Journal Title
XX(X):2–28
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

Determining the likely place of HIV acquisition for migrants in Europe combining subject-specific information and biomarkers data

Nikos Pantazis[1], Christos Thomadakis[1], Julia del Amo[2], Debora Alvarez-del Arco[2], Fiona M Burns[3 4], Ibidun Fakoya[3], Giota Touloumi[1], on behalf of the aMASE and CASCADE study groups

Abstract

In most HIV-positive individuals, infection time is only known to lie between the time an individual started being at risk for HIV and diagnosis time. However, a more accurate estimate of infection time is very important in certain cases. For example, one of the objectives of the aMASE study was to determine if HIV-positive migrants, diagnosed in Europe, were infected pre- or post-migration. We propose a method to derive subject-specific estimates of unknown infection times using information from HIV biomarkers' measurements, demographic, clinical and behavioral data. We assume that CD4 cell count (CD4) and HIV-RNA viral load (VL) trends after HIV infection follow a bivariate linear mixed model. Using post-diagnosis CD4 and VL measurements and applying the Bayes' rule, we derived the posterior distribution of the HIV infection time, whereas the prior distribution was informed by AIDS status at diagnosis and behavioral data. Parameters of the CD4-VL and time-to-AIDS models were estimated using data from a large study of individuals with known HIV infection times (CASCADE). Simulations showed substantial predictive ability (e.g. 84% of the infections were correctly classified as pre- or post-migration). Application to the aMASE study (n=2,009) showed that 47% of African migrants and 67% to 72% of migrants from other regions were most likely infected post-migration. Applying a Bayesian method based on bivariate modeling of CD4 and VL, and subject-specific information, we found that the majority of HIV-positive migrants in aMASE were most likely infected after their migration to Europe.

Keywords
Bayes rule, HIV, Infection, Migrants, Prediction

1 Introduction

Infection with the Human Immunodeficiency Virus (HIV) causes a slow deterioration of the immune system which progressively increases the risk of opportunistic infections and specific malignancies. These typically occur years after HIV acquisition and are referred to as AIDS defining conditions. Due to the long and relatively asymptomatic period between HIV acquisition and AIDS onset, many HIV-positive individuals remain unaware of their infection for many years. In the absence of previous negative HIV tests, a positive test carries little information regarding the timing of the infection. However, knowing the time of infection is important, as it can inform prevention strategies. Prolonged periods of undiagnosed (and thus untreated) HIV infection are associated with high risk of HIV transmission and poorer prognosis.

Knowledge of infection time is of particular importance for HIV-positive migrants as, by comparing it with migration timing, one could infer if an HIV diagnosed migrant was infected pre- or post- migration. Determining the likely place of HIV infection for migrants in Europe is key for designing adequate HIV prevention and testing strategies. The aMASE (Advancing Migrant Access to Health Services in Europe) study[1] was the first European level study which focused on the identification of barriers in HIV prevention, diagnosis and treatment along with the determination of the likely country of HIV acquisition in multiple migrant populations.

The issue of estimating HIV infection time has been already addressed by many researchers but in most cases the focus was on estimating population average parameters (i.e. HIV incidence curves) rather than on making individual level estimation.[2–7]. There are though a few studies that have proposed methods through which the gap between HIV infection and diagnosis can be estimated at the individual level[8–13]. However, the application setting in most of these studies was not only a cohort comprising individuals with unknown HIV infection time (seroprevalent cases) but also a proportion of individuals with known infection times (seroincident cases). The availability of raw data from seroincident cases is crucial for such methods thus they cannot be applied in more general situations.

[1]Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens. Mikras Asias 75, Athens, Greece, 115 27
[2]National Centre of Epidemiology, Instituto de Salud Carlos III, Madrid, Spain
[3]Centre for Sexual Health and HIV Research, Research Department of Infection and Population Health, University College London, London, UK
[4]Royal Free London NHS Foundation Trust, London UK

Corresponding author:
Nikos Pantazis, National and Kapodistrian University of Athens. Mikras Asias 75, Athens, Greece, 115 27.
Phone: +30-210-7462088, Fax: +30-210-7462205
Email: npantaz@med.uoa.gr

Motivated by the need to estimate the unknown HIV infection times of aMASE participants and given the lack of appropriate methods which could be optimally used in this setting, we developed a new procedure for individual-level estimation of HIV infection time in seroprevalent cases. The proposed method uses information on a wide range of subject-specific characteristics, including migration history, and from routine measurements of HIV disease biomarkers. These data are used in conjunction with information on the time trends of such biomarkers after HIV infection, along with their relation with subject-specific covariates. In our application, this information is derived from CASCADE[14], a large multi-cohort study comprising individuals with well estimated dates of infection, in the form of a set of estimated parameters which can be re-used in a large variety of other settings. Thus, the main idea behind the method we propose, is to use results from a model on the distribution of such markers given the time elapsed since HIV infection and reverse the conditioning through Bayes theorem to derive the distribution of the elapsed time given a set of markers' measurements. The required prior distribution of the unknown infection time can be informed by the diagnosis time (i.e. infection must have happened before diagnosis), knowledge about the onset of the HIV epidemic (i.e. infection must have happened after the onset of the epidemic) and the nature of possible routes of transmission (sexual contact or intravenous drug use). Refinements of this prior distribution can be made by taking into account experts' knowledge on the effects of risky behaviors. Our method is partly related to the method proposed by Berman[8] but it also has some similarities with the method proposed by Rice et al.[13] as that method also uses external information on the evolution of routinely measured biomarkers during untreated HIV infection. However, our method handles estimation in a formal way, is more flexible, can be extended to accommodate additional information and explicitly quantifies the uncertainty around the estimated HIV infection times.

In Section 2 we briefly describe the aMASE and CASCADE studies while the proposed method and the estimating procedure are presented in Section 3. Results from the application of the method to simulated and real data from aMASE are presented in Sections 4 and 5, respectively. Finally, in Section 6 we summarize our results and we compare our method with other previously proposed ones discussing its advantages and limitations.

2 Motivating studies

2.1 The aMASE study

The protocol of the aMASE study has been described in detail elsewhere[1]. Briefly, two multi-country cross-sectional studies were carried out, one in the general migrant population and one in migrants diagnosed with HIV (clinical aMASE). In this work we focus on the latter in which HIV-positive migrants were recruited from HIV clinics across 9 European countries from July 2013-July 2015. Patients' inclusion criteria were to be diagnosed with HIV in the last five years, to be over 18 years old, to be living outside the country of birth and residing in one of the nine participating countries for at least

6 months and to be able to complete the questionnaire in one of the 14 languages the questionnaire was available in. aMASE was part of the European network of excellence on HIV research, (www.eurocoord.net). Ethical approval for the aMASE study was received separately in each participating country.

One of the primary goals of the clinical aMASE study was to determine whether HIV acquisition occurred pre- or post-migration by estimating the likely time of HIV infection. A questionnaire was designed together with the community to gather information on risky behaviors and other critical epidemiological variables, along with the migration paths including the dates of arrival in the host country. Using such information one can increase the likelihood of a correct guess regarding the place of HIV infection: for example if an individual, without other high risk behaviors, started injecting drugs at the destination country, it is more likely that infection happened after migration.

A second questionnaire including clinical data was also completed by the research team. This included among others, the date of last negative HIV test when available, whether an AIDS occurrence had happened within 3 months of HIV diagnosis as well as data on the two most relevant biomarkers of HIV infection: the CD4 cell count and the HIV-RNA viral load. CD4 cell count reflects the immunological status of the patient whereas HIV-RNA quantifies the amount of the virus in the blood. Both biomarkers show consistent trends since HIV infection, in the sense that their rate of change is usually smooth over time, and have been subjected to extensive research.

2.2 The CASCADE collaboration

CASCADE is a collaboration of individual HIV cohorts that include subjects with well-estimated dates of HIV seroconversion (i.e. the time in which a person first develops antibodies for HIV). It should be noted that seroconversion usually occurs about 10 days after infection thus hereafter both events will be referred to as HIV infection. In brief, its aim was to combine data from different seroincident HIV cohorts in order to deal with scientific questions that cannot be fully addressed by the individual cohorts. For the majority (85.1%) of the CASCADE subjects, the infection date was estimated by the midpoint between the last documented negative and first positive HIV test date, with the time in between being less than 3 years (Median: 0.91; IQR: 0.45-1.58 years), whereas for the remaining 14.9% of patients other methods, of higher accuracy (e.g. based on laboratory test results), were used [14].

Demographic data such as sex, age, region of origin and mode of infection, along with repeated measurements of CD4 cell count and viral load are available in the CASCADE database. These two markers have been shown to be the most important predictors of disease progression and have been routinely used by clinicians. Since the infection dates are known for the CASCADE patients, we can estimate the biomarkers' trajectories since infection by using appropriate statistical models. In the absence of antiretroviral therapy (ART) and before the AIDS onset, the number of CD4 cells decreases, since CD4 cells constitute the main target of the virus. This decrease has been shown to be approximately linear on the square root or fourth root scale [15]. On the other hand, viral load trends are

not linear over time and are characterized by an exponential-like decay for approximately the first year after infection followed by a slow subsequent increase on the $\log_{10}$ scale [16]. It should be noted that since we focus on estimating markers' evolution during natural history (i.e. the disease evolution from infection to AIDS or death), measurements taken after ART initiation or AIDS onset need to be excluded. Post-ART measurements have to be excluded as ART successfully increases the number of CD4 cells up to almost normal levels, whereas measurements after AIDS onset are excluded since the CD4 decline may accelerate after that point [17]. To appropriately model CD4 cell counts and viral load measurements, one should use methods that account for the correlation in repeated measurements, e.g. linear mixed models (LMM) [18,19]. Also, as these markers are usually moderately correlated, they should ideally be modeled together.

