

A joint back calculation model for the imputation of the date of HIV infection in a prevalent cohort

Patrick Taffé^{1,*,†}, Margaret May² and the Swiss HIV Cohort Study[‡]

¹*Data Center, Swiss HIV Cohort Study, CHUV, Lausanne, Switzerland*

²*Department of Social Medicine, University of Bristol, Bristol, U.K.*

SUMMARY

In studies of the natural history of HIV-1 infection, the time scale of primary interest is the time since infection. Unfortunately, this time is very often unknown for HIV infection and using the follow-up time instead of the time since infection is likely to provide biased results because of onset confounding. Laboratory markers such as the CD4 T-cell count carry important information concerning disease progression and can be used to predict the unknown date of infection. Previous work on this topic has made use of only one CD4 measurement or based the imputation on incident patients only. However, because of considerable intrinsic variability in CD4 levels and because incident cases are different from prevalent cases, back calculation based on only one CD4 determination per person or on characteristics of the incident sub-cohort may provide unreliable results. Therefore, we propose a methodology based on the repeated individual CD4 T-cells marker measurements that use both incident and prevalent cases to impute the unknown date of infection. Our approach uses joint modelling of the time since infection, the CD4 time path and the drop-out process. This methodology has been applied to estimate the CD4 slope and impute the unknown date of infection in HIV patients from the Swiss HIV Cohort Study. A procedure based on the comparison of different slope estimates is proposed to assess the goodness of fit of the imputation. Results of simulation studies indicated that the imputation procedure worked well, despite the intrinsic high volatility of the CD4 marker. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: joint modelling; back calculation; conditional imputation; informative censoring; interval censoring; HIV infection date

*Correspondence to: Patrick Taffé, Coordination and Data Center, Swiss HIV Cohort Study, PAV admin-Bât. Maternité, Av. Pierre-Decker 2, CHUV, 1011 Lausanne, Switzerland.

†E-mail: Patrick.Taffe@chuv.ch

‡*The members of the Swiss HIV Cohort Study:* S. Bachmann, M. Battegay, E. Bernasconi, H. Bucher, Ph. Bürgisser, S. Cattacin, M. Egger, P. Erb, W. Fierz, M. Fischer, M. Flepp, A. Fontana, P. Francioli (President of the SHCS, Centre Hospitalier Universitaire Vaudois, CH-1011-Lausanne), H. J. Furrer (Chairman of the Clinical and Laboratory Committee), M. Gorgievski, H. Günthard, B. Hirschel, L. Kaiser, C. Kind, Th. Klimkait, B. Ledergerber, U. Lauper, M. Opravil, F. Paccaud, G. Pantaleo, L. Perrin, J.-C. Piffaretti, M. Rickenbach (Head of Data Center), C. Rudin (Chairman of the Mother & Child Substudy), J. Schüpbach, R. Speck, A. Telenti, A. Trkola, P. Vernazza (Chairman of the Scientific Board), R. Weber, S. Yerly.

Contract/grant sponsor: Swiss National Science Foundation; contract/grant number: 3345-062041

1. INTRODUCTION

In studies of the natural history of HIV-1 infection, such as assessing the influence of genotype on disease progression or the incubation period of HIV/AIDS, the time scale of primary interest is the time since infection (Figure 1). Unfortunately, this time is very often unknown for HIV infection because many individuals are already HIV-seropositive by the time they enter a research study (prevalent cases), and only for a small proportion of the patients (incident cases) is the date of infection known (these are patients seen by the clinician during primary infection or who have a negative and positive test for HIV infection within a conveniently narrow time interval, usually less than a year, in which case the date of infection is estimated as the mid-point). Within a few weeks after infection has occurred, the CD4 level drops down (primary infection) and then goes up to a maximum (during seroconversion, which occurs within a few months from infection), from where generally a continuous decay takes place. We shall refer to this maximum as the 'set point' and to its date of occurrence as the seroconversion date. We take a pragmatic approach in this paper, and assimilate the date of infection and of seroconversion, since HIV disease usually progresses over many years until the AIDS stage is reached (8–10 years). In most HIV cohort studies, the prevalent sub-cohort is much larger than the incident sub-cohort. Using the follow-up time instead of the time since infection is likely to provide biased results because of onset confounding [1–3]. Onset confounding arises when two groups of patients that have different distributions of times since infection are compared in analyses that use follow-up time. In the presence of onset confounding, estimates of the CD4 T cell or HIV-1 RNA profiles over time are biased and therefore no reliable inference is possible.

