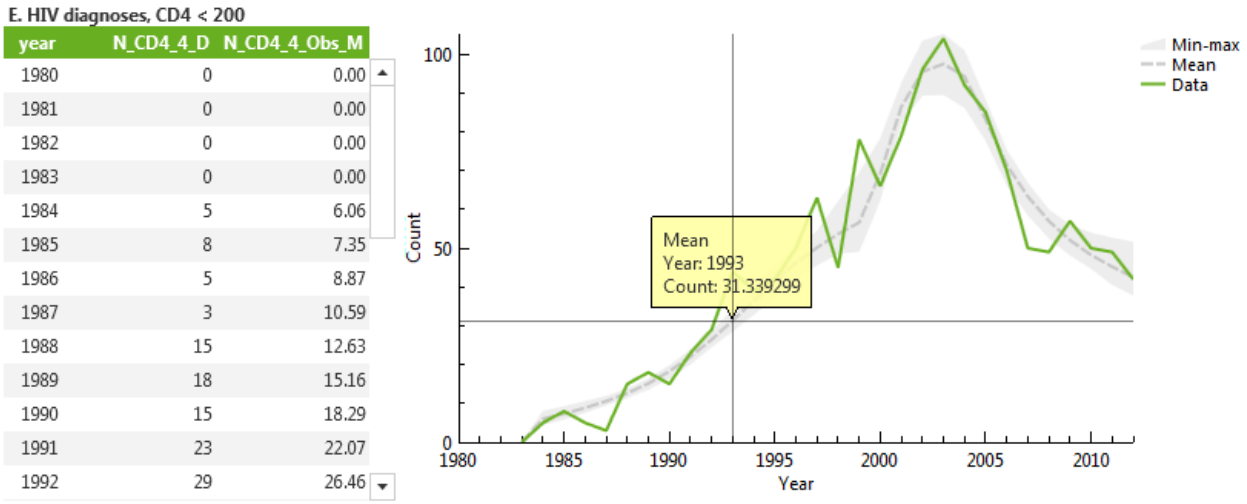
This table shows goodness-of-fit statistics in terms of the deviance of the best-fitting model:

- **Data item**: data item for which the deviance is calculated.
  - `HIV diagnoses, total`: total number of HIV diagnoses
  - `HIV diagnoses, CD4 >=500`: HIV diagnoses with CD4 counts ≥ 500 cells/mm^3.
  - `HIV diagnoses, CD4 350-499`: HIV diagnoses with CD4 counts 350-499 cells/mm^3.
  - `HIV diagnoses, CD4 200-349`: HIV diagnoses with CD4 counts 200-349 cells/mm^3.
  - `HIV diagnoses, CD4 <200`: HIV diagnoses with CD4 counts <200 cells/mm^3.
  - `HIV/AIDS diagnoses`: HIV diagnoses with a concurrent AIDS diagnosis.
  - `AIDS diagnoses, total`: total number of AIDS diagnoses.
  - `Total`: all data items together.
- **Deviance**: value of the deviance function.
- **Observation**: number of observations used in the model fit.

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.

Displaying values

By clicking with the left mouse button near or on top of the lines in the graph will show the value in the nearest calendar year.



