Thembisa version 4.3 user guide

Leigh Johnson
March 2021

Contents

1. Introduction

The Thembisa model is a mathematical model of the HIV epidemic in South Africa, and is important as a source of epidemiological statistics. It is also a demographic projection model and can produce a wide range of South African demographic statistics.

Two versions of the Thembisa model have been created: one programmed in C++, the other programmed in Excel and Visual Basic for Applications (VBA). The two versions produce almost identical results, although there are some small differences. The Excel/VBA version has the advantage of being easier to understand and use, especially for users who are not familiar with C++. However, it has the disadvantage of being slow to run. The Excel version of the model has been made publicly available, for the benefit of individuals who wish to use the model but who have limited programming experience. This user guide explains how to use the Excel version of the model.

2. General points

Colour scheme: All cells that can be changed by the user are formatted in red. Cells that are formatted in black are cells that should not be changed by the user (in the case of numerical values, these are in almost all cases calculated values). It is worth noting that even though many of the red cells can be changed, the user should exercise caution in doing so: some changes could lead to the model producing results that are inconsistent with published South African demographic and HIV statistics. If the user changes the input assumptions from the default values, the results should not be attributed to the Thembisa model without including a clear explanation of the changes that were made to the input parameters (and ideally also some justification for why the changes were made).

Some of the calculated values are formatted in grey. The grey shading indicates that the formula in a particular cell is different from that in the row above, and that the user should therefore not copy down the formula from the previous row. However, as most users should not need to change any of the spreadsheet formulas, the distinction between the grey and clear cells is not important.

Calendar years versus projection years: The Thembisa model works by projecting changes in the population at monthly time steps (one twelfth of a year), starting in the middle of 1985. Although changes are calculated at monthly time steps, results are reported at annual intervals, which run from mid-year to mid-year (in line with conventions in demographic projection). This means that model outputs that are stock variables (e.g. total HIV infections, population size, fraction of HIV-positive individuals receiving ART) are reported at the *middle* of each year. Similarly, model outputs that are flow variables (e.g. new HIV infections, AIDS deaths, total births, infant mortality rate) are calculated over the interval from the middle of the stated calendar year to the middle of the next calendar year. For example, if the number of AIDS deaths in the 1995 column of the 'Results' sheet is 25055, this means that there were 25055 AIDS deaths between mid-1995 and mid-1996.

This point is important not only in understanding the model outputs but also in understanding the model inputs. The input parameters are specified using the same time periods. For example, the starting population size assumptions in columns B and C of the 'Population' sheet specify the population size at the *middle* of 1985. If it is assumed in the 'Rollout' sheet that 54.5% of pregnant women in 2005 received antenatal HIV testing (row 20), that means that of all women who delivered between mid-2005 and mid-2006, 54.5% received antenatal testing. Similarly, if it is assumed in the 'Non-HIV mort' sheet that the infant mortality rate in boys in 1992 was 0.0450, this means that between mid-1992 and mid-1993 a mortality probability of 0.0450 applies during the first year of life.

The only place in the model in which flow variables are presented by calendar year (i.e. from the start of the year to the end of the year) is in the calibration of the model to recorded death data (see the 'Calibration' and 'Death calibration' sheets). We deviate from the mid-year to mid-year convention in this case because the reported death data are published over calendar years.

Age stratification: At ages 10 and older, the population is stratified according to individual age at the start of the projection year (i.e. at the middle of the relevant calendar year). The age refers to the age last birthday (i.e. rounded down, not rounded to the nearest integer). Age is updated only at the end of each projection year, which means that over the course of a particular projection year an individual will continue to be stratified occurring to their age at the start of the projection year even if they have subsequently had their birthday.

Individuals aged 90 and older are not stratified by age. They are assumed to all be subject to the same rate of mortality. The '90' age category in the model should therefore be interpreted as '90 or older'.

Children under the age of 10 are stratified by age in months, since mortality rates change substantially by month, particularly in the first year of life. Because of the monthly stratification, age is updated at monthly intervals (unlike age in adults, whose ages are only updated at annual intervals).

Quoting the model: In general, we ask that users of the Thembisa model cite the most recent journal publication that describes the Thembisa model, when they quote outputs from the Thembisa model. Currently the recommended citation is the following:

Johnson L.F., May M.T., Dorrington R.E., Cornell M., Boulle A., Egger M. and Davies M.A. Estimating the impact of antiretroviral treatment on adult mortality trends in South Africa: a mathematical modelling study. *PLoS Medicine* 2017; 14(12): e1002468.

In some cases it may be more appropriate to quote earlier papers, for example, if you are quoting an earlier version of the Thembisa model, or if the aspect of the model that you are interested in is better described in an earlier publication.

3. Model structure

Most of the calculations in the model are performed in the Excel spreadsheets, but the code for updating the model at monthly time steps is contained in the VBA component of the model. The macro buttons on the first 'Projection' worksheet call the code in the VBA component of the model, and thus run the model.

Within the Excel component of the model, there are a large number of spreadsheets, which serve different functions. There are broadly four types of spreadsheet:

- Input sheets: These contain model assumptions (e.g. 'Adult assumptions' 'Paed assumptions', 'Rollout', 'Fertility' and 'Migration').
- Calculation sheets: These contain calculations of key parameters in the current projection year (e.g. 'Population', 'Sex activity', 'Mixing', 'Condoms' and 'Circumcision').
- Risk group sheets: These contain calculations of changes over the course of a single month in specific sub-populations or risk groups (e.g. 'MHU virgin', 'MHC ST', 'FH SW', 'FL LTH' and 'Child M').
- Results/calibration sheets: These contain the outputs of the model, or show how well the model outputs compare with published epidemiological statistics (e.g. 'Results', 'Calibration', 'Monthly', 'ANC calibration' and 'Death calibration').

A few spreadsheets can be considered both input sheets and calculation sheets (e.g. the 'ART' sheet contains both input parameters that determine CD4 changes after ART initiation and calculations of the CD4 distributions in ART patients).

For the vast majority of model users, the important sheets to focus on are the 'Projection', 'Adult assumptions', 'Rollout' and 'Results' sheets. These are briefly summarized below; a more detailed description of all the spreadsheets is provided in the appendix.

The **'Projection'** sheet contains the controls for running the model. The 'Reset' macro button clears the model of previously-generated results and resets the projection to the starting position (mid-1985). Cell E9 specifies the year to which the user wants to run the model, and the 'Project to specified year' macro button runs the model to this specified year.

The **'Adult assumptions'** sheet contains assumptions about sexual behaviour, HIV transmission, rates of CD4 decline and rates of AIDS mortality. The sheet also contains assumptions about the efficacy of different prevention methods, the prevalence of male circumcision prior to the promotion of male circumcision as an HIV prevention strategy, and factors affecting the frequency of HIV testing. For a more detailed description of the model assumptions, the user is referred to the reports describing the Thembisa model.

The **'Rollout'** sheet contains the assumptions about access to different HIV prevention and treatment services. The rows represent the individual programme elements and the columns represent the different projection years.

The **'Results'** sheet contains the model outputs. The outputs are listed in column A, and in subsequent columns the outputs for each projection year are shown (note that if you have just clicked the 'Reset' button, these columns will be blank). Column B contains the results for

the current projection year (the most recent year for which the model results have been generated). Additional outputs can be added to the Results sheet.