Two other important issues, at least, have to be dealt with when analysing data from an HIV cohort study: differential length-biased sampling and frailty selection [4, 5]. The former arises

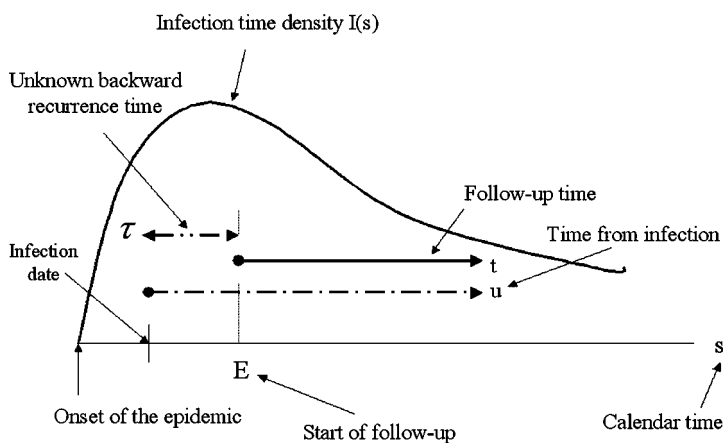
HIV epidemic: prevalent cohort

Figure 1. The date of infection is usually unknown for most of the patients in the Swiss HIV Cohort Study (SHCS) and only the follow-up time is available. However, the time scale of primary interest is usually the time since infection and working with the follow-up time instead is likely to provide biased measures of the effect of factors associated with natural disease progression.

because at the time the individuals are sampled for the prevalent cohort those persons with unusually rapid CD4 decline may have already been selectively removed from the sample (i.e. left truncated), either because they had already reached the disease stage at which AIDS is diagnosed or prematurely died. Therefore, patients in the high risk group have shorter follow-up and are underrepresented and relative risk estimates obtained from prevalent cohorts may be biased towards 1. The frailty selection bias arises because individuals who develop symptoms or are diagnosed with AIDS are at greater risk of loss to follow-up. For those patients who drop out early (i.e. before the scheduled end of study), either because of death or because of disease progression, the number of repeated observations available is associated with the rate of CD4 marker decline and censoring is informative. As a consequence of depletion of frailer individuals, the exposed cohort appears healthier than it really is, biasing the relative risk towards 1. Differential length-biased sampling arises because of frailty selection before the cohort is sampled. It is, however, mitigated by following the cohort over a longer time period, since disparities in the prior durations of infection will have diminishing influence compared with the relatively long observed follow-up period. In studies of the natural history of disease progression, introduction of antiretroviral treatment (ART) induces early exit from study follow-up. This right-censoring process is likely to be informative with respect to the response variable, whereas the left censoring due to staggered entry of patients into the study is assumed to be non-informative and independent.

Depending on the main goal of the study, e.g. assessing the incubation time of AIDS [6–8], estimating the distribution of HIV incidence over time [9–12] or studying disease marker processes [13–17], different methodologies have been developed to deal with the above-mentioned issues. In these studies, either cohorts of seroconverters are considered or the unknown date of infection is treated as a nuisance parameter and the focus is not on estimating that date. There are, however, many good reasons to focus on estimating the unknown date of infection. Firstly, for the clinician in his daily practice, it is convenient to integrate into his judgement of the disease progression an estimate of the duration of the HIV infection, so that the long time trend and the speed of progression may be better evaluated. Indeed, the decision when to start ART may be based on a specific CD4 level, for example, 350 cells/ μL (<http://www.medicinenet.com/script/main/art.asp?articlekey=16494>); however, consideration of the time since infection may also be of importance with patients showing a rapid progression being switched to an ART earlier (personal communication). Secondly, the maximum CD4 level (i.e. the so-called ‘set point’) reached during seroconversion may be of prognostic value and contribute to better understanding of the disease pathogenesis, specifically in genetic studies where genes and mutations may be involved in disease progression, such as CCR5-D32 (see Discussion). However, the set point is usually not observed in a prevalent cohort and an estimate of the date of seroconversion is required to back-project that CD4 level. Thirdly, endpoints based on the time since seroconversion to a specific event of interest, such as reaching a CD4 level of 200 cells/ mm^3 , may be more relevant than considering the time from entry into the study [2]. Finally, having an estimate of the date of infection allows data to be analysed on the appropriate time scale.

