Multi-state Markovian model for estimating HIV incidence from French surveillance data: a simulation study

Charlotte Castel

Santé Publique France and ISPED
Université Paris Est
BPH Inserm U1219

GDR Statistiques et Santé 2019

Supervisors : Mr Yann Le Strat et Mr Ahmadou Alioum
Co-supervisor : Mrs Cécile Sommen

Introduction

Context

- Epidemic still active and difficult to control :
  - Approximately 6,500 HIV-positive discoveries per year, stable since 10 years
  - Seropositivity findings mainly among MSM (44%) and foreign-born heterosexuals (39%)
  - Higher rates in French overseas territories and Ile-de-France
  - Population not knowing their HIV status at high risk of transmission
  - Measures to reduce transmission : condoms, preventive treatments (PreP), screening and rapid treatment

Objectives

- Need to estimate and consolidate 3 epidemiological indicators :
  - Number of people newly infected with HIV
  - Number of people who do not know their HIV status
  - The distribution of the time between infection and diagnosis

Existing methods for estimating HIV incidence in France from HIV mandatory notification

- Back calculation (INSERM U1136 and ISPED) [1] [2] :
  - Use of historical data from clinical stage to diagnosis
  - Joint estimation of the incidence, the distribution of time between infection and diagnosis, and the seropositive population not knowing its status
- Methods from the recent infection test (SpF and ISPED) [3] [4] :
  - Use of HIV serological markers TM and V3
  - Estimated only HIV incidence

1. C. Sommen et al. (2009). In : *Stat Med* 28.
2. L. Marty et al. (2018). In : *J Int AIDS Soc* 21.3.
3. C. Sommen et al. (2011). In : *Biometrics* 67.2.
4. S. Le Vu et al. (2010). In : *The Lancet, Infectious Diseases* 10.10.

HIV diagnoses surveillance data

- HIV mandatory notification and virological surveillance :
  - Since 2003 at Santé Publique France
  - Sociodemographic characteristics
  - Mode of contamination, reason for screening, clinical stage, history of HIV tests, date of possible contamination
  - For the most recent years : CD4, antiretroviral treatments, viral load
  - Virological surveillance : markers of recent infection to distinguish between recently infected and those who have been infected for longer

Simulation design

Simulation design : HIV incidence (1994-2018)

1) 1994-2003 : We generated the number of new HIV infection cases according to the following recurrence relation for $k = 1, ..., 10$:

$$\begin{cases} y_{2004} = \lambda_{2004} \\ y_{2004-k} \sim Poisson(y_{2004-k+1} \times 1.07) \end{cases}$$

2) 2004-2015 : For each year i, we generated the number of new HIV infection cases, noted y_i, according to a Poisson distribution such as $y_i \sim Poisson(\lambda_i)$ with λ_i the number of new HIV infection cases from previous study [5]

3) 2016-2018 : numbers of new HIV infection cases were simulated according to 3 different variation rates over the period : $-$ 5% per year, 0% per year and +5% pear year

5. L. MARTY et al. (2018). In : *J Int AIDS Soc* 21.3.

Simulation design : clinical stage at diagnosis date

- In the HIV mandatory notification, at time of diagnosis, a clinical stage for HIV is assigned to the individual by a medical doctor : primo-infection (PI), asymptomatic (ASY), symptomatic (SYM)) or AIDS
- This distribution does not really vary over time in the HIV mandatory notification, we chose the mean distribution over the period : 8.3% of HIV primo-infection, 61.6 % of HIV asymptomatic, 13% of HIV symptomatic and 17.1% of AIDS
- For each individual, a clinical stage was assigned randomly, under the constraint of respecting this distribution

