Biostatistics (2011), **12**, 4, *pp.* 666–681
doi:10.1093/biostatistics/kxr006
Advance Access publication on April 27, 2011

D

Bayesian evidence synthesis for a transmission dynamic model for HIV among men who have sex with men

A. M. PRESANIS*

MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK
anne.presanis@mrc-bsu.cam.ac.uk

D. DE ANGELIS

MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK and Health Protection Services, Health Protection Agency, 61 Colindale Avenue, London NW9 5EQ, UK

A. GOUBAR

Institut de cancérologie Gustave Roussy, Institut National de la Sante Et de la Recherche Medicale U981, 114 rue Edouard Vaillant, 94805 Villejuif Cedex, France

O. N. GILL

Health Protection Services, Health Protection Agency, 61 Colindale Avenue, London NW9 5EQ, UK

A. E. ADES

School of Social & Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK

SUMMARY

Understanding infectious disease dynamics and the effect on prevalence and incidence is crucial for public health policies. Disease incidence and prevalence are typically not observed directly and increasingly are estimated through the synthesis of indirect information from multiple data sources. We demonstrate how an evidence synthesis approach to the estimation of human immunodeficiency virus (HIV) prevalence in England and Wales can be extended to infer the underlying HIV incidence. Diverse time series of data can be used to obtain yearly "snapshots" (with associated uncertainty) of the proportion of the population in 4 compartments: not at risk, susceptible, HIV positive but undiagnosed, and diagnosed HIV positive. A multistate model for the infection and diagnosis processes is then formulated by expressing the changes in these proportions by a system of differential equations. By parameterizing incidence in terms of prevalence and contact rates, HIV transmission is further modeled. Use of additional data or prior information on demographics, risk behavior change and contact parameters allows simultaneous estimation of the transition rates, compartment prevalences, contact rates, and transmission probabilities.

Keywords: Bayesian; Dynamic transmission model; Evidence synthesis; HIV; Incidence; Prevalence.

*To whom correspondence should be addressed.

1. Introduction

There were an estimated 33 million people worldwide living with human immunodeficiency virus (HIV) in 2007 (UNAIDS and WHO, 2008). Despite huge advances in treatment and prevention, the epidemic is still having a large impact, with unacceptably high numbers of deaths and new infections. In the United Kingdom, it was estimated that 73 300 (68 800–78 500) adults aged 15–59 years were living with HIV in 2007, of whom 28% (24–33%) were still unaware of their infection (Health Protection Agency, 2008). Moreover (Presanis *and others*, 2010), the prevalence of undiagnosed infection among adults aged 15–44 years in England and Wales has not decreased significantly since 2001, indicative of ongoing transmission in some risk groups. There is hence a clear need to quantify HIV incidence and diagnosis rates in order to plan, implement, and evaluate interventions to reduce transmission.

HIV incidence is hard to measure directly. In the past, 2 approaches have been employed to understand the rate of new infection: estimation and simulation. The most common method for incidence estimation in the early years of the epidemic was back-calculation (Brookmeyer and Gail, 1994), using information on AIDS diagnoses and the incubation period. Although the method has developed over the years, to incorporate additional information and to cope with new challenges (e.g. De Angelis *and others*, 1998; Downs *and others*, 2000; Becker *and others*, 2003; Sweeting *and others*, 2005), other methods for estimating incidence have also been explored. Estimation of disease incidence from a series of cross-sectional prevalence surveys is widely established in epidemiology (Keiding, 1991). In the HIV literature, methods making use of the "prevalence = incidence × duration" relationship to estimate HIV incidence from either age- and time-stratified seroprevalence data (Ades and Medley, 1994) or a single cross-sectional survey of individuals (snapshot samples) tested for one or more markers of recent HIV infection (Brookmeyer and Quinn, 1995) have been developed. "Snapshot sampling" methods are increasingly being reconsidered as newly developed laboratory assays to test for immune responses soon after infection continue to appear (e.g. Balasubramanian and Lagakos, 2009; Sweeting *and others*, 2010).

Simulation from mathematical models of the spread of infectious diseases (Anderson and May, 1992) has been the other key tool in investigating HIV incidence. These dynamic transmission models are multistate models where incidence depends on the size of the infected population, that is, prevalence. They are typically variations on a basic epidemic model, the susceptible-infected-removed (SIR) model (see, e.g. Anderson and Garnett, 2000; Becker and Marschner, 2001, for reviews in the HIV and sexually transmitted infection field). Forward simulation from fixed sets of parameters is performed to understand "qualitatively" the effects of various parameter values and/or interventions. Although in recent years, some attempts have been made to move toward a more inferential framework for simple deterministic HIV transmission models (e.g. Alkema *and others*, 2007), fully inferential approaches for more realistic deterministic transmission models are still rare, particularly for HIV (Punyacharoensin *and others*, 2010 unpublished data). Importantly, there is still a gap between the epidemic models used by biomathematicians to represent the evolution of an epidemic, and the inferential approaches statisticians employ to estimate specific aspects of an epidemic. This gap is driven both by a lack of detailed data on epidemics and the consequent differences in motivation between the 2 fields (Solomon and Isham, 2000; Matthews and Woolhouse, 2005).

In this paper, we propose a Bayesian evidence synthesis drawing upon ideas of estimation of incidence from serial prevalence and extending this to inference for a deterministic dynamic transmission model. "Evidence synthesis" refers to a growing and broad class of statistical analyses that combine data from disparate data sources to produce an estimate of key quantities of interest (e.g. Eddy *and others*, 1992; Ades and Sutton, 2006; Jackson *and others*, 2008). Goubar *and others* (2008) proposed a formal evidence synthesis to estimate HIV prevalence in England and Wales in 2001, using a complex probabilistic model. This model, further developed, has been applied to a series of data sets from recent years to estimate the trends in prevalence, particularly of undiagnosed infection (Presanis *and others*, 2010). Here, we use

these estimates of serial prevalence to inform a multistate model of the processes of HIV infection and diagnosis, to estimate incidence and diagnosis rates. We employ an evidence synthesis framework to combine the prevalence data with information on demographics, behavior change, and diagnosis rates. We further develop the multistate model into a nonlinear deterministic dynamic transmission model by parameterizing incidence in terms of prevalence, contact rates, and the probability of transmission given an infectious contact. This is the first time a deterministic epidemic model for HIV has been implemented in a fully Bayesian framework with the explicit aim of drawing formal inferences simultaneously about recent prevalence, incidence, contact rates, and transmission probabilities, based on a synthesis of all available relevant evidence.

We review the prevalence model in Section 2 before describing the multistate model for incidence in Section 3. Section 4 describes the methods we use to assess models. Results from the base incidence model are given in Section 5 before developing the transmission model in Section 6. We end with a discussion in Section 7.

2. PREVALENCE MODEL

The prevalence model has been described in full elsewhere (Goubar *and others*, 2008; Presanis *and others*, 2008; Presanis, 2010), but briefly, the aim was to estimate by region r and risk group g: the proportion of the population in each strata, $\rho_{g,r}$; HIV prevalence, $\pi_{g,r}$; and the proportion of HIV infections that are diagnosed, $\delta_{g,r}$. We synthesized: data from behavioral surveys such as the National Survey of Sexual Attitudes and Lifestyles ("NATSAL" Johnson *and others*, 2001) on the proportions $\rho_{g,r}$; data from unlinked anonymous seroprevalence surveys ("UA surveys" Public Health Laboratory Service *and others*, 2002) on either $\pi_{g,r}$, $\delta_{g,r}$ or prevalence of undiagnosed infection, $\pi_{g,r}(1-\delta_{g,r})$; and finally data from the Survey of Prevalent HIV Infections Diagnosed ("SOPHID" McHenry *and others*, 2000) on the total number $\left(T_r \sum_g \rho_{g,r}\pi_{g,r}\delta_{g,r}\right)$ and risk group composition $\left(\frac{\rho_{g,r}\pi_{g,r}\delta_{g,r}}{\sum_g \rho_{g,r}\pi_{g,r}\delta_{g,r}}\right)$ of diagnosed infections, where T_r is the total population size for region r. Figure 1 is a schematic directed acyclic graph (DAG) showing how we assume the parameters generate the data.

