Appendix

NEW METHOD FOR ESTIMATING HIV INCIDENCE AND TIME FROM INFECTION TO DIAGNOSIS USING HIV SURVEILLANCE DATA: RESULTS FOR FRANCE

Ndawinz JDA[1,2], Costagliola D[1,2,3] and Supervie V[1,2]*

[1]INSERM U943 Paris, F75013, France

[2]UPMC Univ-Paris 6, UMR S943, Paris, F75013, France

[3]APHP, Hôpital Pitié Salpêtrière, Service des maladies infectieuses et tropicales, Paris, F75013, France

*Corresponding author: virginie.supervie@inserm.fr

This appendix includes a detailed description of our modeling approach as well as supplementary table and figure.

Description of the new back-calculation model

Assigning specific test-seeking behaviors to newly diagnosed individuals according to their clinical status at initial diagnosis. For this purpose we partitioned our quarterly data on newly diagnosed individuals according to the clinical status at diagnosis. We thus created two groups. Group 1 consisted of individuals diagnosed during the primary stage of HIV infection $\left(N^{P}\right)$. By definition, individuals belonging to group 1 were diagnosed very early in the course of the infection. We assumed that these individuals decided to be tested because they experienced, and recognized, symptoms of primary HIV infection (PHI), or because they had recently been exposed to HIV. Group 2 consisted of individuals who already had AIDS at initial diagnosis of HIV infection $\left(N^{A}\right)$ and individuals who had neither AIDS nor symptoms of PHI at initial diagnosis $\left(N^{H}\right)$. We assumed that some individuals belonging to group 2 were not tested for HIV before being diagnosed with AIDS while others decided to be tested for other reasons than those of individuals belonging to group 1 (e.g. routine medical examination or onset of symptoms that occur towards the end of the incubation period). Our method does not require that all HIV-infected individuals diagnosed during PHI be classified as such (i.e. in group 1). Indeed, some HIV-infected individuals diagnosed during PHI may not report symptoms of PHI and thus will not be classified in group 1. Furthermore, notification of PHI symptoms could vary over time and between transmission categories. Our method accounts for this by not constraining to zero the probability of testing within months after HIV infection for individuals supposedly diagnosed without AIDS nor symptoms of PHI (see equation (4) in the next section). This probability is determined independently for each transmission category through the estimation of the two unknown parameters of the pre-AIDS HIV testing distribution (see section “Estimating unknown parameters”)

Specifying group-specific distributions of time interval between infection and initial diagnosis. The next step consisted of linking the observed incidence of newly diagnosed cases to the unobserved incidence of HIV infection by specifying the distribution of the interval between infection and initial diagnosis in each of the two groups. For group 2 (individuals diagnosed without symptoms of PHI), we adopted the same approach as Becker *et al.* [1]. Basically, each (unobserved) newly infected individual belonging to group 2 was allocated, independently, a duration T_A from infection to an AIDS diagnosis and a duration T_H from infection to a diagnosis of HIV infection. Thus, a newly infected individual belonging to group 2 could either have AIDS at initial diagnosis, meaning that he/she was not tested for HIV before being diagnosed with AIDS, or have a diagnosis of HIV infection at initial diagnosis, meaning that he/she did not develop AIDS before being tested. Assuming that each individual in group 2, infected during a quarter (3-month period) *s*, is independently assigned an interval between HIV infection and initial diagnosis, leads to:

$$E\left(N_t^A\right)=\sum_s E\left(I_s^{G2}\right)f_{t-s,A} \quad (1)$$

and

$$E\left(N_t^H\right)=\sum_s E\left(I_s^{G2}\right)f_{t-s,H} \quad (2)$$

where $E\left(N_t^H\right)$ (respectively $E\left(N_t^A\right)$) is the mean number of new diagnoses of HIV infection (resp. AIDS) in quarter *t*, $E\left(I_s^{G2}\right)$ is the mean number of new HIV infections in quarter *s* among individuals belonging to group 2, $f_{d,A}=P\left(T_A=d,T_H\geq d\right)$ is the probability that an infected individual has AIDS at initial diagnosis of HIV infection, *d* quarters after contracting the infection, and was not tested for HIV before developing AIDS, and $f_{d,H}=P\left(T_H=d,T_A\geq d\right)$ is the probability that an infected individual is diagnosed with HIV infection, *d* quarters after contracting the infection, without first developing AIDS.