Simulation design : testing behaviours

- Simulated HIV test dates depend on the frequency of testing among diagnosed individuals. Indeed, individuals have different HIV testing behaviours and we distinguished regular testers from non-regular testers
- The proportion of regular testers in the HIV mandatory notification was stable since 2004, we chose the mean distribution of 23% using the following definition [6] :
  - Regular tester : last negative HIV test in 2 years before his positive test date
  - Non regular tester : last negative HIV test more than 2 years before his positive test date
- Primo infection, symptomatic and AIDS : 100% non regular testers
- Asymptomatic : 23% regular testers and 77% non regular testers

6. M. S. JAMIL et al. (2017). In : *Lancet HIV* 4.6.

Simulation design : diagnosis date (1)

- For primo infection stage, symptomatic stage and AIDS stage :
  - Primo-infection stage : time from infection to diagnosis was assumed to be uniform between 2 weeks and 6 weeks.
  - Symptomatic stage : duration from infection to diagnosis was generated according to a cumulative distribution $F_{SYM}(t)$ obtained from previous study [7] giving a median of 4.3 years.
  - AIDS stage : the AIDS incubation was generated according to a Weibull distribution [8] $F_{AIDS}(t) = 1 - \exp -(0.0215t)^{2.516}$ with a median of 10 years.

7. C. SOMMEN et al. (2009). In : *Stat Med* 28.
8. R. BROOKMEYER et al. (1989). In : *Stat Med* 8.1.

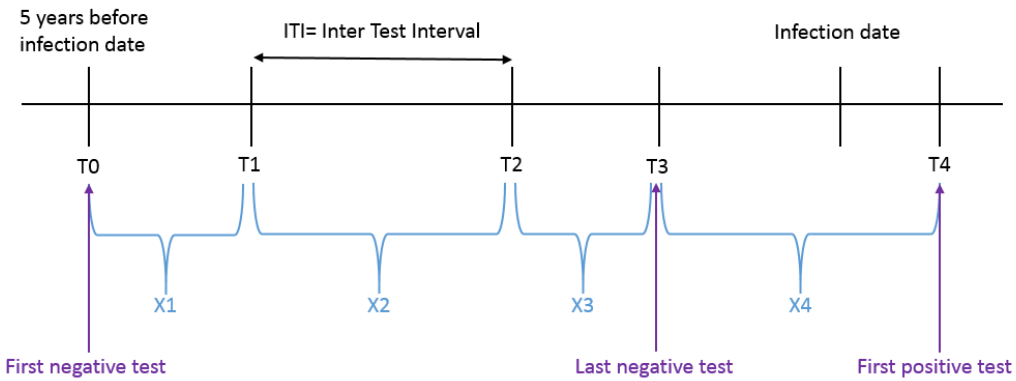
Simulation design : diagnosis date (2)

- For asymptomatic stage :
  - For non regular testers : duration from infection to diagnosis was generated according to a cumulative distribution $F_{ASYM}(t)$ obtained from previous study [9] giving a median of 2.3 years.
  - For regular testers : duration between infection and diagnosis was generated according to a renewal process [10] [11] [12].

9. C. Sommen et al. (2009). In : *Stat Med* 28.
10. J. M. Karon et al. (2008). In : *Stat Med* 27.23.
11. Q. An et al. (2017). In : *Am J Prev Med* 53.3.
12. R. Song et al. (2005). In : *Commun Stat Theory Methods* 34.8.

Simulation design : diagnosis date (3)

Figure 1 : Renewal process for regular testers



- $X \sim exp(\frac{1}{\mu})$ with μ the mean of time between last negative test and first positive test, for regular testers with a positive recent infection test at the diagnosis [13].
- In the HIV mandatory notification $\mu = 9.23$ months, so the renewal process $X \sim exp(\frac{1}{\mu} = 0.11)$.