Laboratory markers such as the CD4 T cell count or HIV-1 RNA plasma level carry important information concerning disease progression and can be used in a model of back calculation to predict the unknown date of infection. Few papers have attempted to estimate the unknown date of infection and they usually use only one determination or they use only incident cases to estimate the parameters of the unknown density of infection times [18–21]. There is, however, considerable intrinsic variability in CD4 levels and a back calculation based on only one CD4 determination per person may result in unreliable estimates. In addition, imputation of the infection date based on

the characteristics of the incident sub-cohort might be biased if incident cases are different from prevalent cases [2, 22]. Therefore, our approach is based on the repeated individual CD4 T cells marker measurements and uses both incident and prevalent cases to impute the unknown date of infection.

At least two general methodologies have been proposed in the biostatistical literature to deal with the selection issues. The most common one is the likelihood-based approach [23, 24], where generally a full parametric specification of the joint distribution of the outcomes and the selection mechanism is specified. A non-likelihood-based approach has also been used [25, 26], in which selection is accounted for by appropriately weighting the data in the estimating equations. In the second approach, inferences are usually at the population level, whereas the first approach seems to be more amenable to subject-specific inferences by the use of random effects. Therefore, in this paper we jointly model the time since infection, the evolution of CD4 over time and the drop-out process using a shared random effects approach [27–31]. More specifically, the CD4 marker process is specified in terms of variation from baseline (i.e. the first available measurement) using linear mixed modelling, and both the unknown backward recurrence time and the time from the first available CD4 measurement to the end of follow-up or drop-out are described by accelerated failure time (AFT) frailty models. The link between the three models is induced by sharing common random effects. Estimation is done by direct maximization of the likelihood function using the Newton–Raphson algorithm and the integrals in the likelihood function are evaluated by the adaptive Gauss–Hermite quadrature formula. We implemented the joint model using the SAS *Proc NLMIXED*, which allows the implementation of user-defined likelihood functions conditional on the random parameters (the SAS code is available upon request).

2. MODELS AND ESTIMATION

2.1. Modelling the backward recurrence time

The backward recurrence time is the time from the start of follow-up, when the first CD4 measurement is available, to the moment the infection occurred. To estimate this time for each individual from the prevalent cohort (and therefore the infection date), we considered the ‘back’ time from each CD4 measurement to the date of infection. However, this is unobserved and, instead, only the ‘back’ time to the upper and lower boundaries of the seroconversion window is measured (Figure 2). Therefore, we considered modelling the elapsed time between each CD4 determination and the unknown infection date as a problem of interval censoring. This model of back calculation uses both the information on disease progression provided by the CD4 marker repeated measurements and the information provided by the last negative and first positive HIV tests.