In Presanis *and others* (2010), the prevalence model is fitted simultaneously to each of K sets of data $Y_{t_k}, k \in \{1,\ldots,K\}$, one referring to each year from 2001 to 2007, so that $K = 7$. Hence, each parameter is indexed also by the series of timepoints $t_k, k \in \{1,\ldots,K\}$, where t_k is defined as the date 31st December 2000 + k. The evidence synthesis is carried out in a Bayesian framework with vague uniform prior distributions for all basic parameters $\rho_{t_k,g,r}, \pi_{t_k,g,r}$, and $\delta_{t_k,g,r}$ and Binomial or Poisson likelihoods (Section 1 of the supplementary material available at *Biostatistics* online).

Restricting attention to one risk group, men who have sex with men (MSM) in England and Wales, we obtain for each timepoint t_k the joint posterior distribution of the proportion of the total population T_{t_k} in each of 4 compartments: not at risk (i.e. not MSM), $1-\rho(t_k)$; susceptible to HIV, $\rho(t_k)\{1-\pi(t_k)\}$; HIV positive but undiagnosed, $\rho(t_k)\pi(t_k)\{1-\delta(t_k)\}$; and HIV positive and diagnosed, $\rho(t_k)\pi(t_k)\delta(t_k)$ (Figure 1 of the supplementary material available at *Biostatistics* online). Denote by $\mathbf{\Pi}(t_k) = \{\rho(t_k), \pi(t_k), \delta(t_k)\}$ the vector of prevalence parameters.

3. A COMBINED INCIDENCE AND PREVALENCE MODEL

The estimates from the prevalence model of the 4 compartment proportions (Table 1 of the supplementary material available at *Biostatistics* online) may be used to infer HIV incidence and diagnosis rates by considering the continuous-time multistate model shown in Figure 2. We have 2 choices: we can consider the incidence model to be the second stage in a 2-stage process, estimating first serial prevalence, then "plugging in" the prevalence estimates to the incidence model; or we can formulate a joint model, combining both into a single synthesis. This allows simultaneous estimation of transition rates and compartment

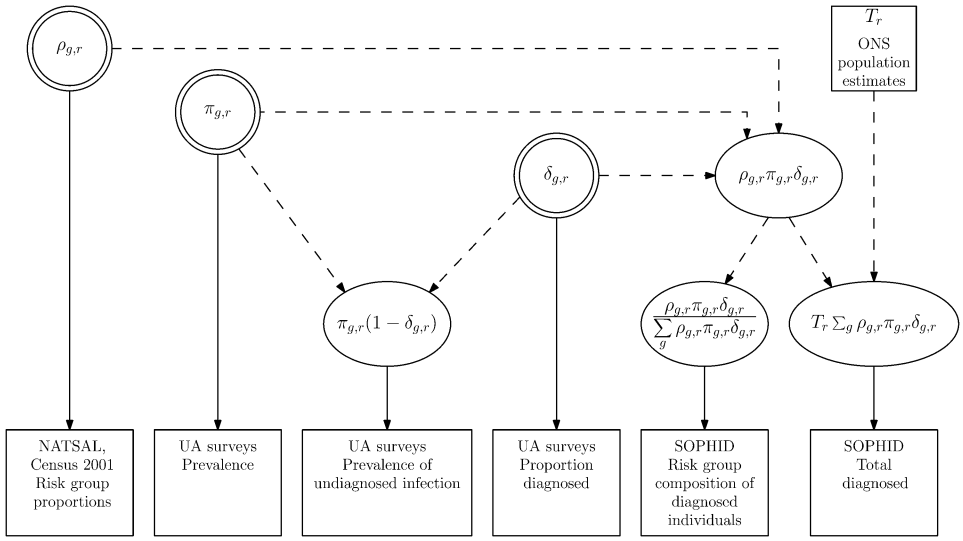

Fig. 1. Schematic DAG for the prevalence model.

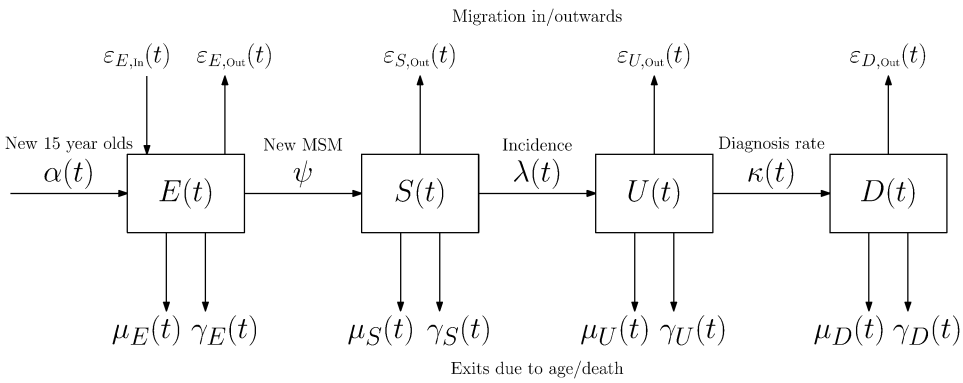

Fig. 2. Markov multistate model describing the male population of England and Wales in terms of risk group (non-MSM E, MSM) and HIV infection states: Susceptible, Undiagnosed, and Diagnosed.

prevalences. We concentrate here on the joint model—for an exposition of the 2-stage model, see Presanis (2010).

Consider the total number of men in England and Wales aged 15–44 years, regardless of risk group. Let E denote the set of non-MSM and $S \cup U \cup D$ the set of MSM, composed of Susceptible (S) + infected Undiagnosed (U) + Diagnosed (D) MSM (Figure 2). For each generic compartment $C \in \{E, S, U, D\}$, denote the number of men in the compartment at time t by $C(t)$ and the proportion of men in the compartment by $c(t) = C(t)/T(t)$, where $T(t) = E(t) + S(t) + U(t) + D(t)$ is the total number of men. Denote by $\boldsymbol{c}(t) = \{e(t), s(t), u(t), d(t)\}$ the vector of compartment proportions at time t, where $e(t) = 1 - \rho(t)$, $s(t) = \rho(t)\{1 - \pi(t)\}$, $u(t) = \rho(t)\pi(t)\{1 - \delta(t)\}$, and $d(t) = \rho(t)\pi(t)\delta(t)$.