13. R. SONG et al. (2005). In : *Commun Stat Theory Methods* 34.8.

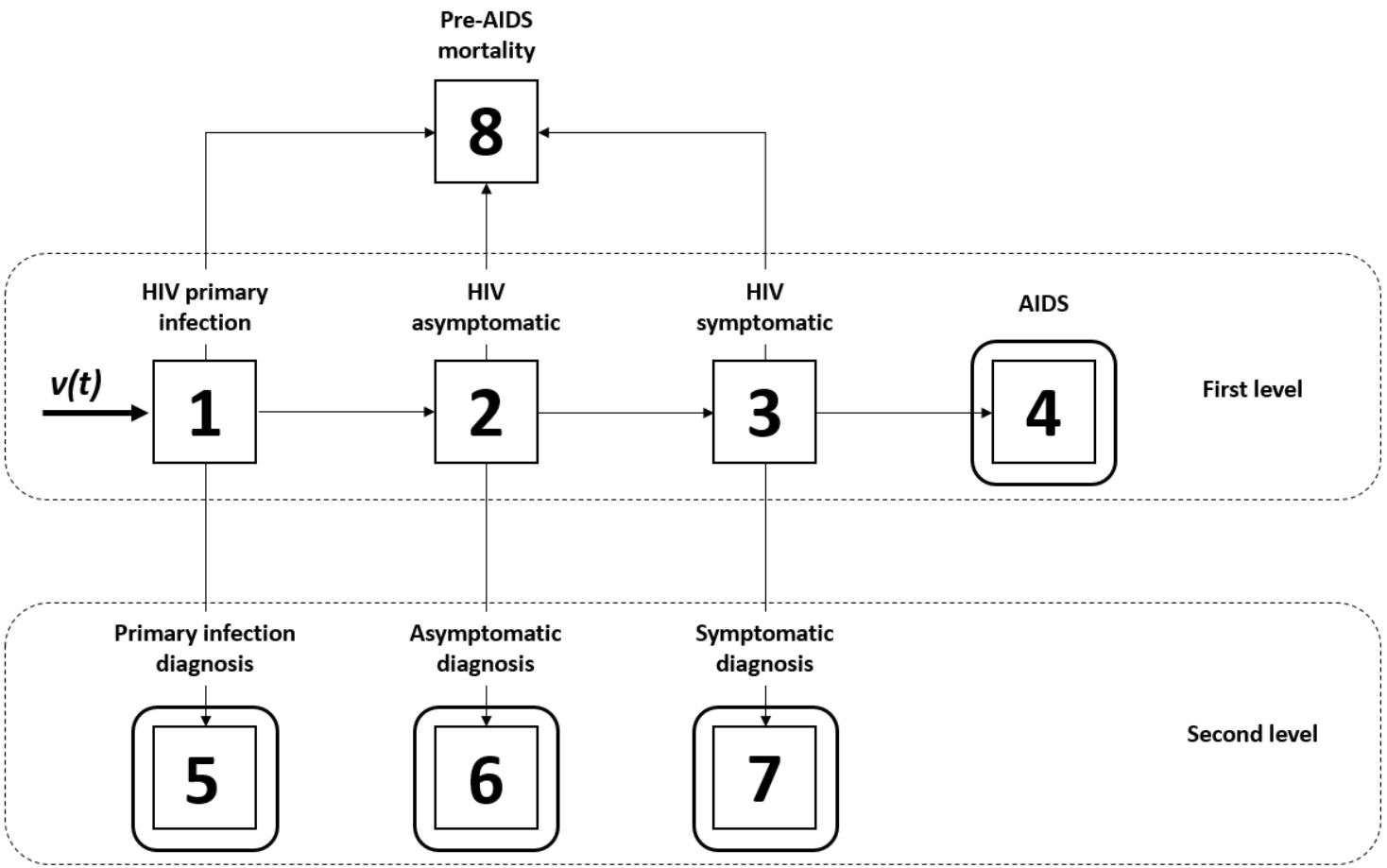
Simulation design : summarize

- We know the theorical incidence
- For each individual, we know :
  - His clinical stage at diagnosis
  - His diagnosis date
  - His infection date
  - His testing behaviours

Multi-state Markov model

Multi-state Markov model

Figure 2 : Multi-state Markov model describing the progression of HIV infection.