3 Proposed method

3.1 Estimating infection time through Bayes Theorem

Let $\mathbf{Y}_i^c = (Y_{i1}^c, \ldots, Y_{in_i^c}^c)^\top$ and $\mathbf{Y}_i^r = (Y_{i1}^r, \ldots, Y_{in_i^r}^r)^\top$ be the CD4 cell count and viral load measurements during the HIV natural history collected at times $\mathbf{t}_i^c = (t_{i1}^c, \ldots, t_{in_i^c}^c)^\top$ and $\mathbf{t}_i^r = (t_{i1}^r, \ldots, t_{in_i^r}^c)^\top$ since HIV infection, respectively, on a subject i. Markers' values are only observed after the HIV diagnosis, thus the time intervals $\mathbf{t}_i^c$ and $\mathbf{t}_i^r$ are generally unknown in seroprevalent cohorts. Let $\mathbf{d}_i^c = (d_{i1}^c, \ldots, d_{in_i^c}^c)^\top$ and $\mathbf{d}_i^r = (d_{i1}^r, \ldots, d_{in_i^r}^c)^\top$ denote the time intervals from the HIV diagnosis to the date on which the two biomerkers are measured, and w_i denote the time gap between HIV infection and diagnosis. Then it follows that $t_{ij}^c = d_{ij}^c + w_i$ and $t_{ij}^r = d_{ij}^r + w_i$, with w_i being unknown but $\mathbf{d}_i^c$ and $\mathbf{d}_i^r$ being always observed. We assume that a bivariate model $f(\mathbf{y}_i^c, \mathbf{y}_i^r | \mathbf{t}_i^c, \mathbf{t}_i^r)$ correctly characterizes the evolution over time of both markers, with its parameters being known. This model will be described in detail later on in this paper.

Given that our ultimate goal is to estimate w_i and $\mathbf{d}_i^c$ and $\mathbf{d}_i^r$ are observed, we express the distribution of the biomarkers conditionally on w_i by simply replacing t_{ij}^c and t_{ij}^r with $d_{ij}^c + w_i$ and $d_{ij}^r + w_i$, respectively. Considering w_i as another unknown quantity within a Bayesian framework, we need to assign a prior distribution for w_i. As it does not make much sense to define w_i on the whole real line, we need to define a date after which subjects start being at risk for HIV infection. Based on prior knowledge on the HIV epidemic, we assumed that the onset of risk for HIV is the maximum of (a) the date of a documented previous HIV negative test, (b) the age of 10 years (since we exclude mother to child transmissions) and (c) the presumed date of the start of the epidemic (i.e. January 1, 1980). Defining u_i as the time interval from risk onset to HIV diagnosis, we assume a uniform prior distribution for w_i over the interval $(0, u_i)$. Markers' measurements, times and their relations are visually depicted in Figure 1.

Given the observed measurements $(\mathbf{y}_i^c, \mathbf{y}_i^r)$, we reverse the problem deriving the posterior distribution of the unknown w_i conditionally on $(\mathbf{y}_i^c, \mathbf{y}_i^r)$. This can be easily carried out through Bayes Theorem. Letting $\mathbf{y}_i^\top = (\mathbf{y}_i^c, \mathbf{y}_i^r)$ be the observed

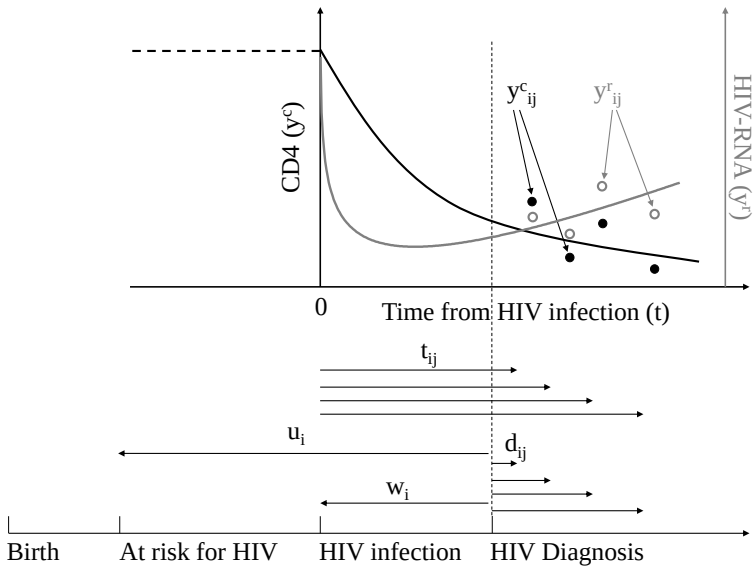

Figure 1. An artificial example for an HIV-positive individual showing the sequence of important relevant events (birth, beginning of being at risk for acquiring HIV, and HIV infection and diagnosis), along with the corresponding average evolution of the two HIV biomarkers (CD4 cell count in black and HIV-RNA viral load in gray) and their measurements after the HIV diagnosis.

measurement of both markers, the posterior distribution of w_i becomes

$$f(w_i|\mathbf{y}_i) = \frac{f(\mathbf{y}_i|w_i)f(w_i)}{\int_0^{u_i} f(\mathbf{y}_i|w_i)f(w_i)dw_i}, \quad 0 < w_i < u_i, \tag{1}$$

where the dependence on the parameters of the measurement model is suppressed for ease of notation and $f(\mathbf{y}_i) = \int_0^{u_i} f(\mathbf{y}_i|w_i)f(w_i)dw_i$ is a normalizing constant. To estimate the infection date, we can use any measure of central tendency such as the posterior mode, the posterior mean and the posterior median.

The posterior mode is defined as $\text{argmax}_w\, f(w_i|\mathbf{y}_i) = \text{argmax}_w\, f(\mathbf{y}_i|w_i)f(w_i)$, as the normalizing constant does not depend on w_i. We used the general-purpose optimizer `optim` in `R`[20] to maximize the logarithm of the posterior distribution. More specifically, we used the BFGS algorithm with lower and upper bounds, with the bounds reflecting the fact that infection must have occurred after the assumed date of starting being at risk for HIV and the HIV diagnosis which corresponds to the "L-BFGS-B" method in `optim`.

The posterior mean is defined as

$$\int_0^{u_i} w_i f(w_i|\mathbf{y}_i)dw_i = \frac{\int_0^{u_i} w_i f(\mathbf{y}_i|w_i)f(w_i)dw_i}{\int_0^{u_i} f(\mathbf{y}_i|w_i)f(w_i)dw_i}.$$

Note that the posterior mean is the ratio of two integrals, where, unlike the posterior mode, the normalizing constant needs to be computed. To calculate these integrals we used quadrature methods that have been implemented within the function `integrate` in `R`[20].

The posterior median M_i is defined as the solution to

$$\int_0^{M_i} f(w_i|\mathbf{y}_i)dw_i = 0.5. \tag{2}$$

We numerically solved (2) by using the root solver `uniroot` in `R` combined with the `integrate` function to calculate the integrals involved[20].

Besides measures of central tendency, the proposed method allows for the estimation of probabilities of infection occurring before or after a specific point in time. For example, for the purposes of the aMASE study, this time point corresponds to an individual's migration date. Thus, let m_i be the time from migration to diagnosis. The objective is to infer whether infection happened in the country of origin ($w_i > m_i$) or in the country of destination ($w_i < m_i$). To quantify such statements in terms of probabilities, we need to derive the posterior probabilities of infection pre- or post-migration; for example, the pre-migration posterior probability is

$$\pi_i = \mathrm{P}\left(w_i > m_i\right) = \int_{m_i}^{u_i} f(w_i|\mathbf{y}_i)dw_i, \tag{3}$$

whereas the probability of infection post migration is $1 - \pi_i$. To calculate these probabilities we again used the `integrate` function in `R`[20].

The methods described up to now deal with estimation of HIV infection times or related probabilities at the individual level. However, there are certain cases in which the interest lies in deriving and comparing population-average parameters such as the mean time gap between HIV infection and diagnosis or the proportion of migrants acquiring HIV post migration. Ignoring the uncertainty of point estimates derived from the posterior distribution $f(w_i|\mathbf{y}_i)$ (e.g. the posterior mean) and carrying out all subsequent analyses treating these estimates as known, is highly likely to lead to over-precise inferences. To account for the uncertainty of point estimates, we propose a multiple imputation approach. That is, for each individual, we simulate K random infection times from his/her posterior distribution, $\omega_i^k \sim f(w_i|\mathbf{y}_i)$. As the posterior distribution of ω_i is defined on a finite interval, random draws from this distribution can be easily obtained through rejection sampling. In this way, K pseudo-complete data set are constructed. Results from any model fitted to each pseudo-complete data set (e.g. a simple normal model to estimate the mean time from infection to diagnosis) can be combined using Rubin's rules[21].

Particularly for the purposes of the aMASE study, the goal is to estimate the probability of post-migration HIV acquisition (overall or by specific characteristics) while taking into account the uncertainty of HIV infection times at the individual level. In the specific example there is a subset of individuals that can be certainly classified as infected pre- or post-migration based on their HIV testing history data. For the remaining individuals, for

whom the posterior distribution of the infection time is derived, we treat infection times as "missing data". Let δ_i be the indicator of a post-migration infection. For those known to have been infected pre or post-migration this is either 0 or 1, respectively. For the remaining, we followed the procedure described above, that is, we simulated K random draws from their posterior distributions, $\omega_i^k \sim f(w_i|\mathbf{y}_i)$. Given the simulated ω_i^k, the indicator variable of post-migration infection can be defined as $\delta_i^k = I(\omega_i^k < m_i)$, based on the kth imputation. Then, a logistic model was fitted to each pseudo-complete data set, with the results combined using Rubin's rules[21].

