A method to estimate the size and characteristics of HIV-positive populations using an individual-based stochastic simulation model

eAppendix

November 2015

Contents

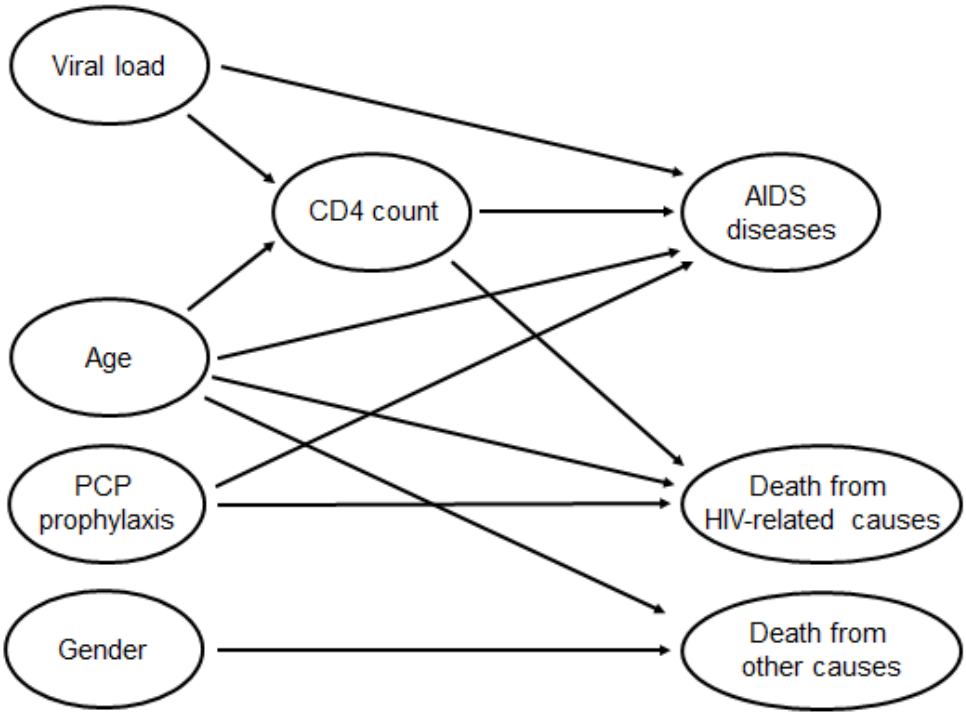
1 Brief description of HIV Synthesis Progression model

The HIV Synthesis Progression model is an individual-based stochastic computer simulation model of HIV progression and the treatment of HIV infection. The current version of the model is Synthesis version 6 (V6).

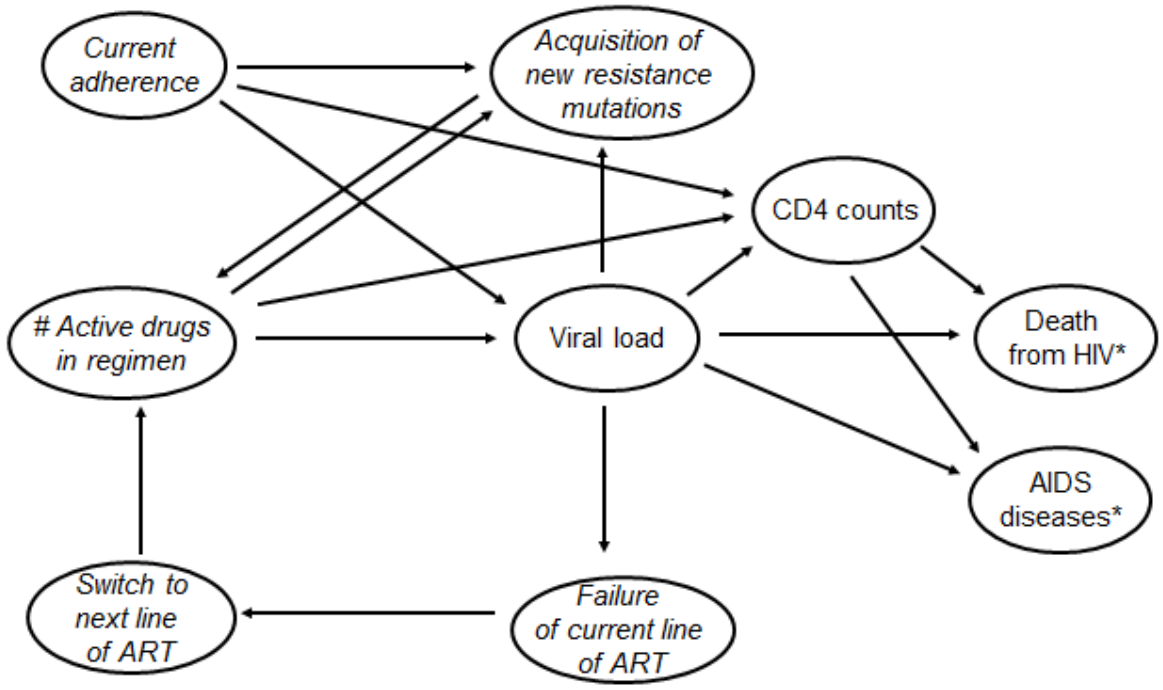
The model was originally developed by Phillips and colleagues to reconstruct the HIV-infected population in the UK(1). It incorporates our understanding of the underlying processes of HIV disease progression and the effect of antiretroviral therapy (ART), based on data from clinical trials and epidemiological data. **eFigure 1** and **eFigure 2** show a simplified influence diagram of the main parameters modelled to describe the natural history and the effect of ART respectively.

In brief, the Synthesis Progression model generates simulated "data" on the progression of HIV infection and effect of ART on simulated persons living with HIV. Each individual in the model is simulated from the time of infection (although for simplicity we do not explicitly model acute changes in viral load and CD4 count around the time of seroconversion) and they are followed until either death, emigration or to a given calendar year of interest. For each simulated person, the model generates variables such as calendar date, calendar year of HIV infection, CD4 cell count, viral load, age and presence of transmitted resistance mutations. The values of these variables are updated every three months in the model. Use of specific antiretroviral drugs, adherence, accumulation of resistance mutations and clinical events are also modelled in order to incorporate the effects of ART. The model has been shown to provide a generally close fit to observed data relating to the natural progression and therapy outcomes(1-3).

eFigure 1: Influence diagram of HIV Synthesis progression model showing variables modelled to describe the natural history of HIV



eFigure 2: Influence diagram of HIV Synthesis progression model showing variables modelled to describe the effect of ART



* also influenced by age and use of PCP prophylaxis

2 Model details

2.1 Documentation

The most up to date documentation for Synthesis V6 can be found here: (3;4)

2.2 Updates made to Synthesis V6 since last published documentation

2.2.1 Interruption of ART without clinic/clinician being aware

The proportion of people who have interrupted, but where the clinic/clinician is not aware of the interruption, is given by *clinic_not_aware_frac*. The proportion used in the updated model is 0.5 (previously 0.3), which relates to the proportion of interruptions where the clinic/clinician is not aware of it at any given point in time. This variable was introduced because data from the literature clearly show that not all people who experience virologic failure have resistance in majority virus; these remaining people are likely to have interrupted ART, with the interruption unknown to the clinician, causing their viral load to rebound(5;6).

2.2.2 Accumulation of resistance mutations

newmut(t) is a probability used to indicate the level of risk of new mutations arising in a given 3 month period. If this chance comes up in a given 3 month period (determined by sampling from the binomial distribution) then there is a probability of resistance mutations to arise if on a given drug.

Note that *newmut(t)* depends on length of time it has been since the individual started their current period of continuous therapy, number of active drugs in current regimen,'effective adherence' in the last 6 months and whether their previous underlying viral load was below or above 4 $\log_{10}$ copies/ml.

The probability of acquiring I50V, I54, L76 and I84 whilst on darunavir is now 0.01 (previously 0.02), conditionally on this chance arising (which depends on the above factors listed). The probability was made lower based on new data showing that darunavir mutations are very rare(7;8).

2.2.3 Effect of being on ART on the occurrence of AIDS/HIV-related deaths and non-HIV-related deaths

The rate of AIDS, defined by the variable, *base_rate*, according to (most recent) CD4 count is multiplied by a further factor of 0.9, 0.85 and 0.6 if on a single drug, 2 drug or 3 drug regimen respectively. The multiplicative factor of 0.6 if on a 3 drug regimen used to be 0.8. This was modified to reflect the low rates of AIDS seen in people on combination therapy.

2.2.4 Non-HIV-related deaths

The rate of death now differs by ART status: if not on ART, then we assume a 2-fold increased rate compared to HIV-negative individuals of all non-HIV causes of death throughout life (based on evidence that HIV infection itself may indicate a raised risk of common clinical conditions). However if on ART, then this increased rate is reduced to 1.3-fold, given that being virally suppressed has been associated with a reduced risk of death(9-11).

3 Model calibration procedure

3.1 What is meant by 'calibration'

Synthesis V6 is a simulation model which depicts the progression of HIV and the effects of ART for a given hypothetical population. The primary outcome of the model is a longitudinal dataset for such a population with a range of variables that describes the course of HIV for each individual. This simulated dataset can be analysed to output outcomes of interest. Simulated populations have different characteristics and therefore different outcomes depending on the values of the input (prior) parameters into the model (parameterisation is described in more detail in section 3.4).

National-level surveillance of HIV/AIDS is conducted in most European countries although there is large variability in the type of data collected and methods of collecting and reporting. Individual data (as opposed to aggregate level data) on HIV and AIDS case-reports and aggregate data on number of HIV tests conducted are routinely submitted to European Centre for Disease Prevention and Control (ECDC) and World Health Organisation (WHO) Regional Office for Europe.

The concept of 'calibration' refers to the process of finding the set of parameter values inputted into the Synthesis model which generates a set of modelled outcomes which are similar to what is actually observed in reality, i.e. the surveillance data.