Let $\alpha(t)$ be the rate of entry of new 15-year-old men into the system and ψ the rate at which men move from E to S. ψ is assumed constant over time. Denote by $\lambda(t)$ the incidence of HIV and $\kappa(t)$ the diagnosis rate. Let $\boldsymbol{\mu}(t) = \{\mu_c(t)\colon c \in \{E, S, U, D\}\}$, $\boldsymbol{\gamma}(t) = \{\gamma_c(t)\colon c \in \{E, S, U, D\}\}$, and

$\boldsymbol{\varepsilon}(t) = \{\varepsilon_{c,j}(t)\colon c \in \{E, S, U, D\}, j \in \{\text{In,Out}\}\}$ be the vectors of mortality rates, exit rates out of the population due to age and migration rates into/out of the population, respectively, where each element refers to a single compartment-specific rate. We assume all inward migration into the population occurs into the state E, for simplicity. Alternative migration assumptions have also been explored (Presanis, 2010). Assume all rates except ψ are piecewise constant with break points at the end of each year, so that, for example, $\lambda(t) = \lambda(t_{k-1})$, if $t_{k-1} \leqslant t < t_k$ for each $k \in \{2, \ldots, K\}$ and hence $\{\lambda(t_k)\colon k \in \{2, \ldots, K\}\}$ is a vector of 6 parameters. Note that as we have information on the compartment prevalences only at the set $\{t_k\colon k \in \{1, \ldots, K\}\}$, that is, on $\{\boldsymbol{c}(t_k)\colon k \in \{1, \ldots, K\}\}$ via the prevalence model, we concentrate on estimating the set of rates $\{\boldsymbol{\theta}(t_k)\colon k \in \{2, \ldots, K\}\}$, where $\boldsymbol{\theta}(t_k)$ denotes the vector of rate parameters at time t_k, namely $\{\alpha(t_k), \psi, \lambda(t_k), \kappa(t_k), \boldsymbol{\mu}(t_k), \boldsymbol{\gamma}(t_k), \boldsymbol{\varepsilon}(t_k)\}$. As each rate is assumed piecewise constant, this gives estimates of the rates at any time t between t_1 and t_K. The prevalence parameters $\boldsymbol{\Pi}(t) = \{\rho(t), \pi(t), \delta(t)\}$ at any time t are defined in terms of $\boldsymbol{c}(t)$:

$$\begin{aligned} \rho(t) &= s(t) + u(t) + d(t), \\ \pi(t) &= \frac{u(t) + d(t)}{s(t) + u(t) + d(t)}, \\ \delta(t) &= \frac{d(t)}{u(t) + d(t)}. \end{aligned} \tag{3.1}$$

The dynamics of the multistate model of Figure 2 may be described by a system of differential equations in terms of the numbers in each compartment:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}E(t) &= \alpha(t) + \varepsilon_{E,\text{In}}(t) - \{\mu_E(t) + \gamma_E(t) + \varepsilon_{E,\text{Out}}(t) + \psi\}E(t), \\ \frac{\mathrm{d}}{\mathrm{d}t}S(t) &= \psi E(t) - \{\lambda(t) + \mu_S(t) + \gamma_S(t) + \varepsilon_{S,\text{Out}}(t)\}S(t), \\ \frac{\mathrm{d}}{\mathrm{d}t}U(t) &= \lambda(t)S(t) - \{\kappa(t) + \mu_U(t) + \gamma_U(t) + \varepsilon_{U,\text{Out}}(t)\}U(t), \\ \frac{\mathrm{d}}{\mathrm{d}t}D(t) &= \kappa(t)U(t) - \{\mu_D(t) + \gamma_D(t) + \varepsilon_{D,\text{Out}}(t)\}D(t), \end{aligned}$$

or in terms of the proportion in each compartment, $e(t)$, $s(t)$, $u(t)$, and $d(t)$ (Section 2 of the supplementary material available at *Biostatistics* online). We proceed using the system of equations in terms of proportions to ensure we work in continuous-state-space rather than a continuous approximation to a discrete-state-space. Our aim is to estimate simultaneously the prevalence ($\boldsymbol{\Pi}(t_k)$) and the incidence ($\boldsymbol{\theta}(t_k)$) parameters, given all the available information.

3.1 *Data on transition rates*

As well as the data used in the prevalence model (Section 2) informing $\boldsymbol{\Pi}(t_k)$ or functions thereof, data are also available to inform the transition rates. Yearly demographic data on men aged 15–44 years in England and Wales are available from the Office for National Statistics ("ONS" Office for National Statistics, 2008a, 2008b, 2008c). These data are the numbers: in the population, $T(t_k)$; entering the population due to ageing, that is, aged 14 years, turning 15 years, $\alpha(t_k)$; leaving the population due to ageing, that is, aged 44 years , turning 45 years, $y^{\text{ONS}}_{T,44}(t_k)$; of deaths, $y^{\text{ONS}}_{T,\text{deaths}}(t_k)$; and entering ($y^{\text{ONS}}_{T,\text{In}}(t_k)$) or leaving ($y^{\text{ONS}}_{T,\text{Out}}(t_k)$) the population through migration. The population sizes we assume refer to the time points

$t_k, k \in \{1, \ldots, K\}$. The numbers moving in and out of the population are assumed to be the number of transitions observed between t_{k-1} and t_k for $k \in \{2, \ldots, K\}$.

Demographic data on MSM as a population are not currently available, so assumptions that susceptible and undiagnosed MSM have the same demographic rates as non-MSM will be required (Section 3.2). Data are instead available on MSM who are Diagnosed with HIV, from SOPHID, an annual cross-sectional survey of individuals with diagnosed HIV infection accessing a health care facility (Table 2 of the supplementary material available at *Biostatistics* online). Finally, data are also available on the yearly number of new diagnoses among MSM (Table 2 of the supplementary material available at *Biostatistics* online) from the HIV/AIDS Patient register (Health Protection Agency, 2006).

3.2 *Model assumptions*

Details on how the transition rate data are assumed generated from the prevalence and incidence parameters are given in Section 2 of the supplementary material available at *Biostatistics* online. Inference proceeds in a Bayesian setting: all parameters except $\psi, \alpha(t_k)$, and $\boldsymbol{\varepsilon}(t_k)$ are given vague prior distributions, independently for each t_k: $k \in \{2, \ldots, K\}$.

A schematic DAG of the combined prevalence and incidence model is shown in Figure 3. Note that the functional relationships hold at any time t, not just the observed time-points $t_k, k \in \{k \in 1, \ldots, K\}$. The proportions $\boldsymbol{c}(t_k)$ at the end of each year $t_k, k \in \{2, \ldots, K\}$ (Figure 3(b)) are defined in terms of the transition rates $\boldsymbol{\theta}(t_{k-1})$ during the year and the initial conditions for the system at the end of 2001, $\boldsymbol{c}(t_1)$ (Figure 3(a)) via the system of differential equations. The prevalence model for MSM is shown in the lower two-thirds of the DAG: the proportion of men who are MSM $\rho(t_k)$, HIV prevalence $\pi(t_k)$, and proportion diagnosed $\delta(t_k)$ are defined in terms of the proportion in each compartment, $\boldsymbol{c}(t_k)$, as shown in Figure 3 and (3.1). These prevalence parameters generate the prevalence data at the bottom of the DAG, whereas the transition rates $\boldsymbol{\theta}(t_{k-1})$ generate the transition data in the top right-hand corner of the DAG.

The combined model may result in a set of estimates of $\boldsymbol{\Pi}(t_k)$ that may potentially (but not necessarily) differ from the prevalence model estimates (Table 1 of the supplementary material available at *Biostatitics* online). We would therefore expect the combined model to allow us to assess whether the data, priors, and model structure from the incidence model are consistent (Lu and Ades, 2006; Presanis *and others*, 2008) with those from the prevalence model. If they are consistent, we would not expect the 2 sets of estimates to differ substantially. If there are differences, we might conclude there is evidence of inconsistency. As the combined model entails simultaneous estimation of both the prevalence and the incidence model parameters, in the presence of inconsistency, we might expect the combined model estimates of $\boldsymbol{\Pi}(t_k)$ to be a compromise between those of the prevalence model and the constraints imposed by the incidence model data, priors, and structure.