3.2 A model for the evolution of CD4 and viral load

In the previous subsection we assumed a model for the evolution of CD4 cell counts and HIV-RNA viral load during the HIV natural history (i.e. before the initiation of ART or AIDS onset). More specifically, this model was assumed to be a bivariate linear mixed model (BLMM) for the forth root CD4 counts and the viral load on the $\log_{10}$ scale of the form

$$\begin{pmatrix} \mathbf{Y}_i^c \\ \mathbf{Y}_i^r \end{pmatrix} = \begin{pmatrix} \mathbf{X}_i^c & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_i^r \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^c \\ \boldsymbol{\beta}^r \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_i^c & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_i^r \end{pmatrix} \begin{pmatrix} \mathbf{b}_i^c \\ \mathbf{b}_i^r \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_i^c \\ \boldsymbol{\epsilon}_i^r \end{pmatrix}, \tag{4}$$

where $\mathbf{X}_i^c$ and $\mathbf{X}_i^r$ are the $n_i^c \times p^c$ and $n_i^r \times p^r$ design matrices associated with the fixed effects $(\boldsymbol{\beta}^c, \boldsymbol{\beta}^r)$ of the two markers, respectively, and $\mathbf{Z}_i^c$ and $\mathbf{Z}_i^r$ are the $n_i^c \times q^c$ and $n_i^r \times q^r$ design matrices associated with the random effects $(\mathbf{b}_i^c, \mathbf{b}_i^r)$ of the two markers for the ith subject. Also, $\boldsymbol{\epsilon}_i^c$ and $\boldsymbol{\epsilon}_i^r$ are the within-subject residuals for both markers respectively, assumed to be normally distributed with zero mean and covariance matrices $\sigma_c^2 \mathbf{I}_{n_i^c}$ and $\sigma_r^2 \mathbf{I}_{n_i^r}$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. To allow for correlation between the two biomarkers, we assume that the random effects $\mathbf{b}_i^c$ and $\mathbf{b}_i^c$ jointly follow the multivariate normal distribution with zero mean and covariance matrix

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}^c & \mathbf{D}^{cr} \\ \mathbf{D}^{rc} & \mathbf{D}^r \end{pmatrix},$$

with the submatrices $\mathbf{D}^c$ and $\mathbf{D}^r$ denoting the covariance matrices of the random effects for the CD4 cell count and viral load levels respectively, and $\mathbf{D}^{rc}$ the covariances between the random effects of each marker. The above assumptions imply that the marginal distribution of the observed measurements over time since HIV infection is the following multivariate normal:

$$\begin{pmatrix} \mathbf{Y}_i^c(\mathbf{t}_i^c) \\ \mathbf{Y}_i^r(\mathbf{t}_i^r) \end{pmatrix} \sim N(\boldsymbol{\mu}_i(\mathbf{t}_i), \mathbf{V}_i(\mathbf{t}_i)), \tag{5}$$

where the mean vector $\boldsymbol{\mu}_i(\mathbf{t}_i)$ and the variance-covariance matrix $\mathbf{V}_i(\mathbf{t}_i)$ are equal to

$$\boldsymbol{\mu}_i(\mathbf{t}_i) = \begin{pmatrix} \mathbf{X}_i^c(\mathbf{t}_i^c) & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_i^r(\mathbf{t}_i^r) \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^c \\ \boldsymbol{\beta}^r \end{pmatrix}, \text{ and}$$

$$\mathbf{V}_i(\mathbf{t}_i) = \begin{pmatrix} \mathbf{Z}_i^c(\mathbf{t}_i^c) & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_i^r(\mathbf{t}_i^r) \end{pmatrix} \mathbf{D} \begin{pmatrix} \mathbf{Z}_i^c(\mathbf{t}_i^c) & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_i^r(\mathbf{t}_i^r) \end{pmatrix}^\top + \begin{pmatrix} \sigma_c^2 \mathbf{I}_{n_i^c} & \mathbf{0} \\ \mathbf{0} & \sigma_r^2 \mathbf{I}_{n_i^r} \end{pmatrix} \Bigg),$$

respectively. To estimate the parameters of this model, we used the CASCADE data in which infection dates are well estimated. In order to make the specific model applicable to the aMASE study participants, we included all covariates which were available in both datasets and could have confounding effects. More specifically the model included adjustments for sex, age at infection evaluated through linear splines with knots at 25, 35 and 45 years, region of birth (Africa, Europe, Asia, America), mode of infection (through sex between men, injection of drugs, heterosexual contact, and through other routes), and calendar year of infection modelled via linear splines with knots on 1/1/1996 and 1/1/2002. Calendar times of infection were incorporated in the model as some studies have shown that CD4 cell count at baseline may decrease over time, while viral load at baseline may increase over time[22,23]. We assumed that the mean CD4 cell count evolution is linear over time on the fourth root scale, with a random intercept and slope to capture the correlation of repeated CD4 measurements. It should be noted that the fourth root transformation resulted in a better fit to the data compared to the more frequently used square root transformation. For the mean viral load evolution, we worked on the $\log_{10}$ scale using a time term along with a $\log(t_{ij}^r + c)$ term. Three random effects were used: a random intercept accounting for variability in baseline viral load measurements and two other random effects associated with the time terms accounting for the variability and correlation in subsequent viral load measurements. The parameter c was estimated after fitting the model over a grid set of values of c. The best model had $c = 0.013$ and turned out to be superior to a model with natural cubic splines. The inclusion of the $\log(t_{ij}^r + c)$ term allowed us to capture the non-linear evolution of HIV-RNA viral load during natural history as viral load tends to reach very high levels soon after the infection, then it drops to some minimum levels after almost a year increasing subsequently but at a much slower rate (see also Figure 1). It needs to be emphasized that the model describes the markers' evolution during the HIV natural history, thus measurements taken after AIDS onset or ART initiation are irrelevant.

As part of a sensitivity analysis, when applying the proposed method to the aMASE study data, the parameters of model (4), were re-estimated after excluding data from CASCADE participants whose infection date was determined by the midpoint method and had relatively wide (i.e. $\geq$1 year) HIV test intervals.

We used the `lme` function of the `nlme` package in `R` to fit the biomarkers model.

3.3 Incorporating additional information on AIDS status

It is well known that HIV infection, in the absence of treatment, leads to development of AIDS within 8-10 years on median[10]. Thus, the presence or absence of AIDS-defining symptoms at HIV diagnosis carries additional information regarding the time gap between HIV infection and diagnosis. For example, a person diagnosed without AIDS symptoms would be more likely to have been infected recently compared to a similar person diagnosed while having already progressed to AIDS. Thus, we can refine our method in order to take into account an individual's AIDS status, given the available epidemiological information on the distribution of time between HIV infection and AIDS onset. We used the CASCADE data truncated on 1/1/1996 to derive this distribution

given sex, region of birth, risk group and age at infection. We truncated data after 1/1/1996, since after that point, effective antiretroviral therapy substantially reducing the probability of developing clinical AIDS became widely available. We used a Weibull proportional hazards model of the form

$$\lambda(t|\mathbf{X}_i^s) = \kappa \exp\left\{\mathbf{X}_i^{s\top}\boldsymbol{\beta}^s\right\} t^{\kappa-1},$$

with t denoting the time since HIV infection, $\mathbf{X}_i^s$ the implied design matrix and κ a parameter controlling the shape of the Weibull distribution. As a result, the survival function induced by this model is $S(t|\mathbf{X}_i^s) = \exp\{-\lambda_i t^\kappa\}$, with $\lambda_i = \exp\left\{\mathbf{X}_i^{s\top}\boldsymbol{\beta}^s\right\}$. The survival estimates from the Weibull model were compared with the Kaplan-Meier estimates yielding very similar results suggesting adequate model fit.

Thus, for a subject known to be alive, AIDS-free and not on therapy until some time d_i since HIV diagnosis, the posterior distribution of w_i becomes

$$f(w_i|\mathbf{y}_i, T_i > d_i + w_i) = \frac{f(\mathbf{y}_i|w_i)S(d_i + w_i|\mathbf{X}_i^s)f(w_i)}{\int_0^{u_i} f(\mathbf{y}_i|w_i)S(d_i + w_i|\mathbf{X}_i^s)f(w_i)dw_i}, \tag{6}$$

where T_i is a latent variable representing the time from HIV diagnosis to AIDS onset.

Suppose now that a subject diagnosed with HIV developed AIDS at some time d_i after HIV diagnosis, while not on therapy within that period of time. For that individual, we know that the time between HIV infection and progression to AIDS is $d_i + w_i$, which means that the posterior distribution of w_i becomes

$$f(w_i|\mathbf{y}_i, T_i = d_i + w_i) = \frac{f(\mathbf{y}_i|w_i)\lambda(d_i + w_i|\mathbf{X}_i^s)S(d_i + w_i|\mathbf{X}_i^s)f(w_i)}{\int_0^{u_i} f(\mathbf{y}_i|w_i)\lambda(d_i + w_i|\mathbf{X}_i^s)S(d_i + w_i|\mathbf{X}_i^s)f(w_i)dw_i}. \tag{7}$$

If the subjects had AIDS at HIV diagnosis, there were no markers' measurements taken during the natural history, that is prior to clinical AIDS development, and thus the distribution of w_i becomes

$$f(w_i|T_i = w_i) = \frac{\lambda(w_i|\mathbf{X}_i^s)S(w_i|\mathbf{X}_i^s)f(w_i)}{\int_0^{u_i} \lambda(w_i|\mathbf{X}_i^s)S(w_i|\mathbf{X}_i^s)f(w_i)dw_i}, \tag{8}$$

We used the `survreg` function of the `survival` package in `R` to fit the model for time-to-AIDS.