The time from infection to AIDS diagnosis, T_A, was assumed to follow a Weibull distribution with a median of 40 quarters (i.e. 10 years) [1, 2]:

$$F_A(t) = 1 - \exp\left[-(0.0215t)^{2.516}\right] \quad (3)$$

The distribution of the rate of pre-AIDS HIV testing was assumed to depend on two unknown parameters that represent uptake of routine testing (ν) and onset of symptoms that occur towards the end of the incubation period (γ):

$$F_H(t) = 1 - \exp\left[-\nu t - \gamma(0.0215t)^{2.516}\right] \quad (4)$$

It is important to realize from equation (4) that we did not constraint to zero the probability of HIV testing within months after HIV infection for individuals supposedly diagnosed without PHI nor AIDS. In consequence, our approach accommodates the situation where some HIV-infected individuals diagnosed during the primary stage of HIV infection do not report symptoms of PHI and thus be classified as not having PHI.

The probabilities in equations 1 and 2, defined with the discrete random variables T_A and T_H, were then specified by:

$$f_{t,H} = [F_H(t+0.5) - F_H(t-0.5)][1 - F_A(t+0.5)] \quad (5)$$

and

$$f_{t,A} = [1 - F_H(t-0.5)][F_A(t+0.5) - F_A(t-0.5)] \quad (6)$$

For individuals diagnosed during the primary stage of HIV infection (group 1), the distribution of the interval from infection to initial diagnosis was assumed to be uniform from 0 to 6 months (i.e. the average length of the primary stage of HIV infection was assumed to be 3 months). Thus:

$$E\left(N_t^P\right) = \sum_s E\left(I_s^{G1}\right) f_{t-s,P} \quad (7)$$

where $E\left(N_t^P\right)$ is the mean number of individuals diagnosed with symptoms of the primary stage of HIV infection in quarter *t*, $E\left(I_s^{G1}\right)$ is the mean number of new HIV infections in quarter *s* among individuals belonging to group 1, and $f_{d,P}$ is the probability that an infected

individual is diagnosed during the primary stage of HIV infection, *d* quarters after contracting HIV.

Estimating unknown parameters. The next step consisted of estimating the unknown parameters of the model. As we split the population into two mutually exclusive groups, the mean number of new HIV infections (parameterized as a four-quarter (i.e., yearly) step-function) was estimated separately for each of the two groups. To estimate the number of new infections $\left(I_s^{G1}\right)$ among individuals belonging to group 1 (i.e. individuals diagnosed with symptoms of PHI), we first assumed that $\left(I_s^{G1}\right)$ are independent Poisson variables, then specified the likelihood of the model by using equation 7, and finally derived the estimates of the mean number of new HIV infections among individuals belonging to group 1 $\left(\hat{I}_s^{G1}\right)$ using the expectation-maximization-smoothing (EMS) algorithm [1]. For group 2, we had to estimate both the unobserved number of new HIV infections $\left(I_s^{G2}\right)$ and the two unknown parameters of the distribution of the rate of pre-AIDS HIV testing (equation 4). As in group 1, we first assumed that $\left(I_s^{G2}\right)$ are independent Poisson variables, then specified the likelihood of the model by using equations 1 and 2, before finally deriving estimates of the two unknown parameters of the distribution of the pre-AIDS HIV testing rate (equation 4) and of the mean number of new HIV infections among individuals belonging to group 2 by using the Newton-Raphson method and the EMS algorithm (see [1] for more details). It was possible to estimate the HIV incidence as well as the two parameters of the distribution of pre-AIDS HIV testing rates without encountering identifiability problems because we used additional information on HIV diagnoses, namely data on the clinical status at initial diagnosis. Finally, by adding together the estimates of the mean number of new infections in each group, $\left(\hat{I}_s^{G1}\right)$ and $\left(\hat{I}_s^{G2}\right)$, we obtained an estimate of the mean number of new HIV infections.

