

Published in final edited form as:

Med Decis Making. 2015 February ; 35(2): 148–161. doi:10.1177/0272989X13493143.

## Calibration of complex models through Bayesian evidence synthesis: a demonstration and tutorial

Christopher Jackson, Mark Jit, Linda Sharples, and Daniela DeAngelis

### Summary

Decision-analytic models must often be informed using data which are only indirectly related to the main model parameters. The authors outline how to implement a Bayesian synthesis of diverse sources of evidence to calibrate the parameters of a complex model. A graphical model is built to represent how observed data are generated from statistical models with unknown parameters, and how those parameters are related to quantities of interest for decision-making. This forms the basis of an algorithm to estimate a posterior probability distribution, which represents the updated state of evidence for all unknowns given all data and prior beliefs. This process calibrates the quantities of interest against data, and at the same time, propagates all parameter uncertainties to the results used for decision-making. To illustrate these methods, the authors demonstrate how a previously-developed Markov model for the progression of human papillomavirus (HPV16) infection was rebuilt in a Bayesian framework. Transition probabilities between states of disease severity are inferred indirectly from cross-sectional observations of prevalence of HPV16 and HPV16-related disease by age, cervical cancer incidence, and other published information. Previously, a discrete collection of plausible scenarios was identified, but with no further indication of which of these are more plausible. Instead, the authors derive a Bayesian posterior distribution, in which scenarios are implicitly weighted according to how well they are supported by the data. In particular, we emphasise the appropriate choice of prior distributions and checking and comparison of fitted models.

### 1 Introduction

Building a decision-analytic model to represent the history of disease and treatment usually involves choices that are based on uncertain information. The magnitude of this uncertainty can be quantified, but improperly quantifying uncertainty can lead to biased decisions as well as wrongly allocating resources for further research [1, 2]. Uncertainties in models are usually best characterised probabilistically: by parameterising the model flexibly [3], characterising each parameter by a data-derived distribution and simulating the resulting probabilistic model to produce a distribution for model outputs. One approach to this is described by the umbrella term of *Bayesian evidence synthesis*, which is a statistical framework for explicitly modelling several related and connected sources of data, and naturally incorporates the uncertainty in model parameters.

In particular, when the current evidence is weak or only indirectly related to the main parameters, it may not be straightforward to accurately represent it in the model. A crude approach to calibrating a model against indirect data is to informally adjust the parameters until predictions of key outcomes visually appear to fit observed data. For example, recent

models of human papillomavirus (HPV) infections [4, 5, 6] adjust parameters governing the transmission and natural history of HPV until models produce estimates of HPV prevalence which are similar to data. A more quantitative approach involves sampling model parameters from plausible ranges, comparing observed data to outputs from the model, and retaining a subset of parameters which satisfy some arbitrary standard of fit to the data, resulting in a range of scenarios, usually with no further indication of which are more plausible. This approach has been used, for example, in models of hepatitis C [7] and HPV [8, 9, 10, 11]. Vanni et al. [12] reviewed the choices involved in model calibration, including the appropriate measure of fit to the data, the algorithm to search for the best-fitting parameters, the standard of fit required to deem a set of parameters plausible, and methods for weighting the retained scenarios in probabilistic sensitivity analysis. The uncertainty about these choices can substantially affect the model outputs [13, 14].

### 1.1 Calibration as Bayesian evidence synthesis

These uncertainties can be accounted for by considering model calibration as a problem of Bayesian evidence synthesis. The main features of such an approach are as follows.

1. Multiple indirectly-related datasets are assumed to have been generated from probability distributions (“statistical models”) with parameters which are related to each other. Consider the simplified example in Figure 1. There are two datasets (A and B), each providing direct information about different parameters, which are known functions of the parameters of interest to decision-making. This is an example of a *directed acyclic graph*, explained in more detail in §3.
2. This graph forms the basis for computing the *posterior probability distribution* of the parameters. This represents the updated state of evidence, from combining prior beliefs and observed data using Bayes’ theorem. We describe this process in more detail in §3.
3. The uncertainty about the parameters of interest, expressed through this distribution, is then propagated through the model to generate uncertainty distributions for the results required for policy-making, such as life-years gained, infections prevented or expected costs (Figure 1).
4. The distributions of the results give appropriate weights to each parameter value, according to how much evidence there is in the data, thus providing a natural way to calibrate the model against the data.

This approach is variously referred to as multiparameter evidence synthesis [15, 16], generalized evidence synthesis [17], or comprehensive decision modelling [18]. This provides a theoretically-grounded solution to the choices discussed by Vanni et al. [12]. The measure of fit to the data is implicit in Bayes’ theorem, and there are standard methods for computing the posterior distribution (§3.4). There is no need to define an arbitrary standard of fit, and calibration is unified with probabilistic sensitivity analysis. For example, Goubar et al. [19] and Presanis et al. [20] synthesised multiple survey datasets in this way to estimate the prevalence of HIV infection. Whyte et al. [21] used these methods to calibrate the parameters of a decision model for colorectal cancer natural history, and in a model for

transmission of HPV, Bogaards et al. [22] used similar methods to estimate transmissibility and resistance to infection.

In §3 we outline the key steps of building a comprehensive Bayesian model for evidence synthesis, calibration and expressing uncertainty, and show how they can be applied to rebuilding a recently-published model for the progression of HPV infections [11], introduced in §2. By improving the characterisation of uncertainty in this model, we reflect more closely the extent of the current evidence and the potential value of further research. A more subtle benefit is that we reduce biases in the model outputs, since they are a non-linear function of the uncertain inputs [2, chapter 4]. The posterior also tells us how well the data confirm or modify our prior beliefs about the input parameters. We emphasise careful assessment of the fit of the model, potential conflicts between different sources of evidence, and the appropriate choice and influence of the prior distribution. Finally we discuss the strengths and challenges of this approach.

## 2 HPV progression model

Infection with HPV types 16 or 18 is associated with about 70% of cervical cancers. To evaluate the long-term benefits of cervical screening and vaccination against HPV, estimates of the natural history of HPV-related disease from initial infection to invasive cancer are required. A model has previously been developed [11] to estimate progression rates of HPV-related disease, through grades of cervical intraepithelial neoplasia (CIN), to cancer. This was combined with transmission and economic models to evaluate policies for HPV vaccination in the UK [10, 23].

For the purpose of our tutorial, we investigate only the progression component. A discrete-time Markov model is used to represent the natural history of a single HPV16 infection. This has a monthly cycle and 9 states, illustrated in Figure 2. The parameters we aim to estimate are the five progression and three regression probabilities  $P_{ij}$  between the disease states  $i, j: j \in \{1, \dots, 6\}$ , conditionally on the infection not clearing naturally. There were no relevant longitudinal data available for the UK at the time of the study from which to estimate these probabilities. Instead, a range of indirect cross-sectional datasets were used, listed in Table 1 together with the model parameters they directly inform. The principal data source is the recruitment phase of the ARTISTIC trial [24], which informs age-specific prevalences of HPV16 and HPV16-related cervical dysplasia. This is supplemented by data from the UK National Health Service cervical cancer screening program (NHSCCSP) [25], cancer registry data, national mortality statistics and published literature, as detailed in Jit et al. [11].

In Jit et al. [11], point estimates of the transition probabilities were calculated from data. The only expression of uncertainty was a discrete set of 54 alternative scenarios for the transition probabilities, otherwise these and many other uncertain parameters were assumed to be fixed. Here we reimplement this model as a Bayesian evidence synthesis. The scenarios and the fixed parameters are replaced by uncertain parameters with smooth prior distributions, which are updated to posterior distributions conditionally on the data. This implicitly weights each of the scenarios by how well they are supported by the evidence, and improves the characterisation of uncertainty.

### 3 Building a Bayesian model for evidence synthesis

#### 3.1 Directed acyclic graphs

In a Bayesian evidence synthesis, the “model” consists of *not only* a mathematical representation of the disease and/or treatment process (Figure 2 in this example), *but also* the network of statistical models and relationships which connect the parameters of that process with observed data and prior information. The basis of this approach is to build a *directed acyclic graph* or *graphical model* to represent these relationships, shown in Figure 3 for the HPV example. Quantities at the start of an arrow are assumed to “generate” the quantities at the end of the arrow; therefore we call these *parents* and *children* respectively. There are three basic types of quantity (or “node”).

- a) data generated from a statistical model, whose parameters are the node’s parents. These are shown as bold boxes in Figure 3.
- b) unknown parameters defined as deterministic functions of other parameters. These are shown as white boxes in Figure 3.
- c) parameters with no parents, thus no arrows directed towards them in the graph. These are shown as dashed boxes in Figure 3. In other words, while they may generate data or be used in the definitions of child parameters, they are not defined themselves as functions of further parameters. These must be given *prior distributions* representing beliefs about their plausible values, as explained further in §3.3. Or if uncertainty about them is negligible, they may be assumed to be constant, for example, mortality rates which are estimated from full-population data in the HPV application.