Likelihood of the model (1)

- Poisson process in discrete time
- Using the same approach as in Aalen et Al.[14], we can then write the expected number of individuals in states 1 to 8 at the time t_i with the vector $E_i = (E_{i,l}), l = 1, ..., 8$, by the relation of recurrence according to P_i and H_i :

$$\begin{cases} E_0 = H_0 \\ E_i = P_i^T E_{i-1} + H_i, i = 1, 2, ..., K \end{cases}$$

- $H_i = \left(h_i = \int_{t_{i-1}}^{t_i} \nu(x)\mathrm{d}x, 0, 0, 0, 0, 0, 0, 0 \right)^T$
- $P_i = (\alpha_{k,l}^i)$ where $\alpha_{k,l}^i$ is the transition probabilities of state k to state l between t_{i-1} and t_i with $k, l = 1, 2, ..., 8$:
  - Homogeneous markov model : P_i is no time dependant
  - Non-homogeneous markov model, P_i is time dependant

14. O. O. Aalen et al. (2007). In : *Stat Med* 16.19.

Likelihood of the model (2)

- Expected number of new positive HIV diagnoses in the week $T_i, i = 1, ..., K$ is expressed by :
  - $e_i^4 = E_{i-1,3}\alpha_{3,4}$ the expected number of individuals entering in AIDS stage (state 4) in week T_i
  - $e_i^5 = E_{i-1,1}\alpha_{1,5}$ the expected number of individuals entering in primo-infection stage (state 5) in week T_i
  - $e_i^6 = E_{i-1,2}\alpha_{2,6}$ the expected number of individuals entering in asymptomatic stage (state 6) in week T_i
  - $e_i^7 = E_{i-1,3}\alpha_{3,7}$ the expected number of individuals entering in symptomatic stage (state 7) in week T_i

Penalized likelihood (1)

- The likelihood of the model can be expressed as follow :

$$L = \prod_{i=H}^{K} \prod_{j=4}^{7} (e_i^j)^{n_i^j} exp(-e_i^j)$$

- Smooth curve, no negative values and low local variations
- The penalized likelihood is : [15] [16]

$$pl = \log(L) - \lambda \int \nu''(u)^2 \mathrm{d}u$$

- Parameters to be estimated are $\alpha_{1,5} = (\alpha_{1,5}^1, \alpha_{1,5}^2, ..., \alpha_{1,5}^K)$, $\alpha_{2,6} = (\alpha_{2,6}^1, \alpha_{2,6}^2, ..., \alpha_{2,6}^K)$, $\alpha_{3,7} = (\alpha_{3,7}^1, \alpha_{3,7}^2, ..., \alpha_{3,7}^K)$ and $\nu(.)$.

15. I.J. GOOD et al. (1980). In : *Journal of the American Statistical Association* 75.369.
16. P. JOLY et al. (1998). In : *Biometrics* 54.

Penalized likelihood (2)

- $\nu(.)$ is approximated by a base of M-splines Cubic functions of order 4 [17] :

$$\widetilde{\nu}(.) = \sum_{j=1}^{Q+2} \theta_j M_j(.)$$

- For a fixed value of λ we try to estimate the vector of parameters $\hat{\Theta}_\lambda = (\hat{\theta}, \hat{\alpha}_{1,5}, \hat{\alpha}_{2,6}, \hat{\alpha}_{3,7})$ which maximizes the penalized log-likelihood
- λ is by a cross-validation approximation and then injected in the penalized likelihood for maximisation with Marquardt algorith [18]