4. How to run the model

Suppose you want to produce model outputs for each year from the start year of the projection (1985) to the 2019 projection year. To do this:

- Clear the model of any previously-generated results. To do this, click the 'Reset' button in the 'Projection' worksheet. (If you go to the Results sheet, you will see that from column C onwards the output table is blank.)
- Enter the year 2019 in cell E9 of the 'Projection' sheet.
- Click the 'Project to specified year' button in the 'Projection' worksheet.

Once the simulation has finished running, you will find in the 'Results' sheet that outputs have been stored for each year from 1985 to 2019 (in columns C to AK).

Note that the model is very slow and it may take as long as an hour for the model to run such a simulation, depending on your PC specifications. The model also requires a lot of memory, so you may find it difficult to work on other tasks on the same PC while the model is running. For this reason, you might choose to run the model before you take a lunch break or before you go home – or alternatively, work on a different device while your PC is running the Thembisa model. A message in the lower toolbar tells you how far the simulation has progressed.

If after having run the model to 2019 you decide that you need the results for another year (2020), the easiest way to generate this additional year of outputs is to click the 'Project one year' button. This is much quicker than resetting the model and running the model to 2020. An alternative (equally quick) way to generate the additional year of outputs is to change cell E9 in the 'Projection' sheet to 2020 and click the 'Project to specified year' button (but don't click the Reset button before doing this or you will lose the results you generated previously and will have to start again from scratch).

5. How to generate non-standard outputs

Most of the outputs that you are likely to need are already calculated in the 'Results' sheet. However, there may be an output you require that is not in the 'Results' sheet. If this is the case, you will need to add a formula in column B of the 'Results' sheet to calculate this additional output. Insert a row at any point in the worksheet between row 3 and the dashed line at the bottom of the worksheet (where you see the comment "Insert additions above this row"). In column A of the new row, type a brief explanation of what the output is. Then in column B of the new row, type a formula for calculating the output. When you run the model, the new output will be calculated and stored for each year that you run the model.

Determining an appropriate formula is not trivial, and it is difficult to give generic guidance. We will present a number of examples below to illustrate the key ideas. You may also find it helpful to look at how other outputs in the 'Results' sheet are calculated, particularly if there is an existing output that is closely related to the output that you are attempting to calculate.

The process of calculating the outputs is tricky because although the model projects changes over monthly time steps, outputs are reported only at annual intervals. The monthly outputs are aggregated and stored in the 'Monthly' sheet, before being used for the annual output calculations in the 'Results' sheet. This means that if you are defining a new output in the 'Results' sheet, you might also need to define a new output in the 'Monthly' sheet.

Example 1: AIDS deaths in untreated adults

The easiest outputs to calculate are those that can be calculated directly from the outputs that have already been defined in the 'Results' sheet. For example, suppose we wish to calculate the annual number of AIDS deaths in HIV-positive individuals who have not received ART. This is just the difference between the total AIDS deaths (row 7) and the AIDS deaths in ART patients (row 9). To calculate the output:

- Insert a row below row 9.
- In column A of the new row, enter a description of the new output (e.g. "AIDS deaths in untreated patients").
- In column B enter the following formula to calculate the output: '=B7-B9'.

This output will now be calculated and stored (in the columns to the right of column B) for each year that you run the model.

Example 2: Total HIV infections by sex

Some outputs can also be calculated directly from the 'Monthly' sheet. For example, suppose we want to know the total number of HIV infections in men and women separately. Currently the total number of infections is reported for men and women combined (row 89), but is not disaggregated by sex. From the formula in cell B89 it should be clear that the formulas for the male and female totals should be '=SUM(Monthly!D326:D416)' and '=SUM(Monthly!E326:E416)' respectively. If you go to the 'Monthly' sheet, you will see that the ranges referred to contain the number of HIV-positive males and females at each age, at the *start* of the current projection year, copied from the 'Population' sheet.

You might think that an alternative way to calculate the output for males would be '=SUM(Population!G6:G96)', since the values in the 'Monthly' sheet are copied from the 'Population' sheet. However, this would not give you the same result as the calculation in the previous paragraph, since the values in the 'Population' sheet only get copied to the 'Monthly' sheet at the *start* of each projection year, whereas the values in the 'Population' sheet get updated at each monthly time step. Since stock variables are always reported at the start of each projection year (or equivalently, at the middle of the corresponding calendar year), you should rather refer to the values in the 'Monthly' sheet.

Example 3: HIV incidence in women aged 15-19

Currently the 'Results' sheet contains a calculation of HIV incidence in women aged 15-24 (see row 81). If you wanted to calculate the HIV incidence rate in women aged 15-19, you would modify the formula in row 81 to:
'=SUM(Monthly!X71:X75)/(SUM(Monthly!C341:C345)-SUM(Monthly!E341:E345))'.

Example 4: Number of new HIV infections in sex workers

This is a more complex output to calculate, because there are no existing outputs in either the 'Results' sheet or the 'Monthly' sheet from which the output can be calculated. This means that in addition to adding the new output in the 'Results' sheet, you also need to add a corresponding row in the 'Monthly' sheet. Follow these steps:

- In the 'Monthly' sheet, insert a row below row 5. Note that this section of the spreadsheet is labelled "Flow variables that are updated monthly". This is the appropriate section of the sheet in which to add the new output, since the number of new infections is a flow variable (not a stock variable) and since the number of new infections gets calculated at monthly time steps.
- In column A of the new row, add an appropriate output description (e.g. "New infections in sex workers").
- In column B of the new row, add the following formula: '=SUM('FH SW'!AV93:AV173)+C6'. The first part of this formula refers to the 'FH SW' sheet for the number of new HIV infections in sex workers in the *current* month; the second part of the formula (C6) is the cumulative number of new infections in sex workers in *previous* months (from the start of the current projection year). At the end of each month, the value in B6 gets copied to C6, so that the value in column C represents the cumulative total.
- Insert a new row in the 'Results' sheet for calculating the new output, and in column B of this new row enter the formula '=Monthly!C6'.

6. How to run counterfactual scenarios

The default parameters in the model describe the actual evolution of the South African HIV epidemic up to the current time. However, we might be interested in answering questions such as "How different would South Africa's demographic profile have been in the absence of HIV?" and "How many life years have been saved by ART?" To answer these questions we need to compare the model estimates of what has actually happened with *counterfactual* scenarios that represent 'what might have been'. For example, calculating the number of life years saved by ART would involve subtracting the estimated number of life years lost due to AIDS in the baseline scenario from that estimated in a counterfactual scenario in which there is no ART provision. The following are the most commonly considered counterfactual scenarios:

- *'No ART' scenario*: To evaluate what would have happened in the absence of any ART provision, set to zero all the ART rollout rates (rows '31-48' of the Rollout sheet). In addition, in the 'Adult assumptions' sheet, set the "Year to which ART initiation numbers are specified" (row 127) to some year beyond the term of your projection year (if you are only running the model up to 2017, it is not necessary to change this parameter from the baseline value of 2017). Similarly, in the 'Paed assumptions' sheet, change the "Year to which ART initiation numbers are specified" (row 42) to some year beyond the term of your projection year
- *'No HCT or ART' scenario*: It is not possible to run a scenario in which there is ART but no HIV counselling and testing (HCT), since individuals can only start ART if they have been tested for HIV and diagnosed positive. So you can only consider a 'no HCT' counterfactual if you also exclude ART in the counterfactual. To do this, follow

the same steps as above. In addition set all values in rows 5, 6, 13, 14, 20 and 22 of the 'Rollout' sheet to zero.