3.4 Using behavioral data to refine the prior distribution based on experts' opinions

Behavioral data collected through a questionnaire in the aMASE study can in principle carry information on the likely place of HIV acquisition be it before or after migration. We incorporated such data based on 6 combinations of questions: 3 in favor of pre-migration infection and 3 in favor of post-migration infection. Prior probabilities of infection pre- or post-migration were assigned to these questions by 5 members of the

research team who have extensive experience in the HIV epidemiology in Europe and in migrant populations. These probabilities are based on epidemiological facts on the natural history of the disease and the possible routes of transmission allowing though for a degree of response bias. Let $odds(Q_i)$ be the odds of post-migration infection assigned to a positive answer to Q_i and 1 otherwise. Then the prior odds of infection post-migration are assigned as $\prod_{i=1}^{6} odds(Q_i)$. Thus, the prior distribution of w_i, as defined in Subsections 3.1-3.3, becomes proportional to a step function of the form

$$f(w_i) \propto \begin{cases} \prod_{i=1}^{6} odds(Q_i), \text{ if } w_i < m_i \\ 1, \text{ if } w_i > m_i \end{cases}$$

Note that if there was not any positive answer to any of the 6 combinations of questions, no prior odds were assigned and the prior distribution remained the same. The questions and the corresponding prior probabilities of post migration infection are presented in Table 1. In case of a discrepancy between the HIV experts, the probability that was closest to 50% was chosen as the prior probability of post migration infection. To visually illustrate each stage the posterior distribution passes through until reaching the final form, we present an example from a hypothetical HIV-positive migrant in Figure 2. In this example we start by assuming a uniform prior distribution over the whole period at risk for HIV. This leads to a 64% prior probability of pre-migration infection. Accounting for the fact that the subject was AIDS-free at diagnosis, the probability of pre-migration infection reduces to 35%. Incorporating also the data from both biomarkers, the shape of the posterior distribution changes drastically, with the posterior probability of pre-migration infection becoming 43%. Finally, if we assume that the prior odds of infection pre-migration are 0.7 based on subject's behavioral data, a gap will appear on the migration date and the probability of pre-migration infection reduces to 35%.

To investigate the influence of HIV experts' beliefs on the results, in a sensitivity analysis, behavioral data were ignored when constructing the prior distribution of w_i.

4 Simulation study

4.1 Design

The performance of the proposed method was evaluated in a simulation study. The study was designed to closely mimic the migrant population included in the aMASE study with respect to certain characteristics such as the date of birth, sex, mode of infection, region of birth, date of diagnosis and migration date. Before starting simulating data, we excluded from the aMASE study all individuals that had external information based on which they could be classified as infected pre- or post-migration. (i.e. a positive HIV test pre-migration or a negative HIV test post-migration).

We simulated 10 000 subjects with their characteristics derived from the joint probability distribution of the above factors in the aMASE study, using the chain rule of probability. More specifically, the dates of birth were simulated using the normal distribution truncated on the minimum and maximum dates observed in the aMASE study; sex was simulated conditionally on the date of birth using a logistic regression

Table 1. Assigned prior probabilities of post-migration HIV infection based on behavioural data as evaluated by 5 members of the study group. In bold the prior probabilities used when applying the method to the aMASE data.

Question	Probability of infection post-migration
Has an AIDS diagnosis within 3 months of HIV diagnosis and arrived in the same year of diagnosis and has no evidence of seroconversion	**30%** 20% 20% 30% 20%
Has not had sex in the country of destination and has never injected drugs	35% **40%** 20% 15% 15%
Has only injected drugs in country of origin	**20%** 20% 20% 20% 20%
Patient with negative self-reported HIV test after year of arrival	80% 80% 80% **70%** 80%
Has only injected drugs in country of destination	80% 80% 80% **75%** 75%
Has had unprotected sex only in country of destination and has never injected drugs	65% **60%** 60% 70% 60%

model; mode of infection through a multinomial regression model given sex and age; region of birth from a multinomial regression model given age, sex and mode of infection; dates of diagnosis, transformed into a value in the interval (0,1), were modelled using beta regression[24], with the mean depending on age, sex, mode of infection and region of birth through a logit link function and assuming a constant dispersion parameter. Finally, migration dates were simulated from a beta regression model also using the logit link function given the remaining covariates.

Parameters regarding all the above-mentioned distributions were derived by applying the corresponding models to the aMASE study data. The time intervals between infection and diagnosis, which are not observable and thus had to be specified beforehand, were simulated from a beta distribution, with its parameters found by trial and error in order for the simulated subjects to be similar to those from the aMASE study with respect to all relevant characteristics.

The CD4 and VL measurements were simulated according to the bivariate model (4), with its parameters derived by applying the model to CASCADE data, as described above. Then we excluded the data prior to HIV diagnosis, as such data cannot be

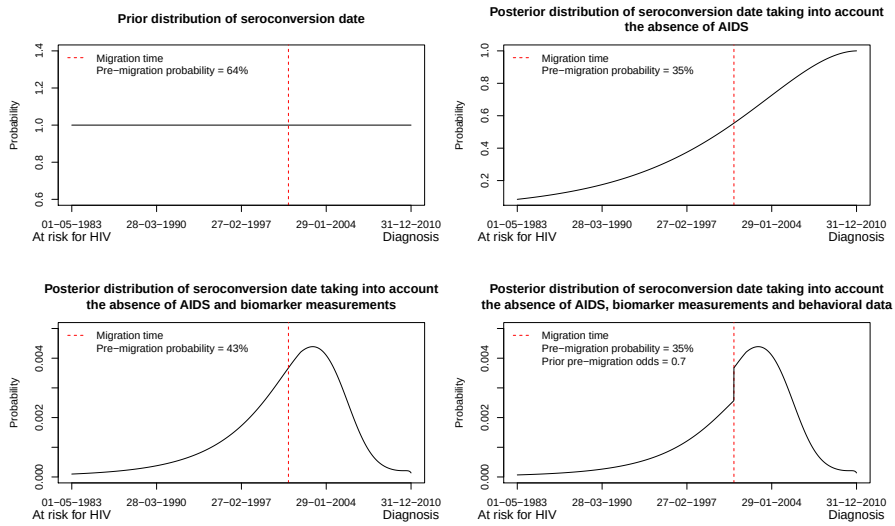

Figure 2. Illustration of the steps followed by the proposed method. Top left: prior distribution of infection date; top right: shape of the posterior distribution using the absence of AIDS; bottom left: shape of the posterior distribution using the absence of AIDS and biomarkers' data; bottom right: final form of the posterior distribution taking into account the absence of AIDS, biomarkers' data and behavioral data.

seen in practice. We simulated a latent time to AIDS or death using a joint model that assumes that the risk for AIDS onset or death depends on the current "true" but unobserved CD4 count[25]. As in the absence of ART it is known that progression to AIDS is mainly influenced by CD4 count, such a model may be a plausible one. The longitudinal submodel of the joint model takes the form of a linear mixed model, i.e.

$$\mathbf{Y}_i(\mathbf{t}_i) = \mathbf{m_i}(\mathbf{t_i}) + \boldsymbol{\epsilon_i}(\mathbf{t_i}),$$

where $\mathbf{Y}_i(\mathbf{t}_i)$ is the vector of the observed CD4 counts on the forth-root scale, and $m_i(t) = \mathbf{X}_i(t)\boldsymbol{\beta}^{joint} + \mathbf{Z}_i(t)\mathbf{b}_i$ is assumed to correspond to the "true" outcome value at the time point t, i.e. the value of the process that would have been observed had the measurement error been eliminated. The survival submodel can then be written as

$$\lambda(t|\mathcal{M}_i(t), \boldsymbol{\omega}_i) = \lambda_0(t) \exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{\omega}_i + \alpha m_i(t)\right\}, \quad t > 0,$$

with $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ denoting the history of the unobserved longitudinal process up to the time point t, the parameter α reflecting the association between the two submodels, $\lambda_0(\cdot)$ the baseline hazard function and $\boldsymbol{\omega}_i$ the vector of baseline covariates (sex, region of origin, risk group and age at infection) with a corresponding vector of regression coefficients $\boldsymbol{\gamma}$. The baseline hazard function was assumed to follow the Weibull distribution. To estimate the parameters of the model we again used

the CASCADE data truncated on 1/1/1996, the date effective ART became available. Maximization of the joint likelihood was carried out through the pseudo adaptive quadrature method[26] which is provided in the package `JM` in `R`[27]. The survival function implied by the joint model is equal to

$$S(t|\mathcal{M}_i(t), \boldsymbol{\omega}_i) = \exp\left\{-\int_0^t \lambda_0(s)\exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{\omega}_i + \alpha m_i(s)ds\right\}\right\}, \tag{9}$$

which cannot be evaluated analytically. To simulate random times from this joint model, one can use inverse transform sampling, which requires solving the equation

$$S(t|\mathcal{M}_i(t), \boldsymbol{\omega}_i) = u \tag{10}$$

in t, where u is a random draw from the uniform distribution in (0,1). To solve (10), we used the `uniroot` function in `R`, along with the `integrate` function to approximate the integrals. It should be noted that when simulating the AIDS or death times, the parameter $\boldsymbol{\beta}$ used in evaluating the survival function (9) was obtained from the bivariate linear mixed model (4) and not from the joint model. Administrative censoring of AIDS or death times was made on December 31, 2015.

After the introduction of ART, the probability of developing AIDS have been substantially reduced since the majority of HIV-positive patients initiate ART prior to AIDS onset. The probability of ART initiation was largely driven by the observed current CD4 count, based on the World Health Organization (WHO) recommendations until 2015. To mimic that mechanism, we applied a discrete survival model to the CASCADE data, regressing the hazard of initiating ART on current CD4 count (on the forth-root scale), sex, region of birth, risk group, age at diagnosis, calendar time of diagnosis and time since diagnosis, using a logit link function. Based on this model, for each simulated participant, we simulated the ART initiation date. AIDS or death times were censored at the ART initiation date, i.e. AIDS or death times occurring after ART initiation were ignored. Similarly, all CD4 and VL measurements were censored at the earliest of the date of ART initiation and the date of AIDS or death.

The simulated patients who developed AIDS prior to migration times were excluded as such patients were also not included in the aMASE study. For the patients who developed AIDS prior to diagnosis but after the migration date and 1/1/2008 (the date after which the aMASE study started collecting data), we set the date of diagnosis equal to the date of progression to AIDS. To summarize the procedures we followed to estimate the infection dates:

- when a subject was AIDS-free up to ART initiation or administrative censoring, we applied equation (6) to estimate the infection date.
- when AIDS occurred prior to ART initiation but there were some CD4 and VL measurements prior to AIDS onset, infection dates were estimated using (7).
- if a subject turned out to have AIDS at diagnosis, we used equation (8) as there were no biomarkers measurements available prior to AIDS development.