3.2 Approximate Bayesian Computation

The model is calibrated using Approximate Bayesian Computation (ABC) methods(12). The model naturally lends itself to working in a Bayesian framework to account for multiple parameter combinations which produce model outputs which fit well to the observed data (instead of converging to a single set of parameters as would be the case in maximum likelihood estimation). ABC involves running the model multiple times where each run of the model is considered one simulation. The outcomes of the model (usually a summary statistic) are then compared against the observed data and sets of parameter values are accepted if sufficiently close (i.e. within an arbitrary tolerance threshold). The posterior distributions of the relevant parameters are then approximated based on the accepted values.

ABC methods are suitable for calibrating simulation models to multiple data sets within tolerance bounds and have the advantage of accounting for parameter uncertainty and parameter correlations. We have chosen to use ABC methods because we wish to explore a wide parameter space and consider as many parameter sets as possible which are adequately consistent with the data and not focus on finding the single parameter set that is most consistent with the data.

3.3 Population to calibrate model to

In order to calibrate the model to a particular population, the parameters which are sampled are those thought to differ between populations or which have a large degree of uncertainty and should not be fixed in each simulation. This means it is necessary to decide and consider the population or sub-population to calibrate the model to. This will depend on whether there are substantial differences predominantly in the incidence of new infections and probability of diagnosis by sub-population. Sub-populations may be specific HIV transmission risk groups, or perhaps a regional

population. Here we use transmission risk groups to define our sub-populations. The greater the number of sub-populations to calibrate to separately the longer the calibration procedure will take because of an increased number of parameters to calibrate. The choices made here are:

- Whether to calibrate to data on the total HIV-positive population
- Whether to break down the whole HIV-positive population into transmission risk groups (e.g. men who have sex with men (MSM), people who inject drugs (PWID), heterosexually-acquired infections)

If calibrating to data on the heterosexual group separately, then there is also the choice to decide whether there are appreciable numbers of infections in people who once lived outside Europe, primarily sub-Saharan Africa, who have then subsequently immigrated to Europe. If there are large numbers of people infected heterosexually in sub-Saharan African and who then subsequently immigrate to the European country of interest, then these infections are modelled separately and the modelling approach described in this manuscript is not appropriate.

3.4 Parameterisation

3.4.1 Types of parameters

In the Synthesis V6 model, parameters can either be fixed (fixed parameter) or varied (variable parameter) per simulation. A fixed parameter is one in which the value does not change from simulation to simulation. A varied parameter on the other hand is one in which the value is sampled from a probability distribution and therefore changes from simulation to simulation.

3.4.2 Parameter values reflecting the intrinsic effects of HIV and ART

As explained in Section 1, the progression model has been shown to provide a generally close fit to observed data relating to the natural progression of HIV and ART outcomes(1-3). Therefore, for the purpose of calibrating the model to a given HIV-positive population (which could be the total population or one particular risk group), we hold parameter values reflecting the intrinsic effects of HIV and ART fixed, thereby becoming part of the model structure. This assumes that these parameter values are the same regardless of the population under consideration.

3.4.3 Parameters which are varied per simulation

There are also a number of variable parameters in cases where there is uncertainty about what values the parameter should take. These parameters take a different value per simulation and are sampled from suitable distributions in each simulation.

In order to calibrate the model to a given HIV exposure group in a given country, the main parameters for which values are sampled are those which describe the incidence (number of new infections per year) and the diagnosis rate (probability of diagnosis in any 3-month period). See more in section 3.4.4.

In addition, other parameter values that may be specific for a given HIV exposure group and country and therefore could also be varied across simulations are: proportion of people who avoid testing for HIV, probability of not being linked to care soon after diagnosis, rate of being lost to follow-up whilst

ART-naïve, rate of being lost to follow-up whilst on ART, rate of re-entry into care after being lost to follow-up, probability of starting ART when eligible, population distribution of underlying levels of ART adherence, rate of ART interruption, rate of re-starting ART after interruption and rate of emigration. See more in section 3.4.5.

3.4.4 Parameters which determine HIV incidence and diagnosis rate

The parameters which determine the incidence and diagnosis rate are given in **eeTable 1**. The distributions presented are an example template, with only suggested values (which may be amended). Specific values are discussed within the relevant population simulated (e.g. section 4 for MSM epidemic in the UK, section 4.2 for pseudo epidemic).

eTable 1: Parameters which determine HIV incidence and diagnosis rate per simulation. Values for *Max1* and *Max2* are described on page 8.

Name	Description of parameter	Prior distribution
i1	Number of infections per year during 1980-84	Beta(2,4)**Max1*
i2	Number of infections per year during 1985-89	Beta(2,4)**Max1*
i3	Number of infections per year during 1990-94	Beta(2,4)**Max1*
i4	Number of infections per year during 1994-99	Beta(2,4)**Max1*
i5	Number of infections per year during 2000-04	Beta(2,4)**Max1*
i6	Number of infections per year during 2005-09	Beta(2,4)**Max1*
i7	Number of infections per year during 2010-13	Beta(2,4)**Max1*
d1	Diagnosis rate per 3 month period during 1984-91	Beta(1,5)**Max2*
d2	Additional absolute change in diagnosis rate from 1984-91 (i.e. diagnosis rate during 1992-99 is *d1+d2*)	Beta(1,50)
d3	Additional absolute change in diagnosis rate from 1992-99 (i.e. diagnosis rate during 2000-08 is *d1+d2+d3*)	Beta(1,50)
d4	Additional absolute change in diagnosis rate from 2000-08 (i.e. diagnosis rate during 2009-13 is *d1+d2+d3+d4*)	Beta(1,50)

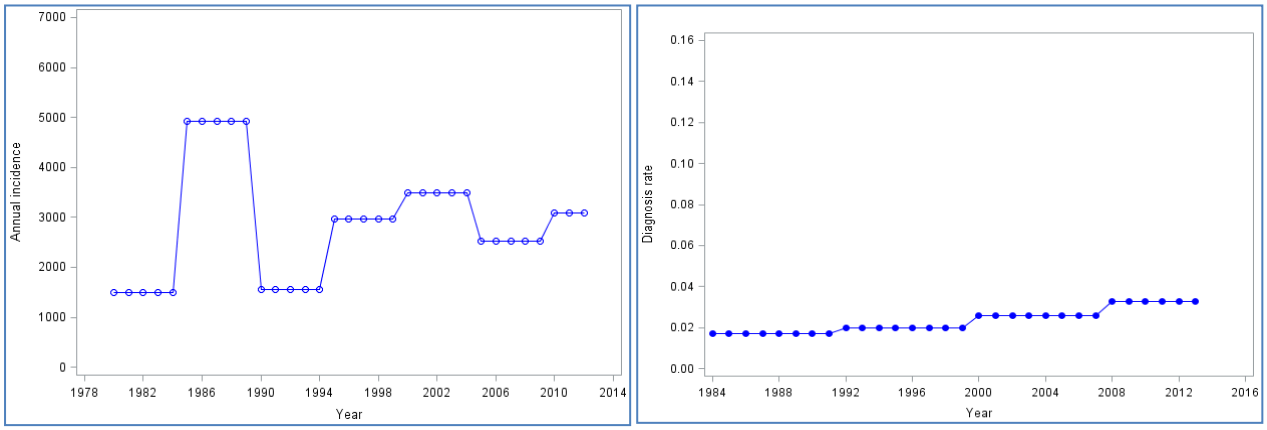
Parameters *i1,i2,…,i7* determine the number of HIV infections per year (incidence) for a given 5 year period. These parameters inform a piecewise constant curve for the HIV incidence. As *i1,i2,…,i7* are all sampled randomly in each simulation, the curve is also randomly generated in each simulation (see **eFigure 3**).

Model variables are updated every 3 months and so the infection is sampled to occur at one of 4 time points in a year (25% probability they will be infected in January-March, April-June, July-September and October-December).

Similarly, parameters *d1, d2, d3* and *d4* determine the diagnosis rate for a given 8 year period. These parameters inform a piecewise constant curve for the HIV diagnosis rate. As *d1, d2, d3* and *d4* are all sampled randomly in each simulation, the curve is also randomly generated in each simulation (see **eFigure 3**). The change in diagnosis rate is sampled as opposed to the absolute diagnosis rate because we thought that the diagnosis rate is likely to change only slightly from one period to the next.

The diagnosis rate denotes the probability with which an HIV-infected individual gets diagnosed with HIV in a given 3-month period, given that they are not in the primary infection phase, are not symptomatic nor have AIDS.

eFigure 3: Example of a randomly sampled incidence curve (left) and diagnosis rate curve (right)



We assume that the minimum value for both the incidence and change in diagnosis rate is 0 in a given period over which the values are constant (5 years for incidence and 8 for diagnosis rate). The following formula is thus used to determine the maximum value of the sample space, where we set the minimum values to be 0, and thus the prior distributions, for the incidence curve and diagnosis rate curve respectively:

$$Max1 = Maximum\ number\ of\ HIV\ infections\ per\ year$$
$$= Highest\ number\ of\ HIV\ diagnoses\ ever\ observed\ in\ one\ year\ \times 3$$

$$Max2 = Maximum\ diagnosis\ rate\ per\ 3\ months$$
$$= \frac{Proportion\ of\ prompt\ presentations\ in\ most\ recent\ year}{4}$$

A prompt presentation is where the person has a CD4 count >350 cells/mm^3 at the time of diagnosis (or within 6 months of diagnosis). This is thought to be a reasonable choice of prior because it is assumed that diagnosis rates have generally increased over time given increases in testing. Although both these formulae were somewhat arbitrarily devised, we wanted to decide on a fixed formula so that the calibration method can be used in any setting. We consider this range is wide enough to capture unlikely but plausibly high incidence or diagnosis rates.

The purpose of having these formulae is to provide a simple method to decide on initial prior distributions. If the above formulae do not seem appropriate for a given setting, the initial parameter distributions for both the incidence curve and diagnosis rate curve could also be decided using a combination of surveillance/observational data and expert opinion. In addition, if there is none or little data to inform the prior distributions, then the prior distributions could always start with a wider distribution.