17. P. JOLY et al. (1998). In : *Biometrics* 54.
18. D. MARQUARDT (1963). In : *SIAM Journal on Applied Mathematics* 11.

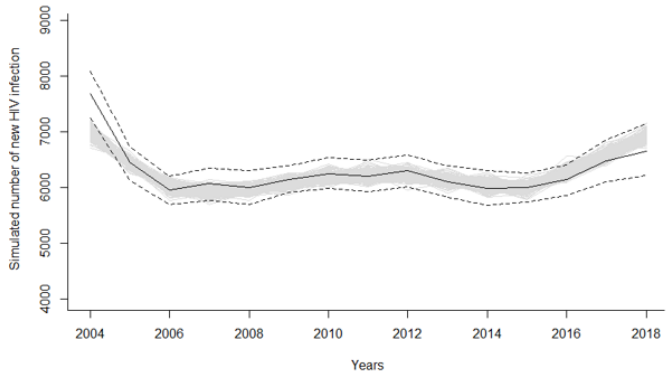
Simulation results

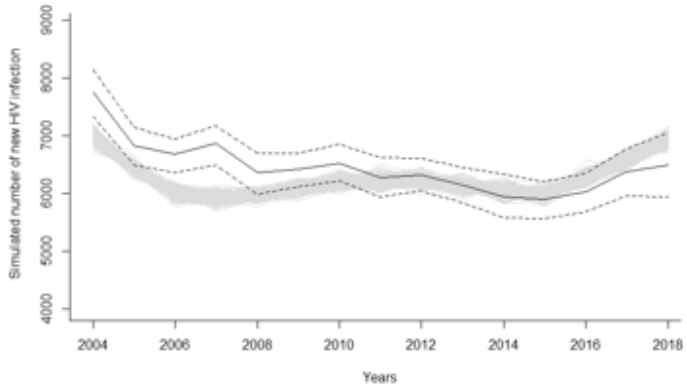
Simulation results

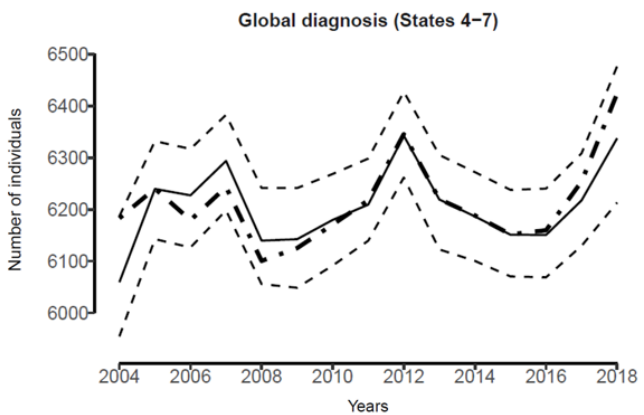
- Results presented after were obtained from 200 simulations
- We present results obtained from the increasing trend for the last 3 years
- On the left we presented results with homogeneous Markov model : only one transition probabilities matrix for the period 2004-2018 (3 transition probabilities)
- On the right we presented results with non-homogeneous Markov model. We defined 5 times periods of 3 years : 5 transition probabilities matrix for the period 2004-2018 (15 transition probabilities)

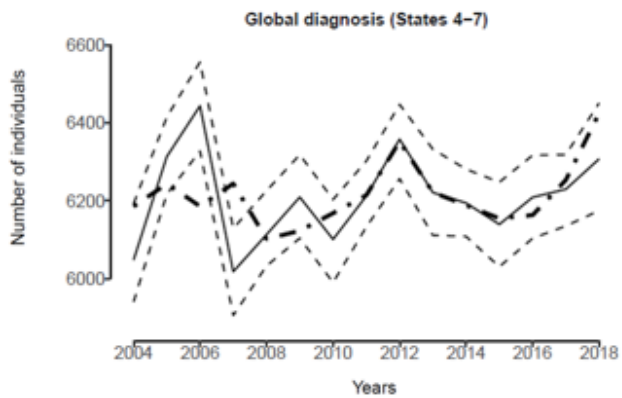
Incidence and global diagnosis

Figure 3 : Estimated incidence and global diagnosis. Homogeneous period at left and non-homogeneous period by 3 years at right