3.3 *Inference*

The likelihood for the joint incidence and prevalence model is $L_\theta \times L_\Pi$, where L_θ denotes the likelihood from the transition rate data and L_Π denotes the prevalence model likelihood (Section 2 of supplementary material available at *Biostatistics* online). Having defined our priors and likelihood, samples from the joint posterior distribution are obtained, using a standard adaptive Metropolis–Hastings algorithm in WinBUGS (Lunn *and others*, 2000). The system of differential equations is solved numerically at each Monte Carlo Markov chain iteration for the current set of parameter values using the Runge–Kutta algorithm in the WBDiff package (WBDiff, 2004) to provide values of the compartment proportions $\boldsymbol{c}(t_k), k \in \{1, \ldots, K\}$ with which to calculate the likelihood.

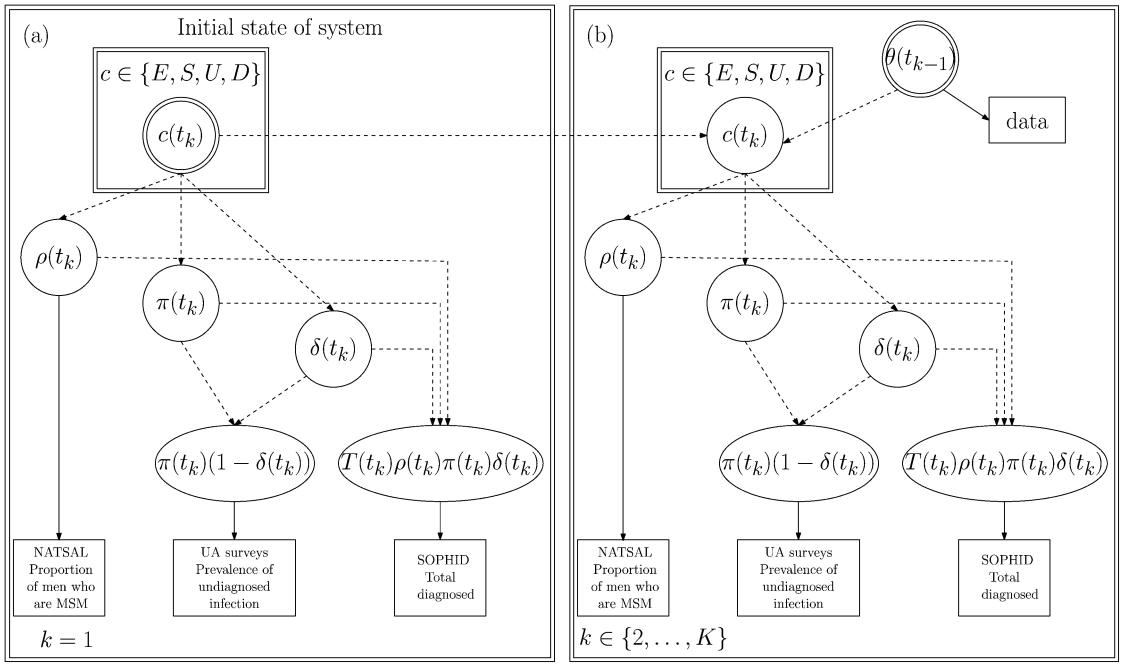

Fig. 3. Schematic DAG of the combined prevalence and incidence models. Squares/rectangles denote nodes that are observed data and circles denote stochastic nodes. Double circles denote nodes with prior distributions, whether diffuse or informative. Solid lines denote distributional dependencies, whereas dashed lines denote functional relationships. Panel (a) gives the initial state of the system at time t_1, while Panel (b) represents the system at subsequent timepoints $t_2, \ldots, t_K$. Data informing the prevalence part of the model are shown at the bottom of the DAG, while data informing the transition rates in the combined model, namely the data of Section 3.1 and Table 2 of the supplementary material available at *Biostatistics* online, are shown in the top right-hand corner of Panel (b).

Table 1. *Number of data points, posterior mean (saturated) deviance, (saturated) deviance evaluated at posterior means, effective number of parameters, and DIC, by model*

Model	n	$\bar{D}$	$D(\bar{\theta})$	p_D	DIC
Prevalence	126	124	1	123	248
Combined base	174	177	21	156	334
Combined base: prevalence data	126	125	16	110	235
Combined transmission	174	176	22	153	329
Combined transmission: prevalence data	126	124	16	108	231

4. MODEL ASSESSMENT

To assess absolute goodness of fit, we compare the posterior mean deviance $\bar{D} = E_{\boldsymbol{\theta}|\boldsymbol{y}}\{D(\boldsymbol{\theta})\} = \int D(\boldsymbol{\theta}, \boldsymbol{y})\, p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$ with the number of observations n (Dempster, 1997; Spiegelhalter *and others*, 2002). When the model is true and under standard regularity conditions, the mean of the sampling distribution of $\bar{D}$ is asymptotically equal to n. If $\bar{D}$ is much larger than n, a lack of fit of the model to the data is suggested. However, how near $\bar{D}$ needs to be from n to constitute goodness of fit is an open question.

Further deviance summaries have been proposed for assessing Bayesian model fit. Denote by $D(\bar{\boldsymbol{\theta}})$, the deviance calculated at the posterior mean of the parameters $\boldsymbol{\theta}$ and define the effective number of

parameters to be $p_D = \bar{D} - D(\bar{\boldsymbol{\theta}})$. p_D is a measure of model complexity. To compare models, we employ the deviance information criterion ("DIC" Spiegelhalter *and others*, 2002), defined, analogously to the Akaike information criterion (Akaike, 1973), as DIC $= D(\bar{\boldsymbol{\theta}}) + 2p_D = \bar{D} + p_D$, that is, a measure of model adequacy penalized by a measure of model complexity.

5. Results

Results from the combined model suggest an increase in incidence (95% posterior probability that $\lambda_{2007} \geqslant \lambda_{2002}$), with small troughs in 2002 and 2005, together with a slightly increasing diagnosis rate over time (Figures 4(a) and (b)). An increasing trend in the risk group proportion ρ over time (Figure 4(c)) is also estimated, in contrast to the prevalence only model, where the trend is constant. The additional data on transitions, priors, and structure of the combined model clearly have a different impact than those of the prevalence model on the posterior distribution of the MSM population. In the combined model, the estimate of ρ is determined by all the transition rates (Figure 3) as well as by the NATSAL data, whereas in the prevalence model, it is influenced only by the NATSAL data of 2001, assumed to provide an unbiased estimate of ρ in all years 2001–2007. This assumption, perhaps implausible, is made to compensate a lack of more recent information on the MSM population size. For ρ to be estimated as increasing, the transition rate data and priors induce net inflows to the set of MSM, $S \cup U \cup D$, that imply a greater rate of increase of MSM than the rate of increase in the total population would suggest. The greater amount of data in the combined model, together with the constraints imposed by the differential equations, also results in tighter posterior distributions for ρ than the prevalence model (Figure 4(c)).