**Example (type a)** In the ARTISTIC data, there are  $y_k$  women diagnosed with HPV16 infection out of  $n_k$  women in several age groups  $k$ .  $y_k$  arise from a binomial distribution with denominator  $n_k$  and some probability  $p^D(t_k)$ , assuming that each woman in the age group with midpoint  $t_k$  has the same probability of being diagnosed with HPV16. In graphical model terminology, this probability and the denominator generate the observed counts, thus are the parent nodes. Note the distinction between (observed) data and (unknown) parameters. The observed prevalence is  $y_k/n_k$ , but the prevalence *parameter* is the unknown probability  $p^D(t_k)$  that an unobserved woman from the same population and age group is diagnosed with HPV16. Instead of including just  $y_k/n_k$  in the model as a constant estimate of the underlying prevalence, we estimate its posterior distribution to account for statistical uncertainty, which depends heavily on the number of women contributing to that estimate.

**Example (type b)** The DNA test for HPV16 used in ARTISTIC has 100% sensitivity, but is not always clinically relevant due to cross-reactivity between HPV types. We therefore express the *diagnosed* prevalence  $p^D(t)$  as a sum of the probability  $p^H(t)$  that a woman is truly HPV16-positive and the chance of a false positive DNA test  $p^{FP}$  multiplied by the probability of being truly HPV16-negative:

$$p^D(t) = p^H(t) + p^{FP}(1 - p^H(t)).$$

$p^{FP}$  is given a prior based on expert belief. Thus the false positive probability and the true HPV16 prevalence are the parents of the diagnosed prevalence in Figure 3.

Once nodes have been defined, they can be used in the definitions of other uncertain quantities required by the model. In this example, the true HPV16 prevalence  $p^H(t)$  is itself assumed to be a non-linear function of age  $t$  (illustrated in Figure 4). This enables us to extrapolate prevalences observed in women aged 20-64 from ARTISTIC to younger (or older) ages  $t$ , which is required since the risk of infection by sexual transmission begins around ages 10–14, and the peak HPV16 prevalence occurs around 20 years. The unknown parameters governing the non-linear curve are given prior distributions, which are updated to posteriors given the diagnosed prevalence data  $y_k$ , giving the fitted curve in Figure 4. Figure 3 shows how these parameters,  $p^H$ ,  $p^D$  and the model generating  $y_k$  are connected through the graph.

### 3.2 Including indirect data

Indirect data can be included by adding extra nodes and arrows in the graph to define how the data are generated from a statistical model, and how the parameters of that model are related to other parameters.

**Example.** The transition probabilities between grades of HPV16-related neoplasia are largely informed by age-specific counts  $y_{jk}^+$  of states  $j$  of cervical dysplasia observed by cytological screening among women diagnosed with HPV16 in ARTISTIC. These states are related to grades of neoplasia through parameters describing the sensitivity and specificity of screening, and their prevalences are also adjusted for the specificity of the test for HPV16. We also incorporate *indirect* data on these cytological state prevalences from the national screening programme. This gives the number of women, by age group, diagnosed in each state, but it is not known whether these women have any type of HPV. To include these in the model, we assume the counts of women (in age group  $k$ ) by state arise from a multinomial distribution defined by a set of probabilities  $q_{jk}$  of occupying each state  $j = 1, \dots, 5$ , and an age-specific denominator. These probabilities are an average of the prevalences among women diagnosed HPV16 positive ( $q_{jk}^+$ ) and negative ( $q_{jk}^-$ ), weighted by the probability of being diagnosed with HPV16 or not, respectively:

$$q_{jk} = p^D(t_k)q_{jk}^+ + (1 - p^D(t_k))q_{jk}^- \quad (1)$$

Thus, knowing the  $q_{jk}$ ,  $q_{jk}^-$  and  $p^D(t_k)$  gives implicit information about the  $q_{jk}^+$  to supplement the direct information from the HPV16-specific count data in ARTISTIC.

The remainder of the graphical model for the HPV example is set out in detail in the online supplement. Briefly, the transition probabilities to diagnosed (squamous cell) cervical cancer are informed by observed incidences from 2004 cancer registry data, adjusted to represent HPV16-related cases in the screened population. The transition probabilities between states of neoplasia are assumed to generate monthly CIN state prevalences and cancer incidences in a hypothetical open cohort of women infected with HPV16. These prevalence and

incidence parameters, after various adjustments, are assumed to generate the observed data. Thus, all evidence informing the model has been included in the graph, either as explicit data, parameters with prior distributions, or constant parameters.

Thus while decision models are typically described as being “populated” with values derived from data, the Bayesian approach also makes the data and its analysis an integral part of the model.

### 3.3 Specifying priors for parameters

Prior distributions must be chosen for quantities with no parents in the graph, just as in standard probabilistic decision modelling [2]. These may be vague or based on substantive information, but must represent our beliefs prior to observing the data included in the graph.

**3.3.1 Vague priors**—If there are sufficient data already in the model to give a precise estimate of a parameter, then a vague prior *within plausible ranges* is reasonable. With sufficient data, the exact choice of prior will not be influential. Though sensitivity analysis is advisable if the choice is uncertain or suspected to be influential (§4).

**Example.** The prevalences of cytological states  $q_{jk}^-$  among women not diagnosed with HPV16 are given vague priors (a uniform Dirichlet distribution, as recommended in Briggs et al. [26]). Since we have strong information about them from the corresponding counts in ARTISTIC, this prior will have little influence on the results.

Parameters should be transformed to a natural scale before being given a prior, to enable beliefs to be expressed intuitively.

**Example.** Since the parameters  $A$  and  $C$  in the prevalence extrapolation model (Figure 4) are difficult to interpret, we use vague priors, within plausible ranges, for transformations of those parameters to intuitive scales: a uniform(0,1) prior for the maximum HPV16 prevalence  $p_{max}$ , a uniform(0,30) prior for the age at this maximum  $t_{max}$ , and also a uniform(10,14) prior for the minimum age of infection  $t_0$ . Assuming the prevalence for the oldest women  $B = 0$  for the purpose of deriving these priors, these give  $C = n/(t_{max} - t_0)$  and  $A = p_{max} \exp(n/(n/C)^{-n})$ . To replace the discrete scenarios used in [11] with a continuous distribution, we place a uniform(1,2) prior on the polynomial order  $n$ . These priors are all very vague compared with the information in the data — see Figure 6 for the corresponding posteriors.

**3.3.2 Informative priors**—If there are no direct data to inform a parameter, informative priors could be derived from published literature or from expert beliefs, ideally formally elicited [27]. When the priors are updated to posteriors, we can assess how much the data confirm or modify our substantive beliefs.

**Example.** We enhance the original analysis of Jit et al. [11] by incorporating prior knowledge about the transition probabilities between grades of neoplasia in the presence of HPV16 ( $P_{ij}; j \geq 6$ ). These are derived from a systematic review of HPV natural history by Insinga et al. [28], which did not include Jit et al. [11]. This presents annual probabilities,

with the number of patients and denominators  $N$  used to obtain them, for four out of the eight transitions. These were converted to monthly probabilities and corresponding monthly counts  $y$  by assuming a constant transition rate within the year. This gives  $\text{Beta}(y + 1, N - y + 1)$  prior distributions for the monthly probabilities [2, 29], illustrated in Figure 5. The other four transition probabilities were given  $\text{uniform}(0,1)$  priors. Any correlations between different parameters should be acknowledged in decision models [30]; we assume prior independence between the probabilities, but if any correlations are plausible, given the data, these will appear in the posterior.

**Example.** For the sensitivities and specificities of cytological screening and the false positive rate of the HPV16 test,  $p^{FP}$ , we used uniform or Beta priors whose bounds or quantiles were chosen to cover the discrete scenarios previously presented in the analysis of Jit et al. [11]. The priors were independent given the lack of published information about any correlations.

### 3.4 Computing the posterior distribution

Once all quantities in the graph have been defined or given priors, the joint posterior distribution of all unknowns will be calculated, which may include quantities of direct interest to a decision maker, such as the incremental net benefit of an intervention. The decision then allows for the parameter uncertainty in all model inputs, and includes the evidence from the calibration data. In other words, probabilistic sensitivity analysis and model calibration are performed simultaneously.