We classified patients as infected post migration when the posterior probability of post-migration infection using all available data was above 50%. The performance of the

proposed method was evaluated through the kappa coefficient of agreement between the true infection status (pre- or post-migration) and the one estimated by the proposed method. The Lin's concordance correlation coefficients[28] between the true times from infection to diagnosis and the corresponding ones based on the posterior mode, mean and median were also estimated. We also evaluated the sensitivity of our method, defined as the probability of suggesting a post-migration infection when the truth was that the subject had been infected in the country of destination, along with the specificity, defined as the probability of suggesting a pre-migration infection when the true infection status was pre-migration.

4.2 Results from the simulated study

Out of the 10 000 subjects, 1800 (18%) were dropped since they developed AIDS or died pre-migration. Results from the application of the proposed method to the simulated study are presented in Table 2. Demographic characteristics of the simulated subjects were similar to those of the aMASE participants, as expected. The proportion of subjects who had developed AIDS at HIV diagnosis was 15.51%, very close to the 15.28%, the corresponding proportion in the aMASE participants that were not definitely classified as infected pre- or post-migration. In addition, the median CD4 counts at diagnosis in the simulated study were 311 cells/μL, again very similar to the median of 288 cells/μL in the aMASE subjects. These data suggest that our simulated study succeeded to closely mimic the subjects included in the aMASE study.

The kappa coefficient of agreement between the true infection status and the one derived by the posterior distribution of the unknown infection date was 0.69, demonstrating substantial agreement. This result was similar when the posterior distribution was based on biomarkers measurements (with or without AIDS onset), whereas estimates based only on the distribution of time from HIV infection to AIDS produced a lower kappa coefficient. The estimated sensitivity and specificity suggest that our method was able to correctly identify a post-migration and pre-migration infection, respectively, in about 85% percent of the cases. The correlation coefficients between the true time gap from HIV infection and diagnosis and the estimated one based on the posterior mode, mean and median were 0.49, 0.55 and 0.56, respectively. However, the correlation was much lower when using the time-to-AIDS distribution only, suggesting that the time from HIV infection to AIDS has large inherent variability and thus cannot be adequately used for estimation purposes in our case. Comparing the results from the posterior mode, mean and median showed that the posterior median provided the most accurate estimates (Table 2). Overall, and using the median of the posterior distribution

of the infection time, the difference between estimated and true time of infection was 0.21 years.

Table 2. Results from a simulated study including 10000 subjects

	All n=8200; 100%	**Biomarkers + Absence of AIDS** n=6692; 81.61%	**Biomarkers + AIDS** n=236; 2.88%	**AIDS at diagnosis** n=1272; 15.51%
Lin's coef. of agreement between true and estimated infection time based on posterior				
Mode	0.49	0.51	0.47	0.03
Mean	0.55	0.62	0.47	0.15
Median	0.56	0.61	0.46	0.15
Mean diff. between estimated and true infection time based on posterior (estimated-true)				
Mode	-1.44	-1.93	-1.31	1.11
Mean	1.02	0.47	1.25	3.88
Median	0.21	-0.32	-1.15	3.27
Kappa coefficient	0.69	0.71	0.74	0.56
Sensitivity[1]	0.84	0.87	0.86	0.71
Specificity[2]	0.85	0.84	0.88	0.93
Proportion correctly specified	0.84	0.86	0.87	0.78

[1] Defined as the proportion of the simulated subjects for whom the proposed method suggests post-migration infection when the true infection status is post-migration.
[2] Defined as the proportion of the simulated subjects for whom the proposed method suggests pre-migration infection when the true infection status is pre-migration.

5 Application to the aMASE study

Out of 2,249 aMASE participants, 2,009 with complete information on critical variables were included in this analysis. Table 3 includes absolute (N) and relative(%) frequencies for the main demographic and clinical characteristics of the aMASE study population. The proportion of women (30.1%) and of those infected through heterosexual contact (45.6%) was higher than those observed in most European cohorts which are usually dominated by men infected through homosexual contact. The mean (SD) age of the study participants was 35.6 (9.7) years and most of them were born in either Africa or South America. Almost half (48.1%) were diagnosed with less than 350 CD4 cells/microL and 12.3% had already developed AIDS or progressed to AIDS soon after diagnosis. For 624 (31.1%) and 129 (6.4%) there was definite evidence for post- or pre-migration HIV infection, respectively based on the timing of their available negative and positive HIV tests. Of the remaining 1,256 individuals 1,056 (84.1%) had CD4 and/or HIV-RNA viral load measurements (1.5 and 1.3 on average) available before ART initiation or AIDS development. Thus, the proposed method was applied to 1,056 individuals using biomarkers' data, AIDS status and behavioral data and 200 individuals using AIDS status and behavioral data whereas for the remaining 753 individuals the timing of HIV infection (pre- or post-migration) was already known and treated as observed data in the relevant multiple imputation analyses. All subsequent results refer to the full sample of

2,009 individuals except for the case of time gaps from HIV infection to diagnosis where they refer to the subset of patients to which we applied the proposed method (n=1,256).

Estimated median time gaps from HIV infection to diagnosis and probabilities of post-migration infection are presented in Table 3. All estimates and 95% CIs shown in this table along with the corresponding tests are based on a multiple imputation approach with $K = 50$ random draws from the posterior distribution of the time gap between HIV infection and diagnosis, as described at the end of section 3.1. Estimates of median times were derived from median regression for clustered data[29] whereas mixed effects logistic regression models were used to derive the probabilities of post-migration infection. In all models, clustering of individuals in the collaborating clinics was taken into account. It is also noteworthy that the distribution of time between HIV infection and diagnosis was positively skewed, in all imputed datasets, thus medians are provided rather than means.

For individuals without definite evidence for pre- or post-migration HIV acquisition (n=1,256), the estimated median (95% CI) time gap between HIV acquisition and diagnosis was 4.58 (3.99, 5.17) years and differed significantly ($p = 0.001$) across migrants from different regions. The corresponding estimated medians (95% CIs) were 4.20 (3.18, 5.22), 5.60 (4.75, 6.45), 5.73 (3.94, 7.52) and 3.52 (2.75, 4.30) for migrants from Europe, Africa, Asia and South America, respectively (Table 3). Differences in the median gap between HIV acquisition and diagnosis were also statistically significant for all factors presented in Table 3 ($p < 0.05$) except for age at diagnosis ($p = 0.060$) and destination country ($p = 0.385$).

Regarding the most likely place of HIV acquisition, there was strong evidence of post-migration infection for 1,110 (55.3%) individuals for whom the corresponding estimated probability was >0.75. For 180 (9.0%) the evidence of post-migration infection was weaker with the corresponding probabilities ranging from 0.5 to 0.75. For the remaining 167 (8.3%) and 552 (27.5%) individuals there was either weak or strong evidence of pre-migration HIV acquisition as the corresponding probabilities were 0.5 to 0.75 or >0.75, respectively. In Table 3 probabilities of post-migration infection are estimated across levels of important demographic and clinical factors and overall. The estimated probabilities of post-migration infection ranged from 0.67 to 0.72 for all migrants except for those born in Africa for whom the corresponding probability was 0.47. Infections through injecting drug use or sexual contact between men were also associated with higher probabilities of post-migration infection compared to those attributed to heterosexual contact. As expected, those who were diagnosed having already progressed to clinical AIDS (or developed AIDS soon after diagnosis) were more likely to be infected for longer times thus their probability of post-migration infection was lower (0.43) compared to the corresponding probability among those who were AIDS free at diagnosis (0.66). Finally, the probability of post migration HIV acquisition increased with higher CD4 cell counts at diagnosis and ranged from 0.43 for those diagnosed with <100 CD4 cells/microL to 0.75 for those with >500 CD4 cells/microL. Differences in these probabilities were statistically significant ($p < 0.001$) for all factors presented in Table 3.

All analyses presented in this section were repeated a) using a subset of the CASCADE data (i.e. excluding participants whose infection date was determined using the midpoint

method and had an HIV test interval of 1 year or more) to estimate the parameters of the bivariate mixed model (4) of Section 3.2 and b) omitting the contribution of the available aMASE behavioral data to the prior distribution of w_i (see Section 3.4). In the first case all main estimates remained practically unaffected. In the second case, the overall median (95% CI) estimated time from infection to diagnosis was slightly higher being 4.82 (4.24, 5.39) vs. 4.58 (3.99, 5.17) years in the main analysis and consequently

the overall probability (95% CI) of post-migration infection became slightly lower being 0.61 (0.56, 0.65) vs. 0.63 (0.57, 0.67) years in the main analysis.

Table 3. Demographic-clinical characteristics of the sample from the aMASE study, estimated median time from HIV infection to diagnosis and estimated probabilities of post-migration HIV infection. Estimates and 95% CIs based on a multiple imputations ($K = 50$) approach.