The top half of Table 1 gives deviance summaries for the prevalence and combined models. Given the different amounts of data involved in the 2 models, the full DICs are clearly not comparable. However, we can compare the DIC of the prevalence model (248) to the DIC contributions of the prevalence data in the combined model (235, row 3 in the table). We find an equal fit to the prevalence data in both models ($\bar{D} = 125$ in the combined model compared to 124 in the prevalence model) but a smaller effective number of parameters in the combined model and hence a smaller DIC. The smaller p_D (110 compared to 123) is due to the fact that in the combined model, the prevalence parameters $\rho(t)$, $\pi(t)$, and $\delta(t)$ are functions of the compartment proportions $\boldsymbol{c}(t)$ and hence are functions of the compartment proportions in 2001, $\boldsymbol{c}(t_1)$. In contrast, the prevalence model has no shared parameters over time. The equal fit to the prevalence data in the prevalence and combined models suggests the differing estimates of trend in ρ in the 2 models do not manifest as a lack of fit to the NATSAL data. This is unsurprising given the uncertainty in the estimates of ρ in the prevalence model compared to the combined model (Figure 4(c)): the posterior distributions from the latter are contained within the range of the prevalence model posterior distributions. We therefore would not expect $\bar{D}$ and DIC to identify any lack of fit to the NATSAL data.

6. Transmission modeling

Explicitly modeling the disease transmission process allows us to understand the relationship between incidence and prevalence, and hence the potential effect of interventions on an epidemic. To do so, we need to examine the relationship between disease incidence and 3 key factors: the prevalence of disease, the contact structure in the population of interest, and the probability of disease transmission given a contact. This allows us to partition incidence by contact group, where these groups may be defined by behavioral risk or diagnosis status, for example. Prevention and diagnosis policies may then be focussed at groups at highest risk of transmission. A first step toward understanding the relationship is simply to estimate the ratio of incidence to prevalence: that is, express $\lambda(t) = \beta(t)\pi(t)$. The ratio $\beta(t)$ is the effective contact rate ("ECR" hereafter), the average number of contacts per unit time a susceptible individual makes that are sufficient for transmission if the contact is with an infectious individual. In the HIV/STI context, factors

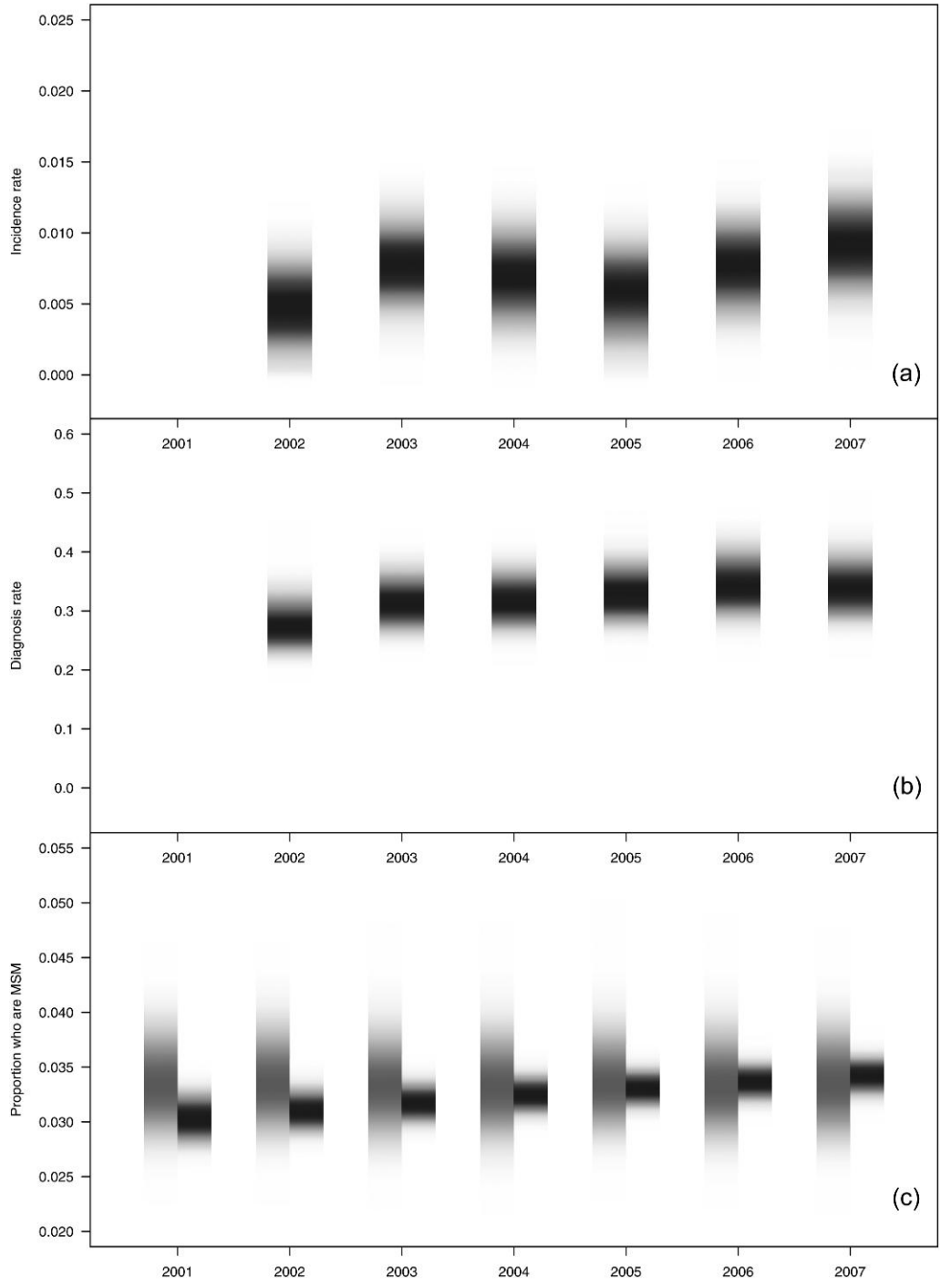

Fig. 4. Posterior distributions from the combined model of incidence (a) and diagnosis (b) rates, and posterior distributions of proportion who are MSM (c), by model: prevalence (grey) and combined (black). Note that these distributions are plotted at the year-end break points only, but the rates are in fact piecewise constant.

determining "sufficiency" for transmission (i.e. affecting $\beta(t)$) include factors affecting both contact rates and transmission given an infectious contact (Anderson and Garnett, 2000). $\beta(t)$ may therefore be further defined (Keeling and Rohani, 2007) as $\chi(t)p_\tau$, where $\chi(t)$ is the average contact rate experienced by a susceptible individual and p_τ is the transmission probability conditional on contact with an infectious individual. Incorporation of information on either of the 2 components $\chi(t)$ and p_τ allows identification of the other.

Once incidence is parameterized in terms of prevalence, the multistate model described so far is a generalization of the classic SIR model, where the susceptibles are MSM in state S, the infectious state I is split into 2 states, U and D, and the removed state R is the population in E (not at risk) plus the population outside of the system.

We assume that 2 groups are mixing effectively with susceptible MSM, both undiagnosed and diagnosed HIV positive MSM, so that incidence due to each group may be estimated. We parameterize $\lambda(t) = \chi(t)p_{\tau_U}\pi(t)\{1-\delta(t)\} + \chi(t)p_{\tau_D}\pi(t)\delta(t)$, where $\chi(t)$ is the average number of new partners a susceptible MSM has per year (so that "contact" is interpreted as a partnership), and p_{τ_U} and p_{τ_D} are the transmission probabilities conditional on contact with U and D, respectively (Section 3 of supplementary material available at *Biostatistics* online). The contact rate is piecewise constant, with break points at each year end $t_k, k \in 2, \ldots, K$ and is given a Gamma(16,4) prior independently for each year, reflecting estimates of $\chi(t)$ from NATSAL. The transmission probabilities have the following prior distributions: $p_{\tau_U} \sim$ Uniform(0, 0.3) and $p_{\tau_D} \sim$ Uniform(0, p_{τ_U}). This model allows for differing transmission probabilities for individuals in U and D since diagnosed individuals may be on treatment, with consequent lower probabilities of transmission given contact. The prior for p_{τ_U} reflects the wide range of estimates of this quantity in the literature (Baggaley, 2006). Denote this the combined "transmission model," as opposed to the combined "base model" of Section 3. This is the simplest possible model for how groups mix, assuming homogeneous mixing of susceptibles with the 2 classes of infectious MSM, that is, that the probability of choosing an infectious partner is proportional to the prevalence in the group. We start with this model as a proof of concept, paving the way for future development to more realistically model mixing and transmission in MSM.