The Beta distribution is proposed for describing the prior distributions because of the great flexibility offered as parameterised by the two shape parameters α and β. The mean, standard deviation and mode for Beta(α,β) is given by α/(α+β), $\sqrt{(\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1))}$, and (α-1)/(α+β-2) respectively. Beta (2,4) is used for the incidence to reflect the fact that the true incidence is more likely to lie in the lower numbers, but we want to be able to sample a much wider sample space. Beta (1,5) is used for parameter d1, for a similar reason, as we expect initial rates of diagnosis to be much lower than in later years. For most settings, which have observed gradually increasing rates of testing, a monotonically non-decreasing rate of diagnosis seems appropriate which is why we set the minimum change in diagnosis possible as 0; in other words, the diagnosis rate can stay the same, or increase by an increment (distributed by Beta(1,50)) from period to period but that it will not decline. If there are specific country conditions that suggest a decline is possible then the parameterization must be changed to allow a decrease as well as an increase.

The choice of parameterisation for the incidence and diagnosis rate curves (i.e. choice of keeping the incidence constant over five years and diagnosis rate constant over eight years) was influenced by some preliminary work. In brief, this work involved using a simple regression approach using pseudo data to decide on the number of parameters and shape of the resulting curves. For the incidence curve, we looked at using 5, 7 and 11 year periods and similarly for the diagnosis rate curve, 4, 5 and 6 year periods. We used a multivariable logistic regression model, where the outcome variable was the calibration score (whether it was less than a certain cut-off) and the dependent variables were the categories representing the different number of parameters used to describe the piecewise incidence curve.

The true incidence and diagnosis rate curves are unknown and for most epidemics there will be a lack of data to inform the prior distributions for these parameters. Therefore, when sampling these parameters it is important to ensure that the parameter space sampled is chosen to be large enough to allow for extremities and that the parameterisation is sufficiently flexible but, given this, as restricted as possible to limit computation time. We use such a crude parameterisation for both the incidence and diagnosis rate parameters so that any sets of parameter values which may calibrate well to the data are not excluded. Also, as explained further in section 3.5, the aim of the calibration procedure is to not estimate these curves per se, but to find sets of curves which generate a modelled population with characteristics similar to that of the observed data.

3.4.5 Other parameter values that may be specific for a given HIV exposure group and country

There are a number of other parameters which are varied per simulation. These are parameters which are likely to vary by sub-population, or perhaps there are few data sources to inform them precisely. These are: *Prop_avoid_testing, prob_loss_at_diag, rate_lost, rate_lost_art, rate_return, prob_art, adh_pattern, rate_inter, rate_restart, rate_emig (***eeTable 2***).* Therefore, we sample these from a distribution for each simulation, in addition to the parameters describing the incidence and diagnosis rate curve. Although the values are sampled from a distribution in each simulation, these prior distributions will be informed from surveillance data or observational studies carried out in the country of interest or from other European studies. The sample space for these parameters will therefore be much narrower, compared to the prior distributions for the incidence and diagnosis rate parameters. Therefore for these nine parameters the aim is not to estimate the plausible range these lie in, but in fact the sampling is done per simulation to reflect the uncertainty associated with the prior distributions chosen.

eTable 2: Other parameters which are sampled per simulation

Parameter name	Description of parameter	Notes
prop_avoid_testing	Proportion of people who are resistant to testing	These people are 10-fold less likely to be diagnosed with HIV (unless they have HIV related symptoms)
prob_loss_at_diag	Probability of not being linked to care soon after diagnosis	This is the probability per 3-month period. These people are then classified as lost to follow-up.
prob_lost	Probability of being lost to follow-up for those not on ART	This is the base probability per 3-month period, for those with average adherence ≥0.8. Those with average adherence 0.5-0.8 and <0.5 have these probabilities multiplied by a factor of 1.5 and 2 respectively
prob_lost_art	Probability of being lost to follow-up given that treatment has been interrupted	This is the base rate per 3-month for those with average adherence ≥0.8. Those with average adherence 0.5-0.8 and <0.5 are 2-times and 3-times more likely to being lost to follow-up.
rate_return	Rate of re-entry into care	This is the base rate per 3-month for those with average adherence ≥0.8. Those with average adherence 0.5-0.8 and <0.5 are 2-times and 3-times less likely to re-enter into care. If the person has an AIDS-defining condition however, they consistently have 80% probability every 3-months of re-entry into care.
prob_art	Probability of starting ART when eligible	If the person has AIDS, then there is a 90% probability of starting ART in any 3-month time period..
adh_pattern	Population distribution of underlying levels of ART adherence	This is a categorical variable. The value of this variable is sampled from (1,2,3,4,5), each representing a different distribution of the adherence levels in a population, such that 1 represents a very good population level of adherence (around 95% of people on treatment with suppressed viral load) to 5 which represents a poor population level of adherence (around 60% of people on treatment with suppressed viral load).
rate_inter	Rate of ART interruption	This is the base probability per 3-month period, for those with average adherence ≥0.8. Those with average adherence 0.5-0.8 and <0.5 have these probabilities multiplied by a factor of 1.5 and 2 respectively. If the person has an ART-related side effect, this rate is further multiplied by a factor of 2. The rate of interruption is further modified by age, such that the younger the person, the more likely they are to interrupt.
rate_restart	Rate of re-starting ART after interruption (for people still in clinical care)	This is the base probability per 3-month period. If symptomatic or an AIDS-defining condition has developed then this probability is multiplied by a factor of 2 and 3 respectively.
rate_emig	Rate of emigration	This is the rate of emigration per 3-month period.

There are a number of other parameters which could potentially differ across populations and are factors known to have an impact on HIV progression, such as the proportion of people with HBV co-infection, proportion of people with HCV co-infection and prevalence of smoking in the exposure group of interest. Whether these parameters are sampled or not will depend on the setting and availability and reliability of data to inform these parameters.

3.5 Calibration procedure

As described in section 3.2, the approach we have used to calibrate the model is based on ABC methods. The calibration procedure used is formed of three stages.

In the first stage, 10,000 sets of parameter values (those listed in **eeTable 1**, i.e. seven for incidence and four for diagnosis rate) are sampled using Latin hypercube sampling(13). We use Latin hypercube sampling to ensure maximum and even coverage of the entire plausible parameter space. The simulation model is then run 10,000 times, each time using one of these sampled sets of parameter values. A different population of HIV-positive people until 2013 are constructed in each simulation (because of different rates of incidence and diagnosis in each simulation). Outcomes of the model up until 2012 are then compared with a range of surveillance (and/or observational) data. We quantify how well the model outputs match the surveillance data, i.e. assessing the goodness-of-fit, by calculating the 'calibration-score'. If the calibration-score deems it a close fit (i.e. within an arbitrarily defined tolerance threshold), then the parameter values are accepted. See section 3.6 for the description of the calibration-score.

The second stage is similar to the first, but the parameters are sampled using simple random sampling. Instead, we look at the distribution of all accepted values for each separate parameter from stage 1, and use the minimum and maximum values to refine our prior distributions (reduce the sample space to improve efficiency of calibration procedure). We then remain in this second stage until enough sets of parameters values calibrate well to the observed data.

In the third stage, further simulations are run using the parameter values which calibrated well in the second set. Due to the stochastic nature of the model, a set of parameter values may fit well to the data in one simulation but not in another. Therefore only the simulations which again calibrate well to the observed data are used to determine the final model outputs. The median values and 5% and 95% centiles of the distribution of these model outputs at each calendar year are considered the point estimate and plausibility range (PR) limits, respectively.

At every stage and for each simulation, the parameters listed in **eeTable 2** are sampled to reflect the uncertainty reflected in the chosen prior distributions.

3.6 Calibration-score

In order to assess how well the model outputs match the surveillance data, the calibration-score is calculated for each simulation. If the calibration-score is within a defined tolerance threshold, then the set of parameter values used for that simulation are accepted. It is defined as the weighted sum

of the deviances of the modelled outputs from the observed data, averaged over the number of years data was available for and for each type of data available.

The calibration-score was derived based on the chi-square test statistic, which is a measure of how close the observed frequencies are to the expected frequencies. If D is the surveillance (and/or observational) data point and M is the corresponding output of the model for the same data point, then the deviance is simply quantified by

$$\frac{|D - M|}{D}$$

We use this definition because it succinctly and simply captures the magnitude of the difference between the surveillance data and modelled outputs.

This deviance is calculated for each data point, which is defined by the calendar year, i, and data point, j (e.g. one data point could be the number of HIV diagnoses reported in 2005, so i=2005 and j=number of reported HIV diagnoses). 'Data items' are described later. For each of the n_d data items, the sum of the deviances is then averaged over the number of years, r_j, that data were available. The un-weighted calibration-score per data item j is then summarised by

$$\text{un-weighted calibration-score, } S_j = \frac{1}{r_j} \sum_{i=y_1}^{y_r} \frac{\left|D_{ij} - M_{ij}\right|}{D_{ij}}$$

Where i=[y_1,…,y_r] represents the calendar year and j=[1,…,n_d] represents the data item.

The calibration-score can be further weighted by factors which describe the confidence in the observed data. Specifically,

$$\text{weighted calibration-score} = \frac{\sum_{j=1}^{n_d} w_j S_j}{\sum_{j=1}^{n_d} w_j}$$

The weights, w_j, should ideally be chosen *a priori* to reflect the confidence, quality or conversely, uncertainty associated with data items. A larger weight would be used for data types which the model should calibrate better to. This would ensure that these data types would contribute more towards the calibration-score. The potential range of weights used was decided to be [0.5,5]. Any data item deemed to have a weight of less than 0.5 because of the lack of certainty or quality should not be included in the calibration-score.

The calibration-score is calculated in a way that it is standardised by the quantity of observed data and the weights used, and therefore the calibration-score for one setting (with a certain range of data items with associated confidence in quality) can be directly compared to that of another setting with a different range of data items and associated confidence in quality. The main limitation is that the weights chosen will be subjective, which is why they are chosen *a priori*.