- *'No MMC' scenario*: To run a scenario in which there are no medical male circumcision (MMC) campaigns, make sure that only zeros are entered in row 78 of the 'Rollout' sheet. In addition, set the "Ultimate annual probability of MMC in boys aged 10-14" (in the 'Adult assumptions' sheet) from 25% to 0. Note that making these changes will not mean that there is no male circumcision; it will only mean that there is no increase in rates of male circumcision relative to what would have been expected in the absence of any campaigns to promote male circumcision as an HIV prevention strategy.
- *'No HCP/condom promotion' scenario*: Suppose we want to consider what would have happened if there had been no HIV communication programmes (HCPs) or condom promotion, and consequently no increase in condom use since the start of the HIV epidemic. To run such a scenario, go to the Condoms sheet. In cell E3 change the formula to just '=InitCondom *'Adult assumptions'!K41'.
- *'No AIDS' scenario*: The easiest way to consider what would have happened in the absence of an AIDS epidemic is to set the "Initial HIV prevalence among FSWs and female high risk group" (K61 of the 'Adult assumptions' sheet) to zero.

Note that all of the changes described above are changes you should make before running the simulation. In other words, make the above changes, then click the Reset button, then click the 'Projected to specified year' button to generate the results for the counterfactual scenario.

It also worth remembering that if you save the model after having generated a counterfactual scenario, you should save it under a different name (preferably a name that is descriptive of the counterfactual scenario), otherwise it may be difficult to get back to the default model assumptions (or even worse, you may forget that the model you've saved doesn't correspond to the default scenario that you saved).

7. How to model future changes to current policy

Although it is not possible to cover all possible changes to current HIV prevention and treatment policy, we have listed here some of the scenarios that model users may be particularly keen to consider, and have in each case explained how to adapt the model to consider these scenarios:

- *Rapid initiation of ART*. Suppose we wish to consider what would happen if from mid-2020 onwards, the fraction of individuals diagnosed with HIV who immediately start ART increases from its default value of 40% to a level of 80%. In the Rollout sheet, we would change the values in the "% of other newly-diagnosed ART-eligible adults who start ART" row (row 47) to 0.80 from 2020 onwards (i.e. change cell AP47 to 0.80, and copy this value across for all columns up to column BJ). Make a similar change to row 46 (for patients with opportunistic infections). Note that the fraction is already at 95% for pregnant women, so no changes are required in row 45.
- *Intensified ART adherence support*. Suppose we wish to consider what would happen if from mid-2020 onwards, the fraction of ART patients who are virally suppressed (viral load <400 copies/ml) increases from its default value of 92.6% to a level of 95%. In the Rollout sheet, we would change the values in the "% of adults starting

ART with CD4 <200 who have VL <400 copies" row (row 49) to 0.95 from 2020 onwards.

- *Early infant male circumcision*. Suppose we wish to consider what would happen if from mid-2020 onwards, the fraction of male neonates who are circumcised neonatally increases from its default value of 10.5% to a level of 21%. In the Rollout sheet, we would change the values in the "% of neonates circumcised" row (row 84) to 0.21 from 2020 onwards. Note that because of the delay between the circumcision and the start of sexual activity, this intervention would have no impact on HIV incidence in the short term.
- *Pre-exposure prophylaxis (PrEP) for youth*. Suppose we wish to consider what would happen if from mid-2020 onwards, PrEP were offered to all youth aged 15-24, and there was an annual rate of PrEP initiation of 20 per 100 person years in this age group in sexually-active high-risk individuals. In the Rollout sheet, we would change the values in rows 56, 57, 62 and 63 to 0.2 from 2019 onwards.

8. How to generate province-specific outputs

The default model assumptions relate to South Africa as a whole. To generate province-specific estimates, we use the same model, but with different input parameters. Most of the province-specific outputs that you are likely to need are contained in the model output file that is available for download. However, there may be some non-standard outputs that you want to calculate at a provincial level, which are not in this output file. If this is the case, you should see the previous section ("How to generate non-standard outputs") for instructions on how to add formulas to the 'Results' sheet for the new outputs that you want to calculate. In addition, you should follow these steps for the purpose of getting the province-specific inputs:

- Download the 'AssumptionsProv for THEMBISA 4.3v2.xlsm' file (or whatever the most recent version is), and save it to the same folder as that in which you've saved the Thembisa model that you're working with. (It's very important that the Thembisa model and 'AssumptionsProv' workbook be in the same folder, otherwise the macros in the latter won't work.)
- The 'AssumptionsProv for THEMBISA 4.3v2.xlsm' file contains the input parameters for all provinces (as well as the country as a whole). Make sure you have opened this file and the version of the Thembisa model into which you want to copy the province-specific parameters.
- In the 'Initializer' sheet of the 'AssumptionsProv' workbook, enter the name of the file into which you want to copy province-specific assumptions in cell D3.
- From the drop-down list in this sheet, select the province for which you want to generate results. (You can also select 'SA' if you previously copied in the results for a particular province and now want to re-enter the parameters for the country as a whole.)
- Click the "Copy assumptions to selected model" button on the right-hand side. This will copy all of the parameters for the province that you selected over into the version of the Thembisa model that you selected.
- You can now close the 'AssumptionsProv' workbook. Click the Reset button in the Thembisa model before starting the projection and then proceed with running the Thembisa model in the same way as before.

This process will obviously only give you the results for one province. If you want the results separately for all nine provinces, you need to go through the same steps (apart from the first and second) for each province.

9. Trouble shooting

We have listed here some of the common problems that people encounter when trying to run the model, and the corresponding solutions.

Nothing happens when I click on the macro buttons.

This is usually because you haven't enabled the macros. When you open the spreadsheet, you will see a yellow bar just below the top of the screen, with a button that says 'Enable macros'. You need to click on this 'Enable macros' button before any of the macros will work.

The model takes an extremely long time to run and the results look like garbage.

This can happen if you've set the default calculation mode to 'Automatic'. The model only produces sensible results when the calculation mode is 'Manual'. This should be the default when you run the model, but if for some reason it isn't, the best way to correct this is to go to the 'Formulas' tab, and where you see 'Calculation options', select 'Manual' from the drop-down list.

The model runs very slowly, and/or a run-time error message appears before the projection has completed, and/or the PC freezes while the model is running.

Check which version of Excel you are using. There are known problems with Excel 2013 and Excel 2016 (and possibly newer versions), which cause problems with other spreadsheet models (not just Thembisa). If you attempt to run the model in Excel 2013 or 2016, the model will run very slowly, and the run-time error messages appear to occur randomly (i.e. sometimes they will appear and sometimes they won't). If you encounter problems, try to run the model again. Even if the model appears to run fine, please check that the results you get are roughly consistent with the published outputs in the ProvOutput file (bearing in mind that the ProvOutput file is generated using the C++ version of the Thembisa model, which is not identical to the Excel version).