The graph shows how the underlying parameters of interest — in Figure 3, the transition probabilities  $P_{ij}$  between disease states  $i, j$  — can be informed simultaneously from several data sources. Learning about the posterior of a parameter can occur “both ways” along arrows. For example (§3.2) the posterior distribution of the HPV16-positive cytological state prevalences  $q_{jk}^+$  is derived directly from the corresponding ARTISTIC data  $y_{jk}^+$ , but is also influenced by its three parent parameters and their own ancestors and descendants, and by its other children, the HPV16-unknown cytological state prevalences  $q_{jk}$ , which are inferred from the national screening programme. The posterior distribution of the transition probabilities thus depends directly on its prior, but also indirectly on all data and on the priors of all other parameters.

**3.4.1 Markov Chain Monte Carlo**—Bayes’ theorem states that the joint posterior distribution (probability density)  $p(\theta/x)$  of the set of unknowns in the model  $\theta$  given all data  $x$  is proportional to the (joint) prior  $p(\theta)$  multiplied by the sampling distribution of data given parameters  $p(x/\theta)$  (often called the *likelihood*). However the constant of proportionality, and summaries of the resulting posterior, are generally too complex to be calculated directly. Instead, the graphical model structure gives a basis for an iterative *Markov Chain Monte Carlo (MCMC)* algorithm to generate samples from the posterior distribution. The distribution can be expressed as the product  $\prod_{v \in \theta} p(v|pa[v])$  of all distributions of individual nodes  $v$ , each conditional on its parents  $pa[v]$  [31].

- Initial values are chosen for all  $v$ .

- New values are then sampled from the *full-conditional* distributions  $p(v|\cdot)$  of each node  $v$  in turn, where  $\cdot$  indicates the current values of all nodes other than  $v$ . Each full-conditional distribution can be simplified as the product of the prior distribution and the distributions of all children of  $v$ :

$$p(v|\cdot) = p(v|pa[v]) \prod_{w \in pa[w]} p(w|pa[w]).$$

Through the distributions of the children (the likelihood) any data directly or indirectly generated by  $v$  will contribute to the posterior of  $v$ .

- Each iteration consists of one sample from the full set of nodes  $\theta$ . After an initial sequence of “burn-in” iterations, the algorithm will converge (under weak conditions, see Gilks et al. [31]) such that the samples will eventually be drawn from the true joint posterior distribution.

**3.4.2 The BUGS language and software**—The BUGS language [32] allows graphical models to be expressed in a text format. In brief, each node’s definition, as a deterministic or random function of its parent parameters, corresponds to a language statement. For example, the definitions of  $y_k$  and  $p^D(t_k)$  in §3.1 are written as

```
for (k in 1:Nage){
  y[k] ~ dbin(pD[k], n[k])
  pD[k] <- pH[k] + pFP*(1 - pH[k])
}
```

The software then constructs the graphical model internally, and chooses and implements appropriate methods to draw random numbers from the full-conditional distribution of each node. An extensive guide to Bayesian modelling using the BUGS language and software is given by Lunn et al. [29], and Welton et al. [16] give a practical guide to its use in health decision modelling. The full BUGS model code representing the definitions in this example is provided in an online supplement. For this application, we use the JAGS software for BUGS language interpretation and computation [33]. Implementation in WinBUGS or OpenBUGS [32] would have been equally feasible.

## 4 Model checking and sensitivity analysis

Although a Bayesian graphical model is a natural framework for evidence synthesis, it can still involve many assumptions which should be questioned.

- A model can be assessed by checking and comparing the fit of model predictions to the data used to build it [34], then improving the model if necessary (“internal” [35] or “dependent” [36] validation).
- Where more than one dataset informs a quantity of interest, any potential inconsistency or conflict must be investigated [16, 20].



- We recommend that any relevant data are included in the graphical model, rather than held back for external validation. Judgement is needed here, since modifying the model to accommodate increasingly less relevant evidence will make it increasingly cumbersome and prone to misspecification. If it is uncertain whether some data are relevant, perhaps due to differences in population characteristics or clinical practice, sensitivity analysis should be undertaken.
- When the evidence informing a particular part of the model is weak, so that different reasonable choices of prior distribution may affect the results, these should be compared in sensitivity analysis. Sensitivity analyses may also show the relative influence of the prior and the data on the conclusions.
- Relative fit between alternative Bayesian models can be compared with the posterior mean deviance [as in 20] or the deviance information criterion (DIC), which is an estimate of the ability to predict a replicate dataset [37, 38].

Note that the Bayesian approach does not eliminate the need for discrete sensitivity analyses for the parts of the model that are poorly informed by data, but the number of scenarios can be greatly reduced if these are replaced wherever possible by smooth priors, as in the HPV example.

In the next sections, we explain how these checks were carried out in the example, and led to refinements of the model. In Figure 5, the prior and posterior distributions of the five state progression and regression probabilities are illustrated under the model as described, and four other assumptions explained below. This also shows the estimates from the scenarios considered in Jit et al. [11] — note that there are some extreme scenarios which are implausible, thus have negligible weight in the posterior. The only substantial posterior correlation ( $\rho > 0.5$ ) was between the progression/regression probabilities to/from CIN3. The only substantial inverse correlation ( $\rho < -0.5$ ) was between the CIN3→pre-cancer and pre-cancer→diagnosed cancer transition probabilities, due to there being few observations in the pre-cancer state.

#### 4.1 Model checking and accommodating conflicts

Firstly in the HPV example, we graphically compare the posterior distribution for the cytological state prevalences  $q_{jk}^+$ ,  $q_{jk}^-$  to the corresponding observed prevalences from the ARTISTIC data. The posterior from the model as described so far (labelled “conflicted model”) is estimated not only from ARTISTIC, but also from the national screening programme through Equation (1). Figure 7 illustrates that this model does not fit the ARTISTIC data well, as the observed prevalences are only just within the 95% posterior credible limits, both for women with and without HPV16 diagnoses. This suggests that the ARTISTIC and national data are providing conflicting information to this part of the model, such that the two datasets do not agree on the proportions of women in the same age group being in various states.

We accommodate this conflict by assuming that the odds of being in states CIN1 or higher for a HPV16-negative woman in the ARTISTIC data is a constant multiplier of the corresponding odds in the national data. This constant is estimated as part of the model. This produces a much better fit in Figure 7 (model labelled “base case”). This supposes that women selected for ARTISTIC have slightly different prevalences of non-HPV16-related neoplasia from the general population. The prevalences of HPV16-related disease might also be different, but we assume these are the same given that the adjustment we have made produces an adequate fit. All subsequent results we present are from this model unless stated otherwise. Both the progression probability to CIN1 and the corresponding regression rate are changed after resolving this conflict (Figure 5). Other datasets are fitted reasonably well by the posterior distributions (not shown).

Note that the posterior distributions for  $q_{jk}^+$  (bottom half of Figure 7) are smooth functions of age, because the prior information about these prevalences, which comes indirectly from all other data sources in the model via the Markov transition probabilities, “shrinks” each individual data point towards the prior mean.

#### 4.2 Sensitivity to data inclusion

There is also doubt about whether the national screening data are worth including at all, given that state prevalences by HPV16 presence are observable from ARTISTIC alone, and HPV16 presence is not recorded in the national data. While including it may improve the precision of inferences from ARTISTIC alone, there is the risk of other implicit, conflicting information “feeding back” through the graph and affecting other parameter estimates, for example the HPV16 prevalences. We therefore compare the results under a third model (labelled “No NHSCCSP data”), where these data and the HPV16-negative data from ARTISTIC (which are only required to understand the contribution of non-HPV16 related dysplasia to the prevalences of dysplasia in the national data) are excluded. The estimated transition probabilities do not substantially change, see Figure 5. However, the posterior distribution of the maximum HPV16 prevalence (Figure 4) is shifted upwards by a couple of percentage points, accounting better for the observations from around age 20. This suggests that the national data are feeding back to this part of the model.

Since we would expect the national screening programme to better represent the distribution of states than ARTISTIC which is a smaller, regionally-biased sample of women consenting to participate, our base case results are those which do include the national data. In general, if the model is used to evaluate a policy for the population, any relevant full population data should be included, particularly since the highly selected populations typically included in randomised trials may not represent the patients of interest.

#### 4.3 Sensitivity to prior assumptions

As well as the two alternative models which use the national screening data differently, we performed a further two sensitivity analyses. The first investigates how much the posterior distributions of the transition probabilities were affected by the strong priors, compared to the data. We downweighted the contribution of the prior distributions by making them substantially weaker, arbitrarily multiplying their variances by 10. Note that completely

vague Uniform(0,1) priors on all probabilities were not practicable, since they resulted in MCMC convergence failure, but more importantly they would not have really represented our prior beliefs in this case — for instance values over 0.5 for the monthly progression probability would be deemed highly unrealistic beforehand. Under the weaker priors, the posterior distributions of the CIN1 →CIN2 and CIN2 →CIN1 probabilities became more widely dispersed. The posterior means of the other parameters were unchanged, and the posterior variances were only slightly inflated, suggesting that the stronger prior merely adds precision to the study data rather than shifting estimates away from the data. The exception is the regression rate from CIN1 to (CIN-free) HPV16, where the study data appear to support lower values of this probability than the prior. The base case model compromises between prior and data through Bayes' theorem.