Once the (individual) conditional density of infection time has been estimated for each CD4 determination, the date of infection can be estimated by mean or median conditional imputation, though for further utilization and to account for the uncertainty a multiple imputation procedure may also be envisaged. For practical purposes, when there are several CD4 determinations per individual, the mean (or a weighted mean, see Discussion) or median of the individual’s estimated times is used as the estimated time since infection. In addition, a 95 per cent prediction interval can be calculated using the conditional quantile function (see Section 2.4). It is, however, unclear how the multiple prediction intervals generated from each CD4 measurement should be combined, and one possibility is to consider the lowest and highest boundaries provided by the different intervals,

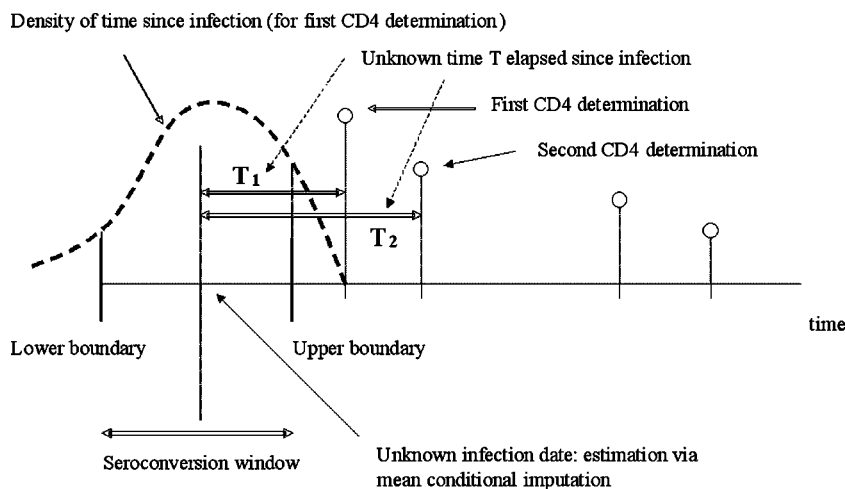


Figure 2. For each patient a seroconversion window was defined by the last documented negative and first available positive HIV tests. If no documented negative test was available, then the lower bound for the window was either January 1st 1980 if older than 14 years at that date or the date of the 14th birthday. However, if the patient was infected by clotting factors, then the date of birth was used. The upper bound was defined as the first available positive test date. If no such date was available, then either the registration date into the cohort or the date of first CD4 measurement was used.

though this approach may be too conservative and the width of the interval dramatically increases (see Discussion).

To determine the relationship between the CD4 count and the time since infection we identified incident patients or seroconverters. These patients are seen by the clinician during primary HIV-1 infection or have a documented negative and positive HIV test within a conveniently narrow seroconversion window (we considered a maximum length of one year, though this is only for convenience and these patients could be handled as interval censored). The date of infection was imputed using the mid-point of the interval. Figure 3 depicts the typical profile of CD4 counts over the course of the disease.

Our formulation of the model of backward recurrence time was based on the observation that, post seroconversion, the square root of CD4 evolves approximately linearly over time [14, 15], though other transformations have sometimes been considered, such as logarithm, cube or fourth root [13, 32]. Note that a simple scatter plot of the transformed CD4 *versus* time may be misleading for assessing the shape of the relationship because frailty selection tends to make the curve less steep with time, and a more formal approach is required (i.e. joint modelling). Under the assumption that the set point (intercept) and the rate of CD4 depletion (slope) are bivariate normal, a linear mixed model applies, i.e. $\varphi(\text{CD4}_i) = (\beta_0 + \beta_{0i}) + (\beta_1 + \beta_{1i})t_i + \varepsilon_i$, with φ appropriately chosen and $\varepsilon_{ij} \cong N(0, \sigma_\varepsilon^2)$. By inverting the relationship between CD4 level and time, a model of time elapsed since infection and current CD4 level is obtained. Taking the logarithm of time and considering a Taylor expansion of low order (e.g. one or two) leads to an AFT bivariate frailty model. The Taylor expansion, however, results in complicated functions of the two random parameters, and the inverse of the slope may be problematic if the patient is a long-term non-progressor with a