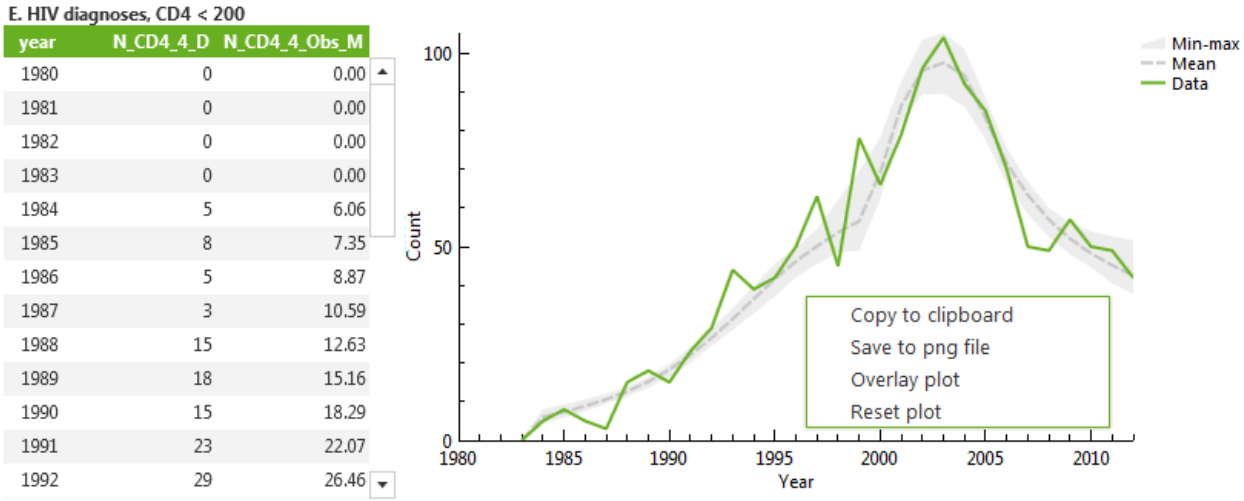
E. HIV diagnoses, CD4 < 200

year	N_CD4_4_D	N_CD4_4_Obs_M
1980	0	0.00
1981	0	0.00
1982	0	0.00
1983	0	0.00
1984	5	6.06
1985	8	7.35
1986	5	8.87
1987	3	10.59
1988	15	12.63
1989	18	15.16
1990	15	18.29
1991	23	22.07
1992	29	26.46

Copying, saving, and overlaying plots

By clicking with the right mouse button in a graph a menu will be opened:

- **Copy to clipboard**: This will copy the graph to the clipboard so that it can be pasted in e.g. Microsoft Word.
- **Save to png file**: Save the graph as a PNG file on your computer.
- **Overlay plot**: Overlay the current plot with the corresponding plot (data and best-fitting model) from another population in the `population list`.
- **Reset plot**: Clear the overlay plot and go back to the plot for the currently selected population.

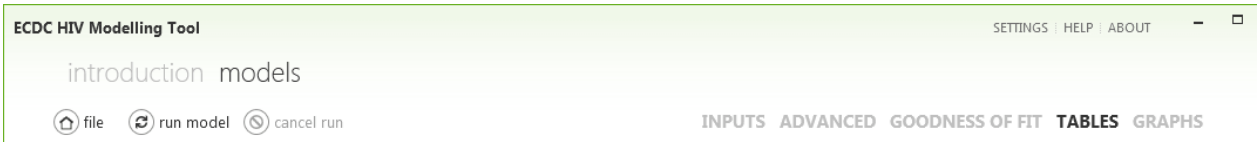


Notes

- Graphs will include confidence intervals if during running of the Incidence Method `number of iterations` is set to a value larger than zero under `Bootstrap` in the `advanced` window.
- After running the Incidence Method, all graphs will be available as PNG files in the output folder that was specified in Step 1 of the wizard.

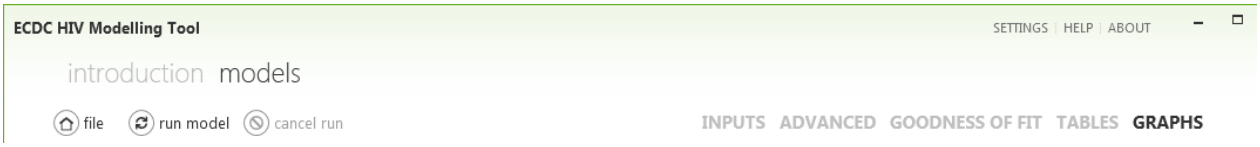
7.2 Tables

The `tables` window shows a table with all the data items and model outcomes for the Incidence Method. The names of the columns are explained in the Appendix.



7.3 Graphs

The `graphs` tab displays outcomes of the Incidence Method and the London Method.



Each of the graphs labelled A, B etc. shows a table with the following columns

- `year`: calendar year.
- model outcomes: estimated model outcome according to the best-fitting model, e.g. annual number of infections or time between infection and diagnosis.

The names of the columns are the same as those used in the output CSV files. A further explanation of the names is given in the Appendix.

On the right of each section is a figure showing

- estimated model outcome according to the best-fitting model (line).
- 95% confidence interval (band, labelled `Min-max`).

Results are shown for the selected population in the `population list`.