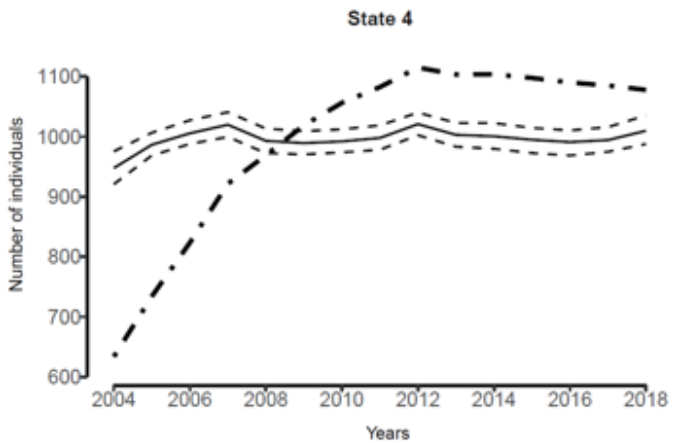


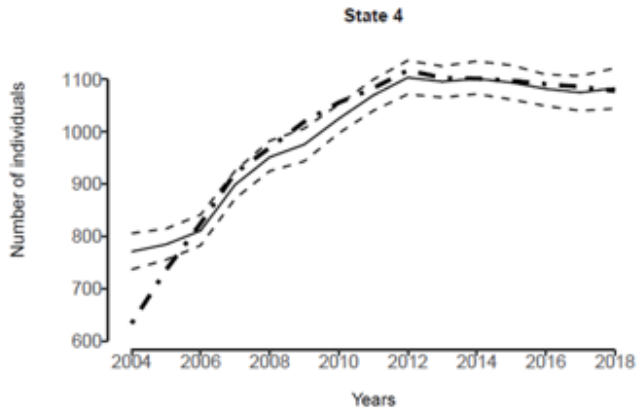


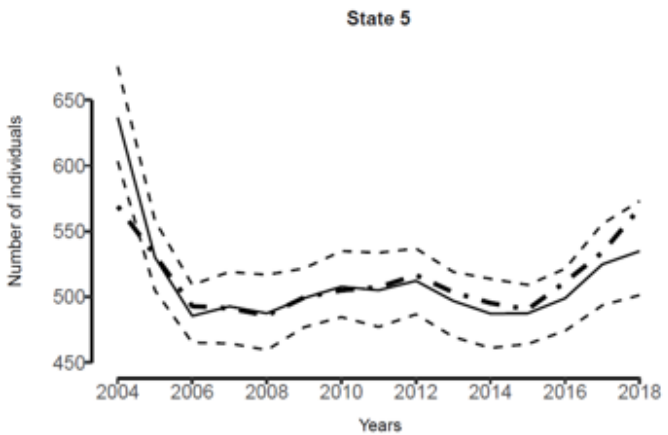


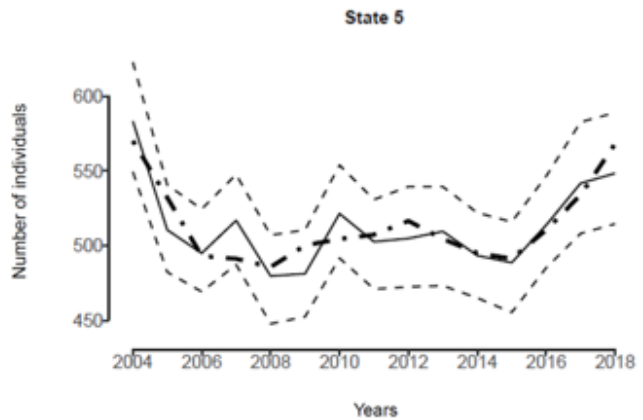
Estimated diagnosis for state 4 and state 5