Estimates of the overall (non-stationary) distribution of the time interval between infection and initial diagnosis. Using the group-specific estimates of the number of new HIV infections and the group-specific distributions of the interval between infection and initial diagnosis, we calculated the overall distribution of the interval between infection and initial diagnosis as follows:

$$M_{s,d} = \hat{I}_s^{G1} f_{d,P} + \hat{I}_s^{G2} \left(f_{d,A} + f_{d,H} \right) \quad (8)$$

where $M_{s,d}$ is the calculated number of individuals infected in quarter *s* who are initially diagnosed *d* quarters after being infected. It is important to note that, although the group-specific distributions are stationary, the overall distribution $\left(M_{s,d}\right)_d$ varies with time since two of its arguments ($\hat{I}_s^{G1}$ and $\hat{I}_s^{G2}$) vary with time. Changes in test-seeking behaviors over time will affect these two arguments and will thus be accounted in our model. For instance, if individuals test for HIV more often due to new guidelines promoting early HIV diagnosis, then more individuals will be diagnosed in early stages of the HIV infection. Hence, the number of people diagnosed with symptoms of PHI (i.e. group 1) will increase, while the size of group 2 (consisting of individuals diagnosed without symptoms of PHI) will decrease accordingly. As a result, the number of new infections among individuals belonging to group 1 ($\hat{I}_s^{G1}$) will increase, the number of new infections among individuals belonging to group 2 ($\hat{I}_s^{G2}$) will decrease, and the interval between infection and initial diagnosis will become shorter. Thus, our method allows for accounting and tracking for changes in test-seeking behaviors over calendar time without having to estimate time-calendar specific diagnosis probabilities as done in previous back-calculation models [3].

Estimates and their precision. Missing data were estimated by using multiple imputation techniques. Ten imputed datasets were generated by using the missing-at-random assumption. From each of the ten datasets we generated 200 bootstrap samples [4] and thus

obtained 2000 bootstraps samples. Numbers were then adjusted for under-reporting and the reporting delay. The under-reporting rate and reporting delay are estimated every year by the "Institut de Veille Sanitaire" (InVS), which manages the HIV and AIDS registries [5]. No variability deriving from the under-reporting and reporting delay was included in the model since we corrected the data using single point estimates (mean values). Figure S2 shows the contribution of each source of uncertainty (missing data versus under-reporting and reporting delay) in the dataset. We then run our model using these 2000 simulated datasets and calculated mean estimates and 95% confidence intervals (CI).

Calculation of the percentage of late diagnosis among individuals diagnosed with HIV infection in year T. We defined late diagnosis as a diagnosis occurring more than 8 years (32 quarters) after the infection, 8 years being the mean time required for the CD4 cell count to fall below the critical threshold of 200/mm^3 [6]. The number of individuals with a late HIV diagnosis in year *T* ($\hat{L}_T$) was calculated as follows:

$$\hat{L}_T = \sum_{j=1}^{4} \sum_{s=1}^{t_j-32} \hat{I}_s^{G2} \left(f_{t_j-s,A} + f_{t_j-s,H} \right)$$

where t_1, t_2, t_3 and t_4 are the 4 quarters of year *T*, $\hat{I}_s^{G2}$ is the estimated number of new infections in quarter *s* among individuals belonging to group 2, $f_{d,A}$ is the probability that an infected individual will have AIDS at initial diagnosis *d* quarters after contracting the infection (see equation 6), and $f_{d,H}$ is the probability that an infected individual will be AIDS-free at initial diagnosis of HIV infection (see equation 5), *d* quarters after contracting the infection.