6.1 *Results*

Posterior distributions by model for incidence, prevalence, ECRs, contact rate, and transmission probabilities are shown in Figure 5. Uncertainty is greater for the base model (red) than for the transmission model (blue) due to the greater constraints imposed by mechanistically modeling the transmission process and the informative prior distributions on the contact and transmission parameters. Note that the transmission model tends to smooth the temporal trends in incidence and prevalence slightly in comparison to the base model, and both smooth the trend in prevalence compared to the prevalence model. Both models suggest a slight increase in incidence over time, with the transmission model suggesting this increase is driven by both new infections due to the sustained prevalence of undiagnosed infection and by incidence due to diagnosed individuals (Figure 5).

In terms of deviance (Table 1), we find very little difference in the absolute fit to the data ($\bar{D} = 176$ in the transmission model compared to 177 in the base model). The effective number of parameters is slightly lower in the transmission ($p_D = 153$) than the base model ($p_D = 156$), as the increase in parameters due to introducing $\chi(t)$ and p_{τ_U} is offset by a greater reduction due to parameterizing incidence as a function of $\chi(t)$ and p_{τ_U} and to the fact that their priors are informative. The DIC is therefore also slightly lower (329 in the transmission model compared to 334 in the base model).

The transmission probability p_{τ_U} is constrained by the definition of incidence and by p_{τ_D} being a lower bound to lie between $\lambda(t)/(\chi(t)\pi(t))$ and $\lambda(t)/(\chi(t)\pi(t)(1-\delta(t)))$. As the contact rate $\chi(t)$ has a fairly tight informative prior, giving a posterior distribution suggesting on average 4 new partners per

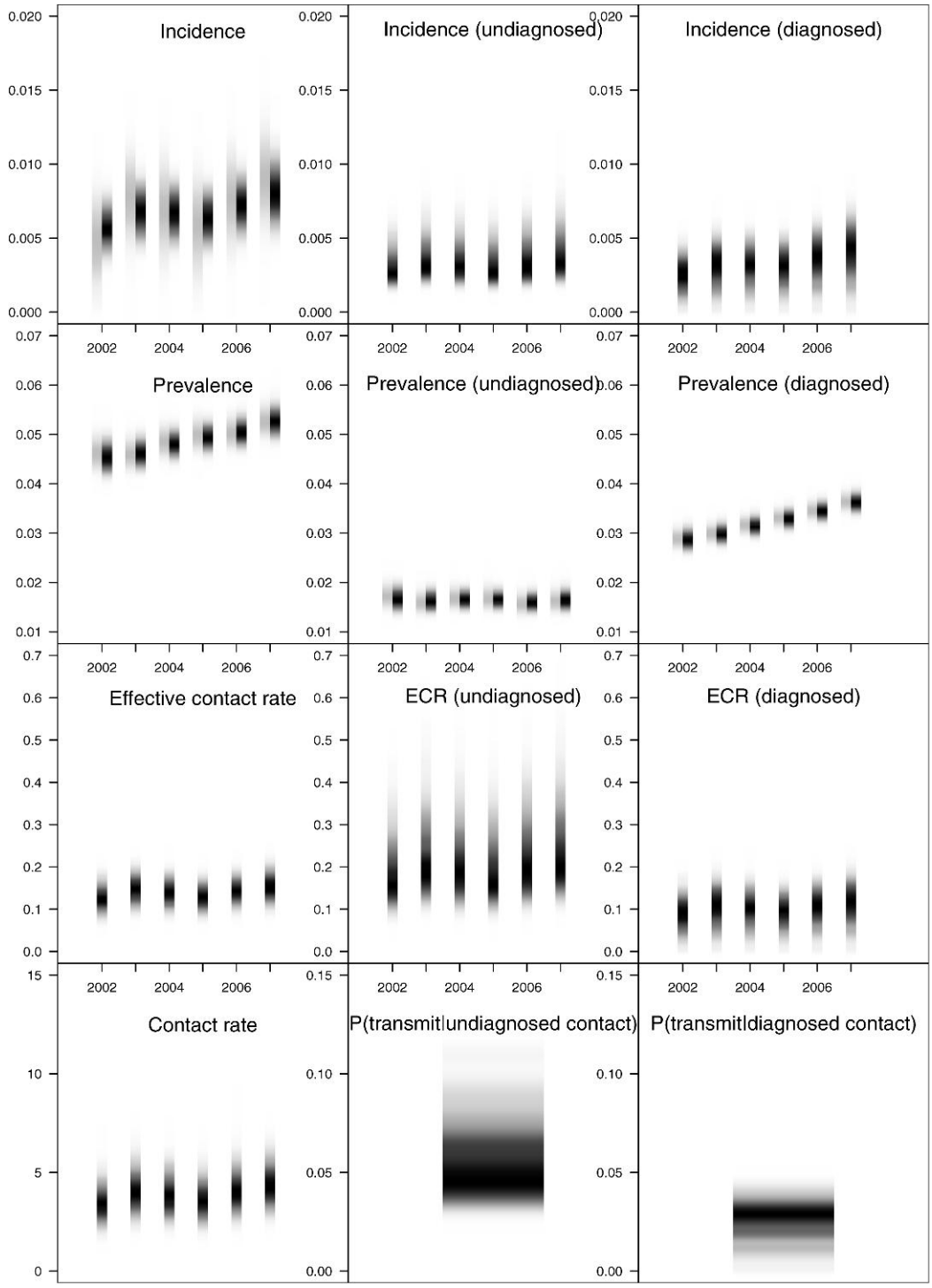

Fig. 5. Posterior distributions from the 2 combined models: base (grey), and transmission (black).

susceptible MSM per year, and as incidence, prevalence (both diagnosed and undiagnosed) and hence the ECRs are all well identified (Figure 5), the transmission probability conditional on an undiagnosed partnership has a tight posterior distribution relative to its Uniform(0,0.3) prior. We estimate that 5% (3–9%) of partnerships with undiagnosed MSM result in transmission. This estimate is in the lower range of per-partnership probability estimates reviewed by Baggaley (2006) and should be interpreted with caution, as the transmission model presented here uses a simplistic mixing assumption and does not account for behavior (e.g. condom use).

7. Discussion

We have estimated HIV incidence from serial prevalence estimates, together with demographic and behavior change data, using a multistate model in a Bayesian evidence synthesis framework. We have, in addition, embedded a (deterministic) transmission dynamic model within this framework by parameterizing incidence in terms of prevalence, contact rates, and transmission probabilities. Deterministic transmission dynamic models for HIV have historically been simulation-based, with any attempts to quantify the uncertainty in the resulting "estimates" focused largely on scenario-type sensitivity analyses. Only relatively recently has a need been recognized for models that are both formally fitted to data and for which the uncertainty in the data sources is fully accounted. Although some steps have been made in this direction, relatively few fully Bayesian deterministic transmission dynamic models have so far appeared, even outside the HIV literature (Cancre *and others*, 2000). Alkema *and others* (2007) describe a susceptible-infected model for HIV implemented in a "Bayesian melding" framework, designed to make short-term projections of HIV prevalence using a sampling importance resampling algorithm. The authors also estimate the effective contact rate, assuming this to be constant over time, but do not estimate HIV incidence. Moreover, the authors fit the susceptible-infected model to prevalence data,without explicitly modeling HIV prevalence.