The results I'm getting aren't consistent with those in the ProvOutput file.

This might be because you didn't follow the instructions closely. A common mistake when using the AssumptionsProv workbook together with Thembisa is to first click the Reset button in Thembisa, then copy the assumptions in AssumptionsProv over to Thembisa and then run the projection. The correct sequence is to *first* copy the assumptions in AssumptionsProv over to Thembisa, *then* click the Reset button in Thembisa, and finally run the projection. Also bear in mind that the ProvOutput file is generated using the C++ version of the Thembisa model, which is not identical to the Excel version, and thus exact consistency with the ProvOutput file shouldn't be expected.

If you encounter other problems, please inform Leigh Johnson (Leigh.Johnson@uct.ac.za).

10. Caveats

Models are powerful tools, but in the wrong hands they can lead to bad conclusions. Users should note the following points to avoid the worst of these pitfalls.

Projections of the future

Making projections of what will happen in the future is difficult. Scientists generally produce projections of what they expect to happen if current trends continue and there are no major changes in government policies or external 'game changing' events. In the same spirit, the Thembisa projections of what we expect in the future are based on the prevention and treatment programmes that are currently in place in South Africa, assuming trends in access to services and trends in sexual behaviour similar to those observed in the recent past. However, the future may turn out to be very different if, for example, an effective HIV vaccine were developed, or if patterns of sexual behaviour were to change substantially. Even the recent past can be difficult to 'predict', because the data on which we base our projections are often a few years out of date. For example, HIV incidence estimates are usually based indirectly on HIV prevalence data. This means that if HIV prevalence data are not available beyond 2017, HIV incidence estimates from the model can only be considered to have been determined from empirical data in the period up to 2017; beyond that, model estimates of HIV incidence are only a reflection of the assumed continuation of past trends. In the output file we have tried to distinguish between those model outputs that can be considered 'calibrated' to data (formatted in black) and those that are projections beyond the period for which we have data (formatted in grey). Because the data used in calibration are less recent for some outputs than others, the grey cells don't all start in the same year. Users should exercise caution in quoting these outputs that are formatted in grey, as they are *projections* based on past trends, and we lack recent data to confirm or validate these estimates.

Confidence intervals

It is also worth noting that even when we have recent data for the purpose of calibrating the model, model estimates may still be imprecise, and there may be wide ranges of uncertainty around the model estimates. In the Excel output file, which is produced using the C++ version of the Thembisa model, we have included 95% confidence intervals around the model outputs. However the Excel version of the model does not produce confidence intervals, because the Excel model is slow to run, and the generation of 95% confidence intervals requires many thousands of simulations, which would be impractical when using the Excel version of the model.

Changes to default parameters and formulas

In general, users should exercise extreme caution when changing the default parameters in the Thembisa model. Changing the default parameters can lead to the model producing implausible estimates. Before changing a parameter, make sure you have read and understood the Thembisa working paper and the justification given for the relevant parameter value, or the explanation of the method used to set the model parameter. It is also important to make sure that after you have changed a parameter value, the model still produces results that make sense. At a minimum, you should check the calibration sheets ('CD4 calibration', 'ANC calibration', 'HSRC calibration' and 'Death calibration') to make sure that the model fits to the data are not worse than they were before you made the change.

The same comments apply to changes to default formulas in the Thembisa model. Do not modify the existing formulas unless you are confident that you understand the logic behind the formulas.

Checking output consistency

If you need to run the Thembisa model, you should check that the outputs you obtain are roughly consistent with the published outputs in the ProvOutput file. This is important because although the vast majority of users are able to run the model without experiencing any problems, some users have reported problems with the model either producing implausible results or producing error messages. This appears to happen particularly with the newer versions of Excel (Excel 2013 and to a lesser extent Excel 2016). We have made several changes to the macros to minimize the risk of such errors, but as a safety precaution we recommend that you check that your outputs are consistent with the published outputs in the ProvOutput file. If there is an inconsistency, the likely explanation is that something went wrong when you ran the model – try to run the model again and see if the problem resolves. If you are not able to resolve the problem, please let us know.

Appendix: Worksheets in the Thembisa model

This appendix lists the worksheets in the Thembisa model in order of appearance, and briefly explains what each spreadsheet does and how it is structured.

Projection

This sheet contains the controls for running the model. The 'Reset' macro button clears the model of previously-generated results and resets the projection to the starting position (mid-1985). Cell E9 specifies the year to which the user wants to run the model, and the 'Project to specified year' macro button runs the model to this specified year.

Adult assumptions

This sheet contains assumptions about sexual behaviour, HIV transmission, rates of CD4 decline and rates of AIDS mortality. The sheet also contains assumptions about the efficacy of different prevention methods, the prevalence of male circumcision prior to the promotion of male circumcision as an HIV prevention strategy, and factors affecting the frequency of HIV testing.

The sheet starts with the assumptions about sexual behaviour. These are divided into four groups of assumptions: (a) rates of short-term partnership formation, sexual debut, and sexual mixing between high and low risk groups; (b) commercial sex; (c) coital frequencies and condom use; and (d) the effect of HIV and knowledge of HIV status on sexual behaviour.

The second part of the sheet contains the assumptions about heterosexual HIV transmission. The first parameter specified is the initial HIV prevalence in high risk women in 1985; although this is not strictly a transmission parameter, it is the parameter that 'seeds' the epidemic. This is followed by assumptions about HIV transmission probabilities per sex act, and the effect of risk group, HIV stage, ART and condom use on HIV transmission probabilities.

The third part of the sheet contains the assumptions about HIV disease progression and HIV-related mortality. Most of these parameters are stratified by CD4 stage (acute infection, post-acute CD4 500+, CD4 350-499, CD4 200-349 and CD4 <200). Mortality rates in treated patients are stratified by baseline CD4 category (not current CD4 count) and time since first ART initiation.

The fourth part of the sheets contains assumptions about interventions: male circumcision, PrEP and vaginal microbicides. Note that most of the intervention assumptions are actually specified in the 'Rollout' sheet; only the parameters that are not time-dependent are specified in the 'Adult assumptions' sheet.

After the intervention assumptions, several demographic parameters are specified: the effects of HIV on fertility and the fraction of births that are male births.

Finally, the MSM assumptions are specified. These include assumptions about proportions of men who begin sexual activity as MSM, proportions of their partners who are men, age mixing patterns, and probabilities of transmission per sex act.

Paed assumptions

This worksheet contains the assumptions about mother-to-child transmission (MTCT) of HIV and paediatric HIV survival. The sheet begins with the assumptions about the MTCT rates that would be expected in the absence of any interventions to prevent MTCT. These are followed by the assumptions about rates of HIV survival in the absence of ART (HIV-positive children are classified as having early disease or late disease prior to ART initiation), and assumptions about the effect of ART on HIV mortality in children.

The next section of the spreadsheet contains the prevention of MTCT (PMTCT) parameters. Note that most of the PMTCT assumptions are actually specified in the 'Rollout' sheet; only the parameters that are not time-dependent are specified in the 'Paed assumptions' sheet.

Lastly, a number of breastfeeding assumptions are specified. Note that although there is mention here of the effect of breastfeeding on non-AIDS mortality, this effect is not currently included in the model.