Graphs can be copied and saved as explained here.

7.3.1 Incidence Method

The following graphs are available for the Incidence Method:

A. HIV infections per year

Estimated number of HIV infections in each calendar year.

B. Time to diagnosis, by year of infection

Estimated average time between infection and diagnosis by year of infection if diagnosis probabilities would remain the same as in the year of infection.

C. Time to diagnosis, by year of diagnosis

Estimated average time between infection and diagnosis by year of diagnosis. This is the average duration patients have been infected by the time they are diagnosed.

D. 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 (gray), and the number living with undiagnosed HIV (blue).

E. Proportion undiagnosed of all those alive

Percentage of individuals with undiagnosed HIV amongst those living with HIV. This percentage is equal to the ratio of the blue and the green line in graph .

7.3.2 London Method

The following graphs are available for the London Method:

A. Number of undiagnosed, CD4 <200

Estimated number of individuals living with undiagnosed HIV and a CD4 count <200 cells/mm^3.

B. Number of undiagnosed, CD4 <350

Estimated number of individuals living with undiagnosed HIV and a CD4 count <350 cells/mm^3.

7.4 CSV files

Results of the Incidence Method and the London Method are written to CSV files. These files are stored in the output folder that was specified in Step 1 of the wizard.

For each population in the `population list` a separate set of CSV files are created.

A detailed description of the variables in each file is given in the Appendix.

Incidence Method

The following CSV files are created for the Incidence Method:

- **name_Result_main**: data and model outcomes for the best-fitting model.
- **name_Result_main_ConfIntervals**: estimates of the confidence intervals for a selection of model outcomes.
- **name_Result_BS**: data and model outcomes for each bootstrap run.
- **name_Param_BS**: internal and estimated parameters and goodness-of-fit statistics for the best-fitting model and each bootstrap run.

Here, *name* is the name of the selected population in the `population list`.

London Method

The following CSV files are created for the London Method:

- **name_Result_LM_1**: estimates of the number living with undiagnosed HIV and CD4 count below 200 or below 350 cells/mm^3, based on HIV/AIDS diagnoses.
- **name_Result_LM_2**: estimates of the number living with undiagnosed HIV and CD4 count below 200 or below 350 cells/mm^3, based on HIV/AIDS diagnoses and HIV diagnoses with HIV-related symptoms.

Here, **name** is the name of the selected population in the `population list`.

8 Appendix

8.1 Result_main

The CSV file `name_Result_main` contains the data, model fits to these data, and model outcomes for the Incidence Method. Model fit and outcomes distinguish between true and observed. True refers to the total number of diagnoses in a calendar year, while observed is the observed number of diagnoses taking into account, for instance, missing CD4 counts at the time of diagnosis.

Items with an asterisk (*) are the outcomes that are compared with the observed data to find the best-fitting model.

General	
`run`	ID of model fit (0: main model; 1: bootstrap)
`year`	calendar year
HIV diagnoses, total	
`N_HIV_M`	true annual number of HIV diagnoses (model)
`Cum_HIV_M`	true annual number of HIV diagnoses, cumulative (model)
`N_HIV_Obs_M*`	observed annual number of HIV diagnoses (model)
`N_HIV_D`	observed annual number of HIV diagnoses (data)
HIV diagnoses by CD4 count	
`N_CD4_1_M`	true annual number of HIV diagnoses CD4 ≥ 500 (model)
`N_CD4_1_Obs_M_NoW`	observed annual number of HIV diagnoses CD4 ≥ 500 if no missing CD4 counts (model)
`N_CD4_1_Obs_M*`	observed annual number of HIV diagnoses CD4 ≥ 500 (model)
`N_CD4_1_D`	observed annual number of HIV diagnoses CD4 ≥ 500 (data)
`N_CD4_2_M`	true annual number of HIV diagnoses with CD4 350-499 (model)
`N_CD4_2_Obs_M_NoW`	observed annual number of HIV diagnoses CD4 350-499 if no missing CD4 counts (model)
`N_CD4_2_Obs_M*`	observed annual number of HIV diagnoses CD4 350-499 (model)
`N_CD4_2_D`	observed annual number of HIV diagnoses CD4 350-499 (data)
`N_CD4_3_M`	true annual number of HIV diagnoses with CD4 200-349 (model)
`N_CD4_3_Obs_M_NoW`	observed annual number of HIV diagnoses CD4 200-349 if no missing CD4 counts (model)
`N_CD4_3_Obs_M*`	observed annual number of HIV diagnoses CD4 200-349 (model)
`N_CD4_3_D`	observed annual number of HIV diagnoses CD4 200-349 (data)
`N_CD4_4_M`	true annual number of HIV diagnoses with CD4 <200 (model)
`N_CD4_4_Obs_M_NoW`	observed annual number of HIV diagnoses CD4 <200 if no missing CD4 counts (model)
`N_CD4_4_Obs_M*`	observed annual number of HIV diagnoses CD4 <200 (model)