Therefore, given the formula shown above, the property of the calibration-score means that the lower the calibration-score, the smaller the deviance between the modelled data and observed data and thus the better the model calibrates to the data. The aim of the fitting method is therefore to find sets of parameter values which achieve low calibration-scores.

Examples of observed data (termed ‘data items’) which could be used to calibrate the model include for a given year:

- Number of HIV diagnoses by year
- Number of AIDS diagnoses by year
- Number of simultaneous HIV/AIDS diagnosis by year
- Number of deaths by year
- CD4 count at diagnosis (median CD4 count, or proportion with CD4 count <200 or <350 cells/mm^3)
- Proportion of diagnosis which were recently acquired infections
- Number of people seen in care
- Number of people seen in care and on ART

In the process of developing this calibration procedure, we judged that a simulation with a calibration-score <0.25 demonstrated that the modelled data were fairly comparable to the observed data. Simulations with a calibration-score <0.2 demonstrated an even closer comparison. In this particular application of the method, we have therefore specified that the tolerance threshold of a well-calibrated model should have a calibration-score of <0.2 Note that a simulation with calibration-score <0.2 can be interpreted as the average deviation of the modelled outcomes from the observed data across all data items is <20%. In a setting with large quantities of data to calibrate to, we consider that it would be rare to observe a calibration-score <0.1 because it would be unlikely to be able to simultaneously calibrate to all sources of data, because sometimes the data themselves are inconsistent with each other.

Referring to stage 1, as described in the beginning of this section (3.5), the tolerance threshold used here will be greater than 0.2, because from experience it has been very unlikely to achieve parameter sets which calibrate well, purely from 10,000 samples. Also the aim is to narrow the prior distribution without risking exclusion of parameter values that could be part of low calibration-score parameter sets. The threshold chosen here will vary by setting. It may be as low as 0.3, 0.5 or even as high as 1.0.

In the second stage of the calibration process however, this is where we remain until enough simulations are completed with calibration-score <0.2. If this never happens for a given calibration-score, then it is likely that there are inconsistencies between the data sources. So either, the calibration-score can be re-defined (different data items or weights after re-consideration of possible biases in data) or rather than using a tolerance threshold of 0.2, this may have to be increased to 0.3 or higher.

The parameter sets used in stage 3 will be those which achieved a calibration-score<0.2 in the second stage.

The final model outputs are based on the simulations from stage 3 which again achieve a calibration-score<0.2

3.7 Additional features of calibration procedure

During each simulation, the calibration-score will be calculated in five year intervals. If it becomes clear during the simulation, that it will not be possible, based on the current calibration-score, to

calibrate well to the observed data, then the simulation will be terminated prematurely. These criteria may differ by setting, but an example criterion could be that the calibration-score is already > 0.5, i.e. the modelled outputs are more than 50% away from the observed data on average. This has the advantage that we save on computing resources to be used for other runs.

Re-constructing the HIV-positive population for a setting may involve a vast number of people – as an individual-based model, this means greater computing resources and time are required. For settings with more than 10,000 individuals thought to be HIV-positive in 2012, we choose a random sample of these people, and simulate only a proportion of them. The modelled outcomes are then multiplied back up to represent the full population. The calibration-score is calculated from this full population data.

In the calibration process, we use a process by which simulations are kept if the calibration score is less than the threshold (e.g. 0.2), and throw away those which are greater than this threshold. This is a standard approach in ABC. This means that in the second stage of the calibration process, simulations with a calibration score of 0.19 are less likely to be carried forward into the third stage compared with simulations with a calibration score of 0.01. Nevertheless, once the final sets of parameter values are chosen for use in the third stage, we do give all simulations equal weight, no matter what the calibration score was in the previous stage. We recognise that there are other ways in which to choose the parameter sets. One alternative method is to consider weighting the ones with lower calibration score more highly, however this is not undertaken within this paper.

4 MSM epidemic in the UK

To reconstruct the HIV epidemic in MSM in the UK, we simulate a random 1/10th of the number of infections which took place (predominantly to make the simulations more manageable). Although this depends on the actual incidence sampled, this corresponds to approximately 4,000 to 10,000 people. We assume that the HIV epidemic began in 1980, with first diagnoses in 1984, and follow-up until the end of 2013.

4.1 Method

4.1.1 Choice of prior distributions

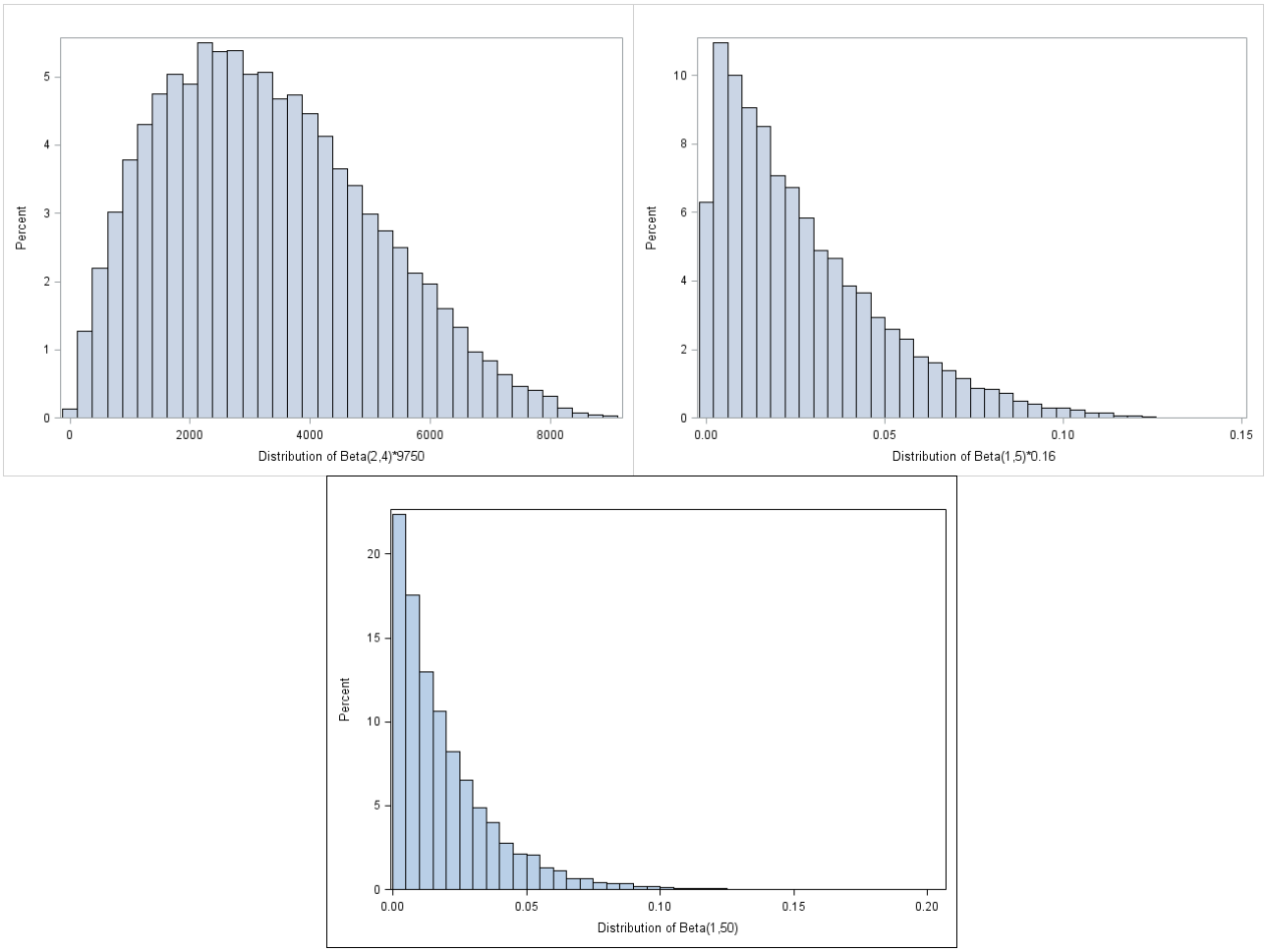
The prior distributions for the incidence and diagnosis rate parameters which are used are shown in **eTable 3** and **eFigure 4**. The prior distributions were informed using the formulae presented in section 3.4.4.

eTable 3: Prior distribution used to simulate MSM epidemic in the UK

Parameter	Prior distribution	Posterior distribution [90% plausibility range]
i1	Beta(2,4)*9750	[822,2205]
i2	Beta(2,4)*9750	[951,3275]
i3	Beta(2,4)*9750	[706,2600]
i4	Beta(2,4)*9750	[596,2257]
i5	Beta(2,4)* 9750	[1394,3942]
i6	Beta(2,4)* 9750	[1960,4567]
i7	Beta(2,4)* 9750	[2324,5438]
d1	Beta(1,5)*0.16	[0.003,0.027]
d2	Beta(1,50)	[0.013,0.038]*
d3	Beta(1,50)	[0.020,0.056]*
d4	Beta(1,50)	[0.026,0.061]*

* These posterior distributions actually refer to the absolute diagnosis rate for that time period, as opposed to the additional change.

eFigure 4: Histogram of prior distributions; parameters i1-i7 (top left), parameter d1 (top right), parameters d2-d4 (bottom)