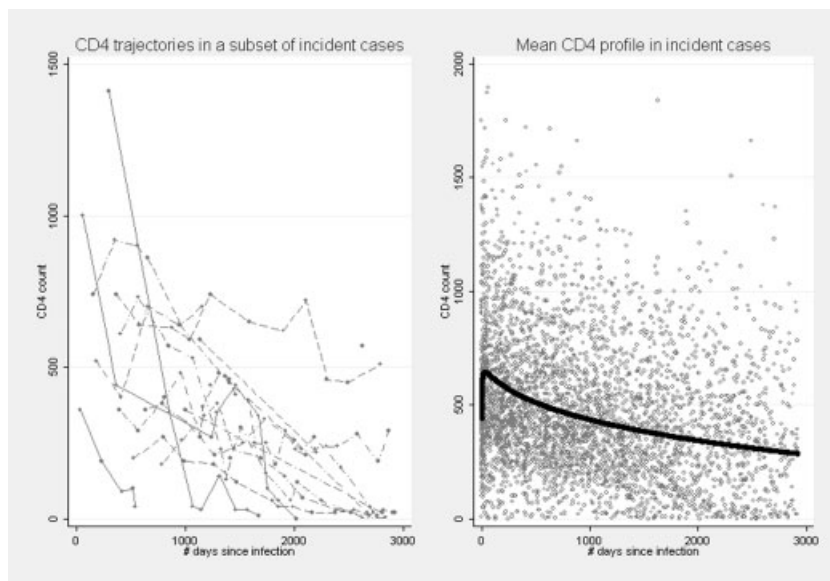


Figure 3. Scatter plot of CD4 counts (left) along with a population-averaged curve (right) in incident cases. Note that the first dramatic CD4 drop during primary infection is usually not observable. However, it is followed by a sharp increase as a result of a powerful initial reaction of the immune system (during seroconversion) and a maximal ‘set point’ is attained from where a monotonic decay takes place.

0 slope. Therefore, we considered a simpler (linear) formulation with the first frailty accounting for the individual intercept and the second for the multiplicative slope factor.

Let T_{ij} represent the ‘back’ time from the j th CD4 measurement from individual i to the date of infection. Only for incident patients is T_{ij} observed and for prevalent cases it is interval censored by the lower L_{ij} and upper U_{ij} boundaries of the seroconversion window. Therefore, for most patients T_{ij} is only known to belong to the interval $[L_{ij}; U_{ij}]$ and we shall use the indicator Δ_i taking value 1 if interval censored and 0 otherwise. It is assumed that censoring and infection times are independent. Denoting the square root of CD4 as y_{ij} , the model can be expressed (*model 1*) as

$$\log(T_{ij}) = (\gamma_0 + \beta_{0i}) + (\gamma_1 + \gamma_2(\beta_1 + \beta_{1i}))y_{ij} + \sigma_\omega^2 \omega_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n \quad (1)$$

where $[\beta_{0i} \ \beta_{1i}]' \cong N(0, 0, \sigma_{\beta_0}^2, \sigma_{\beta_1}^2, \sigma_{\beta_0\beta_1})$ represent the two frailties, with $\sigma_{\beta_0}^2, \sigma_{\beta_1}^2, \sigma_{\beta_0\beta_1}$ being the variance–covariance parameters, and $\gamma_0, \gamma_1, \gamma_2, \sigma_\omega^2$ and β_1 the regression parameters to be estimated. If the error term ω_{ij} is independently and identically $N(0, 1)$ distributed, with $\omega_{ij} \perp [\beta_{0i} \ \beta_{1i}]'$, then a log-normal frailty model is obtained, and a Weibull model results if instead an exponential distribution is postulated. Other models such as the generalized Gamma with three parameters can be used, particularly to assess nested simpler two parameters distributions, such as the log-normal or the Weibull distributions. The frailties take into account the correlation between repeated measurements. Note that $(\gamma_0 + \beta_{0i})$ and $(\gamma_1 + \gamma_2(\beta_1 + \beta_{1i}))$ represent simpler linear functions of the unobserved CD4 intercept and the slope for individual i . In addition, β_0 does not appear here, since the time of infection is unobserved. In this model, the ‘back’ time depends both on the CD4

level and on the rate of depletion. Therefore, two individuals with the same CD4 value but with different slopes will have differing backward time estimates.