Figure 4 : Estimated number of diagnosis. Homogeneous period at left and non-homogeneous period by 3 years at right



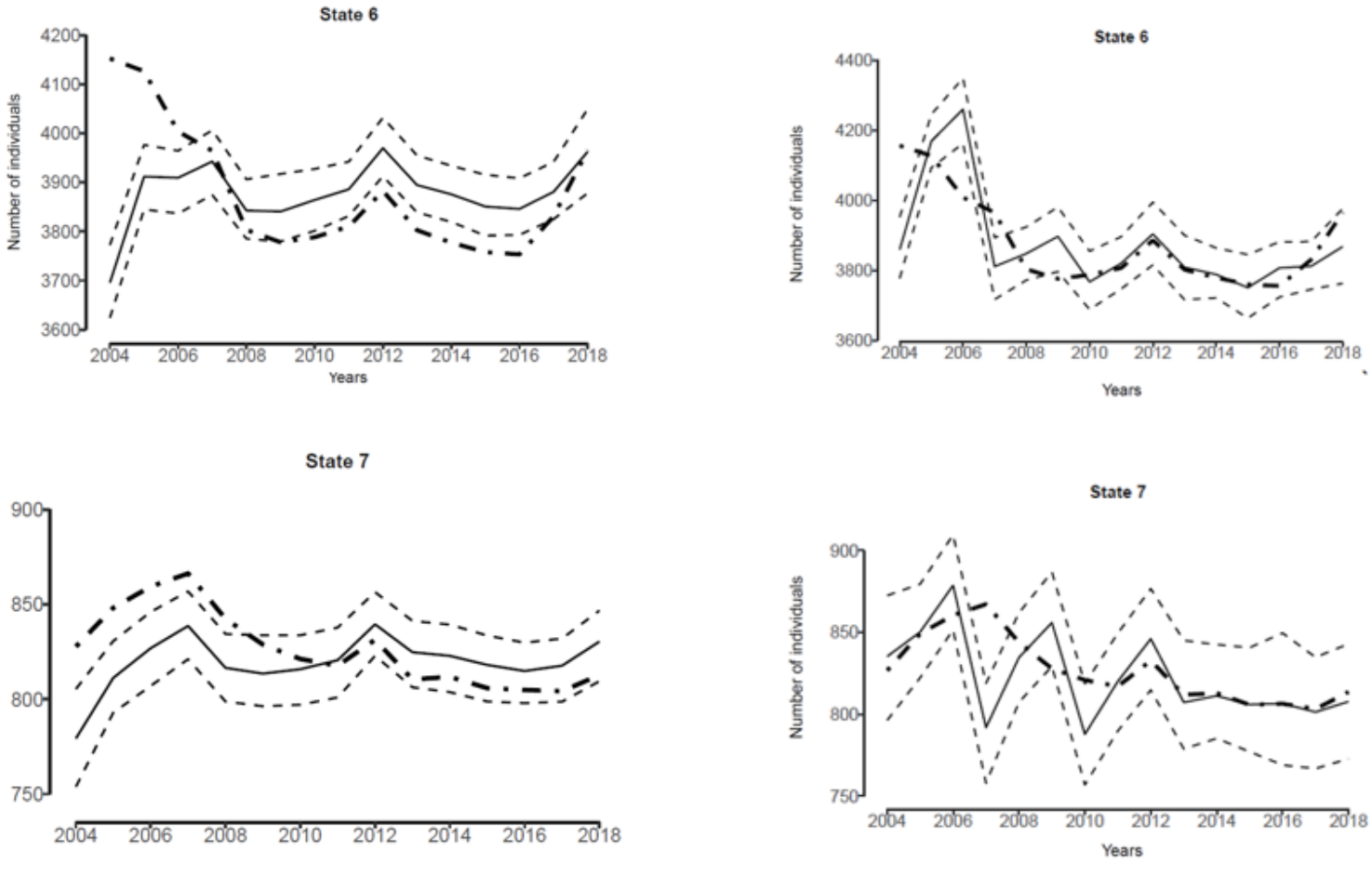






Estimated diagnosis for state 6 and state 7

Figure 5 : Estimated number of diagnosis. Homogeneous period at left and non homogeneous period by 3 years at right