The other key uncertain parameters describe the accuracy of the cytological screening tests. The information about these came from a review of a very heterogeneous range of studies. Therefore we performed another analysis where all sensitivities and specificities are fixed at 100%, one of the extreme scenarios considered in Jit et al. [11]. The posterior distributions for many of the transition probabilities (Figure 5) changed, confirming that they are sensitive to the screening accuracies. In particular, the posterior variances are smaller when these probabilities are fixed. The fit of this model is also much poorer than the base case, judging from DIC and plots (not shown). Information about the screening accuracies also comes, very indirectly, from the data (through Figure 3). This results in posteriors (Figure 6) which are reasonably precise compared to the diffuse priors, and indicate the strength of evidence in the data for each value within the prior bounds. There is a moderate posterior correlation ( $\rho = 0.43$ ) between the sensitivity and specificity of detecting CIN1, but no other notable correlation. The specificity of the test to identify HPV16 DNA had been given a fairly strong prior (median 0.9975%), which is slightly modified by the data (posterior median 0.9984%).

## 5 Discussion

Bayesian graphical modelling is a useful framework for including all policy-relevant evidence in a decision model, even evidence which is only indirectly related to the main parameters. We have outlined the key steps of this approach, and demonstrated how we used it to obtain posterior distributions for progression and regression probabilities of HPV16-related cervical neoplasia. The posterior gives a more accurate reflection of the available evidence than the scenario analyses used in previous HPV models, by also including the evidence about how plausible each scenario was, and reducing bias due to ignoring some parameter uncertainties. A single posterior distribution is also easier to interpret.

These methods are complex, but necessarily so in the HPV example, and only require a similar amount of programming to the original implementation. The principles of Bayesian statistical modelling are widely applicable in health policy evaluation [17, 16]. The BUGS language and software can provide the posterior distribution of any Bayesian model in principle, though models with very large numbers of unknowns, such as this one, require custom extensions to the software to be written in more low-level programming languages. Whyte et al. [21] also describe a similar implementation of a Bayesian health economic model using Visual Basic within Excel.

The HPV model could be extended as in Jit et al. [10, 23] to incorporate an infection and economic model to evaluate policies for HPV vaccination. This was originally based on over a thousand alternative scenarios which met a certain standard of fit, but with no measure of relative plausibility between them. A Bayesian reimplementation to formally characterise this uncertainty would be expected to have a similar computational burden to the original approach, as was the case for the progression-only component. However this is likely to raise more uncertainties due to weak evidence, for example, about transmission, infection, and natural clearance, and potential conflicts with other data sources in the model. It is therefore unclear how much benefit would be gained from greater statistical formality in this example. As a graphical model becomes more complex, there is greater potential for erroneous information from one part of the model to indirectly give bias in another part. There is ongoing research on controlling the propagation of information in graphs through restrictions variously termed “cutting feedback” [39] or “modularisation” [40].

No statistical method can eliminate uncertainty in a model, since evidence is always limited. A decision model should synthesise all relevant evidence, but judgements must always be made about what evidence is sufficiently relevant or strong. Potential conflicts between two sources of data on the same quantity should be investigated and explained, leading to refinements in the model [20]. In general, for more complex models, careful validation and sensitivity analysis become more important (§4). This was demonstrated in the HPV example, where, after examining plots of model fit, some model assumptions were relaxed to accommodate the population screening data.

Weak evidence about some component of a Bayesian model will result in sensitivity of the results to the prior distribution. Conversely, if prior sensitivity is detected, this indicates areas where stronger data or further research are required. In the HPV model, for example, there was substantial uncertainty around the accuracy of cervical screening. While the prior that we used was based on the best available information, sensitivity analysis showed that these parameters were influential. Although a posterior distribution gives a better representation of the data than a range of scenarios, a limited number of sensitivity analyses are still useful to show how the results would be affected if the evidence were to change. If this model had been employed for decision-making, formal “value of information” methods [see, e.g. 41] could be used to predict for which parameters more research would give the greatest expected benefits.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

This work was supported by the Medical Research Council [Unit Programme numbers U015232027, U105260566]. The funding source had no role in the study which was the authors’ independent work.

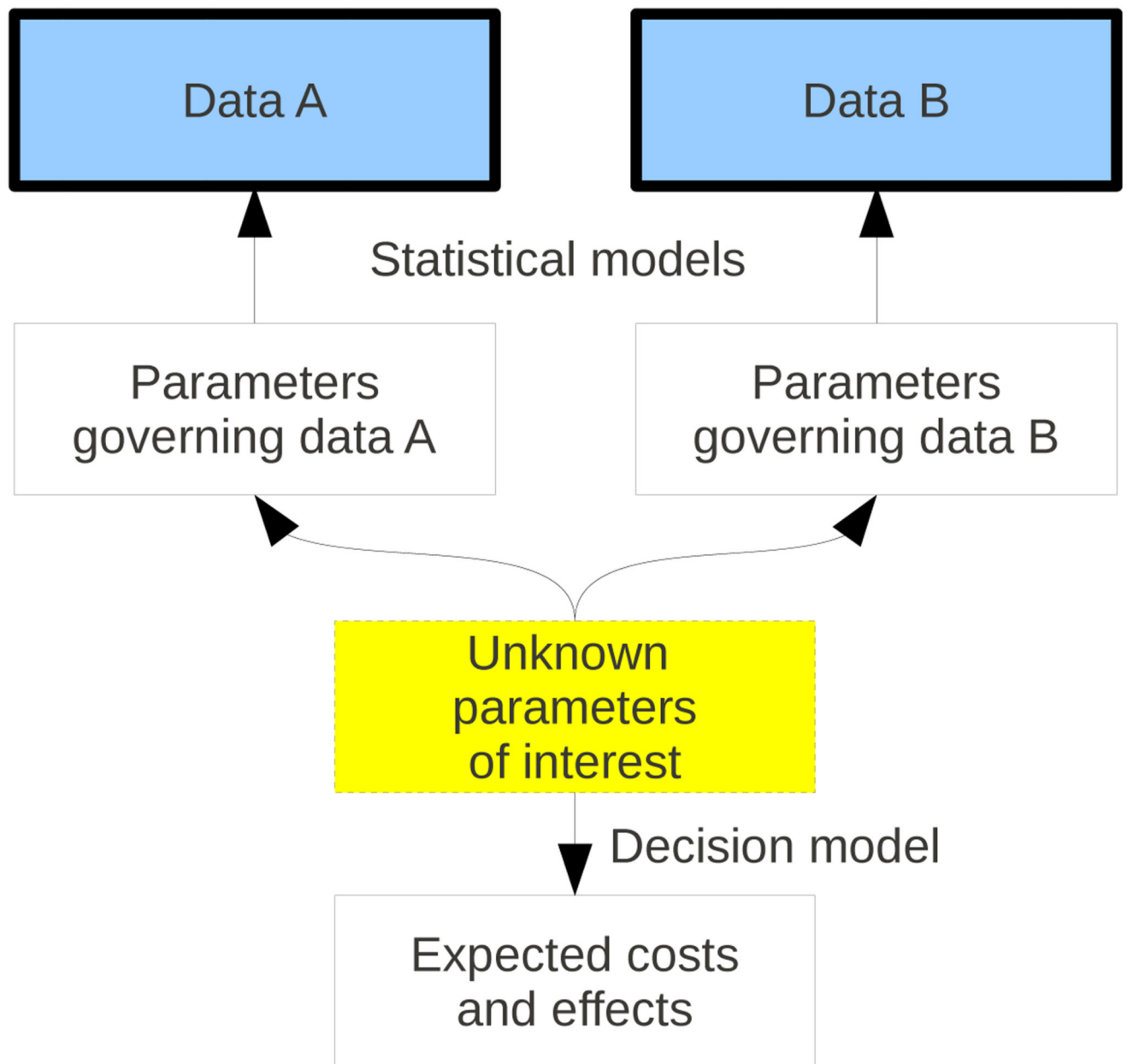
## References

1. Claxton K, Sculpher M, Drummond M. A rational framework for decision-making by the National Institute for Clinical Excellence (NICE). *Lancet*. 2002; 360(9334):711–715. [PubMed: 12241891]