2.2. Modelling the CD4 marker process

A popular model adopted for modelling the decline in CD4 (or transformed CD4) is the linear mixed model, particularly when the focus is on individual profiles [32]. Keeping with the linearity assumption of the relationship between the square root of CD4 and time, and as the date of infection is usually unobserved, we considered a model (*model 2*) of differences [14]:

$$\Delta y_{ij} = (\beta_1 + \beta_{1i})t_{ij} + \varepsilon_{ij} - \varepsilon_{i1} \quad (2)$$

with $\Delta y_{ij} = y_{ij} - y_{i1}$, where y_{i1} represents the baseline value of the square root of CD4 count, $t_{ij} = (t_{ij}^* - t_{i1}^*)$ the time measured on the follow-up scale whereas t_{ij}^* is unobservable and represents the time measured from the date of infection, $(\beta_1 + \beta_{1i})$ the slope for individual i , $\varepsilon_{ij} \cong N(0, \sigma_\varepsilon^2)$ the independently and identically distributed (iid) error term and ε_{i1} the residual for period 1. Again, we adopt the common assumption $\beta_{1i} \cong N(0, \sigma_{\beta_1}^2) \perp \varepsilon_{ij}$, although this orthogonality assumption would be questionable if on the chosen scale the linear model (in time) was not appropriate, e.g. some important regressors were omitted from the regression model [33, 34]. Note that β_{0i} , the random intercept, does not appear here but appears in (1) since a model of differences was considered.

2.3. Modelling the drop-out process

The literature on modelling the drop-out process to account for informative right censoring (or non-ignorable missingness) is very abundant and different methodologies have been proposed to accommodate different study designs [27]. When there is staggered entry and visits do not take place at preset fixed and equally spaced time intervals, a convenient approach is to use survival analysis models to describe the time to early termination of follow-up [30, 31] and link this process with the marker process by common random parameters. Let T_{Di} denote the time from the first CD4 determination to the occurrence of an event that causes subject i to exit the study early, such as the start of ART or death (though the case of right censoring due to death raises tricky issues and, here, we mainly focused on the introduction of ART, see Discussion), and δ_i be the drop-out indicator taking value 1 if exit was premature and 0 otherwise. The relevant observed length of time in the study is given by $T_{Oi} = \min(T_{Di}, T_{Ei})$, where T_{Ei} denotes the scheduled follow-up time. Note in our case that we allow the time scale to start at the first CD4 determination available since the date of infection is unobserved and adjust for the baseline value. Missing appointments are assumed to be non-informative. We model T_{Di} using the following AFT log-normal frailty model (*model 3*):

$$\log(T_{Di}) = \alpha_0 + \alpha_1 y_{i1} + \alpha_2 \beta_{1i} + \sigma_u^2 u_i \quad (3)$$

where $y_{i1} \perp \beta_{1i}$ represents the baseline square root of CD4 value, β_{1i} the frailty accounting for the heterogeneity in the depletion rates of CD4 T cells between individuals, $\alpha_0, \alpha_1, \alpha_2$ and σ_u^2 parameters to be estimated and $u_i \cong N(0, 1) \perp \beta_{1i}$ the iid error term. The drop-out process is linked to the marker and the backward recurrence time processes through the commonly shared frailty β_{1i} . Clearly, the higher the depletion rate or the lower the baseline CD4 value, the more likely the individual has dropped out by time t and the shorter the follow-up time.

2.4. Estimation procedures

2.4.1. *Estimation of the fixed parameters.* Consider the trivariate vector:

$$\beta_i = [\beta_{0i} \ \beta_{1i} \ \varepsilon_{i1}]' \cong N \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\beta_0}^2 & & \\ \sigma_{\beta_0\beta_1} & \sigma_{\beta_1}^2 & \\ 0 & 0 & \sigma_{\varepsilon}^2 \end{pmatrix} \right\}$$