Rollout

This sheet contains the assumptions about access to different HIV prevention and treatment services. The rows represent the individual programme elements and the columns represent the different projection years. Parameters are grouped into the following programme headings: HIV counselling and testing, PMTCT, ART, PrEP, vaginal microbicides and medical male circumcision.

Column BL in this sheet calculates the parameter values for the current projection year.

Results

This sheet contains the model outputs. The outputs are listed in column A, and in subsequent columns the outputs for each calendar year are shown (note that if you have just clicked the 'Reset' button, these columns will be blank). Column B contains the results for the current projection year (the most recent year for which the model results have been generated). Additional outputs can be added to the Results sheet, but make sure that if you insert any rows for additional outputs, or enter any formulas for new outputs, you do so above the line that says "Insert additions above this row".

Results are grouped into seven sections: mortality statistics, other demographic output, HIV incidence statistics, HIV prevalence statistics, antiretroviral treatment and unmet need, prevention indicators, and CD4 calibration output.

Calibration

This sheet compares the model projections with empirical data. An important point to note is that parts of the calibration sheet (particularly the recorded death data, the HIV prevalence

data in key populations and the CD4 data) only make sense when one is using the national version of the model; for the provincial versions of the model, the calibration data are structured differently and are therefore not comparable with the model in the same way as the national calibration data. All of the empirical data and model outputs are organized according to the year to which the estimates relate (which are shown in the column headings).

The sheet begins with the antenatal survey estimates of HIV prevalence. This is followed by the modelled estimates of HIV prevalence in pregnant women, after adjustment for antenatal survey bias (since the antenatal surveys represent only the women using public sector antenatal services). This is followed by household survey estimates of HIV prevalence in the general population, and the model estimates of HIV prevalence in the corresponding age categories. The corresponding graphs are in the 'ANC calibration' and 'HSRC calibration' sheets, towards the end of the workbook.

The next part of the sheet shows the recorded numbers of deaths, followed by the modelled numbers of deaths in each calendar year, and the modelled numbers of deaths after adjusting for incomplete death recording (i.e. adjusting the modelled numbers to show the numbers of deaths that we would expect to be reported for a given level of under-reporting). The corresponding graphs are in the 'Death calibration' sheet.

The next part of the sheet compares model and survey estimates of HIV prevalence in sex workers. It is important to note that none of the surveys are nationally representative, and given the known variability in HIV prevalence between regions within South Africa, we would expect many of the survey estimates to differ substantially from the model estimate of the national average. The graph on the right hand side compares the modelled and survey estimates of HIV prevalence in sex workers.

The next part of the sheet compares CD4 distributions in cross-sectional surveys of HIV-positive adults with the corresponding model estimates. Two types of comparison are presented: surveys of the general adult population and surveys of pregnant women (the former includes individuals on ART, the latter generally excludes women who were already on ART prior to their first antenatal visit). As with the surveys of HIV prevalence in sex workers, it is important to note that none of the surveys are nationally representative, and hence the observed CD4 distributions might be expected to differ substantially from the national average predicted by the model in some cases. The corresponding graphs are in the 'CD4 calibration' sheet.

The next part of the sheet compares the model estimates of the numbers of individuals tested for HIV with estimates of the actual numbers of HIV tests performed in the public and private sectors of South Africa. The corresponding graph is in the 'HCT calibration' sheet.

Finally, the model estimates of HIV prevalence in each 5-year age group are compared with the results of the HSRC household surveys and the 2016 DHS. The corresponding graphs are in the 'HSRC calibration' sheet.

Population

This sheet summarizes the profile of the population in the current year, by age, sex and HIV stage. In addition, this sheet shows the assumed initial population pyramid at the start of the

simulation (in 1985) in cells B6:C96, and the age- and sex-specific adjustments that are applied to the average HIV prevalence in high risk women aged 15-49 at the start of the simulation (cells B119:C153).

The table in cells E6:P96 contains the following age- and sex-specific totals: total HIV-positive, total on ART, total HIV-diagnosed, total sexually experienced and total married. This table gets copied and pasted to the bottom of the 'Monthly' sheet at the start of each projection year, so that key indicators can be calculated from the stored values. In cells G114:H194 the total numbers of HIV-diagnosed individuals are calculated (note that these are not calculated for children under the age of 10).

The table in cells R16:BI97 shows the total numbers of males stratified by age and HIV stage. The values in this table are calculated by summing across the relevant male risk group sheets. The same comments apply to cells BK16:DB35, except that these calculations relate only to males who are not yet sexually experienced (virgins). Note that some of these males may be HIV-positive, having acquired HIV at birth or after birth, through breastfeeding.

The tables in cells R114:BI195 and BK114:DB133 follow the same format, but apply to females.

Sex activity

This sheet contains the calculations of the rates of relationship formation, divorce, commercial sex activity and partner age preferences. Columns B-C show the rates of marriage and columns D-E show the rates of divorce. (Note that following conventions in demographic research, 'marriage' includes cohabiting relationships, and the term 'divorce' here should similarly be understood to include the break-up of cohabiting relationships.)

The tables in columns G-L show the calculation of the initial sexual behaviour profile of the population at the start of the simulation, in 1985 (females are at the top of the sheet and males below). This is an important intermediate calculation in the distribution of the starting population between the different risk groups. Columns N-Q show the corresponding sexual behaviour profiles in the current time step.

In cells R6:R86, we calculate the relative rates at which short-term partnerships are formed by single females who are sexually experienced, and in the column to the right we calculate the expected number of new short-term partnership formed by these women over a period of one year. In cells U6:DB86 we calculate the expected fraction of these new partnerships that are formed with men of each age, and in cells Z88:DB88 we calculate the total number of new short-term partnerships stratified by the age of the male partner. Using these outputs we calculate the rates at which heterosexual men acquire new short-term partners at each age (cells R93:R173) and the fraction of female partners at each age (cells Z93:DB173). Rates of male short-term partnership formation are thus calculated to be consistent with female rates of short-term partnership formation.

Cells U93:V173 calculate the male rates of sex worker contact and the expected total numbers of male contacts with sex workers per annum. Cells GT6:GT86 calculate the fraction of sex workers at each age.

In cells DD11:DD86 we calculate the total number of married/cohabiting women. To the right of this (cells DF11:GO86) we calculate the fraction of husbands at each age, and based on this we calculate in cells DF88:GO88 the expected numbers of married males at each age (note that these may differ from the model estimates of the actual numbers of married men at each age, since they are based on the desired partner age preferences). In the table below this, we calculate the fraction of wives at each age, for men of a given age.

In cells GQ11:GS86 we calculate the annual probability that a married individual becomes widowed due to their partner dying from non-AIDS causes, taking into account the previously-calculated partner age distributions. Columns GW-GZ then calculate the annual probability of divorce or widowhood (combining AIDS and non-AIDS mortality).

Cells GR93:GU113 calculate the annual probability of sexual debut.

Cells T180:DB260 calculate the patterns of age mixing in MSM relationships.

Mixing

This sheet calculates rates of mixing between the high and low risk groups. These calculations are performed separately for people entering short-term and long-term relationships. In cells F11:I86, we adjust the marriage rates specified in the 'Sex activity' sheet so that they apply only to individuals who are sexually experienced, and not to virgins.