	Median (95% CI) years from infection to diagnosis[1]	**Probability (95% CI)** of post-migration infection[2]
Sex		
Female (n=605; 30.1%)	5.26 (4.46, 6.06)	0.52 (0.45, 0.58)
Male (n=1404; 69.9%)	4.19 (3.51, 4.87)	0.68 (0.63, 0.72)
Age at diagonosis (years)		
<25 (n=256; 12.7%)	3.46 (2.34, 4.58)	0.51 (0.43, 0.60)
25-34 (n=811; 40.4%)	4.26 (3.44, 5.07)	0.58 (0.51, 0.64)
35-44 (n=606; 30.2%)	5.12 (4.13, 6.12)	0.66 (0.60, 0.72)
45+ (n=336; 16.7%)	5.47 (4.19, 6.74)	0.75 (0.68, 0.81)
Mode of infection		
Sex between men (n=994; 49.5%)	3.26 (2.61, 3.92)	0.71 (0.66, 0.76)
Injecting Drug Use (n=40; 2.0%)	5.47 (2.54, 8.40)	0.67 (0.47, 0.83)
Sex between men and women (n=917; 45.6%)	5.55 (4.84, 6.26)	0.55 (0.50, 0.61)
Other (n=58; 2.9%)	5.77 (2.90, 8.64)	0.57 (0.41, 0.71)
Destination country		
Belgium (n=232; 11.5%)	4.23 (2.84, 5.62)	0.42 (0.30, 0.55)
Greece (n=175; 8.7%)	5.76 (4.35, 7.17)	0.63 (0.51, 0.74)
Germany (n=29; 1.4%)	8.03 (3.86, 12.20)	0.53 (0.30, 0.76)
Italy (n=55; 2.7%)	5.64 (3.33, 7.95)	0.27 (0.13, 0.48)
Netherlands (n=109; 5.4%)	4.15 (2.43, 5.88)	0.72 (0.57, 0.82)
Portugal (n=170; 8.5%)	4.86 (2.90, 6.82)	0.66 (0.54, 0.76)
Spain (n=685; 34.1%)	4.03 (3.08, 4.99)	0.71 (0.65, 0.77)
Switzerland (n=174; 8.7%)	4.57 (3.06, 6.08)	0.46 (0.35, 0.58)
United Kingdom (n=380; 18.9%)	4.96 (3.31, 6.60)	0.67 (0.58, 0.75)
Region of origin		
Europe (n=469; 23.3%)	4.20 (3.18, 5.22)	0.71 (0.65, 0.76)
Africa (n=682; 33.9%)	5.60 (4.75, 6.45)	0.47 (0.41, 0.53)
Asia (n=181; 9.0%)	5.73 (3.94, 7.52)	0.67 (0.57, 0.75)
S. America (n=677; 33.7%)	3.52 (2.75, 4.30)	0.72 (0.66, 0.77)
AIDS within 3 months of diagnosis		
No (n=1762; 87.7%)	3.89 (3.37, 4.41)	0.66 (0.61, 0.70)
Yes (n=247; 12.3%)	10.04 (8.22, 11.85)	0.43 (0.34, 0.52)
CD4 cell count at diagnosis (cells/microL)		
<100 (n=301; 15.0%)	9.21 (7.95, 10.48)	0.43 (0.34, 0.52)
100-199 (n=243; 12.1%)	6.11 (4.95, 7.27)	0.55 (0.47, 0.64)
200-349 (n=422; 21.0%)	3.92 (3.10, 4.73)	0.63 (0.56, 0.70)
350-499 (n=379; 18.9%)	3.07 (2.28, 3.86)	0.70 (0.63, 0.76)
500+ (n=590; 29.4%)	2.43 (1.81, 3.06)	0.75 (0.69, 0.81)
NA (n=74; 3.7%)	4.90 (1.47, 8.33)	0.43 (0.30, 0.57)
HIV-RNA viral load at diagnosis (copies/mL)		
<500 (n=146; 7.3%)	6.69 (4.09, 9.30)	0.40 (0.30, 0.51)
500-9,999 (n=297; 14.8%)	3.57 (2.55, 4.59)	0.67 (0.59, 0.74)
10,000-99,999 (n=774; 38.5%)	3.72 (3.06, 4.37)	0.69 (0.63, 0.74)
100,000-999,999 (n=539; 26.8%)	5.64 (4.64, 6.63)	0.61 (0.54, 0.68)
1,000,000+ (n=137; 6.8%)	8.09 (5.82, 10.36)	0.63 (0.51, 0.73)
NA (n=116; 5.8%)	5.20 (2.26, 8.13)	0.46 (0.34, 0.58)
Total (n=2009; 100.0%)	4.58 (3.99, 5.17)	0.63 (0.57, 0.67)

[1] Estimates and 95% CIs based on median regression models for clustered data and provided only for individuals without definite evidence for pre- or post-migration HIV infection based on timing of positive and/or negative HIV tests.

[2] Estimates and 95% CIs based on mixed effects logistic regression models.

6 Conclusions

Motivated by the need to determine the most likely place of HIV acquisition among migrants diagnosed with HIV in Europe, we developed a method which can provide subject-specific estimates of the time gap between HIV infection and diagnosis. The core of the method is based on applying the Bayes theorem in order to reverse the conditioning of the distribution of biomarkers of HIV infection given the time since HIV infection and derive the distribution of time since HIV infection given a set of biomarkers' measurements. More specifically, we first estimated the parameters of a bivariate linear mixed model for the evolution of the CD4 cell count and HIV-RNA viral load during untreated HIV infection and prior to the onset of clinical AIDS. To estimate these parameters, as precisely as possible, we used data from the CASCADE study which maintains the largest database of HIV infected individuals with well estimated dates of infection. The data we used to fit the bivariate mixed model were derived from 19,788 individuals contributing 125,195 CD4 cell count measurements, 106,160 HIV-RNA viral load measurements along with a very rich set of demographic and clinical characteristics. Treating these parameters as known and having one or more measurements of one or both biomarkers of interest (i.e. CD4 cell count and HIV-RNA viral load) of an AIDS free and untreated individual we were able to derive the distribution of the time elapsed from HIV infection to diagnosis. The prior distribution of this time gap was initially set to a uniform one, ranging from a point in time corresponding to the onset of risk for acquiring HIV up to the date of HIV diagnosis. This prior distribution was being updated by taking into account the absence of AIDS symptoms and experts' knowledge related to presence or absence of risky behaviors while the individual was at the country of origin or at the destination country.

The proposed method was initially assessed in a simulation study. The simulated data were generated under realistic scenarios regarding the evolution of CD4 cell count and HIV-RNA viral load and the risk of developing AIDS or dying. Times of infection, migration and diagnosis were also simulated along with demographic-clinical covariates. The simulation parameters were derived from the application of the relevant models to the CASCADE and aMASE data with the ultimate goal being to mimic as close as possible the data of the aMASE study. Results from the application of the proposed method to the simulated data showed that although the correlation between the true infection times and the estimated ones was moderate (Spearman's $\rho = 0.46$) the overall rate of correct classification of HIV infections as pre- or post migration was 84.4% with the corresponding Cohen's κ coefficient being 0.69. In addition, based on the median of the posterior distribution of infection time, the mean difference between the true and the estimated time of infection was as low as 0.21 years, indicating that the estimated infection dates were very close to the true ones.

Applying the method to the aMASE study population, we were able to use additional information from the questionnaire data (i.e. sexual behaviour, injecting drug use and HIV testing history) in conjunction with experts' opinion on their effects on the probabilities of pre- or post-migration HIV acquisition. Results from this application showed that the majority of migrants diagnosed in the participating European countries

were more likely to have been infected in the destination rather than the originating country. This trend was mostly apparent among migrants from Asia, America and Europe while for those originating from Africa, the probabilities of post-migration infection were slightly lower than those of pre-migration infection. It is interesting though that, according to these results, post-migration infection is more frequent than it was believed in many European countries [30] and our results are more compatible with those reported by Rice et al [13]. The overall estimated median time between HIV infection and diagnosis for those without definite evidence for pre- or post-migration HIV infection based on their HIV testing history was 4.6 years with the corresponding mean being 6.2 years. The difference between the mean and median estimates was due to the right skewed shape of the corresponding distribution which was obvious in all imputed datasets. Given this asymmetry, the choice of the median as a measure of location is rather preferable. It is also noteworthy, that when we used a subset of the CASCADE data, using more strict criteria regarding the accuracy in the determination of HIV seroconversion, to estimate the parameters of the model for the evolution of CD4 cell count and HIV-RNA viral load, the main results remained unaffected. Additionally, omitting the contribution of HIV experts to inform the prior using the available behavioral data, resulted in only minor changes of the main estimates.

The proposed method provides a unified method to use all available sources of information in order to optimise our estimation regarding the time between HIV infection and diagnosis and consequently the most likely place of HIV acquisition for migrants. In the absence of repeated HIV tests or HIV recency testing [31], the only way to make meaningful estimation of the timing of HIV infection is to rely on known aspects of the natural history of the disease and behavioral data. The HIV natural history aspects that we used in the proposed method are related to the gradual decline of the CD4 cell count and the non-linear evolution of HIV-RNA with time after HIV infection. The behavioral data used in our method are related to the possible modes of HIV infection among adults, that is, unprotected sexual contacts and needle sharing among injecting drug users.

CD4 cell count decline during natural history of HIV infection has been used in all previously proposed methods which provide subject-specific estimates of the unknown infection time [8–13] but the respective HIV-RNA evolution has only rarely been used [11]. Unlike most of the aforementioned methods, the proposed method does not require data from individuals with known infection times; it only requires the estimates of the fixed and random parameters from the fit of the bivariate model on the CASCADE data, which can be easily shared and are provided in the supplementary appendix. More importantly, our method uses parameters for the CD4 cell count and HIV-RNA viral load evolution which vary according to many crucial covariates. For example the method of Rice et al. [13] uses different equations for the CD4 cell count decline according to the age of an individual at diagnosis and his/her race. However, besides age and race, the evolution of CD4 cell count post HIV infection depends on many other factors including for example sex, mode of infection [32,33] and even calendar time of infection [22,23]. Similar findings hold for the HIV-RNA viral load evolution thus, in order to improve the estimation of an individual's unknown infection time, one should ideally take into account both his/her CD4 cell count and HIV-RNA viral load levels at diagnosis along with his/her

demographic-clinical characteristics that are shown to be related with the evolution of these markers during untreated HIV infection. We should note though, we did not have the chance to use another important predictor of CD4 cell count and HIV-RNA viral load in our method which is the HIV subtype[34] as this information was available for only a part of the CASCADE study participants.

Another advantage of the proposed method is that it can easily accommodate information from multiple measurements of CD4 cell count and/or HIV-RNA viral load, irrespectively of their timing, as long as they are taken while the individual is AIDS free and untreated. This advantage is expected to be less important for new diagnoses though as current guidelines recommend immediate initiation of treatment after diagnosis[35,36] but is still relevant for individuals diagnosed in previous years which tended to have more than one CD4 cell count and/or HIV-RNA viral load available before initiating treatment.