In this paper, on the other hand, we have shown the feasibility of simultaneously estimating HIV prevalence, incidence and contact rates, and transmission probabilities. We have estimated a rise in incidence in MSM, due to both undiagnosed and diagnosed infection, with posterior probability that incidence in 2007 is greater than in 2002 of 95%. We have further estimated that this rise is due to both diagnosed and undiagnosed infectious MSM, albeit this finding relies on somewhat simplistic behavioral and mixing assumptions.

The combined and prevalence models we compared were fitted to different sets of data. In trying to understand the value of adding further data and potentially further parameters to model the extra data, the DIC may not be employed directly, due to the differing amounts of information. We therefore compared only the contributions to DIC of those data points that are common to all models under examination. While informally this technique allows us to compare relative goodness of fit of different models to the specific common subsets of data, it is unclear how to formally evaluate which model is "best." Inevitably, plausibility and usefulness in a public health setting must be taken into account. Model choice in this case study (Presanis, 2010) has been informally driven by considerations of both credibility of the underlying migration, contact, and transmission structures and assessment of when differences in DIC become sufficiently large to discriminate models. Understanding the DIC's distribution is the subject of ongoing research (Spiegelhalter *and others*, 2002; Celeux *and others*, 2006; Plummer, 2008), in particular for complex probabilistic models. This case study suggests the DIC may not be the most useful tool for model selection in this context, so that further work is required to explore other model comparison methods.

Our dynamic transmission model based on evidence synthesis is clearly a first step in the HIV application: there are many extensions to be explored. First, work is in progress to elaborate the contact and transmission structure (Presanis, 2010), by considering more realistic contact patterns, such as preferential

mixing, and stratification by behavioral risk, as results are likely to be sensitive to these model assumptions. Additionally, there is a growing body of evidence on antiretroviral treatment and treatment resistance (UK Group on Transmitted HIV Drug Resistance, 2005; Brown *and others*, 2009) that may be used to inform probabilities of transmission from diagnosed to susceptible individuals.

Second, further work is required to model the MSM population realistically. The combined model results in different estimated trends for the proportion of men who are MSM than the prevalence model, with the combined model suggesting a greater rate of increase in the MSM population than in general. Additionally, compared to the prevalence model, the estimated trends in π and δ in the incidence models are somewhat smoothed (results not shown), due perhaps to the dependence over time in dynamic models not allowing for abrupt changes from year to year. An increase in ρ over time is not implausible (Mercer *and others*, 2004), whereas our assumption that NATSAL provides unbiased information for each year, necessitated by lack of recent evidence, is perhaps unrealistic. We have explored various migration assumptions and have assessed the different models through comparisons of the DIC contributions of data common to all models under consideration (results not shown; Presanis, 2010). We have chosen to concentrate on models assuming outward migration occurs from all states, with inward migration occurring only into the non-MSM group, as differences in DIC between the different migration models are small. Migration and population data for MSM are currently limited but further data are being collected (Jolozo *and others*, 2010; National Centre for Social Research, 2010), giving potential for further work modeling the MSM population.

Third, the models presented here make a number of prior and likelihood assumptions, to which results may be sensitive. The Binomial and Poisson likelihoods assumed in both the prevalence and combined models do not allow for any overdispersion in the data, and while the deviance summaries indicate no lack of fit to the data, so that at first glance, there is little evidence of overdispersion, further sensitivity analyses are in progress. Investigation of the contribution of each piece of evidence (data, priors, or model assumptions) to the inference is also underway.

Other possibilities include extending the model to other risk groups, stratifying by heterosexual or parenteral transmission, gender, age, region of birth, and region of residence, for example. Transmission and contact rates could also be allowed to vary by time since infection. The piecewise constant transition rates could instead be allowed to vary smoothly over time, through the use of splines for example. While there are many possibilities for developing this work, we have nevertheless demonstrated an important step toward the joint modeling of HIV prevalence, incidence, and the transmission mechanism linking the 2 in a fully inferential framework.

SUPPLEMENTARY MATERIALS

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors thank Tim Chadborn, Brian Rice, and Caterina Hill (HPA) for supplying data and valuable discussion. We also thank David Spiegelhalter for helpful discussion and Chris Jackson for help with the density strip plots. *Conflict of Interest:* None declared.

FUNDING

Medical Research Council (G0600675 to A.M.P., U.1052.00.007 to D.D.A.); Health Protection Agency (D.D.A.); University of Bristol from funds made available by the Medical Research Council (A.E.A.).

REFERENCES

ADES, A. E. AND MEDLEY, G. F. (1994). Estimates of disease incidence in women based on antenatal or neonatal seroprevalence data: HIV in New York city. *Statistics in Medicine* **13**, 1881–1894.

ADES, A. E. AND SUTTON, A. J. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **169**, 5–35.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. and Csaki, F. (editors), *Proceedings of the 2nd International Symposium on Information Theory*. Budapest, Hungary: Akademiai Kiado, pp. 267–281.

ALKEMA, L., RAFTERY, A. E. AND CLARK, S. J. (2007). Probabilistic projections of HIV prevalence using Bayesian melding. *Annals of Applied Statistics* **1**, 229–248.

ANDERSON, R. M. AND GARNETT, G. P. (2000). Mathematical models of the transmission and control of sexually transmitted diseases. *Sexually Transmitted Diseases* **27**, 636–643.

ANDERSON, R. M. AND MAY, R. M. (1992). *Infectious Diseases of Humans: Dynamics and Control*, 1st edition. New York & Oxford: Oxford University Press.

BAGGALEY, R. F. (2006). The impact of antiretroviral use in resource-poor settings: insights from mathematical models, [PhD. Thesis]. London: Imperial College.

BALASUBRAMANIAN, R. AND LAGAKOS, S. W. (2009). Estimating HIV incidence based on combined prevalence testing. *Biometrics* **66**, 1–10.

BECKER, N. G., LEWIS, J. J. C., LI, Z. AND MCDONALD, A. (2003). Age-specific back-projection of HIV diagnosis data. *Statistics in Medicine* **22**, 2177–2190.

BECKER, N. G. AND MARSCHNER, I. C. (2001). Advances in medical statistics arising from the AIDS epidemic. *Statistical Methods in Medical Research* **10**, 117–140.

BROOKMEYER, R. AND GAIL, M. H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Monographs in Epidemiology & Biostatistics. New York & Oxford: Oxford University Press.

BROOKMEYER, R. AND QUINN, T. C. (1995). Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *American Journal of Epidemiology* **141**, 166–172.

BROWN, A. E., MURPHY, G., RINCK, G., CLEWLEY, J. P., HILL, C., PARRY, J. V., JOHNSON, A. M., PILLAY, D. AND GILL, O. N. (2009). Implications for HIV testing policy derived from combining data on voluntary confidential testing with viral sequences and serological analyses. *Sexually Transmitted Infections* **85**, 4–9.

CANCRE, N., TALL, A., ROGIER, C., FAYE, J., SARR, O., TRAPE, J.-F., SPIEGEL, A. AND BOIS, F. (2000). Bayesian analysis of an epidemiologic model of Plasmodium falciparum malaria infection in Ndiop, Senegal. *American Journal of Epidemiology* **152**, 760–770.

CELEUX, G., FORBES, F., ROBERT, C. P. AND TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis* **1**, 651–674.