We then estimated the percentage of late diagnosis among individuals diagnosed in year *T* ($\hat{p}_T$) by dividing $\hat{L}_T$ by the estimated number of individuals newly diagnosed in year *T*:

$\hat{p}_T = \dfrac{\hat{L}_T}{\sum_{j=1}^{4} \hat{N}_{t_j}^A + \hat{N}_{t_j}^H + \hat{N}_{t_j}^P}$ where $\hat{N}_{t_j}^A$, $\hat{N}_{t_j}^H$ and $\hat{N}_{t_j}^P$ were calculated with equations 1, 2 and 7.

We wrote a computer program to run the model.

Implementation of the method. Our method has been implemented in the C programming language, and is not currently available in a user-friendly format. In principle, we are willing to consider collaborations with individual countries to implement the method.

References

1. Becker NG, Lewis JJ, Li Z, McDonald A. **Age-specific back-projection of HIV diagnosis data**. *Stat Med* 2003;**22**:2177-2190.
2. Brookmeyer R, Goedert JJ. **Censoring in an epidemic with an application to hemophilia-associated AIDS**. *Biometrics* 1989;**45**:325-335.
3. Sweeting MJ, De Angelis D, Aalen OO. **Bayesian back-calculation using a multi-state model with application to HIV**. *Stat Med* 2005;**24**:3991-4007.
4. Efron B, Tibshirani RJ. **An Introduction to the Bootstrap**. London: Chapman and Hall/CRC; 1994.
5. Semaille C, Cazein F, Pillonel J, Lot F, Le Vu S, Pinget R*, et al.* **Four years of surveillance of recent HIV infections at country level, France, mid 2003 - 2006: experience and perspectives**. *Euro Surveill* 2008;**13**.
6. Lodi S, Phillips A, Touloumi G, Geskus R, Meyer L, Thiébaut R*, et al.* **Proportion of Individuals Likely to Need Treatment for CD4 Thresholds <200,<350, and <500 cells/µL**. *17th Conference on Retroviruses and Opportunistic Infections*. San Francisco; 2010.
7. Le Vu S, Le Strat Y, Barin F, Pillonel J, Cazein F, Bousquet V*, et al.* **Population-based HIV-1 incidence in France, 2003-08: a modelling analysis**. *Lancet Infect Dis* 2010;**10**:682-687.
8. Anonymous. **Premiers resultat de l'Enquête Presse Gay 2004**. *Saint-Maurice: Institut de Veille Sanitaire* 2005.
9. Wiessing L, Nardone A. **Ongoing HIV and viral hepatitis infections in IDUs across the EU, 2001-2005**. *Euro Surveill* 2006;**11**:E061123 061122.
10. UNAIDS. **Epidemiological Fact Sheet on HIV and AIDS for France**. 2008.
11. Desenclos J, Costagliola D, Commenges D, Lellouche J. **La prévalence de la séropositivité VIH en France**. *Bull Epidemiol Hebd* 2005;**11**:41-44.

Table S1. Estimated size of the populations at risk of HIV infection in France

	Estimated population size [7]	Estimated prevalence (%)	Estimated size of the population at risk of HIV infection
Men who have sex with men	329500	13.0 [8]	286665
Non French-national heterosexual women	1739760	Not available	~1739760[a]
Non French-national heterosexual men	1884740	Not available	~1884740[a]
French-national heterosexual women	18363590	Not available	~18363590[a]
French-national heterosexual men	18848440	Not available	~18848440[a]
Injecting drug users	81000	12.2 [9]	71118

[a]At the end of 2007, ~140000 persons were living with HIV in France [10], ~40% of whom (i.e. 56000) were non injecting drug user heterosexuals [11]. Since the size of the heterosexual population is very large compared to the estimated number of heterosexuals living with HIV, we assumed that the number of heterosexuals at risk of HIV infection is equal to the number of heterosexuals in the general population.

Figure S1. Estimated annual number of new infections and 95% bootstrap confidence intervals (-----) in France. [a]MSM: men who have sex with men (all nationalities); [b]IDUs: injecting drug users (all nationalities, both sexes).