Discussion

Discussion and perspective

- The model presented here is a new useful tool for estimating the incidence of HIV infection using all informations from the clinical stages of French data.
- Need to consolidate the simulation
- The model will then be expanded to include biomarkers of recent infection and take into account any relevant information provided by mandatory reporting of HIV.

Perspective : Application of the method in the HIV mandatory notification database :

- Global estimation of the incidence
- Estimation of the incidence by transmission groups
- Estimation of the incidence by geographical area

Acknowledgement

Supervisors : Mr Yann Le Strat et Pr Ahmadou Alioum
Co-supervisor : Mrs Cécile Sommen
Mr Edouard Chatignoux, biostatistician at Santé publique France
Mrs Françoise Cazein, epidemiologist at Santé publique France
Mr Pierre Joly, researcher in biostatistics at ISPED
Mrs Cécile Proust Lima, researcher in biostatistics at ISPED
Mrs Viviane Phillips, biostatistician engineer at ISPED

- AALEN, O. O. et al. (2007). « A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment : application to AIDS prediction in England and Wales ». In : *Stat Med* 16.19, p. 2191-2210.
- AN, Q. et al. (2017). « Estimated HIV Inter-test Interval Among People at High Risk for HIV Infection in the U.S ». In : *Am J Prev Med* 53.3, p. 355-362.
- BROOKMEYER, R. et al. (1989). « Statistical methods for short-term projections of AIDS incidence ». In : *Stat Med* 8.1, p. 23-34.
- GOOD, I.J. et al. (1980). « Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data ». In : *Journal of the American Statistical Association* 75.369, p. 42-56.

- JAMIL, M. S. et al. (2017). « Effect of availability of HIV self-testing on HIV testing frequency in gay and bisexual men at high risk of infection (FORTH) : a waiting-list randomised controlled trial ». In : *Lancet HIV* 4.6, e241-e250.
- JOLY, P. et al. (1998). « A penalized likelihood approach for arbitrarily consored and truncated data : application to age-specific incidence of dementia ». In : *Biometrics* 54, p. 203-212.
- KARON, J. M. et al. (2008). « Estimating HIV incidence in the United States from HIV/AIDS surveillance data and biomarker HIV test results ». In : *Stat Med* 27.23, p. 4617-33.
- LE VU, S. et al. (2010). « Population-based HIV-1 incidence in France, 2003–08 : a modelling analysis ». In : *The Lancet, Infectious Diseases* 10.10, p. 682-87.
- MARQUARDT, D. (1963). « An algorithm for least-squares estimation of non linear parameters ». In : *SIAM Journal on Applied Mathematics* 11, p. 431-441.

- MARTY, L. et al. (2018). « Revealing geographical and population heterogeneity in HIV incidence, undiagnosed HIV prevalence and time to diagnosis to improve prevention and care : estimates for France ». In : *J Int AIDS Soc* 21.3, e25100.
- SOMMEN, C. et al. (2009). « A multistate approach for estimating the incidence of human immunodeficiency virus by using HIV and AIDS french surveillance data ». In : *Stat Med* 28, p. 1554-1568.
- SOMMEN, C. et al. (2011). « Estimation of the distribution of infection times using longitudinal serological markers of HIV : implications for the estimation of HIV incidence ». In : *Biometrics* 67.2, p. 467-75.
- SONG, R. et al. (2005). « Distribution of renewal variables when renewal process reaches a special event ». In : *Commun Stat Theory Methods* 34.8, p. 1813-1819.