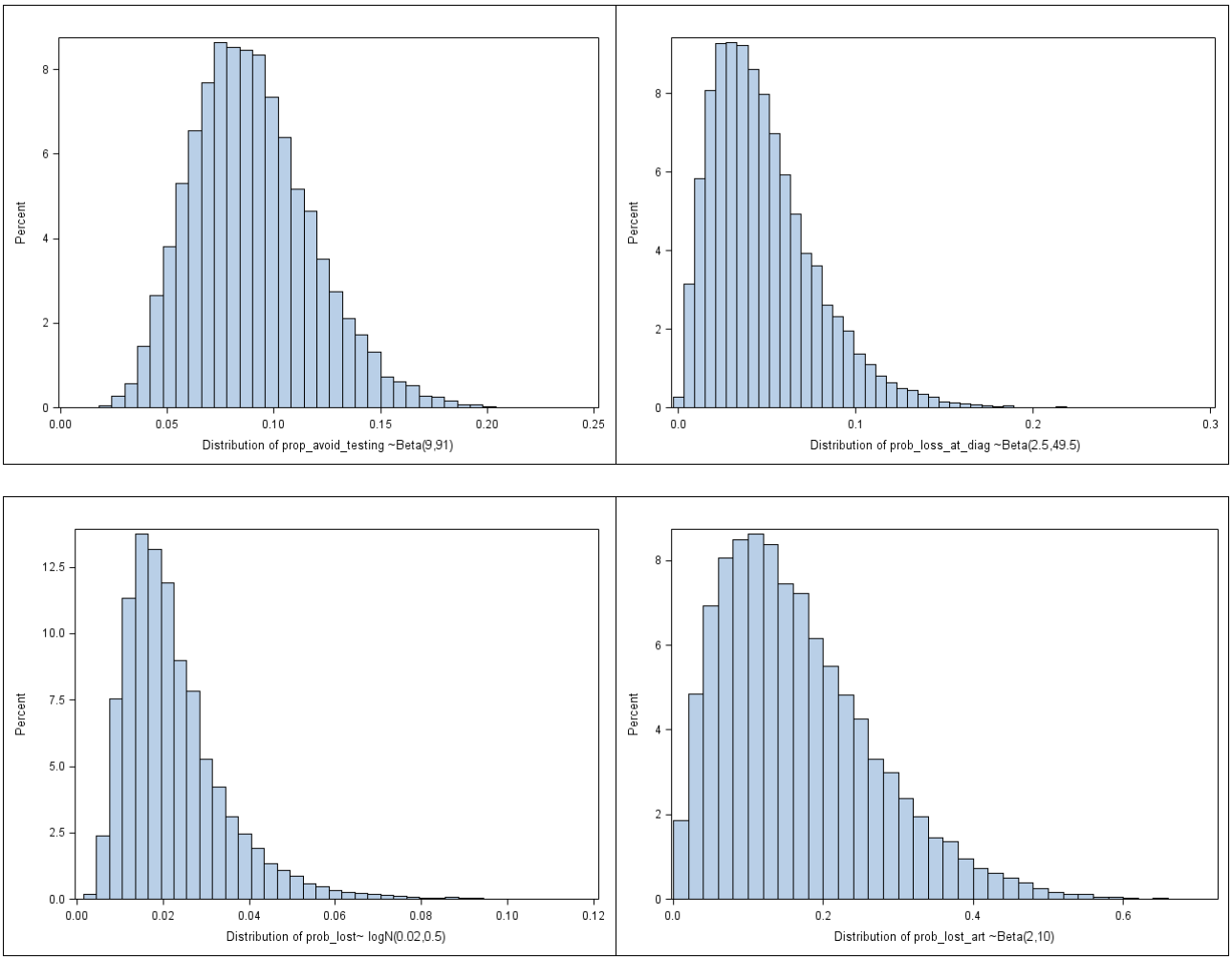
4.1.2 Other sampled parameters

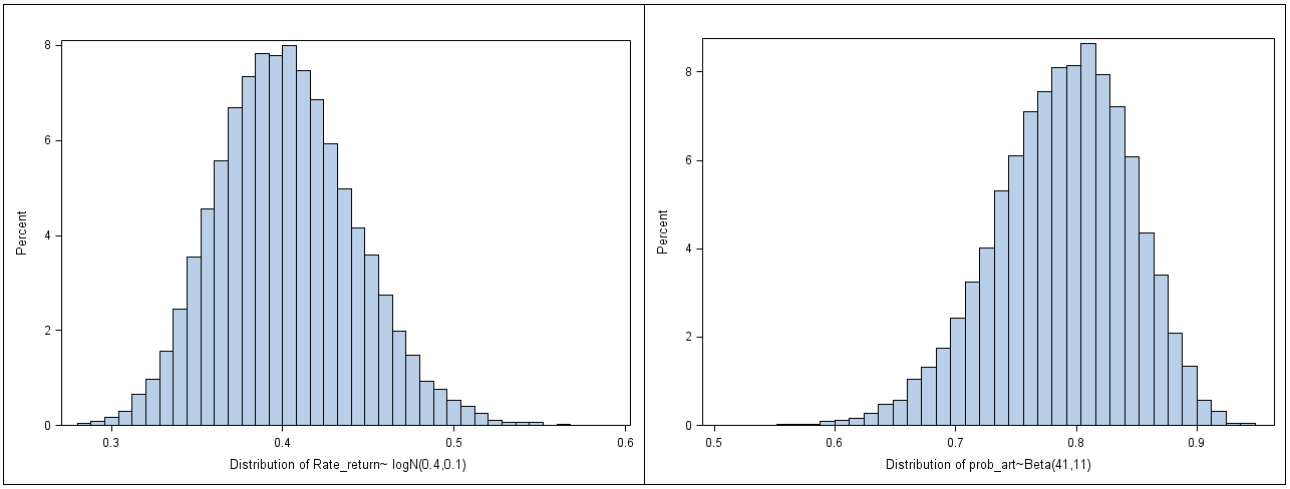
The prior distributions for the other parameters which are sampled per simulation are given in **eTable 4**.These distributions were derived from observational studies carried out in the UK and were informed by the data given in **eTable 5**.

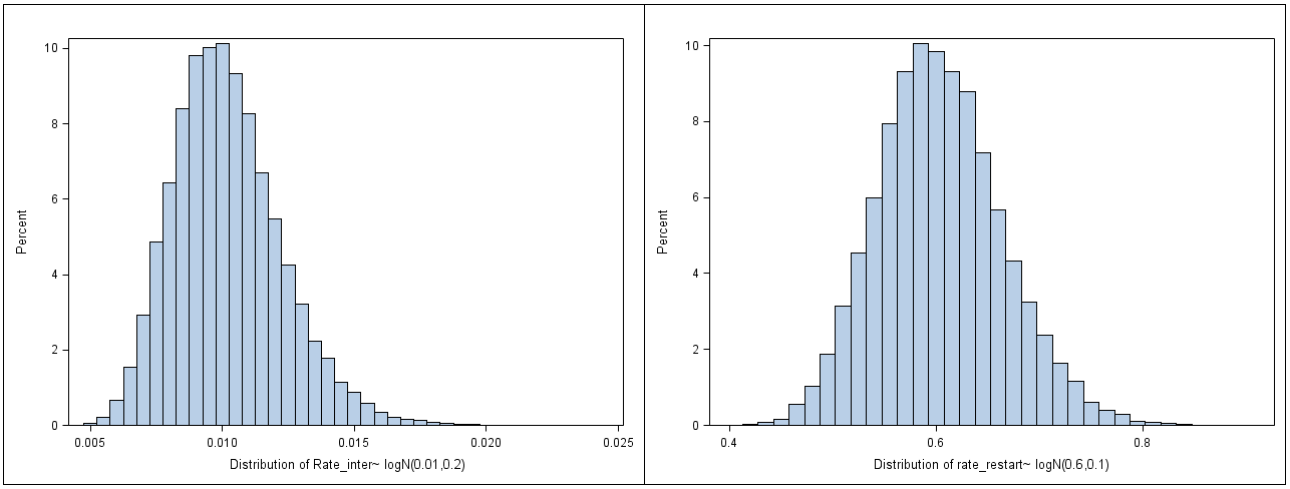
eTable 4: Parameters sampled and the distribution sampled from. All rates expressed per 3 months unless stated otherwise.

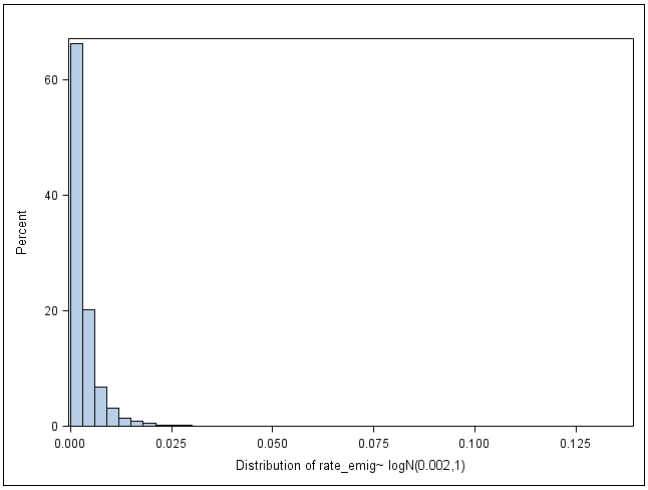
Parameter	Prior distribution*	Mode (Beta) or Mean (LogNormal) of prior distribution	Posterior distribution (90% plausibility range)
Prop_avoid_testing	Beta(9,91)	0.08	(0.039,0.136)
prob_loss_at_diag	Beta(2.5,49.5)	0.03	(0.011,0.105)
prob_lost	Log N(0.02,0.5)	0.02	(0.009,0.026)
prob_lost_art	Beta(2,10)	0.1	(0.025,0.406)
Rate_return	Log N(0.4,0.1)	0.4	(0.300,0.530)
Prob_art	Beta(41,11)	0.8	(0.695,0.873)
Adh_pattern	1: 50%, 2: 50%	N/A	1: 44%, 2: 56%
Rate_inter	Log N(0.01,0.2)	0.01	(0.007,0.014)
Rate_restart	Log N(0.8,0.1)	0.8	(0.701,0.801)
Rate_emig	Log N(0.002,1)	0.002	(0.0004,0.007)

* Probability distributions of parameters shown below









4.1.3 Number of simulations and computing resources

The number of simulations in the first stage of the calibration process was 10,000. Simulations were terminated prematurely if the total number of AIDS case reports in 1986-1990 or 1991-1995 were more than 50% greater than actually observed. Similarly, simulations were also terminated prematurely if the total number of HIV case reports in 1996-2000, 2001-2005 or 2006-2011 were

more than 50% greater than actually observed. In the second stage, the model was ran until there were at least 100 parameter sets with a calibration-score <0.2 (i.e. parameters were accepted only if the calibration-score was within the tolerance threshold of 0.2). In the third stage, we simulated a further 1000 runs (random sampling with replacement, one of the 100 parameter sets).