2. Briggs, A.; Sculpher, M.; Claxton, K. Decision modelling for health economic evaluation. *Handbooks in Health Economic Evaluation*. Oxford University Press; Oxford: 2006.
3. Jackson CH, Bojke L, Thompson SG, Claxton K, Sharples LD. A framework for addressing structural uncertainty in decision models. *Medical Decision Making*. 2011; 31(4):662–674. [PubMed: 21602487]
4. Kulasingam SL, Benard S, Barnabas RV, Largeton N, Myers ER. Adding a quadrivalent human papillomavirus vaccine to the UK cervical cancer screening programme: A cost-effectiveness analysis. *Cost Effectiveness and Resource Allocation*. 2008; 6(4)
5. Barnabas RV, Laukkanen P, Koskela P, Kontula O, Lehtinen M, Garnett GP. Epidemiology of HPV 16 and cervical cancer in Finland and the potential impact of vaccination: mathematical modelling analyses. *PLoS Medicine*. 2006; 3(5):e138. [PubMed: 16573364]
6. Kohli M, Ferko N, Martin A, Franco EL, Jenkins D, Gallivan S, Sherlaw-Johnson C, Drummond M. Estimating the long-term impact of a prophylactic human papillomavirus 16/18 vaccine on the burden of cervical cancer in the UK. *British Journal of Cancer*. 2006; 96(1):143–150. [PubMed: 17146475]
7. Salomon JA, Weinstein MC, Hammitt JK, Goldie SJ. Empirically calibrated model of hepatitis C virus infection in the United States. *American Journal of Epidemiology*. 2002; 156(8):761–773. [PubMed: 12370165]
8. Goldhaber-Fiebert JD, Stout NK, Salomon JA, Kuntz KM, Goldie SJ. Cost-effectiveness of cervical cancer screening with human papillomavirus DNA testing and HPV-16, 18 vaccination. *Journal of the National Cancer Institute*. 2008; 100(5):308–320. [PubMed: 18314477]
9. Kim JJ, Kuntz KM, Stout NK, Mahmud S, Villa LL, Franco EL, Goldie SJ. Multiparameter calibration of a natural history model of cervical cancer. *American Journal of Epidemiology*. 2007; 166(2):137–150. [PubMed: 17526866]
10. Jit M, Choi Y-H, Edmunds WJ. Economic evaluation of human papillomavirus vaccination in the United Kingdom. *British Medical Journal*. 2008; 337(a769)
11. Jit M, Gay N, Soldan K, Choi YH, Edmunds WJ. Estimating progression rates for human papillomavirus infection from epidemiological data. *Medical Decision Making*. 2010; 30(1):84–98. [PubMed: 19525483]
12. Vanni T, Karnon J, Madan J, White RG, Edmunds WJ, Foss AM, Legood R. Calibrating models in economic evaluation: a seven-step approach. *Pharmacoeconomics*. 2011; 29(1):35–49. [PubMed: 21142277]
13. Taylor DCA, Pawar V, Kruzikas DT, Gilmore KE, Sanon M, Weinstein MC. Incorporating calibrated model parameters into sensitivity analyses: Deterministic and probabilistic approaches. *Pharmacoeconomics*. 2012; 30(2):119–126. [PubMed: 22149631]
14. Karnon J, Vanni T. Calibrating models in economic evaluation: a comparison of alternative measures of goodness of fit, parameter search strategies and convergence criteria. *Pharmacoeconomics*. 2011; 29(1):51–62. [PubMed: 21142278]
15. Ades AE, Sutton AJ. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2006; 169(1):5–35.
16. Welton, NJ.; Sutton, AJ.; Cooper, NJ.; Abrams, KR.; Ades, AE. *Evidence Synthesis for Decision Making in Healthcare*. Wiley; Chichester: 2012.
17. Spiegelhalter, DJ.; Abrams, KR.; Myles, JP. *Bayesian approaches to clinical trials and health-care evaluation*. Wiley; Chichester, U.K: 2004.
18. Cooper NJ, Sutton AJ, Abrams KR, Turner D, Wailoo A. Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach. *Health Economics*. 2004; 13(3):203–226. [PubMed: 14981647]
19. Goubar A, Ades AE, DeAngelis D, McGarrigle CA, Mercer CH, Tookey PA, Fenton K, Gill ON. Estimates of human immunodeficiency virus prevalence and proportion diagnosed based on Bayesian multiparameter synthesis of surveillance data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2008; 171(3):541–580.
20. Presanis AM, De Angelis D, Spiegelhalter DJ, Seaman S, Goubar A, Ades AE. Conflicting evidence in a Bayesian synthesis of surveillance data to estimate human immunodeficiency virus

- prevalence. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2008; 171(4): 915–937.
21. Whyte S, Walsh C, Chilcott J. Bayesian calibration of a natural history model with application to a population model for colorectal cancer. *Medical Decision Making*. 2011; 31(4):625–641. [PubMed: 21127321]
  22. Bogaards JA, Xiridou M, Coupé VMH, Meijer CJLM, Wallinga J, Berkhof J. Model-based estimation of viral transmissibility and infection-induced resistance from the age-dependent prevalence of infection for 14 high-risk types of human papillomavirus. *American Journal of Epidemiology*. 2010; 171(7):817. [PubMed: 20231211]
  23. Jit M, Chapman R, Hughes O, Choi Y-H. Comparing bivalent and quadrivalent human papillomavirus vaccines: economic evaluation based on transmission model. *British Medical Journal*. 2011; 343:d5775. [PubMed: 21951758]
  24. Kitchener HC, Almonte M, Wheeler P, Desai M, Gilham C, Bailey A, Sargent A, Peto J. HPV testing in routine cervical screening: cross sectional data from the ARTISTIC trial. *British Journal of Cancer*. 2006; 95(1):56–61. [PubMed: 16773068]
  25. Department of Health. England: 2005–2006. Cervical Screening Programme, England, 2005–2006. Available from <http://www.ic.nhs.uk/pubs/csp0506>
  26. Briggs AH, Ades AE, Price MJ. Probabilistic sensitivity analysis for decision trees with multiple branches: use of the Dirichlet distribution in a Bayesian framework. *Medical Decision Making*. 2003; 23(4):341–350. [PubMed: 12926584]
  27. O'Hagan, A.; Buck, C.; Daneshkhah, A.; Eiser, JR.; Garthwaite, P.; Jenkinson, D.; Oakley, J.; Rakow, T. *Uncertain Judgements: Eliciting Experts' Probabilities*. Statistics in Practice. Wiley; 2006.
  28. Insinga RP, Dasbach EJ, Elbasha EH. Epidemiologic natural history and clinical management of Human Papillomavirus(HPV) Disease: a critical and systematic review of the literature in the development of an HPV dynamic transmission model. *BMC Infectious Diseases*. 2009; 9(1):119. [PubMed: 19640281]
  29. Lunn, D.; Jackson, C.; Best, N.; Thomas, A.; Spiegelhalter, D. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. CRC Press / Chapman and Hall; 2012.
  30. Ades AE, Claxton K, Sculpher M. Evidence synthesis, parameter correlation and probabilistic sensitivity analysis. *Health Economics*. 2006; 15(4):373–381. [PubMed: 16389628]
  31. Gilks, WR.; Richardson, S.; Spiegelhalter, DJ. *Markov Chain Monte Carlo in Practice*. Chapman and Hall; London: 1996.
  32. Lunn DJ, Thomas A, Best NG, Spiegelhalter DJ. Win-BUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*. Oct; 2000 10(4):325–337.
  33. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Hornik, Kurt; Leisch, Friedrich; Zeileis, Achim, editors. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*; Vienna, Austria. 2003. <http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/> ISSN 1609-395X
  34. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. second edition. Chapman and Hall; London: 2003.
  35. Kim LG, Thompson SG. Uncertainty and validation of health economic decision models. *Health Economics*. 2009; 19(1):43–55. [PubMed: 19206080]
  36. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group, part 4. *Medical Decision Making*. 2012; 32:733–743. [PubMed: 22990088]
  37. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*. 2002; 64(4):583–639.
  38. Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics*. 2008; 9(3): 523–539. [PubMed: 18209015]
  39. Lunn D, Best N, Spiegelhalter D, Graham G, Neuenschwander B. Combining MCMC with sequential PK/PD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*. 2009; 36(1): 19–38. [PubMed: 19132515]