Condoms

This sheet calculates the rates of condom use in the current projection year, stratified by age, sex and relationship type (columns E-J). It also calculates coital frequencies in marital relationships (columns B-C) and calculates differences in levels of sexual activity between different HIV stages (columns L-Y).

Circumcision

This sheet calculates the rates of male circumcision that would be expected in the absence of campaigns to promote male circumcision as an HIV prevention strategy (most of which would be traditional male circumcision), as well as the rates of medical male circumcision (MMC) due to campaigns in the current year. Cells B9:C99 calculate the rates that would be expected in the absence of campaigns. Columns E-L calculate the rates of MMC due to campaigns in the current year, taking into account assumed age differences in the acceptability/uptake of MMC. Finally, in columns N-R we combine the previous calculations to obtain a total annual probability of being circumcised, either traditionally or medically.

PrEP + VM

This sheet calculates the rates at which individuals start pre-exposure prophylaxis (PrEP), vaginal microbicides (VM) or a programme of regular HIV counselling and testing (HCT). The interventions are treated as if they are mutually exclusive, since vaginal microbicides would not be recommended for women who are already taking PrEP orally, and vice versa. Individuals receiving PrEP or microbicides are assumed also to receive regular HIV testing,

so that they are diagnosed soon after acquiring HIV (at which point they would discontinue PrEP/VM to avoid acquiring drug resistance).

Transmission

This sheet calculates the average rates at which HIV is transmitted sexually, aggregating the calculations from the different risk groups. The first part of the sheet (columns B-AC) calculates the average transmission probability per short-term partnership, given the risk group and sex of the HIV-negative partner, the risk group of the partner, the type of short-term relationship (same-sex or heterosexual), the type of antiretroviral prophylaxis that the HIV-negative partner is receiving (PrEP, microbicide or none) and the circumcision status if the susceptible partner is male.

The second part of the sheet (columns AE-AP) calculates the average transmission probability per year in long-term partnerships (marriages and cohabiting relationships), given the risk group and sex of the HIV-negative partner, the risk group of the partner, and the type of antiretroviral prophylaxis that the HIV-negative partner is receiving (PrEP, microbicide or none).

In columns AR-AT we calculate the average probability of transmission from clients to sex workers (note that these calculations are not stratified by age because it is assumed that there are no age preferences in sex worker-client relationships).

Finally, in columns AV and AW we calculate the relative levels of biological susceptibility to HIV in men and women at different ages. This is to take into account that women below the age of 25 appear to be at heightened risk of HIV due to factors such as cervical ectopy.

Progression

This sheet calculates the AIDS mortality rates and the rates at which adults move between different CD4 categories prior to ART initiation. Note that we also calculate the average rate at which ART has been initiated in the previous three projection years (cells F3:F4) because this is assumed to determine the untreated mortality rate in the CD4 <200/µl category (if there is a high rate of ART initiation, one would expect that most individuals in this category start ART before their CD4 count drops to the very low levels at which the mortality risk is greatest).

The first table (cells B11:I91) calculates the monthly rates at which adults move between CD4 categories. The second table (cells K11:N91) calculates the monthly AIDS mortality rates ('AIDS mortality' includes mortality that is HIV-related, even if it is not formally classified as an AIDS-defining cause of death). Finally, columns P-Y calculate the probability that someone who moves between CD4 stages in the current month moves to the next stage in the same month, i.e. the probability of two transitions occurring in the same month, conditional upon one transition occurring. Note that untreated HIV-positive individuals can only progress to more advanced stages of HIV immune suppression, and no allowance is made for return to an earlier stage of HIV disease in the absence of treatment.

Testing

This sheet calculates monthly rates of HIV testing, by age. The rates that apply to sexually experienced individuals are stratified by sex (males in columns B-M, females in columns N-Z), HIV status, HIV testing history and CD4 category (if HIV-positive and undiagnosed). It is assumed that youth who are not yet sexually experienced would have a lower rate of HIV testing than in sexually experienced youth, but that their rates of testing would be similar to those in adults if they have HIV-related symptoms; these monthly rates of HIV testing are stratified by age, sex and HIV status/CD4 stage in columns AA-AL.

ART start

This sheet follows a similar structure to the 'Testing' sheet, but calculates the monthly rates at which HIV-positive get diagnosed *and* initiate ART immediately after diagnosis (allowing for the slight delay that may occur in practice because of delays in assessment of ART eligibility and patient preparation prior to ART commencement). Rows 8-10 show the fractions of newly-diagnosed individuals who are assumed to start ART immediately (taking into account the ART eligibility in the relevant year as well as the assumed rates of linkage), for each of the three modes through which it is assumed that HIV testing can occur. These are applied to the assumed monthly rates of HIV testing in the rows below. Rates in sexually experienced adults are stratified by sex (males in columns B-P, females in columns Q-AE), HIV testing history and CD4 category. As in the 'Testing' sheet, separate calculations are performed for youth who are virgins (columns AF-AO).

In columns AQ-BA we calculate the rates of ART initiation in adults who have been diagnosed HIV-positive but who did not start ART immediately after HIV diagnosis. This calculation is performed by subtracting the number of ART initiations that occurred immediately after diagnosis from the total number of ART initiations (cells AT8:AU8). The rates of ART initiation at CD4 <200 are calculated by calculating the roots of a quadratic polynomial (calculations in cells AQ17:AT99). If numbers starting ART have not been specified in the current year, the above calculations are ignored and the calculation of the rate of ART initiation instead depends on the assumed average delay between diagnosis and ART initiation if ART is not started immediately after diagnosis (cells AQ99, AS99). Finally, in cells AV12:BA23 we calculate the rates of ART initiation in the higher CD4 categories, taking into account that individuals in earlier stages of disease often do not initiate ART as quickly as individuals in advanced stages of HIV infection.

ART

This sheet contains calculations of current CD4 distributions in adult patients who are on ART, and rates of mortality in patients on ART. This sheet also contains assumptions about the fraction of ART patients who are temporarily interrupting ART, stratified by time since first initiating ART (cells G5:S5). It is worth noting that most of the model outputs showing 'numbers on ART' are net of these temporary interrupters, i.e. excluding patients who temporarily discontinue ART. It is also worth noting that 'temporary' discontinuers may die while discontinuing therapy, and some of the temporary discontinuers may therefore be considered permanent discontinuers.

Columns G-L calculate CD4 trajectories in ART patients at each integer duration since first ART initiation, stratified according to their baseline CD4 count. The calculations are performed on the assumption that the CD4 distribution at each ART duration is gamma-distributed, with the gamma distribution parameters ('alpha' and 'beta') calculated from assumed means and coefficients of variation. In columns N-S, the same calculations are performed, but stratified by half-integer duration since first ART initiation.

Columns V-AO calculate monthly rates of HIV mortality in men on ART and columns AQ-BJ calculate monthly rates of HIV mortality in women on ART. As in the 'Progression' sheet, mortality rates are adjusted to take into account the rates of ART initiation in recent years (if rates of ART initiation are very high, we would expect that most individuals who start ART in the CD4 <200/μ category do so when their CD4 count is towards the upper end of this interval, and thus have relatively low mortality rates). Columns BL-BS calculate the probability that an individual who starts ART dies from HIV-related causes in the month of starting ART.