The results in the manuscript are based on the 742 runs (out of the 1000 runs in the final stage), where the calibration-score was again <0.2 (there were 258 runs where the calibration-score was >0.2 in these sets of simulations).

When submitting jobs to the computing cluster, we are required to estimate and request for sufficient computing resources which the simulation will use, including wall-clock time, RAM size and temporary file sizes. For each simulation run, we allocated a wall-clock time of 2 hours within the computing cluster. This is the maximum time that we allow the simulation to run for, so that the correct computing resources can be allocated within the cluster. In reality however, each simulation should take no longer than a few minutes on a fast node. The 20,000 simulations which were submitted, therefore took a maximum of 40,000 computing hours. However, all of these simulations did not take the full 2 hours and many simulations were also terminated prematurely due to the fact that it was already clear they would not provide a good fit.

eTable 5: Data used to inform choice of distributions to sample each parameter listed in eTable 4.

Parameter	Data used to inform parameter distribution	Value in our example (MSM, UK)	Data source in our example (MSM, UK)
Prop_avoid_testing	Probability of uptake of HIV tests when offered	88-90% of MSM who were offered a test in STI clinics in England took one.	*Health Protection Agency. Time to test for HIV: Expanded healthcare and community HIV testing in England. Interim Report. December 2010*
prob_loss_at_diag	Proportion of CD4 measured within 3 months of diagnosis	97% for all adult patients in the UK.	*Health Protection Agency. HIV in the UK: 2012 Report. London: Health Protection Services, Colindale. November 2012.*
prob_lost	Proportion still in care after 12 months from diagnosis	86% for all adult patients in the UK.	*Health Protection Agency. HIV in the UK: 2012 Report. London: Health Protection Services, Colindale. November 2012.*
prob_lost_art	Proportion in care among adults seen for care in the last 12 months	96% for all adults seen for care in 2010 and also seen in 2011.	*Health Protection Agency. HIV in the UK: 2012 Report. London: Health Protection Services, Colindale. November 2012.*
Rate_return	Mean length of time spent interrupting ART in people seen in clinics	Of those with suppressed viral load (<50 copies/ml) in UK CHIC, median duration of each interruption was 4.4 (IQR: 1.9-10.1) months.	*Bansi LK et al. Are previous treatment interruptions associated with higher viral rebound rates in patients with viral suppression? AIDS 22(3): 349-356, 2008*
Prob_art	Proportion of ART-naïve people with CD4 count<350 cells/mm^3	9% in all adults in UK CHIC cohort collaboration.	*Kober C, Johnson M, Fisher M, et al. Non-uptake of highly active antiretroviral therapy among patients with a CD4 count < 350 cells/μL in the UK. HIV Med. 2012 Jan;13(1):73-8.*
Adh_pattern	Proportion with viral load <50 copies/ml within 12 months of starting ART	87% for all adult patients in the UK.	*Health Protection Agency. HIV in the UK: 2012 Report. London: Health Protection Services, Colindale. November 2012.*
Rate_inter	Proportion of ART-experienced people who are currently on ART	5.3% and 2.7% respectively interrupted (discontinued all drugs for >2 weeks) during the year in 2003 and 2009 in Royal Free cohort.	*Smith CJ et al. Frequency of treatment interruptions over calendar time: The impact of results from the SMART study. 2011, HIV Med 12(Suppl 1):84*
Rate_restart	Mean length of time spent interrupting ART in people seen in clinics	Of those with suppressed viral load (<50 copies/ml) in UK CHIC, median duration of each interruption was 4.4 (IQR: 1.9-10.1) months.	*Bansi LK et al. Are previous treatment interruptions associated with higher viral rebound rates in patients with viral suppression? AIDS 22(3): 349-356, 2008*
Rate_emig	Estimate of rate of emigration (per 100,000) in the general population (but by risk group if available)	In UK population (63 million), around 300,000-400,000 people emigrate each year.*	*Annual Report on Migration and International Protection Statistics for United Kingdom 2008.* *Rice BD et al. Loss to Follow-Up Among Adults Attending Human Immunodeficiency Virus Services in England, Wales,and Northern Ireland. 2011, Sex Trans Inf 38(8):685-90*

* It is possible that the association between loss to follow-up and black-African ethnicity, acquiring HIV-infection abroad, and having a recent diagnosis can be explained by migrants moving to the United Kingdom leaving shortly after receiving an HIV diagnosis. Their emigration may be voluntary or involuntary, temporary or permanent

4.2 Using different parameterisation for the incidence curve

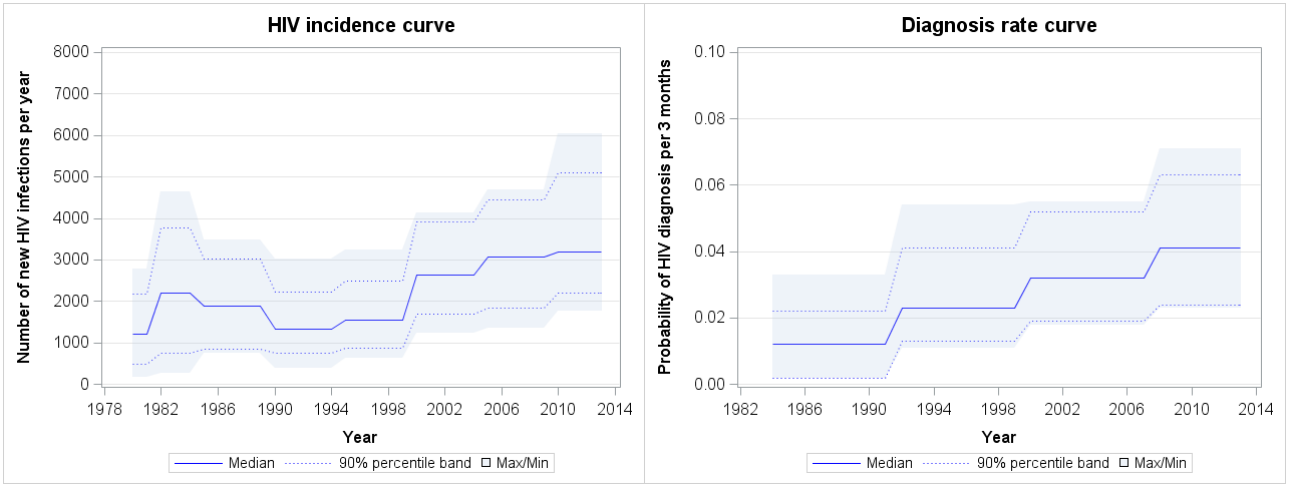
We also looked at parameterising the incidence curve differently, given that the incidence curve in our main results did not reflect a peak in the early to mid-1980s as seen in other MSM epidemics in Western Europe(14-16). Some of this peak may be an artefact of an increase in the number of tests conducted in this period; however it is difficult to disentangle exactly how much was a due to the rise in the number of infections and how much due to an increase in testing.

For the figures in the circulated manuscript, we used 5-year fixed values, each representing 1980-4, 1985-89, 1990-4 etc. For the results below, we use the same calibration method, but instead split the first period into 1980-1 and 1982-4 (i.e. sample one additional parameter).

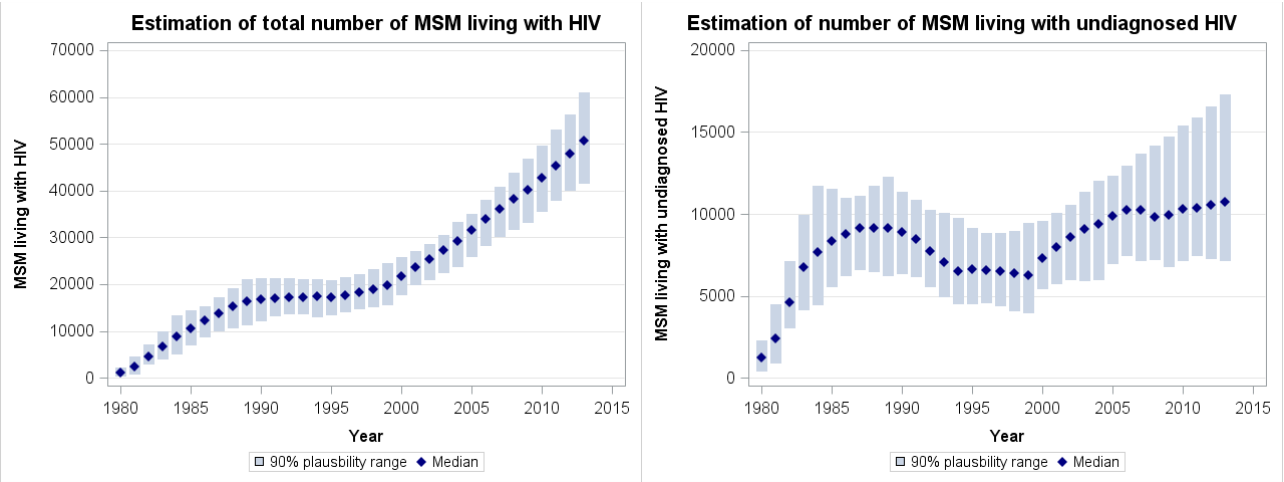
eFigure 5 shows the re-parameterised incidence curve (and corresponding diagnosis rate curve). The peak in incidence which is seen using other back-calculation type models is not obvious looking at the median value, although the 90% percentile band shows that it is plausible.

The main outcomes of interest, total number living with HIV and number living with undiagnosed HIV are shown in **eFigure 6**. For the most recent decade in particular, the estimates are similar to the main results in the manuscript. We consider that this example also demonstrates the repeatability of the calibration method.

eFigure 5: Re-parameterised estimated incidence and diagnosis rate amongst MSM in the UK



eFigure 6: Estimates of the total number of MSM living with HIV in the UK and total number of MSM living with undiagnosed HIV, by calendar year, using re-parameterised incidence parameters. Columns and aste bars: Modelled median and 90% PR.