40. Liu F, Bayarri M, Berger J. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*. 2009; 4(1):119–150.
41. Welton NJ, Ades AE, Caldwell DM, Peters TJ. Research prioritization based on expected value of partial perfect information: a case study on interventions to increase uptake of breast cancer screening. *Journal of the Royal Statistical Society, Series A*. 2008; 171(4):807–841.
42. Munoz N, Bosch FX, Castellsague X, Díaz M, de Sanjose S, Hammouda D, Shah KV, Meijer CJLM. Against which human papillomavirus types shall we vaccinate and screen? The international perspective. *International Journal of Cancer*. 2004; 111(2):278–285. [PubMed: 15197783]
43. Redburn JC, Murphy MFG. Hysterectomy prevalence and adjusted cervical and uterine cancer rates in England and Wales. *BJOG: An International Journal of Obstetrics & Gynaecology*. 2001; 108(4):388–395. [PubMed: 11305546]
44. Nanda K, McCrory DC, Myers ER, Bastian LA, Hasselblad V, Hickey JD, Matchar DB. Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities. *Annals of Internal Medicine*. 2000; 132(10):810–819. [PubMed: 10819705]
45. Arbyn M, Bergeron C, Klinkhamer P, Martin-Hirsch P, Siebers AG, Bulten J. Liquid compared with conventional cervical cytology: a systematic review and meta-analysis. *Obstetrics & Gynecology*. 2008; 111(1):167. [PubMed: 18165406]

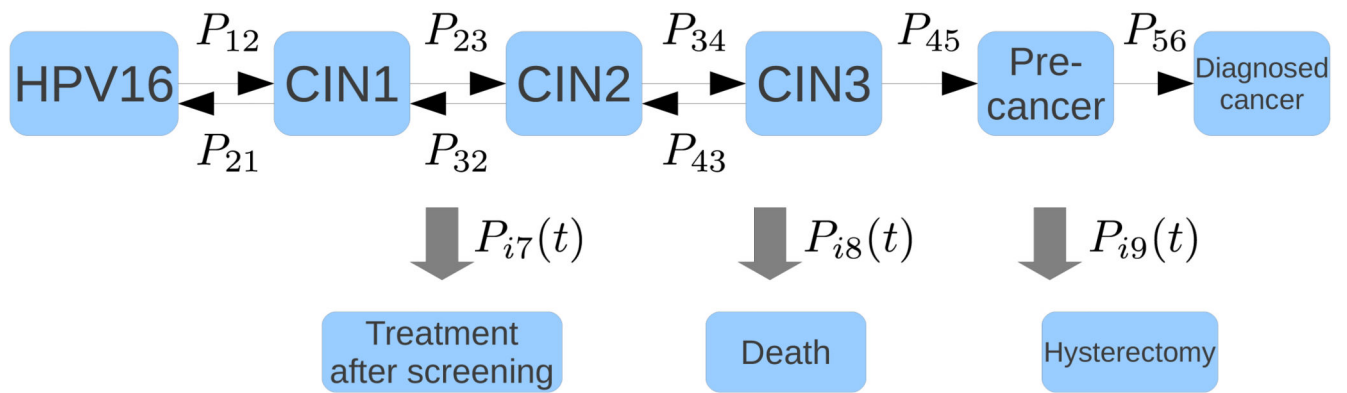


**Figure 1:**

Illustration of decision model calibration as a directed acyclic graph. “Nodes” at the start of an arrow (*parents*) are assumed to “generate” the nodes at the end of the arrow (*children*).

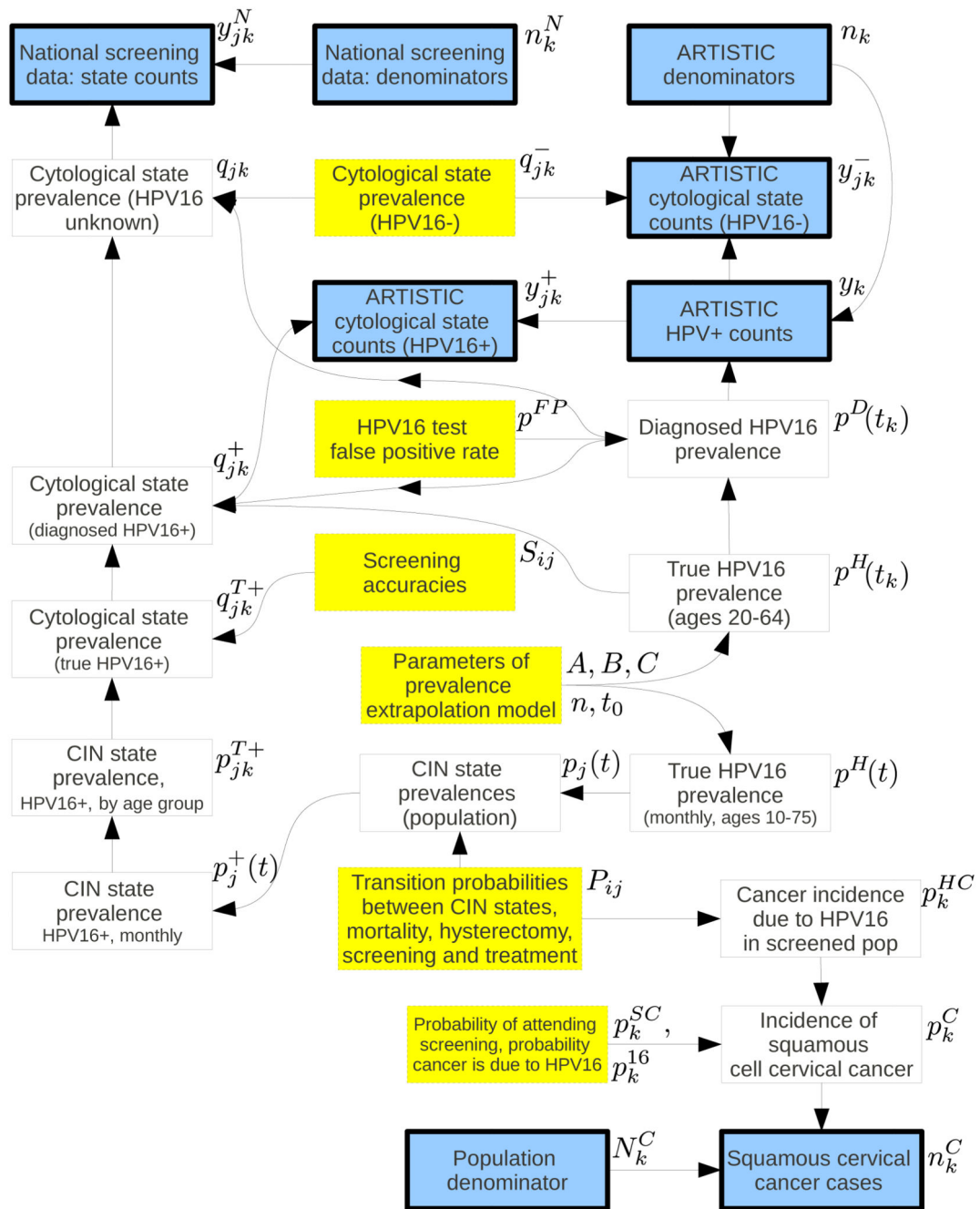
Bold-bordered boxes represent observed data. The white boxes are quantities which are defined as explicit functions of other parameters. The dashed boxes represent quantities which cannot be defined this way, thus are given prior distributions (see Section 3.1).