Figure S2. Annual number of newly diagnosed cases of HIV infection per exposure category. Lines with circles represent raw data collected by the French National Institute for Public Health Surveillance (InVS), plain lines represent data corrected for missing entries and dashed lines represent data corrected for missing entries, underreporting, and the reporting delay. [a]all nationalities; [b]all nationalities, both sexes.

Figure S1

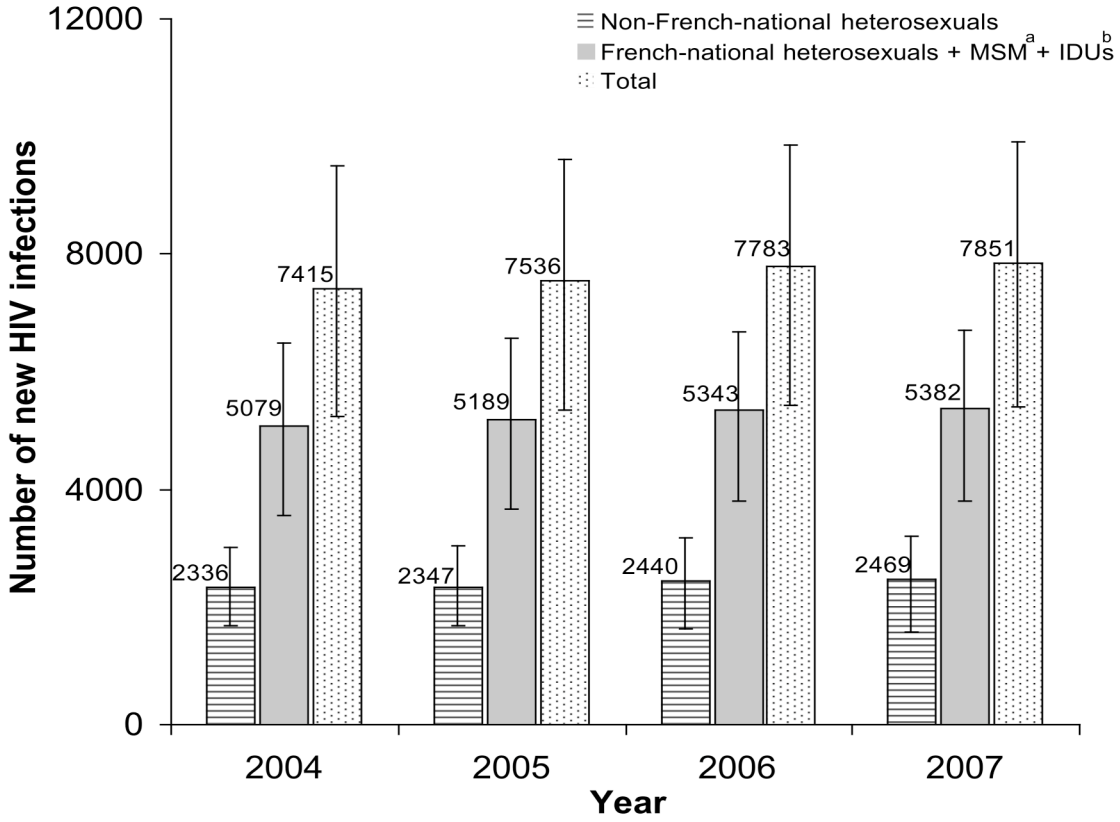
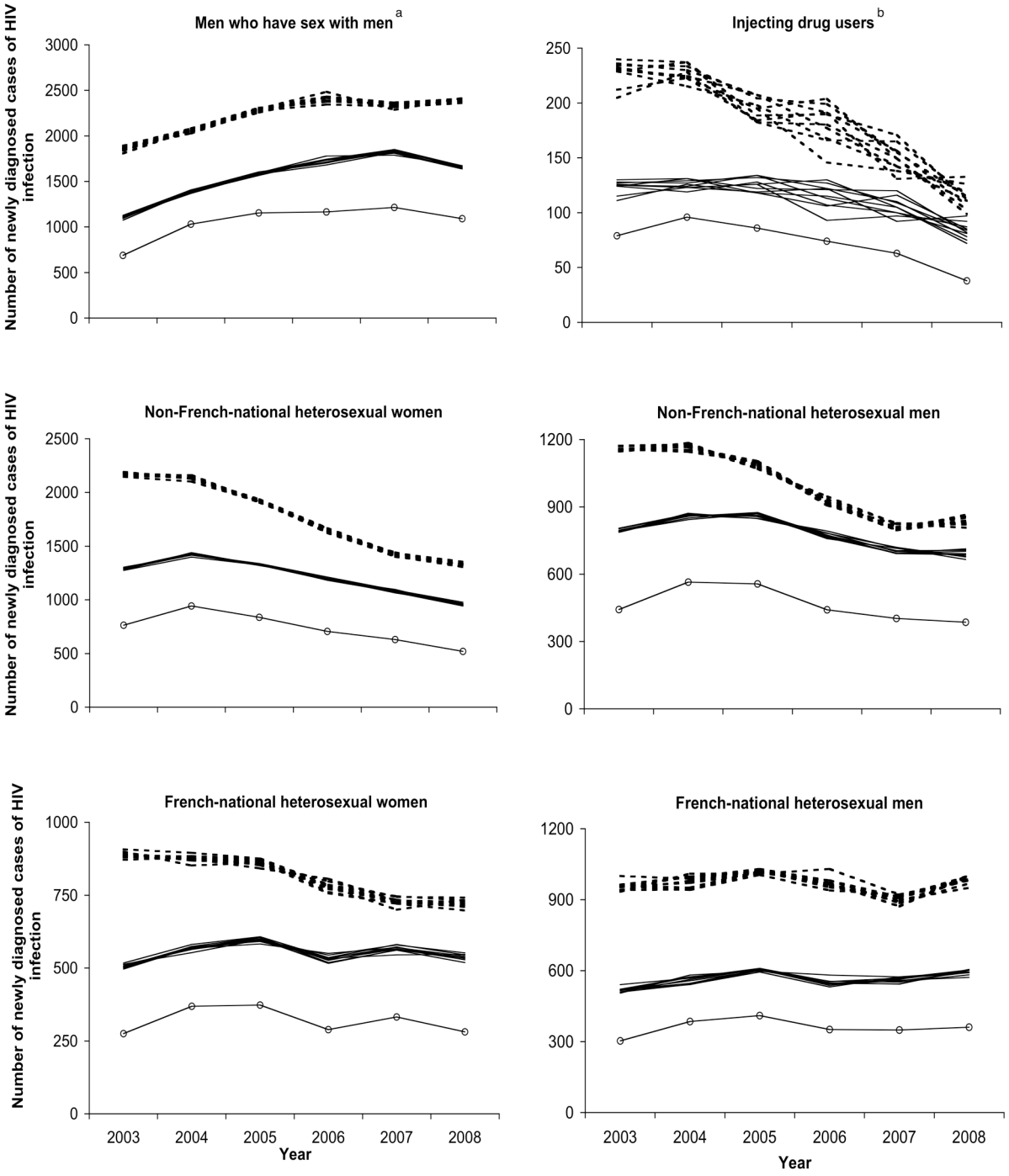

12000
8000
4000
0
Number of new HIV infections
Non-French-national heterosexuals
French-national heterosexuals + MSM[a] + IDUs[b]
Total
2336
5079
7415
2347
5189
7536
2440
5343
7783
2469
5382
7851
2004
2005
2006
2007
Year

Figure S2


Men who have sex with men a
Injecting drug users b
Non-French-national heterosexual women
Non-French-national heterosexual men
French-national heterosexual women
French-national heterosexual men
Number of newly diagnosed cases of HIV infection
Year
2003
2004
2005
2006
2007
2008