`N_CD4_4_D`	observed annual number of HIV diagnoses CD4 <200 (data)
AIDS diagnoses, total	
`N_AIDS_M*`	true annual number of AIDS diagnoses (model)
`N_AIDS_D`	observed annual number of AIDS diagnoses (data)
AIDS diagnoses, total	
`N_HIVAIDS_M`	true annual number of HIV/AIDS diagnoses (model)
`N_HIVAIDS_Obs_M*`	observed annual number of HIV/AIDS diagnoses (model)
`N_HIVAIDS_D`	observed annual number of HIV/AIDS diagnoses (data)
Death	
`N_Dead_D`	observed annual number of deaths among diagnosed individuals (data)
`Cum_Dead_D`	observed annual number of deaths among diagnosed individuals, cumulative (data)
`Cum_Und_Dead_M`	number of AIDS-related deaths among undiagnosed HIV-infected individuals by the end of the year, cumulative (model)
Migration	
`N_Emig_D`	observed annual number of diagnosed individuals who migrated out of the country (data)
`Cum_Emig_D`	observed annual number of diagnosed individuals who migrated out of the country, cumulative (data)
HIV infections	
`Cum_Inf_M`	number of infections by the end of year, cumulative (model)
`N_Inf_M`	annual number of infections (model)
`Cum_Inf_D`	number of infections by the end of year, cumulative (data)
`N_Inf_D`	observed annual number of infections (data)
Diagnosis probabilities and time to diagnosis	
`delta1`	diagnosis rate parameter, CD4 $\geq$ 500
`delta2`	diagnosis rate parameter, CD4 350-499
`delta3`	diagnosis rate parameter, CD4 200-349
`delta4`	diagnosis rate parameter, CD4 <200
`t_diag`	estimated average time between infection and diagnosis by year of infection *when diagnosis probabilities remain the same as in the year of infection*
`D_Avg_Time`	estimated average time between infection and diagnosis by year of diagnosis, i.e. the average duration patients have been infected by the time they are diagnosed
`t_diag_i`	fraction ($\times$ 1000) of individuals diagnosed in the i^{th} calendar year among those infected in calendar year year
People living with HIV	

N_Alive	total number of HIV-positive individuals who are still alive. This number is equal to the total number of infected individuals (Cum_Inf_M) minus the observed number of individuals who died (Cum_Dead_D) minus the estimated number of individuals who died before being diagnosed with HIV (Cum_Und_Dead_M).
N_Alive_Diag_M	total number of diagnosed HIV-positive individuals who are still alive
Undiagnosed population (at the end of the year)	
N_Und	number of undiagnosed individuals, total
N_Und_Inf_p	proportion undiagnosed individuals of those ever infected
N_Und_Alive_p	proportion undiagnosed individuals of those still alive
N_Und_PrimInf	number of undiagnosed individuals, primary infection
N_Und_CD4_1_M	number of undiagnosed individuals, CD4 ≥ 500
N_Und_CD4_2_M	number of undiagnosed individuals, CD4 350-499
N_Und_CD4_3_M	number of undiagnosed individuals, CD4 200-349
N_Und_CD4_4_M	number of undiagnosed individuals, CD4 <200
N_Und_HIVAIDS_M	number of undiagnosed individuals, AIDS
Undiagnosed population (at the end of year, infected in the same year)	
N_Und_T_1	number of undiagnosed individuals, total
N_Und_T_1_p	proportion of N_Und
N_Und_PrimInf_T_1	number of undiagnosed individuals, primary infection
N_Und_CD4_1_T_1	number of undiagnosed individuals, CD4 ≥ 500
N_Und_CD4_2_T_1	number of undiagnosed individuals, CD4 350-499
N_Und_CD4_3_T_1	number of undiagnosed individuals, CD4 200-349
N_Und_CD4_4_T_1	number of undiagnosed individuals, CD4 <200
N_Und_CD4_5_T_1	number of undiagnosed individuals, AIDS
Undiagnosed population (at the end of year, infected for 1 to 4 years)	
N_Und_T_2	number of undiagnosed individuals, total
N_Und_T_2_p	number undiagnosed, proportion of N_Und
N_Und_PrimInf_T_2	number of undiagnosed individuals, primary infection
N_Und_CD4_1_T_2	number of undiagnosed individuals, CD4 ≥ 500
N_Und_CD4_2_T_2	number of undiagnosed individuals, CD4 350-499
N_Und_CD4_3_T_2	number of undiagnosed individuals, CD4 200-349
N_Und_CD4_4_T_2	number of undiagnosed individuals, CD4 <200
N_Und_CD4_5_T_2	number of undiagnosed individuals, AIDS
Undiagnosed population (at the end of year, infected for 5 years or more)	
N_Und_T_3	number of undiagnosed individuals, total