In cells B63:R67 we calculate the proportion of ART patients who are on second-line ART, stratified by baseline CD4 count time since ART initiation (in years). Below this, in cells B70:N76 we calculate the fraction of ART patients who are virally suppressed (again stratified by baseline CD4 count) and the associated reductions in HIV transmission probabilities.

At the bottom of the sheet (cells G78:K80), we specify the average multiples by which the true mortality rates in ART patients exceed the mortality rates estimated using IeDEA-SA data (the main data source for the default mortality assumptions).

Non-HIV mort

This sheet contains the assumptions about non-HIV mortality rates (the mortality rates that apply in HIV-negative individuals). In HIV-positive individuals, any mortality in excess of these non-HIV mortality rates is assumed to be attributable to HIV.

Columns B-E contain the non-HIV mortality rates that apply in the current projection year. Because mortality tables conventionally present mortality probabilities according to individuals' exact age, but the model groups individuals according to their age at last birthday at the start of the year, it is necessary to convert mortality rates by exact age (columns B-C) to mortality rates by age last birthday (columns D-E). In the period up to 2015, the mortality rates are calculated from tables derived from past mortality statistics in South Africa (columns G-AK for males and AM-BQ for females). After 2015, non-HIV mortality rates are assumed to decline exponentially towards some 'ultimate' mortality level; these ultimate rates and exponential decay factors are shown in columns BS-BV.

Columns BY-CA show the life expectancy at the age of death, calculated according to the West level 26 life table. These are used only for the purpose of calculating the number of life years lost due to AIDS (row 14 in the 'Results' sheet).

Fertility

This sheet calculates rates of fertility in HIV-negative women and adjustments to fertility in HIV-positive women. Columns B-AG show the observed age-specific fertility rates in South Africa (up to 2016). These are in effect average fertility rates across HIV-negative and HIV-positive women; in column AI we calculate from these the fertility rates that apply in the current year to HIV-negative women who are sexually experienced. If the current projection year is beyond 2016, the HIV-negative fertility rate is extrapolated from the rate estimated in 2016, assuming an exponential decline in fertility towards some 'ultimate' level of fertility. The ultimate rates and exponential decay factors are shown in columns AL and AN.

Columns AQ-AU calculate the relative rates of fertility by HIV disease stage. In HIV-negative women, the adjustment factors are set to 1, while in HIV-positive women the relative fertility rates are less than or equal to 1, to represent the effect of fertility impairment associated with HIV (the extent of the impairment being dependent on the degree of immune deficiency).

Columns AW-BL calculate total births, stratified by women's HIV status and stage of disease. These calculations are used for the purpose of calibration to the age-specific HIV prevalence data in the antenatal surveys, and also for the purpose of calculating levels of mother-to-child transmission (in the 'Births' sheet).

Migration

This sheet shows the net numbers of in-migrants (immigrants less emigrants). The numbers over the 1985-2016 period are shown in columns E-AJ (males) and AL-BQ (females). Columns B and C calculate migration adjustment factors by dividing the number of net in-migrants in the current year by the population size in the current year. These multiplicative adjustment factors are applied to the entire population at the end of year to represent the effect of immigration/emigration on the population size. In years after 2016, the formulas in columns B and C assume a gradual trend towards zero in the number on net in-migrants, extrapolating from the last estimate in 2016.

In columns BS-BV we specify the effect of differences in HIV status between immigrants and emigrants on the HIV profile of the population. In the national model, for simplification, these adjustment factors are set to 1 (effectively implying that migration into and out of South Africa does not significantly alter HIV prevalence), but in the provincial version of the model these ratios can depart substantially from 1 (for example, due to people from high-prevalence provinces migrating into low-prevalence provinces). Columns BX-CA combine these HIV effects with the multipliers in columns B and C.

Risk group sheets (MHU virgin, MHC virgin, MHU ST, MHC ST, MHU STM, MHC STM, MHU LTH, MHC LTH, MHU LTL, MHC LTL, MLU virgin, MLC virgin, MLU ST, MLC ST, MLU STM, MLC STM, MLU LTH, MLC LTH, MLU LTL, MLC LTL, FH virgin, FH ST, FH SW, FH LTH, FH LTL, FL virgin, FL ST, FL LTH, FL LTL)

The risk group sheets all follow the same format, and we therefore describe the generic structure. The name of the sheet indicates the model risk group:

- Sheets starting with 'M' and 'F' relate to males and females respectively.

- The second letter in the name indicates whether the individuals in the class are 'high-risk' (H) or 'low-risk' (L).
- For males, the third letter in the name indicates whether they are circumcised (C) or uncircumcised (U).
- The letters after the space indicates the individual's relationship status: 'virgin' (never had sex), 'ST' (only engaging in short-term heterosexual relationships, i.e. not married/cohabiting), 'STM' (engaging in same-sex relationships and short-term heterosexual relationships), 'SW' (sex worker – only relevant to women), 'LTH' (in a long-term/cohabiting relationship with a high-risk partner) or 'LTL' (in a long-term/cohabiting relationship with a low-risk partner).

In each of the risk group sheets, the table in cells B6:AT86 shows the numbers of adults by age (rows) and HIV stage/HIV testing history/use of antiretroviral prophylaxis (columns). These are the numbers that apply at the *start* of the current month. In cells AV6:BG86 we calculate the average probability that an individual in the risk group transmits HIV to a partner (stratified according to the type of prevention method that the partner is using). In the case of short-term partnerships, these are calculated as probabilities per partnership, but in the case of long-term relationships, these are calculated as probabilities per period, and in the case of sex worker-client interactions, these are calculated as probabilities per sex act. In cells BH6:BJ86 we calculate the monthly probability of acquiring HIV, and in cells BM6:BN86 these monthly probabilities are converted into monthly rates for previously-tested and untested individuals (the calculation is not strictly necessary, since the rates are assumed to be the same for both).

The bottom half of the sheet calculates the change in the HIV profile over the course of the month, i.e. the bottom half of the sheet calculates transitions between HIV stages over the course of the current month and calculates the HIV profile at the end of the month. The profile at the end of the month is calculated in cells B93:AT173. New HIV infections are calculated in cells AV93:AV173. Numbers starting ART for the first time are calculated in cells AX93:BA173 and numbers of AIDS deaths are calculated in cells BC93:BD173.

Note that there are some deviations from this general structure in individual risk group sheets. For example, in the 'MHU virgin' and other 'virgin' sheets, there are no calculations of the probability of transmitting or acquiring HIV, and calculations are performed only over the 10-30 age range (since all individuals are assumed to be sexually experienced by the time they reach age 30). In the male high-risk sheets, numbers of male contacts with sex workers are calculated in column BK (men in the low-risk group are assumed not to have contact with sex workers). Finally, in the 'FH SW' sheet, numbers of women starting and retiring from sex work are calculated in cells BF93:BH173.