The proposed method is also flexible enough regarding the amount of evidence, beyond biomarkers, one may want to use when estimating the time of HIV infection. In our application of the proposed method to the aMASE study data, we took advantage of the additional information on risk behaviors of the study participants, combined with expert opinion, in order to refine the prior distribution of the time gap between HIV acquisition and diagnosis. Another possible source of information that could be used in conjunction with the proposed method stems from phylogenetic methods and the concept of a molecular clock for HIV[37].

However, there is an inherent drawback of our method which is a shared drawback with all other methods that are using observed CD4 counts, which stems from the within-subject variability of this marker. Even though, the population average of CD4 cell counts has a well-defined decline during untreated HIV infection, individual measurements tend to be noisy and this affects the individual level estimation of the unknown HIV infection times. This was evident in the simulation study where although rates of correct classification were high, the correlation between the true and estimated times was moderate. Nevertheless, still the time difference between the true and estimated time of infection was on average quite small (0.21 years). Similar arguments hold for the HIV-RNA viral load values. However, our method takes into account all sources of variability in markers' values (i.e. between and within subjects) which are reflected to the shape of the posterior distribution of the elapsed time between infection and diagnosis quantifying thus the uncertainty of our estimates. Another limitation of our method is that it is computationally intensive as it depends on numerical approximations of probability and cumulative density functions in order to derive the posterior probabilities of pre- and post-migration infection. This deterred us from taking into account the uncertainty in the parameters estimated through the CASCADE data (through a Monte Carlo based approach for example) as it would multiply the time required to estimate the unknown infection time of each individual. At its current form, the algorithm required on average 7 seconds per individual on an PC using an AMD 3.4GHz CPU and 8 GBs of memory. Finally, as in all similar methods, there is always a concern regarding the compatibility of the population which was used to estimate the parameters of the model with the target population where the model will be applied. For our specific case, the parameters of the model were estimated using data from individuals with well known estimated

dates of HIV infection (i.e. seroconverters) whereas the target population comprises individuals with unknown infection times. However, a previous study by Lodi et al.[38] showed that estimates of HIV progression derived from seroconverters are likely to hold more generally for the HIV-positive population.

To conclude, we believe that although estimation of the time between HIV infection and diagnosis and/or the determination of the likely place of HIV acquisition for migrants is a difficult task, the method we propose provides a unified and formal but also practical way to effectively utilise information from routinely measured biomarkers, demographic and clinical characteristics and behavioral data in order to derive reliable estimates along with a clear measure of their uncertainty.

Acknowledgements

This study would not be possible without the entire EuroCoord Work Package 14 collaboration.

A.1. aMASE

The aMASE study team are: Aerssens A, Aguado M, Alimi B, Álvarez.D, Anagnostou O, Anderson J, Antoniadou A, Arando M, Barberà MJ, Barros H, Barthélemy A, Belda-Ibáñez J, Bertisch B, Bil J, Blanco JR, Block K, Boesecke C, Boura M, Burgos J, Burns FM, Cabo J, Calabuig E, Campbell L, Cardoso O, Claudia W, Clumeck N, Colucci A, Corrao S, Cuellar S, Cunha J, Daikos G, Darling K, del Amo J, del Romero J, Dellot P, Dixneuf M, Domingo P, Dronda F, Ebeling F, Engelhardt A, Engler B, Fakoya I, Farrell J, Fehr J, Feijó M, Fernández E, Fernández García E, Fernandez T, Fortes AL, Fox J, Garcia de Olalla P, García F, Gargalianos-Kakolyris P, Gennotte AF, Germano I, Gilleran G, Gilson R, Goepel S, Gogos HA, Gómez Sirvent JL, Gountas I, Gregg A, Gutiérrez F, Gutierrez MM, Hermans I, Iribarren JA, Knobel H, Koulai L, Kourkounti S, La Morté C, LeCompte T, Ledergerber B, Leonidou L, Ligero MC, Lindergard G, Lino S, Lopes MJ, Lopez Lirola A, Louhenapessy M, Lourida G, Luzi AM, Maltez F, Manirankunda L, Martín-Pérez A, Martins L, Masía M, Mateu MG, Meireles P, Mendes A, Metallidis S, Mguni S, Milinkovic A, Miró JM, Mohrmann K, Monge S, Montero M, Mouhebati T, Moutschen M, Müller M, Murphy C, Nöstlinger C, Ocaña I, Okumu-Fransche S, Onwuchekwa G, Ospina JE, Otiko D, Pacheco P, Palacios R, Paparizos V, Papastamopoulos V, Paredes V, Patel N, Pellicer T, Peña A, Petrosillo N, Pinheiro A, Poças J, Portillo A, Post F, Prestileo F, Prestileo T, Prins M, Prins P, Protopapas K, Psichogiou M, Pulido F, Rebollo J, Ribeirinho A, Río I, Robau M, Rockstroh JK, Rodrigues E, Rodríguez M, Sajani C, Salavert M, Salman R, Sanz N, Schuettfort G, Schüttfort G, Schwarze- Zander C, Serrão R, Silva D, Silva V, Silverio P, Skoutelis A, Staehelin C, Stephan C, Stretton C, Styles F, Sutre AF, Taylor S, Teixeira B, Thierfelder C, Touloumi G, Tsachouridou O, Tudor K, Valadas E, van Frankenhuijsen M, Vázquez M, Velasco Arribas M, Vera M, Vinciana P, Volny-Anne A, Voudouri N, Wasmuth JC, Wengenroth C, Wilkins E, Young L, Yurdakul S, Zafra Espinosa T, Zuure F

A.2. CASCADE

CASCADE Steering Committee: Julia Del Amo (Chair), Laurence Meyer (Vice Chair), Heiner C. Bucher, Geneviève Chêne, Osamah Hamouda, Deenan Pillay, Maria Prins, Magda Rosinska, Caroline Sabin, Giota Touloumi.

CASCADE Co-ordinating Centre: Kholoud Porter (Project Leader), Ashley Olson, Andrea Cartier, Lorraine Fradette, Sarah Walker, Abdel Babiker.

CASCADE Clinical Advisory Board: Heiner C. Bucher, Andrea De Luca, Martin Fisher, Roberto Muga

CASCADE Collaborators:Australia PHAEDRA cohort (Tony Kelleher, David Cooper, Pat Grey, Robert Finlayson, Mark Bloch) Sydney AIDS Prospective Study and Sydney Primary HIV Infection cohort (Tony Kelleher, Tim Ramacciotti, Linda Gelgor, David Cooper, Don Smith); Austria Austrian HIV Cohort Study (Robert Zangerle); Canada South Alberta clinic (John Gill); Estonia Tartu Ülikool (Irja Lutsar); France ANRS CO3 Aquitaine cohort (Geneviève Chêne, Francois Dabis, Rodolphe Thiebaut), ANRS CO4 French Hospital Database (Dominique Costagliola, Marguerite Guiguet), Lyon Primary Infection cohort (Philippe Vanhems), French ANRS CO6 PRIMO cohort (Marie-Laure Chaix, Jade Ghosn), ANRS CO2 SEROCO cohort (Laurence Meyer, Faroudy Boufassa); Germany German HIV-1 seroconverter cohort (Osamah Hamouda, Karolin Meixenberger, Norbert Bannert, Barbara Bartmeyer); Greece AMACS (Anastasia Antoniadou, Georgios Chrysos, Georgios L. Daikos); Greek Haemophilia cohort (Giota Touloumi, Nikos Pantazis, Olga Katsarou); Italy Italian Seroconversion Study (Giovanni Rezza, Maria Dorrucci), ICONA cohort (Antonella d'Arminio Monforte, Andrea De Luca.) Netherlands Amsterdam Cohort Studies among homosexual men and drug users (Maria Prins, Ronald Geskus, Jannie van der Helm, Hanneke Schuitemaker); Norway Oslo and Ulleval Hospital cohorts (Mette Sannes, Oddbjorn Brubakk, Anne-Marte Bakken Kran); Poland National Institute of Hygiene (Magdalena Rosinska); Spain Badalona IDU hospital cohort (Roberto Muga, Jordi Tor), Barcelona IDU Cohort (Patricia Garcia de Olalla, Joan Cayla), CoRIS-scv (Julia del Amo, Santiago Moreno, Susana Monge); Madrid cohort (Julia Del Amo, Jorge del Romero), Valencia IDU cohort (Santiago Pérez-Hoyos); Sweden Swedish InfCare HIV Cohort, Sweden (Anders Sönnerborg); Switzerland Swiss HIV Cohort Study (Heiner C. Bucher, Huldrych Günthard, Alexandra Scherrer); Ukraine Perinatal Prevention of AIDS Initiative (Ruslan Malyuta); United Kingdom Public Health England (Gary Murphy), UK Register of HIV Seroconverters (Kholoud Porter, Anne Johnson, Andrew Phillips, Abdel Babiker), University College London (Deenan Pillay); African cohorts: Genital Shedding Study (US: Charles Morrison; Family Health International, Robert Salata, Case Western Reserve University, Uganda: Roy Mugerwa, Makerere University, Zimbabwe: Tsungai Chipato, University of Zimbabwe); International AIDS Vaccine Initiative (IAVI) Early Infections Cohort (Kenya, Rwanda, South Africa, Uganda, Zambia: Matt A. Price, IAVI, USA; Jill Gilmour, IAVI, UK; Anatoli Kamali, IAVI, Kenya; Etienne Karita, Projet San Francisco, Rwanda).

A.3. EuroCoord

EuroCoord Executive Board: Fiona Burns, University College London, UK; Geneviève Chêne, University of Bordeaux, France; Dominique Costagliola (Scientific Coordinator), Institut National de la Santé et de la Recherche Médicale, France; Carlo Giaquinto, Fondazione PENTA, Italy; Jesper Grarup, Region Hovedstaden, Denmark; Ole Kirk, Region Hovedstaden, Denmark; Laurence Meyer, Institut National de la Santé et de la Recherche Médicale, France; Heather Bailey, University College London, UK; Alain Volny Anne, European AIDS Treatment Group, France; Alex Panteleev, St. Petersburg City AIDS Centre, Russian Federation; Andrew Phillips, University

College London, UK, Kholoud Porter, University College London, UK; Claire Thorne, University College London, UK.