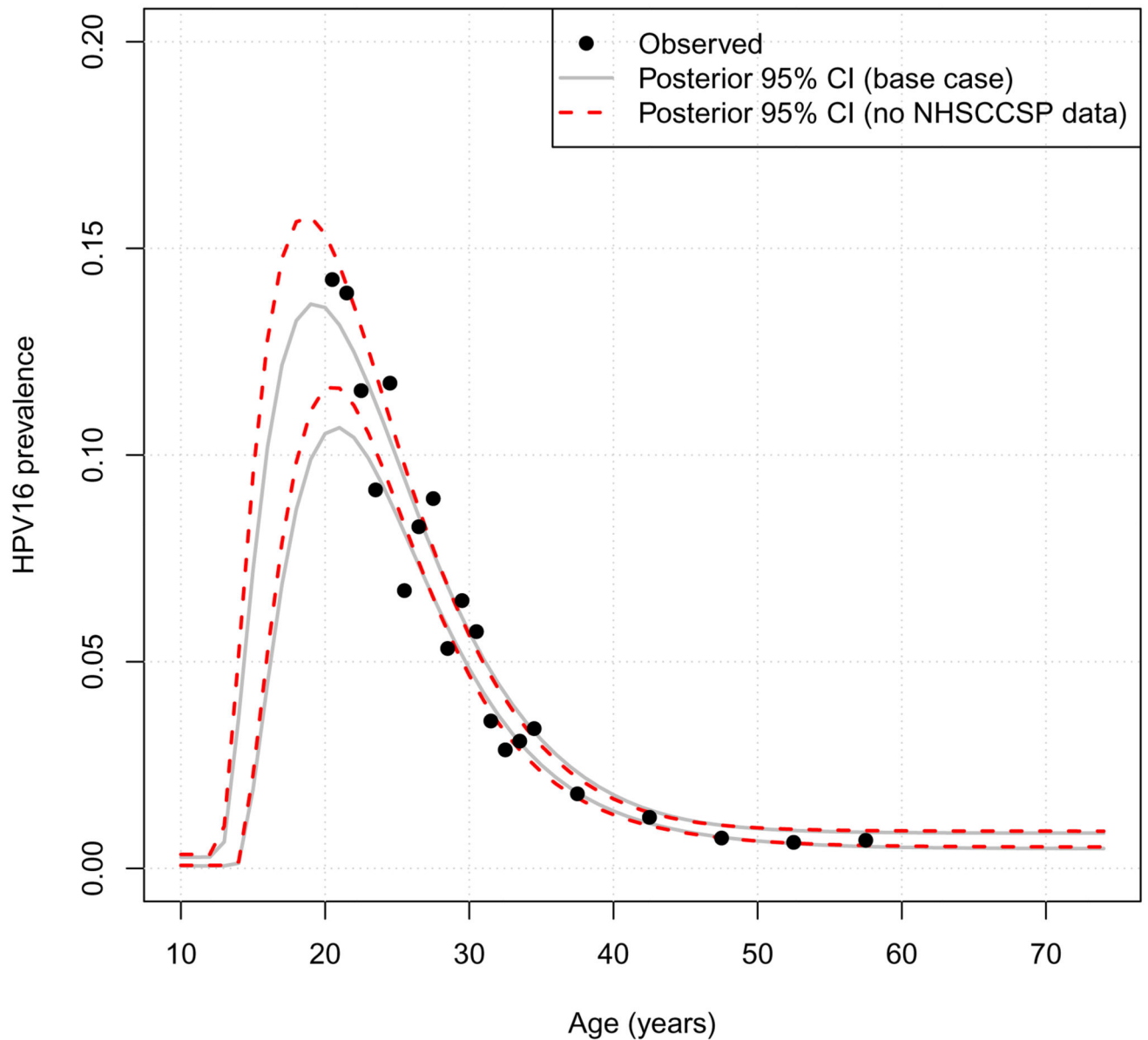


**Figure 2:**

Markov model for natural history of HPV16 infection – states of cervical neoplasia and permitted monthly transitions with associated probabilities. Progression to treatment, death and hysterectomy is allowed from all states up to pre-cancer.

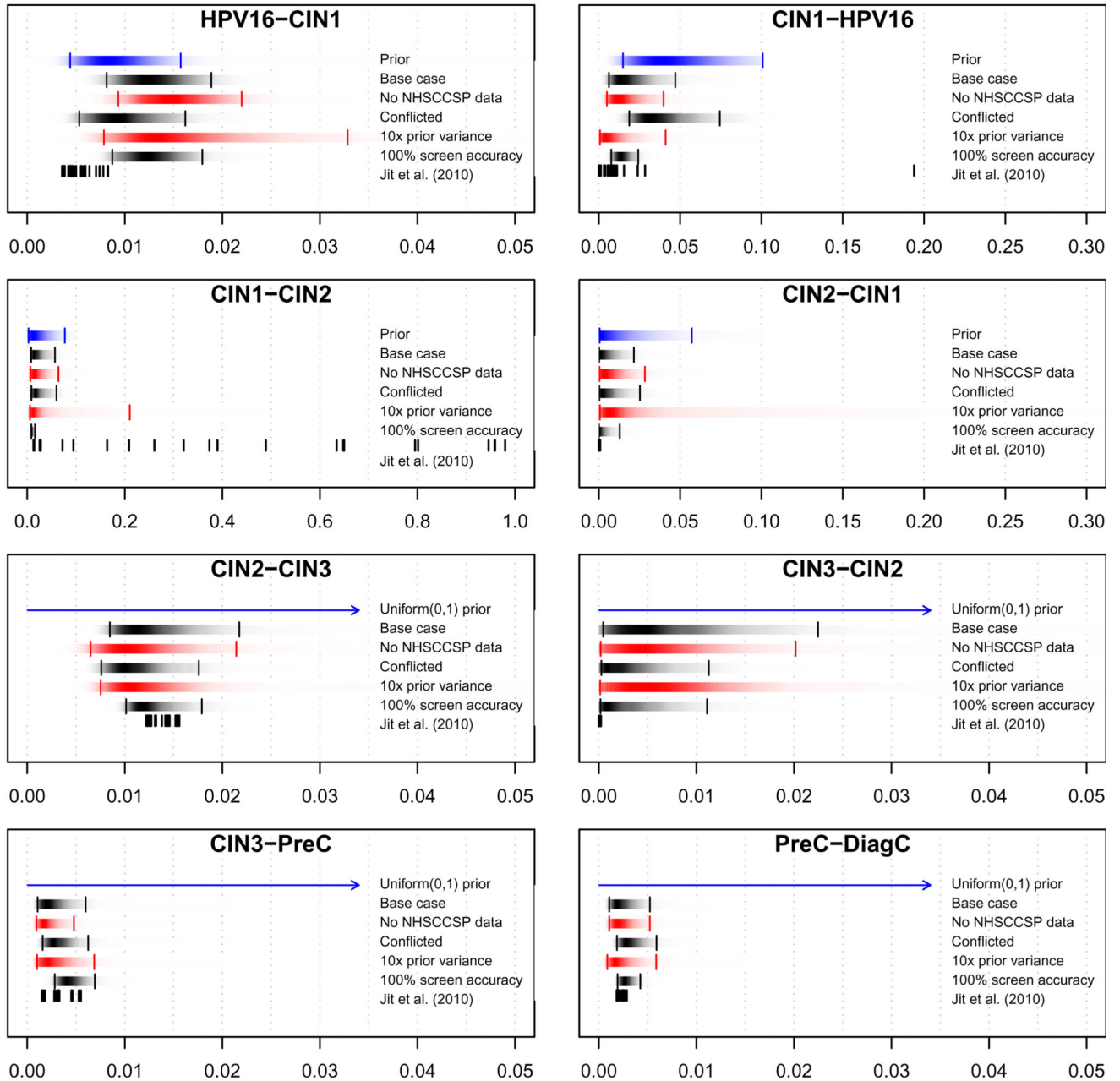


**Figure 3:** Directed graphical model for evidence synthesis to estimate transition probabilities between states of HPV16 infection. Notation as in Figure 1.