Births

This sheet calculates the numbers of births to mothers in different HIV stages, and estimates the total number of children who acquire perinatally (at or before the time of birth). Women are grouped according to whether they were HIV-positive but not on ART at the time of their first antenatal visit (rows 4-10), already on ART prior to their first antenatal visit (rows 11-13) or HIV-negative at the time of their first antenatal visit (rows 14-19). The model also allows for changes in HIV status and knowledge of HIV status over the course of pregnancy. Thus columns E-F represent the changes in knowledge of HIV status after the first antenatal

visit, columns H-I represent the changes in HIV status and knowledge of HIV status between the first antenatal visit and 34 weeks gestation (around the time of the second antenatal HIV test), and columns K-L represent the HIV stage and knowledge of HIV status at delivery. Finally, in column O we calculate the number of infants who acquire HIV perinatally and the number who are HIV-negative although their mothers are positive. Note that although all the calculations in this sheet are expressed as annual totals, the final results are divided by 12 in order to obtain births in the current month.

Child rates

This sheet calculates the non-HIV mortality rates and rates of transition between different HIV stages in children. It also calculates rates of breastfeeding, since this is relevant in calculating postnatal transmission rates. Columns A-I calculate the profile of the child population (including the fraction of children who are being breastfed) at the start of the simulation (in 1985). Columns K-L calculate the non-HIV mortality rates in the current year (these calculations are different from those in the 'Non-HIV mort' sheet because the rates are calculated by month of age). Columns N-S calculate rates of breastfeeding separately for mothers who are HIV-negative (or who have undiagnosed HIV infection) and mothers who have been diagnosed HIV-positive; it is assumed that some HIV-diagnosed mothers practise exclusive breastfeeding (EBF), but EBF is assumed to be negligible in HIV-negative and undiagnosed HIV-positive mothers.

Similar to the calculations in columns AQ-AT of the 'ART start' sheet, columns U-X calculate the rates at which children start ART if they did not start ART at the time of early infant diagnosis (EID). This involves subtracting the number of children starting ART immediately after EID from the total number of children starting ART (cell W147) and then dividing by the number of previously-diagnosed ART-eligible children.

Columns Z-AD calculate the rates of progression from early disease to late disease and the rates of HIV-related mortality, in the absence of ART.

Columns AF-BG calculate the 'dependent' probabilities of transition between health states, i.e. the probabilities that apply after taking into account the competing risks when there is more than one way to exit a given health state.

Columns BI-BL calculate the migration adjustment factors (analogous to the adult calculations in columns BX-CA in the Migration sheet).

Columns BN-BW calculate monthly rates of HIV testing in children, as well as the monthly probability of testing and starting ART soon after diagnosis. Calculations are stratified by HIV disease stage and timing of HIV acquisition (since this affects test sensitivity). Average monthly rates of testing in early HIV disease, stratified by year of age rather than month of age, are calculated in BY3:CA15.

Cells BY17:CI45 calculate the cumulative numbers of ART initiations in children on ART by treatment duration and by age, separately for children who started ART in early and late disease. These values are used in calculating the fraction of ART-experienced children who are currently still on ART (accounting for ART interruptions) and the fraction who are on

second-line ART, in columns CK and CL respectively (both fractions depend on the time since first ART initiation).

Child M and **Child F**

These two sheets calculate the changes in HIV profile in boys (Child M) and girls (Child F) aged 0-10, over the course of a single month. The two sheets follow the same format. They are analogous to the adult risk group sheets, but children are stratified by age in months (not years).

The table in cells B7:X139 shows the HIV profile at the start of the month. HIV-negative children (Columns B-L) are stratified according to their mother's HIV status, their mother's level of engagement in HIV care and the type of breastfeeding they are receiving, as all of these factors affect the mother-to-child transmission risk. HIV-positive children (columns M-X) are stratified according to whether they are diagnosed and on ART, their mode of HIV acquisition (perinatal or postnatal) and their stage of HIV disease (current disease stage if untreated, or disease stage at the start of ART if they were previously on ART). The table in cells Z7:AV139 calculates the HIV profile of the child population at the *end* of the current month (note that there are no calculations in the range Z43:AI139) because it is assumed that no child is breastfed for more than 35 months.

The table in cells AX7:BH139 calculates numbers of events over the course of the current month (new infections, AIDS deaths, non-AIDS deaths, children progressing from early disease to late disease, children starting ART, etc.).

At the bottom of this sheet (rows 143-150), we calculate the number of children leaving the paediatric population to enter the 'adult' (10+) risk group sheets. These calculations are only relevant at the end of each projection year, when the values in these cells get copied to the age 10 rows in the relevant 'virgin' sheets (all children are initially assumed to be virgins at age 10). Although children are conventionally defined as ages 0-14 (or sometimes ages 0-17), we use the age 10 cut-off because it is easier to model early sexual debut (before age 15) if 10-14 year olds are grouped together with adults.

Monthly

This is an intermediate sheet for calculating the outputs that appear in the 'Results' sheet. There are five sections in the sheet (each separated by dashed lines). The first sheet is for calculating 'flow variables' (counts of numbers of events or numbers of people entering a particular health state) that are updated on a monthly basis. Column B calculates the flow variable in the current month and adds it to the cumulative total from previous months (column C). At the start of each year, the cumulative total in column C is set to zero.

The second section is similar to the first, but it relates more specifically to flow variables that are age-specific; outputs are grouped in columns (not rows) and the cumulative totals from previous months appear on the right hand side of the sheet.

The third section calculates age-specific outputs from the second section. These calculations only get used at the end of each projection year, so it is not necessary to 'store' the intermediate/cumulative calculations in previous months, as we do in the first two sections.

The fourth section calculates stock variables (cross-sectional measures) that are updated annually. Column B calculates the values at the current time and column C stores the values that were calculated at the *start* of the projection year (i.e. even though the calculations in column B get updated at monthly time steps, it is only the values at the start of the projection year that we're interested in).

The fifth and final section stores the population profile at the start of the projection year (this is copied and pasted from cells E6:P96 in the 'Population' sheet).

CD4 calibration

This sheet shows the model fit to cross-sectional survey estimates of the fraction of HIV-positive adults in different CD4 categories. The dots represent the survey estimates (with 95% confidence intervals represented by error bars) and the solid lines represent the model estimates. Note that because none of the surveys are nationally representative, we would not expect to see close agreement between the model estimates and the survey estimates in all cases.

ANC calibration

This sheet shows the model fit to the annual antenatal survey estimates of HIV prevalence in pregnant women. The dots represent the survey estimates (with 95% confidence intervals represented by error bars) and the solid lines represent the model estimates. Confidence intervals are not shown for the years prior to 1997 because the surveys in these years did not take into account the survey design effects when reporting confidence intervals.

HCT calibration

This sheet shows the model fit to the reported numbers of individuals tested for HIV (combining estimates for the private and public sectors). The dots represent the survey estimates and the solid lines represent the model estimates.

HSRC calibration

This sheet shows the model fit to HIV prevalence estimates from national household surveys conducted by the Human Sciences Research Council (HSRC) in 2002, 2005, 2008 and 2012, by the Wits Reproductive Health Research Unit (RHRU) in 2003, and by the South African Department of Health/Medical Research Council as part of the 2016 DHS. The dots represent the survey estimates (with 95% confidence intervals represented by error bars) and the solid lines represent the model estimates.

Death calibration

This sheet shows the model fit to the annual reported numbers of deaths in South Africa, stratified by sex and five-year age group. The dots represent the recorded numbers of deaths and the solid lines represent the model estimates, after adjusting for under-reporting.