N_Und_T_3_p	number undiagnosed, proportion of N_Und
N_Und_PrimInf_T_3	number of undiagnosed individuals, primary infection
N_Und_CD4_1_T_3	number of undiagnosed individuals, CD4 ≥ 500
N_Und_CD4_2_T_3	number of undiagnosed individuals, CD4 350-499
N_Und_CD4_3_T_3	number of undiagnosed individuals, CD4 200-349
N_Und_CD4_4_T_3	number of undiagnosed individuals, CD4 <200
N_Und_CD4_5_T_3	number of undiagnosed individuals, AIDS

Goodness of fit	
LL_HIV	deviance function total HIV diagnoses
LL_CD4_1	deviance function CD4 ≥ 500
LL_CD4_2	deviance function CD4 350-499
LL_CD4_3	deviance function CD4 200-349
LL_CD4_4	deviance function CD4 <200
LL_HIVAIDS	deviance function HIV/AIDS diagnoses
LL_AIDS	deviance function AIDS diagnoses

8.2 Result_BS

The CSV file `name_Result_BS` has the same structure as `name_Result_main`. This file contains bootstrap replicates of the data and model outcomes for each bootstrap run.

8.3 Result_main_ConfIntervals

The CSV file `name_Result_main_ConfIntervals` contains estimates of the confidence intervals for a selection of model outcomes. Confidence intervals are updated after every bootstrap run. Suffixes LB and UB represent the lower and upper boundary of the confidence interval.

8.4 Param_BS

General	
run	ID of model fit (0: main model; 1: bootstrap)
runtime	approximate time to complete the run
yr_knots_add	not used

Internal parameters	
mu	not used
alphaP	rate of progression from acute to chronic infection
q_1	rate of progression from CD4 ≥ 500 to CD4 350-499
q_2	rate of progression from CD4 350-499 to CD4 200-349
q_3	rate of progression from CD4 200-349 to CD4 <200
q_4	rate of progression from CD4 <200 to AIDS
q_5	rate of progression from AIDS to death
q_6	not used
f_1	proportion with CD4 ≥ 500 directly after primary infection
f_2	proportion with CD4 350-499 directly after primary infection