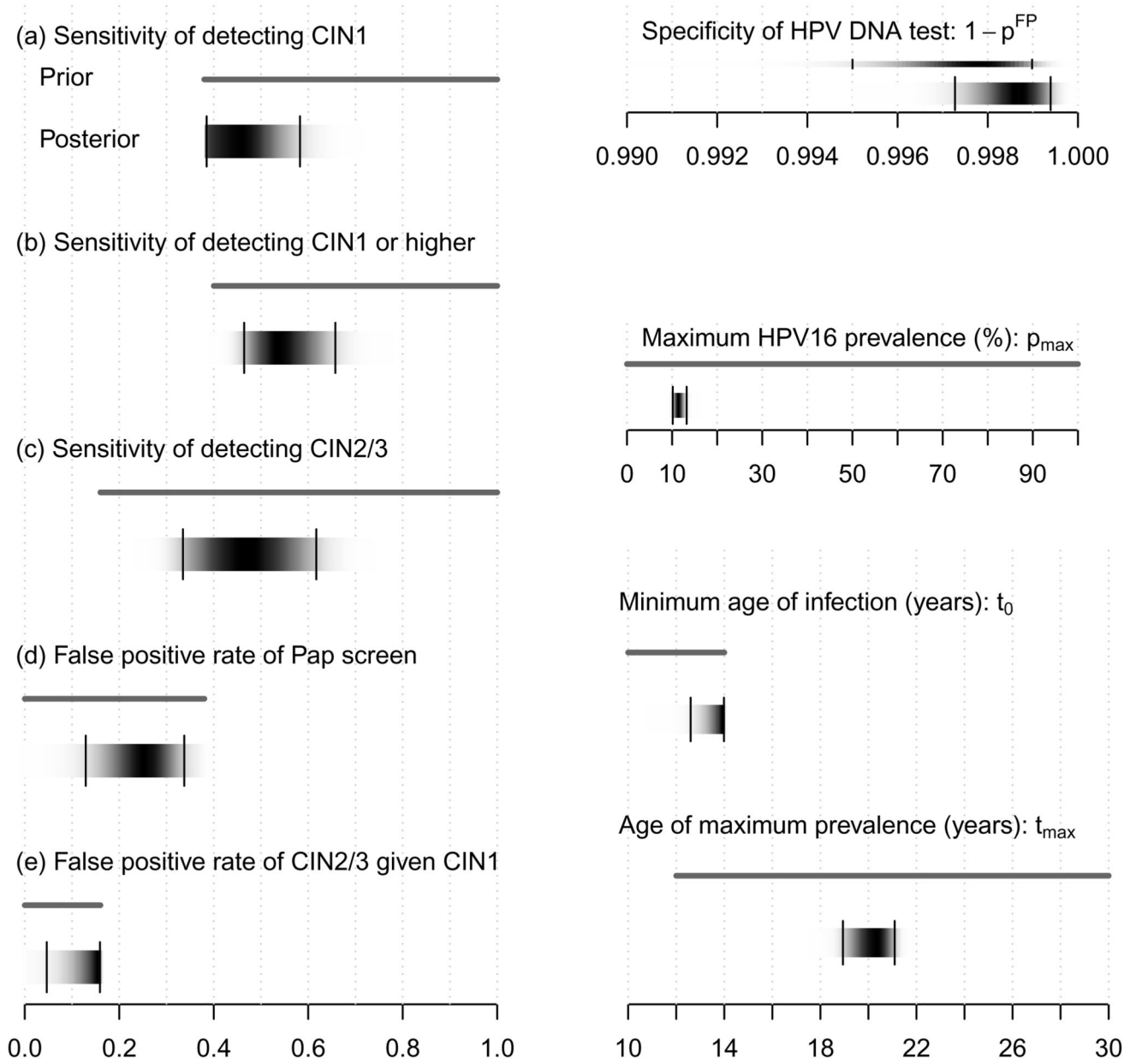


**Figure 4:**

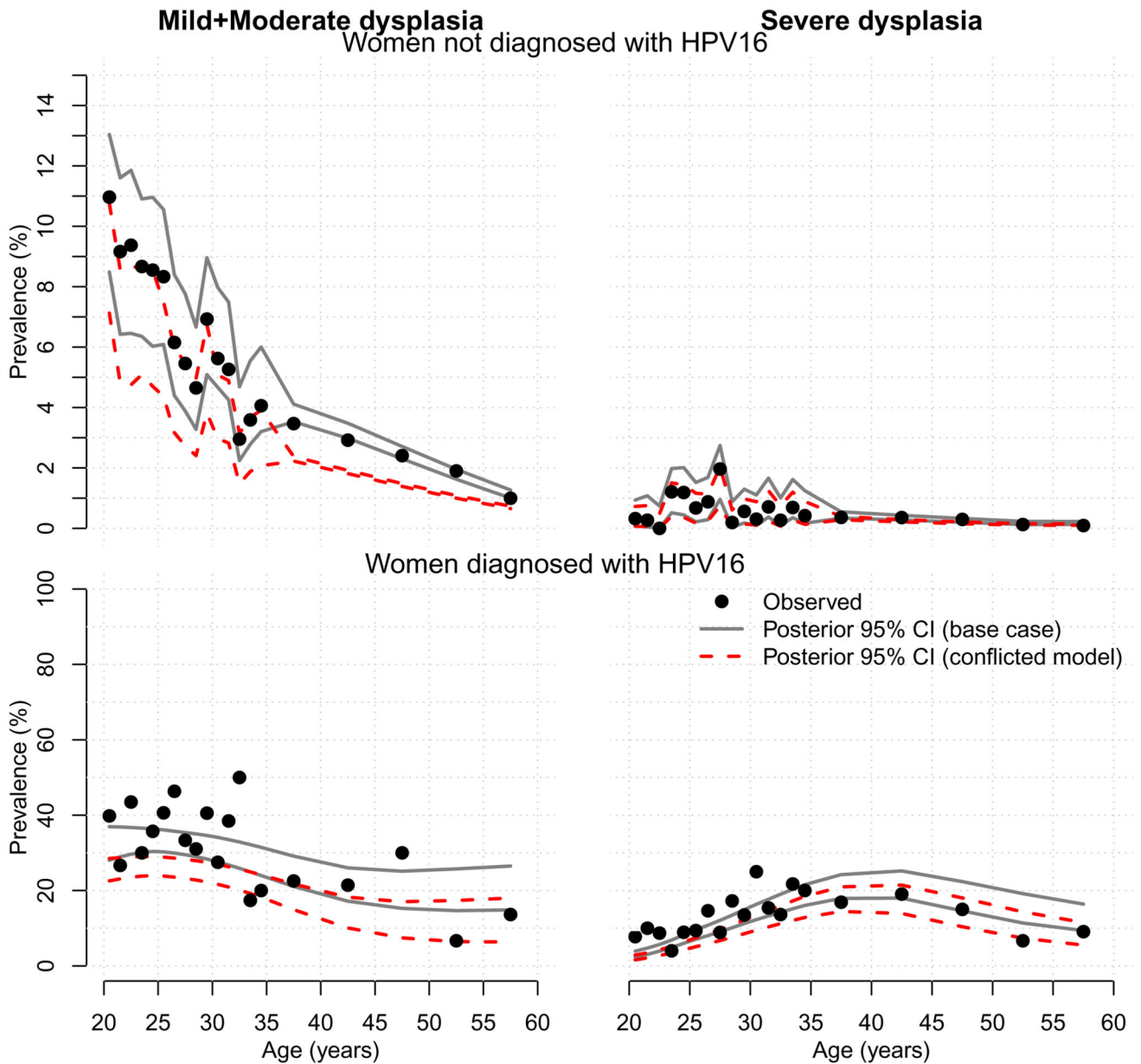
Prevalences of diagnosed HPV16 infection among women in ARTISTIC (2006) data and extrapolation to younger women — observed and model posterior 95% credible intervals. Since the specificity is over 99%, this also illustrates the approximate trajectory of the true prevalence,  $p^H$ , related to age  $t$  as  $p^H(t) = B + \exp(-C(t - t_0))(A(t - t_0)^n - B)$  (see Section 3.3).



**Figure 5:** Prior and posterior distributions of monthly transition probabilities between states of HPV16 infection (note some axis limits are extended), compared with scenarios from Jit et al. (2010). The darkness of each strip is proportional to the posterior density, fading to white at zero density, with 95% credible limits.



**Figure 6:** Prior distributions (thin lines) and posterior distributions (thicker strips with darkness proportional to probability density, and 95% credible limits) for cytological screening accuracies, HPV16 test specificity and parameters of the model for extrapolating HPV16 prevalence.



**Figure 7:** Prevalences of cervical dysplasia among women in ARTISTIC (2006) data by HPV16 diagnosis — observed data and posterior distributions (median and 95% credible interval) under base case and conflicted models.

**Table 1:**

HPV model parameters and associated data sources. See Figure 3 for how these are connected in a graphical model.

Parameter Informed by direct data	Data source	Age groups $k$
Cytological state $j$ prevalence in age group $k$ (HPV16 positive women): $q_{jk}^+$	ARTISTIC( $y_{jk}^+, y_k$ )	Annually 20–35, 5-yearly 35–55, 55–64
Cytological state prevalence (HPV16 negative women): $q_{jk}^-$	ARTISTIC( $y_{jk}^-, n_k - y_k$ ) ARTISTIC ( $y_k, n_k$ )	Under 20, 5-yearly 20–74, 75+ Yearly 10–89
HPV16 prevalence at age $t_k$ : $p^H(t_k)$	NHS cervical cancer screening programme( $y_{jk}^N, n_k^N$ )	<35, 35–49, 50 Screening rates 0–20, 5 yearly to 80 Yearly 10–90
Cytological state prevalence (HPV presence unknown): $q_{jk}$	Office for National Statistics, England	10–20, 5 yearly to 80, 80–90
Squamous cell cervical cancer incidence: $p_k^C$	[2004] ( $n_k^C, N_k^C$ )	
HPV type distribution in cervical cancer: $p_k^{16}$	Munoz et al. [42]	
Screening and treatment rates: $P_T(t)$	Various, see [11]	
Mortality rates at age $t$ : ( $P_{\delta}(t)$ , common to all CIN states $j$ )	Office for National Statistics, England and Wales [2003]	
Hysterectomy rates (for reasons unrelated to cervical disease) $P_{\theta}(t)$	Redburn and Murphy [43]	
Informed by indirect data above and given informative priors Prior source		
Transition probabilities $P_{ij}$ between CIN states $i, j: j = 6$	Insinga et al. [28] Expert belief [see 11]	Independent of age
Specificity of HPV16 test: $1 - p^{FP}$	Nanda et al. [44], Arbyn et al. [45]	
Accuracy of cervical screening: $S$		
Key intermediate quantities — defined as functions of other parameters		Monthly 10–90
HPV16 prevalence at age $t$ under 20 years (predicted): $p^H(t)$		Monthly 10–90
CIN state $j$ prevalence: $p_j(t)$		