EuroCoord Council of Partners: Jean-Pierre Aboulker, Institut National de la Santé et de la Recherche Médicale, France; Jan Albert, Karolinska Institute, Sweden; Silvia Asandi, Romanian Angel Appeal Foundation, Romania; Geneviève Chêne, University of Bordeaux, France; Dominique Costagliola (chair), INSERM, France; Antonella d'Arminio Monforte, ICoNA Foundation, Italy; Stéphane De Wit, St. Pierre University Hospital, Belgium; Peter Reiss, Stichting HIV Monitoring, Netherlands; Julia Del Amo, Instituto de Salud Carlos III, Spain; José Gatell, Fundació Privada Clínic per a la Recerca Bíomèdica, Spain; Carlo Giaquinto, Fondazione PENTA, Italy; Osamah Hamouda, Robert Koch Institut, Germany; Igor Karpov, University of Minsk, Belarus; Bruno Ledergerber, University of Zurich, Switzerland; Jens Lundgren, Region Hovedstaden, Denmark; Ruslan Malyuta, Perinatal Prevention of AIDS Initiative, Ukraine; Claus Møller, Cadpeople A/S, Denmark; Kholoud Porter, University College London, United Kingdom; Maria Prins, Academic Medical Centre, Netherlands; Aza Rakhmanova, St. Petersburg City AIDS Centre, Russian Federation; Jürgen Rockstroh, University of Bonn, Germany; Magda Rosinska, National Institute of Public Health, National Institute of Hygiene, Poland; Manjinder Sandhu, Genome Research Limited; Claire Thorne, University College London, UK; Giota Touloumi, National and Kapodistrian University of Athens, Greece; Alain Volny Anne, European AIDS Treatment Group, France.

EuroCoord External Advisory Board: David Cooper, University of New South Wales, Australia; Nikos Dedes, Positive Voice, Greece; Kevin Fenton, Public Health England, USA; David Pizzuti, Gilead Sciences, USA; Marco Vitoria, World Health Organisation, Switzerland.

EuroCoord Secretariat: Silvia Faggion, Fondazione PENTA, Italy; Lorraine Fradette, University College London, UK; Richard Frost, University College London, UK; Andrea Cartier, University College London, UK; Dorthe Raben, Region Hovedstaden, Denmark; Christine Schwimmer, University of Bordeaux, France; Martin Scott, UCL European Research & Innovation Office, UK.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article

Funding

This project has received funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration under EuroCoord grant agreement no. 260694.

Ibidun Fakoya is funded by a Doctoral Research Fellowship from the National Institute for Health Research, United Kingdom. The views expressed in this paper are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health.

Supplemental material

Parameters' estimates from the required models fitted to CASCADE data along with computer code to apply the method are available at the publisher's website.

References

1. Fakoya I, Alvarez-Del Arco D, Monge S et al. Advancing Migrant Access to Health Services in Europe (AMASE): Protocol for a Cross-sectional Study. *JMIR Research Protocols* 2016; 5(2): e74.
2. Downs AM, Heisterkamp SH, Brunet JB et al. Reconstruction and prediction of the HIV/AIDS epidemic among adults in the european union and in the low prevalence countries of central and eastern europe. *AIDS* 1997; 11(5): 649–662. DOI:10.1097/00002030-199705000-00013.
3. Becker NG, Lewis JJC, Li Z et al. Age-specific back-projection of HIV diagnosis data. *Statistics in Medicine* 2003; 22(13): 2177–2190. DOI:10.1002/sim.1406.
4. Sweeting MJ, Angelis DD and Aalen OO. Bayesian back-calculation using a multi-state model with application to HIV. *Statistics in Medicine* 2005; 24(24): 3991–4007. DOI:10.1002/sim.2432.
5. Karon JM, Song R, Brookmeyer R et al. Estimating HIV incidence in the united states from HIV/AIDS surveillance data and biomarker HIV test results. *Statistics in Medicine* 2008; 27(23): 4617–4633. DOI:10.1002/sim.3144.
6. Brookmeyer R. Measuring the HIV/AIDS epidemic: Approaches and challenges. *Epidemiologic Reviews* 2010; 32(1): 26–37. DOI:10.1093/epirev/mxq002.
7. Ndawinz JD, Costagliola D and Supervie V. New method for estimating HIV incidence and time from infection to diagnosis using HIV surveillance data. *AIDS* 2011; 25(15): 1905–1913. DOI:10.1097/qad.0b013e32834af619.
8. Berman SM. A stochastic model for the distribution of HIV latency time based on T4 counts. *Biometrika* 1990; 77(4): 733. DOI:10.2307/2337096.
9. Geskus RB. On the inclusion of prevalent cases in HIV/AIDS natural history studies through a marker-based estimate of time since seroconversion. *Statistics in Medicine* 2000; 19(13): 1753–1769. DOI:10.1002/1097-0258(20000715)19:13⟨1753::aid-sim487⟩3.0.co;2-f.
10. Muñoz A, Carey V, Taylor JMG et al. Estimation of time since exposure for a prevalent cohort. *Statistics in Medicine* 1992; 11(7): 939–952. DOI:10.1002/sim.4780110711.
11. Drylewicz J, Commenges D and Thiebaut R. Maximum a posteriori estimation in dynamical models of primary HIV infection. *Statistical Communications in Infectious Diseases* 2012; 4(1). DOI:10.1515/1948-4690.1040.
12. Ahmed SE and Reid N (eds.) *Empirical Bayes and Likelihood Inference*. Springer New York, 2001. DOI:10.1007/978-1-4613-0141-7.
13. Rice B, Elford J, Yin Z et al. A new method to assign country of HIV infection among heterosexuals born abroad and diagnosed with HIV. *AIDS* 2012; 26(15): 1961–1966. DOI:10.1097/QAD.0b013e3283578b80.
14. Pantazis N, Touloumi G, Meyer L et al. The impact of transient combination antiretroviral treatment in early HIV infection on viral suppression and immunologic response in later treatment. *AIDS* 2016; 30(6): 879–888.
15. Taylor JM and Law N. Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? *Statistics in Medicine* 1998;

17(20): 2381–2394.

16. Pantazis N, Touloumi G, Walker A et al. Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2005; 54(2): 405–423.
17. Geskus RB, Prins M, Hubert JB et al. The HIV RNA setpoint theory revisited. *Retrovirology* 2007; 4(1): 65. DOI:10.1186/1742-4690-4-65.
18. Lindstrom MJ and Bates DM. Newton-raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* 1988; 83(404): 1014. DOI:10.2307/2290128.
19. Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; 38(4): 963. DOI:10.2307/2529876.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.
21. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.
22. Pantazis N, Porter K, Costagliola D et al. Temporal trends in prognostic markers of HIV-1 virulence and transmissibility: an observational cohort study. *The Lancet HIV* 2014; 1(3): e119–e126. DOI:10.1016/s2352-3018(14)00002-2.
23. Gras L, Geskus RB, Jurriaans S et al. Has the rate of CD4 cell count decline before initiation of antiretroviral therapy changed over the course of the dutch HIV epidemic among MSM? *PLoS ONE* 2013; 8(5): e64437. DOI:10.1371/journal.pone.0064437.
24. Cribari-Neto F and Zeileis A. Beta regression in R. *Journal of Statistical Software* 2010; 34(2): 1–24.
25. Wulfsohn MS and Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997; 53(1): 330. DOI:10.2307/2533118.
26. Rizopoulos D. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis* 2012; 56(3): 491–501. DOI:10.1016/j.csda.2011.09.007.
27. Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* 2010; 35(9): 1–33.
28. Lawrence I and Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; : 255–268.
29. Parente PM and Silva JMS. Quantile regression with clustered data. *Journal of Econometric Methods* 2016; 5(1): 1–15.
30. Fakoya I, Álvarez-Del Arco D, Woode-Owusu M et al. A systematic review of post-migration acquisition of hiv among migrants from countries with generalised HIV epidemics living in europe: implications for effectively managing HIV prevention programmes and policy. *BMC Public Health* 2015; 15(1). DOI:10.1186/s12889-015-1852-9.
31. Murphy G and Parry J. Assays for the detection of recent infections with human immunodeficiency virus type 1. *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin* 2008; 13(36).

32. Touloumi G, Pantazis N, Babiker AG et al. Differences in HIV RNA levels before the initiation of antiretroviral therapy among 1864 individuals with known HIV-1 seroconversion dates. *AIDS* 2004; 18(12): 1697–1705. DOI:10.1097/01.aids.0000131395.14339.f5.
33. Touloumi G. Differences in cd4 cell counts at seroconversion and decline among 5739 hiv-1-infected individuals with well-estimated dates of seroconversion. *Journal of Acquired Immune Deficiency Syndromes* 2003; 34(1): 76–83. DOI:10.1097/00126334-200309010-00012.
34. Touloumi G, Pantazis N, Pillay D et al. Impact of HIV-1 subtype on CD4 count at HIV seroconversion, rate of decline, and viral load set point in European seroconverter cohorts. *Clinical Infectious Diseases* 2013; 56(6): 888–897. DOI:10.1093/cid/cis1000.
35. Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents 2017; Department of Health and Human Services. Available at http://aidsinfo.nih.gov/contentfiles/lvguidelines/AdultandAdolescentGL.pdf. Accessed: June 27, 2017.
36. European AIDS Clinical Society. Eacs guidelines v.8.2 2017; Http://www.eacsociety.org/guidelines/eacs-guidelines/eacs-guidelines.html. Accessed: June 27, 2017.
37. Leitner T and Albert J. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci USA* 1999; 96(19): 10752–10757.
38. Lodi S, Phillips A, Touloumi G et al. CD4 decline in seroconverter and seroprevalent individuals in the precombination of antiretroviral therapy era. *AIDS* 2010; 24(17): 2697–2704.