Conditionally on $\beta_i = [\beta_{0i} \ \beta_{1i} \ \varepsilon_{i1}]'$ the three processes and the observations are independent. Therefore, a convenient formulation of the joint likelihood is

$$\begin{aligned} L(\theta|y, T_O, T) &= \prod_{i=1}^n [L_i(\theta|y_i, T_{O_i}, T_i)] \\ &= \prod_{i=1}^n \left[\int_{\mathfrak{R}^3} \left\{ \prod_{j=2}^{n_i} f_{\Delta y}(\Delta y_{ij} | \varepsilon_{i1}, \beta_{1i}) \right\} f_{T_D}(T_{D_i} | \beta_{1i})^{\delta_i} [1 - F_{T_D}(T_{E_i} | \beta_{1i})]^{1-\delta_i} \right. \\ &\quad \left. \times \left\{ \prod_{j=1}^{n_i} f_T(T_{ij} | \beta_{0i}, \beta_{1i})^{\Delta_i} [S_T(L_{ij} | \beta_{0i}, \beta_{1i}) - S_T(U_{ij} | \beta_{0i}, \beta_{1i})]^{1-\Delta_i} \right\} f_{\beta}(\beta_i) d\beta_i \right] \quad (4) \end{aligned}$$

where f , F and S , respectively, denote the density, cumulative and survival functions and the subscripted index denotes the random variable to which they refer. The vector θ represents the collection of fixed parameters in the linear predictors of the three processes and variance-covariance components. Specifying, for example, the conditioning distribution as $f_{\beta}(\beta_i) = f_{\varepsilon}(\varepsilon_{i1}) f_{\beta_1}(\beta_{1i} | \beta_{0i}) f_{\gamma_0}(\beta_{0i})$ makes this formulation of the likelihood readily implementable in a built-in routine for non-linear mixed models such as *Proc NLMIXED* from SAS. In contrast to other estimation methods for non-linear mixed models based on linearized-pseudo models [35], in *Proc NLMIXED* the maximum likelihood estimate $\hat{\theta}$ is obtained by maximizing (4), after numerical integration over β_i by adaptive Gauss-Hermite quadrature formula. The marginal likelihood function can then be conveniently passed to a maximization algorithm such as the Newton-Raphson gradient method.

2.4.2. *Empirical Bayes estimation of the random effects.* From the joint density of the data and random effects $f(y_i, T_{O_i}, T_i, \beta_i | \theta)$ one can derive the posterior conditional distribution of β_i :

$$g(\beta_i | y_i, T_{O_i}, T_i; \theta) = \frac{f(y_i, T_{O_i}, T_i, \beta_i | \theta)}{L_i(\theta | y_i, T_{O_i}, T_i)} \quad (5)$$

where

$$\begin{aligned} f(y_i, T_{O_i}, T_i, \beta_i | \theta) &= \left\{ \prod_{j=2}^{n_i} f_{\Delta y}(\Delta y_{ij} | \varepsilon_{i1}, \beta_{1i}) \right\} f_{T_D}(T_{D_i} | \beta_{1i})^{\delta_i} [1 - F_{T_D}(T_{E_i} | \beta_{1i})]^{1-\delta_i} \\ &\quad \times \left\{ \prod_{j=1}^{n_i} f_T(T_{ij} | \beta_{0i}, \beta_{1i})^{\Delta_i} [S_T(L_{ij} | \beta_{0i}, \beta_{1i}) - S_T(U_{ij} | \beta_{0i}, \beta_{1i})]^{1-\Delta_i} \right\} f_{\beta}(\beta_i) \quad (6) \end{aligned}$$

The empirical Bayes estimate $\hat{\beta}_i$ of β_i is obtained by considering the mean of the posterior distribution (5) as follows:

$$\hat{\beta}_i = \int_{-\infty}^{\infty} \beta_i g(\beta_i | y_i, T_{0i}, T_i; \hat{\theta}) d\beta_i$$

where the maximum likelihood estimate $\hat{\theta}_i$ has been substituted for θ .