DE ANGELIS, D., GILKS, W. R. AND DAY, N. E. (1998). Bayesian projection of the acquired immune deficiency syndrome epidemic. *Journal of the Royal Statistical Society (Series C): Applied Statistics* **47**, 449–498.

DEMPSTER, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing* **7**, 247–252.

DOWNS, A. M., HEISTERKAMP, S. H., RAVÀ, L., HOUWELING, H., JAGER, J. C. AND HAMERS, F. F. (2000). Back-calculation by birth cohort, incorporating age- specific disease progression, pre-AIDS mortality and change in European AIDS case definition. European Union Concerted Action on Multinational AIDS Scenarios. *AIDS* **14**, 2179–2189.

EDDY, D. M., HASSELBLAD, V. AND SHACHTER, R. (1992). *Meta-Analysis by the Confidence Profile Method*. London: Academic Press.

GOUBAR, A., ADES, A. E., DE ANGELIS, D., MCGARRIGLE, C. A., MERCER, C., TOOKEY, P., FENTON, K. AND GILL, O. N. (2008). Estimates of HIV prevalence and proportion diagnosed based on Bayesian multi-parameter synthesis of surveillance data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **171**, 541–580.

HEALTH PROTECTION AGENCY (2006). Health Protection Agency Centre for Infections unpublished HIV diagnoses surveillance tables 2001–2006. http://www.hpa.org.uk/Topics/InfectiousDiseases/InfectionsAZ/HIV/NewHIVDiagnoses/.

HEALTH PROTECTION AGENCY (2008). HIV in the United Kingdom: 2008 report. *Technical Report*. London: HIV & STI Department, Health Protection Agency. http://www.hpa.org.uk/Publications/InfectiousDiseases/HIVAndSTIs/0811hivUK/.

JACKSON, C., RICHARDSON, S. AND BEST, N. (2008). Studying place effects on health by synthesising individual and area-level outcomes. *Social Science & Medicine* **67**, 1995–2006.

JOHNSON, A. M., MERCER, C. H., ERENS, B., COPAS, A. J., MCMANUS, S., WELLINGS, K., FENTON, K. A., KOROVESSIS, C., MACDOWALL, W., NANCHAHAL, K. *and others* (2001). Sexual behaviour in Britain: partnerships, practices, and HIV risk behaviours. *Lancet* **358**, 1835–1842.

JOLOZO, T., EVANS, J., O'BRIEN, R. (2010). Measuring sexual identity: an evaluation report. *Technical Report*. Newport, UK: Office for National Statistics. http://www.ons.gov.uk/about-statistics/measuring-equality/equality/sexual-identity-project/measuring-sexual-identity–an-evaluation-report.pdf.

KEELING, M. J. AND ROHANI, P. (2007). *Modeling Infectious Diseases in Humans and Animals*, 1st edition. Princeton, NJ: Princeton University Press.

KEIDING, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **154**, 371–412.

LU, G. AND ADES, A. E. (2006). Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101**, 447–459.

LUNN, D. J., THOMAS, A., BEST, N. AND SPIEGELHALTER, D. (2000). WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.

MATTHEWS, L. AND WOOLHOUSE, M. (2005). New approaches to quantifying the spread of infection. *Nature Reviews Microbiology* **3**, 529–536.

MCHENRY, A., MACDONALD, N., SINKA, K., MORTIMER, J. AND EVANS, B. G. (2000). National assessment of prevalent diagnosed HIV infections. *Communicable Disease and Public Health* **3**, 277–281.

MERCER, C. H., FENTON, K. A., COPAS, A. J., WELLINGS, K., ERENS, B., MCMANUS, S., NANCHAHAL, K., MACDOWALL, W. AND JOHNSON, A. M. (2004). Increasing prevalence of male homosexual partnerships and practices in Britain 1990–2000: evidence from national probability surveys. *AIDS* **18**, 1453–1458.

NATIONAL CENTRE FOR SOCIAL RESEARCH (2010). The Third National Survey of Sexual Attitudes and Lifestyles. London: National Center for Social Research. http://www.natcen.ac.uk/study/natsal.

OFFICE FOR NATIONAL STATISTICS (2008a). Death registrations 2001–2007. Source: Office for National Statistics licensed under the Open Government Licence v.1.0. http://www.statistics.gov.uk/statbase/product.asp?vlnk=15096.

OFFICE FOR NATIONAL STATISTICS (2008b). Long-Term International Migration (MN Series) 2001–2007. Source: Office for National Statistics licensed under the Open Government Licence v.1.0. http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=507.

OFFICE FOR NATIONAL STATISTICS (2008c). Mid-year Population Estimates 2001-2007. Source: Office for National Statistics licensed under the Open Government Licence v.1.0. http://www.statistics.gov.uk/statbase/Product.asp?vlnk=15106.

PLUMMER, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics* **9**, 523–539.

PRESANIS, A. M. (2010). Evidence synthesis methods to estimate disease prevalence, incidence and transmission, with application to HIV, [PhD. Thesis]. MRC Biostatistics Unit, University of Cambridge, Cambridge.

PRESANIS, A. M., DE ANGELIS, D., SPIEGELHALTER, D. J., SEAMAN, S., GOUBAR, A. AND ADES, A. E. (2008). Conflicting evidence in a Bayesian synthesis of surveillance data to estimate HIV prevalence. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **171**, 915–937.

PRESANIS, A. M., GILL, O. N., CHADBORN, T. R., HILL, C., HOPE, V., LOGAN, L., RICE, B. D., DELPECH, V. C., ADES, A. E. AND DE ANGELIS, D. (2010). Insights into the rise in HIV infections in England and Wales, 2001 to 2008: a Bayesian synthesis of prevalence evidence. *AIDS* **24**, 2849–2858.

PUBLIC HEALTH LABORATORY SERVICE, INSTITUTE OF CHILD HEALTH AND SCOTTISH CENTRE FOR INFECTION AND ENVIRONMENTAL HEALTH (2002). The prevalence of HIV and Hepatitis infection in the United Kingdom 2001: annual report of the Unlinked Anonymous Prevalence Monitoring Programme. *Technical Report*. London: Health Protection Agency Centre for Infections.

SOLOMON, P. J. AND ISHAM, V. S. (2000). Disease surveillance and data collection issues in epidemic modelling. *Statistical Methods in Medical Research* **9**, 259–277.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. AND VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **64**, 583–639.

SWEETING, M. J., DE ANGELIS, D. AND AALEN, O. O. (2005). Bayesian back-calculation using a multi-state model with application to HIV. *Statistics in Medicine* **24**, 3991–4007.

SWEETING, M. J., DE ANGELIS, D., PARRY, J. AND SULIGOI, B. (2010). Estimating the distribution of the window period for recent HIV infections: a comparison of statistical methods. *Statistics in Medicine* **29**, 3194–3202.

UK GROUP ON TRANSMITTED HIV DRUG RESISTANCE. (2005). Time trends in primary resistance to HIV drugs in the United Kingdom: multicentre observational study. *British Medical Journal* **331**, 1368.

UNAIDS AND WHO (2008). 2008 report on the global AIDS epidemic. Annex: HIV and AIDS estimates and data, 2007 and 2001. *Technical Report*. Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS (UNAIDS). http://www.unaids.org/en/dataanalysis/epidemiology/2008reportontheglobalaidsepidemic/.

WBDIFF (2004). WinBUGS differential interface—worked examples. *Technical Report*. London: Department Epidemiology and Public Health, Imperial College.

[Received August 13, 2010; revised February 16, 2011; accepted for publication February 17, 2011]