`f_3`	proportion with CD4 200-349 directly after primary infection
`f_4`	proportion with CD4 <200 directly after primary infection
`f_5`	proportion with AIDS directly after primary infection
`f_6`	not used
`d4_fac`	extra rate due to non-AIDS symptoms
Modifiable and estimated parameters	
`t_i`	boundaries of the time intervals for diagnosis probabilities
`theta_i`	estimated parameters associated with the HIV incidence curve
`beta_i`	estimated parameters associated with the diagnosis probabilities
Maximum likelihood	
`r_AIDS`	dispersion parameter, AIDS diagnoses
`r_AIDSPos`	dispersion parameter, HIV/AIDS diagnoses
`r_Pos`	dispersion parameter, HIV diagnoses, total
`r_PosCD4`	dispersion parameter, HIV diagnoses, by CD4 count
Number of observations used in the model fit	
`N_LL_HIV`	number of observations, HIV diagnoses total
`N_LL_CD4_1`	number of observations, HIV diagnoses CD4 ≥ 500
`N_LL_CD4_2`	number of observations, HIV diagnoses CD4 350-499
`N_LL_CD4_3`	number of observations, HIV diagnoses CD4 200-349
`N_LL_CD4_4`	number of observations, HIV diagnoses CD4 <200
`N_LL_HIVAIDS`	number of observations, HIV/AIDS diagnoses
`N_LL_AIDS`	number of observations, AIDS diagnoses
Goodness-of-fit	
`LL_HIV`	deviance function, HIV diagnoses total
`LL_CD4_1`	deviance function, HIV diagnoses CD4 ≥ 500
`LL_CD4_2`	deviance function, HIV diagnoses CD4 350-499
`LL_CD4_3`	deviance function, HIV diagnoses CD4 200-349
`LL_CD4_4`	deviance function, HIV diagnoses CD4 <200
`LL_HIVAIDS`	deviance function, HIV/AIDS diagnoses
`LL_AIDS`	deviance function, AIDS diagnoses
`LL_Smooth_1`	not used
`LL_Smooth_2`	not used
`LL_Poisson`	deviance function, Poisson distribution
`LL_Total`	deviance function, total
`X2_Total`	χ^2 statistic, calculated as `LL_Total` divided by the number of observations used in the fit minus the number of estimated parameters
`X2_Pearson`	χ^2 statistic, calculated as the sum of the squared difference of each data point and the corresponding model estimate divided by the variance

`X2_Pearson_n`	χ^2 statistic, calculated as `X2_Pearson` divided by the number of observations used in the fit minus the number of estimated parameters

8.5 Result_LM

The CSV file `name_Result_LM_1` contains results of the London Method based HIV/AIDS diagnoses. The file `name_Result_LM_2` contains results based on HIV/AIDS diagnoses and HIV diagnoses with HIV-related symptoms.

General	
`year`	calendar year
Estimates based on selected population only	
`w_missing`	correction factor for symptomatic diagnoses with missing CD4 count
`N_200_D`	observed annual number of symptomatic diagnoses CD4 <200 (data)
`N_200_M`	estimated number of individuals CD4 <200 (model)
`N_200_M_lcl`	estimated number of individuals CD4 <200, lower bound (model)
`N_200_M_ucl`	estimated number of individuals CD4 <200, upper bound (model)
`N_350_D`	observed annual number of symptomatic diagnoses CD4 <350 (data)
`N_350_M`	estimated number of individuals CD4 <350 (model)
`N_350_M_lcl`	estimated number of individuals CD4 <350, lower bound (model)
`N_350_M_ucl`	estimated number of individuals CD4 <350, upper bound (model)
Estimates based on all populations in the model	
`w_ALL_missing`	correction factor for symptomatic diagnoses with missing CD4 count
`N_200_ALL_D`	observed annual number of symptomatic diagnoses CD4 <200 (data)
`N_200_ALL_M`	estimated number of individuals CD4 <200 (model)
`N_200_ALL_M_lcl`	estimated number of individuals CD4 <200, lower bound (model)
`N_200_ALL_M_ucl`	estimated number of individuals CD4 <200, upper bound (model)
`N_350_ALL_D`	observed annual number of symptomatic diagnoses CD4 <350 (data)
`N_350_ALL_M`	estimated number of individuals CD4 <350 (model)
`N_350_ALL_M_lcl`	estimated number of individuals CD4 <350, lower bound (model)
`N_350_ALL_M_ucl`	estimated number of individuals CD4 <350, upper bound (model)

References

1. 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.

2. Lodwick RK, Nakagawa F, van Sighem A, Sabin CA, Phillips AN. Use of surveillance data on HIV diagnoses with HIV-related symptoms to estimate the number of people living with undiagnosed HIV in need of antiretroviral therapy. *PLoS ONE*. 2015;10(3). doi:10.1371/journal.pone.0121992.

3. Centers for Disease Control and Prevention. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *Clinical Infectious Diseases*. 1992;41(RR-17):802-810. doi:10.1093/clinids/17.4.802.