4.3 Sensitivity analysis

4.3.1 Method

We conducted a sensitivity analysis to evaluate the impact of using different sets of weights given to the calibration-score and of using different calibration-score tolerance thresholds. Based on the outputs from the second stage of the calibration procedure, we have simulated 8 further scenarios, in addition to the main results presented within the main manuscript (i.e. 9 scenarios presented here in total).

Three different sets of weights and three different tolerance thresholds were considered in a factorial structure. The different sets of weights which were considered were labelled weight A, B and C. These sets of weights retain the same ordering of weights (if a data item X has the smallest weight in set A, then it shall also have the smallest weight in set B and C) but we have altered the relative differences between the weights. The allocation of weights to the different data items are summarised in the table below.

Data item	A	B	C
Number of HIV diagnoses, 1997-2012	1	2	5
Number of first AIDS diagnoses, 1980-1996	1	2	5
Median CD4 count at diagnosis, 1980-1996	0.5	0.5	1
Median CD4 count at diagnosis, 1997-2012	1	2	5
Proportion of diagnoses which were in recently acquired infections, 2009-2012	0.5	0.5	1
Number seen for care, 1998-2012	0.5	0.5	1
Number seen for care and on ART, 1999-2012	0.5	0.5	1

The three tolerance thresholds for the calibration-score which were considered were 0.18, 0.20 and 0.23. We resampled from the chosen parameter sets to generate the final results. A further 100, 300

and 1,000 simulations were performed by resampling the parameter sets chosen for the thresholds 0.18, 0.20 and 0.23 respectively.

The outputs which were compared between the 9 scenarios were the estimated number of people living with HIV in 2013, as well as the estimated number of infections per year between 2010 and 2013.

4.3.2 Results

eTable 6 summarises the results from the sensitivity analyses. Note that the row with tolerance threshold 0.20 and set of weight A is the same as the main result presented in the main manuscript. These results show that the simulations are not very sensitive to changes in the calibration score tolerance threshold or allocation of weights to data items. The greater the tolerance threshold however, the wider the 90% plausibility range.

eTable 6: Sensitivity analysis results

Calibration-score tolerance threshold	Set of weight used	Number of parameter sets included	Median modelled outcomes (90% plausibility range)	
			Estimated number of people living with HIV	Estimated incidence for 2010-2013
0.18	A	42	51,200 (41,400-60,000)	3,440 (2,470-5,030)
	B	40	50,500 (40,100-60,700)	3,400 (1,890-5,620)
	C	42	50,000 (38,400-60,400)	3,340 (2,320-5,030)
0.20	A	121	51,000 (41,400-61,000)	3,270 (2,310-5,380)
	B	118	50,000 (40,100-61,200)	3,360 (1,760-5,440)
	C	125	50,600 (40,500-60,400)	3,430 (1,890-5,570)
0.23	A	359	47,900 (38,100-61,500)	3,420 (1,810-5,570)
	B	354	48,000 (35,700-59,700)	3,100 (1,530-5,410)
	C	360	48,100 (36,800-61,200)	3,230 (1,530-5,620)

5 Using pseudo data

To demonstrate the calibration procedure and to show how well the procedure would work in a setting with varying amounts of data to fit to, we also simulate a hypothetical epidemic where the incidence rate and diagnosis rate are fully known. The pseudo data were also generated using Synthesis model. Such pseudo data was used to provide the true incidence, diagnosis rate and outcomes from the hypothetical epidemic, and thus these are fully known, which then allows us to use the method described above to see how closely it is able to reconstruct the true epidemic. Although this approach is somewhat circular, in that the same model is used to generate the epidemic and to analyse it, it provides a useful means of being able to compare our calibration method in the presence of differing levels of data availability. We conceived three scenarios of data availability, high, medium and low:

	Simulated data availability (weights used in calibration-score given in brackets)		
	High	**Medium**	**Low**
Number of HIV diagnoses	1985-2012 (1)	1996-2012 (1)	2011-2012 (1)
Number of first AIDS diagnoses	1985-2012 (1)	1996-2012 (1)	
Number of deaths	1985-2012 (1)		
Median CD4 count at diagnosis	1990-2012 (1)		2011-2012 (1)
Proportion of diagnoses where CD4 count <200 cells/mm^3	1990-2012 (1)		
Proportion of diagnoses which were in recently acquired infections	2009-2012 (1)		
Number seen for care	1998-2012 (1)		
Number seen for care and on ART	1998-2012 (1)	2000-2012 (1)	

Data are available on a per year basis, i.e. there are 28 data points (1985 to 2012 inclusive) to calibrate for the number of HIV diagnoses in the high simulated data availability scenario. The ‘high’ data availability scenario was based on a setting with a well-established surveillance system for HIV which has been used since the first infections were reported. The ‘medium’ data availability scenario was based on a setting which did not start conducting HIV surveillance until the cART era. The ‘low’ data availability scenario was based on a setting with no HIV surveillance until very recently.

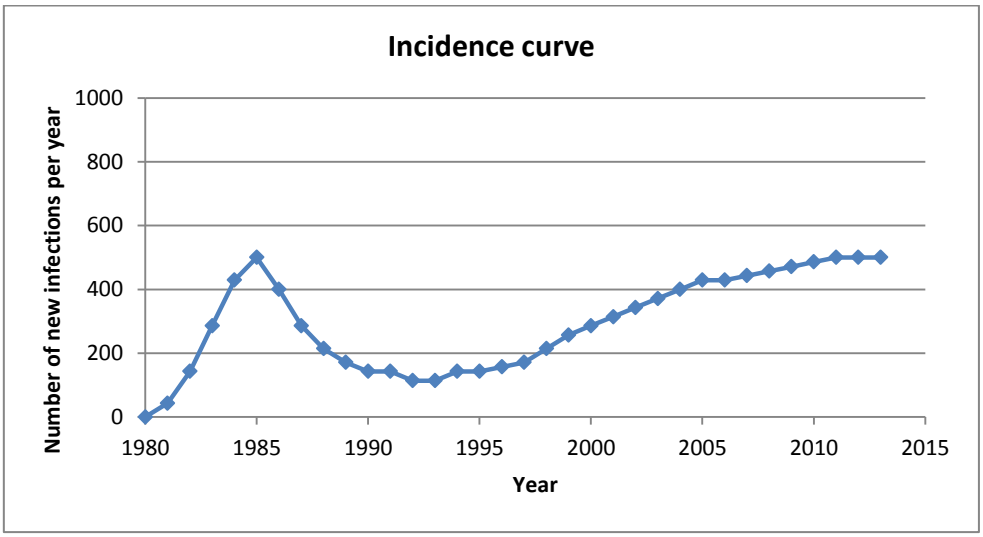
5.1 Choice of epidemic

The hypothetical epidemic is loosely based on that thought to have occurred in MSM in Western Europe.

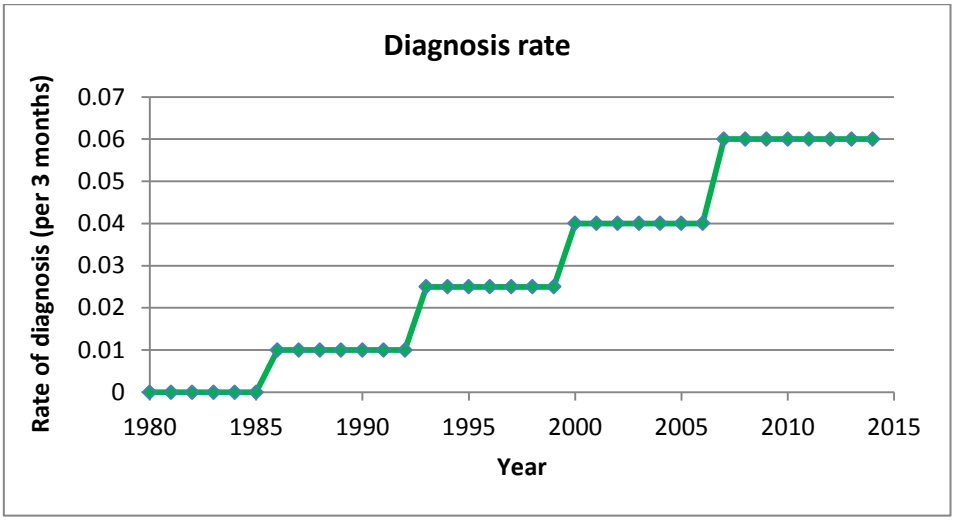
The incidence curve (**eFigure 7**) depicts the number of infections over calendar time. In Western European countries, there is thought to have been an early peak in incidence in mid-1980s, plateauing in the mid-1990s and gradually increasing again. The total number of infections from 1980 to 2014 was 10,000.

Diagnoses of HIV start from 1985 onwards. We assume that the diagnosis rate is monotonically increasing, i.e. entirely non-decreasing (**eFigure 8**).

eFigure 7: Number of infections per year in hypothetical epidemic



eFigure 8: Rate of diagnosis per 3 months in hypothetical epidemic



5.2 Method

5.2.1 Choice of prior distributions

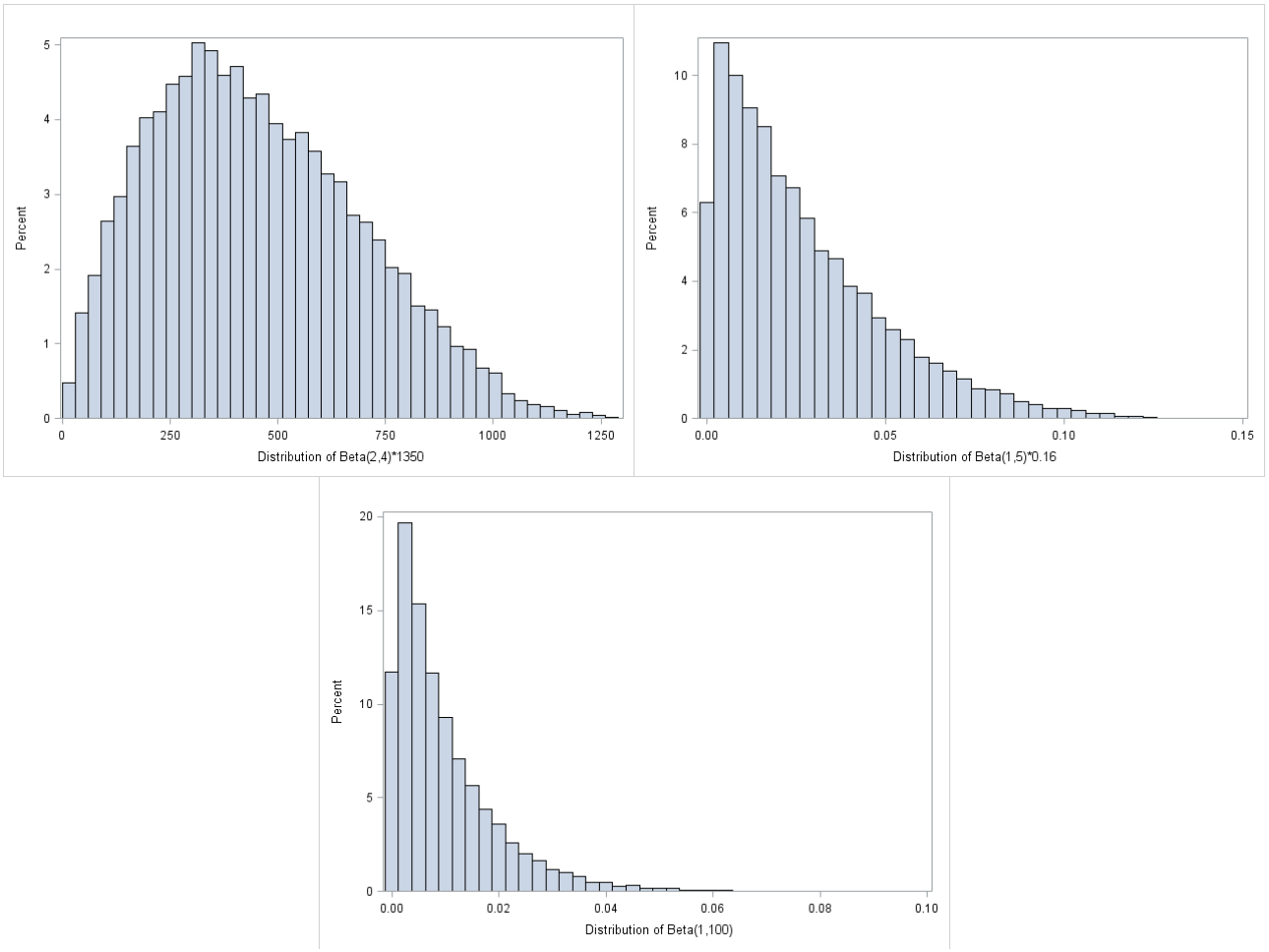
The prior distributions for the incidence and diagnosis rate parameters which are used are shown in **eTable 7** and **eFigure 9**. Priors were chosen as we would when these parameters are unknown. As the epidemic which we are trying to reconstruct has the same incidence and diagnosis rate, the priors used for all three data availability scenarios (high, medium and low) are the same.

In this example using pseudo data, we chose not to sample the other parameters listed in **eTable 2** for simplicity. These parameters are those which describe the average behaviour of the population. For any given setting, the values would be informed by observational data from that setting or similar. We saw in Section 4 for the MSM example that the posterior distributions did not differ hugely from the prior distributions. These parameters themselves are unlikely to alter the incidence and diagnosis rate parameters greatly.

eTable 7: Prior distributions used for all scenarios

Parameter	Prior distribution
i1	Beta(2,4)*1350
i2	Beta(2,4)*1350
i3	Beta(2,4)*1350
i4	Beta(2,4)*1350
i5	Beta(2,4)*1350
i6	Beta(2,4)*1350
i7	Beta(2,4)*1350
d1	Beta(1,5)*0.16
d2	Beta(1,100)
d3	Beta(1,100)
d4	Beta(1,100)

eFigure 9: Histogram of prior distributions; parameters *i1-i7* (top), parameter *d1* (middle), parameters *d2-d4* (bottom)



5.2.2 Number of simulations

The number of simulations in the first stage of the calibration process was 10,000. Similarly to the example using MSM data from the UK, we decided to aim for at least 100 parameter sets with

calibration-score <0.2 for each data availability scenario in the second stage of the calibration procedure. For this illustrative example the third stage of the calibration was not implemented, nor were any simulations terminated prematurely.

The three different data availability scenarios required a different number of simulations in total to achieve 100 parameter sets with calibration-score<0.2. The number of simulations to be run will depend on how quickly we can find 100 such parameter sets. The final set of results presented are based on the 100 smallest calibration-scores were used. The smallest calibration-score which was achieved amongst these 100 simulations were 0.153, 0.138 and 0.001 respectively for 'high', 'medium' and 'low' data availability. This illustrates that the more data there are to calibrate the model to, the harder it is to find a smaller calibration-score. Although the 'low' data availability situation led to the smallest calibration-score, looking at the plausibility range presented in Figure 5 in the main manuscript, we can deduce that a small calibration-score does not necessarily mean that it captures the underlying epidemic well, but in fact just calibrates very closely to the small amount of data available. So in other words, while the calibration-scores are in some sense comparable between situations with a large amount of data to fit to or little, it is important to bear in mind that a calibration-score of 0.18, say, based on a large amount of data is more likely to be accurately capturing an underlying epidemic than a simulation with a calibration-score of 0.07 based on little data. Although this means that the calibration-score itself is somewhat hard to interpret, it is still a useful concept because it indicates that the model was calibrated to the observed data with a given error margin. In the most ideal situation, the model will be calibrated to a range of observed data with the smallest calibration-score possible. This will vary by setting however, as if the data within the surveillance system are inconsistent with each other, given the model then this indicates either bias in one or both observed data sources or model misspecification. If these cannot be resolved (and the basic underlying model cannot be changed just to fit to one set of country data and it should only be done if model-specification is consistently indicated over multiple country data calibrations procedures have been performed), such a conflict would mean that the calibration-score threshold may have to be larger than desired.

Reference List

(1) Phillips AN, Sabin C, Pillay D, Lundgren JD. HIV in the UK 1980-2006: Reconstruction using a model of HIV infection and the effect of antiretroviral therapy. HIV Med 2007;8(8):536-46.

(2) Bansi L, Sabin C, Delpech V, Hill T, Fisher M, Walsh J et al. Trends over calendar time in antiretroviral treatment success and failure in HIV clinic populations. HIV Med 2010;11(7):432-8.

(3) Nakagawa F, Lodwick RK, Smith CJ, Smith R, Cambiano V, Lundgren JD et al. Projected life expectancy of people with HIV according to timing of diagnosis. AIDS 2012;26(3):335-43.

(4) Phillips AN, Cambiano V, Nakagawa F, Brown AE, Lampe F, Rodger A et al. Increased HIV Incidence in Men Who Have Sex with Men Despite High Levels of ART-Induced Viral Suppression: Analysis of an Extensively Documented Epidemic. PLoS One 2013;8(2):e55312.

(5) Hamers RL, Wallis CL, Kityo C, Siwale M, Mandaliya K, Conradie F et al. HIV-1 drug resistance in antiretroviral-naive individuals in sub-Saharan Africa after rollout of antiretroviral therapy: a multicentre observational study. Lancet Infect Dis 2011;11(10):750-9.

(6) Hoffmann CJ, Charalambous S, Sim J, Ledwaba J, Schwikkard G, Chaisson RE et al. Viremia, Resuppression, and Time to Resistance in Human Immunodeficiency Virus (HIV) Subtype C during First-Line Antiretroviral Therapy in South Africa. Clin Infect Dis 2009;49(12):1928-35.

(7) Lathouwers E, Kambili C, Haddad M, Paquet A, De Meyer S, Baugh B. Trends in darunavir resistance-associated mutations and phenotypic resistance: US, 2006 to 2012. 20th Conference on Retroviruses and Opportunistic Infections. March 3-6, 2013. Atlanta. Abstract 590. 2014.

(8) Orkin C, DeJesus E, Khanlou H, Stoehr A, Supparatpinyo K, Lathouwers E et al. Final 192-week efficacy and safety of once-daily darunavir/ritonavir compared with lopinavir/ritonavir in HIV-1-infected treatment-naïve patients in the ARTEMIS trial. HIV Med 2013 January 1;14(1):49-59.

(9) Phillips A, Baker J, Lundgren J. Are antiretrovirals enough for people living with HIV? Lancet 2013;382(9903):1466-7.

(10) Rodger A, Lodwick R, Schechter M, Deeks S, Amin J, Gilson R et al. Mortality in Patients with Well-controlled HIV and High CD4 Counts in the cART Arms of the SMART and ESPIRIT Randomized Clinical Trials Compared to the General Population. 2012.

(11) Freiberg MS, Chang CC, Kuller LH, Skanderson M, Lowy E, Kraemer KL et al. HIV Infection and the Risk of Acute Myocardial Infarction. Jama Internal Medicine 2013;173(8):614-22.

(12) Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian Computation in Population Genetics. Genetics 2002 December 1;162(4):2025-35.

(13) McKay MD, Beckman RJ, Conover WJ. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics 1979 May 1;21(2):239-45.

(14) Sommen C, Alioum A, Commenges D. A multistate approach for estimating the incidence of human immunodeficiency virus by using HIV and AIDS French surveillance data. Stat Med 2009;28(11):1554-68.

(15) Bezemer D, de Wolf F, Boerlijst MC, van Sighem A, Hollingsworth TD, Fraser C. 27 years of the HIV epidemic amongst men having sex with men in the Netherlands: an in depth mathematical model-based analysis. Epidemics 2010;2(2):66-79.

(16) Artzrouni M. Back-calculation and projection of the HIV/AIDS epidemic among homosexual/bisexual men in three European countries: Evalution of past projections and updates allowing for treatment effects. European Journal of Epidemiology 2004;19(2):171-9.