2.4.3. Conditional imputation of the date of infection. An estimate \hat{T}_{ij} of the ‘back’ time from the j th CD4 measurement from individual i to the date of infection can be obtained by mean (or median) conditional imputation using the lower L_{ij} and upper U_{ij} boundaries of the seroconversion window as follows:

$$\hat{T}_{ij} = \hat{E}(T_{ij} | L_{ij} < T_{ij} \leq U_{ij})$$

With the model postulated in equation (1) we have

$$E(T_{ij} | L_{ij} < T_{ij} \leq U_{ij}) = e^{(\mu_{ij} + \sigma_{\omega}^2/2)} \frac{[\Phi((\log(U_{ij}) - \sigma_{\omega}^2 - \mu_{ij})/\sigma_{\omega}) - \Phi((\log(L_{ij}) - \sigma_{\omega}^2 - \mu_{ij})/\sigma_{\omega}))]}{[\Phi((\log(U_{ij}) - \mu_{ij})/\sigma_{\omega}) - \Phi((\log(L_{ij}) - \mu_{ij})/\sigma_{\omega}))]}$$

where the linear predictor $\mu_{ij} = (\gamma_0 + \beta_{0i}) + (\gamma_1 + \gamma_2(\beta_1 + \beta_{1i}))y_{ij}$ and Φ denotes the standard Normal cumulative function. Alternatively, the median estimate of the conditional distribution can be used and is given by the quantile function

$$Q_p(T_{ij} | L_{ij} < T_{ij} \leq U_{ij}) = \exp \left[\sigma_{\omega} \Phi^{-1} \left\{ p \Phi \left(\frac{\log(L_{ij}) - \mu_{ij}}{\sigma_{\omega}} \right) + (1 - p) \Phi \left(\frac{\log(U_{ij}) - \mu_{ij}}{\sigma_{\omega}} \right) \right\} + \mu_{ij} \right]$$

with p set to 0.5. A 95 per cent prediction interval can be calculated from each CD4 determination using 2.5 and 97.5 per cent percentiles.

3. APPLICATION

3.1. Data selection

We applied our method to data from the Swiss HIV Cohort Study (SHCS) to estimate the unknown date of infection in a prevalent sub-cohort. The Swiss HIV cohort has been described elsewhere (<http://www.shcs.ch/>). In brief, the SHCS is an ongoing cohort with over 14 000 patients enrolled to date, and follow-up scheduled twice annually, but with intermediate laboratory data up to every 3 months. In the SHCS, dual therapy became available in 1992. Patients registered before that year had no access to effective ART and censoring of follow-up was mainly due to death or withdrawal. Therefore, to focus on informative censoring due to the introduction of therapy we considered patients registered after 1992.

All patients having two or more CD4 determinations before the initiation of an ARTs including two or more antiviral drugs were selected ($n=4217$). It was hypothesized that monotherapy with either AZT or DDI did not influence the CD4 course over the long incubation period of AIDS [2]. No reliable estimate of the date of infection can be obtained in patients having less than two CD4 determinations, since two points at least are required for estimating an individual intercept and slope. Patients infected at birth were excluded because their pattern of disease progression

may not be similar to those infected as adults ($n = 138$). This selection process may, at first, seem likely to bias inferences. Recall, however, that the focus is not at the population level but rather at the individual, and the density of infection time for the population is not of interest here, and only for those individuals where it is possible will a date of infection be estimated.

One hundred and seventy nine patients were seen by their clinician during primary infection and, therefore, their date of infection was known. For 108 patients the seroconversion window was less than one year and their date of infection was imputed by mid-point. Consequently, there were 287 incident and 3930 (93 per cent) prevalent cases. Figure 3 shows that during primary infection the CD4 level abruptly falls (this is usually unobserved) as a result of the HIV-1 virus killing the CD4 T cells, then rises again as a consequence of a powerful reaction of the immune system (during seroconversion) and reaches a maximum set point from where a monotone decay takes place. Only the part of the trajectory with CD4 measurements after the set point provides a one-to-one correspondence between the CD4 count and the time since infection. Therefore, all CD4 counts measured within four months after the infection occurred in incident cases were discarded.

3.2. Results

Results of the estimation of the individual slope (IS) and date of infection are presented separately in two subsections, since estimation of the CD4 depletion rate requires only models 2 and 3, whereas model 1 is particularly useful for estimating the unknown date of infection. Recall that the speed of disease progression depends on both the CD4 slope and the set point, and both parameters are clinically relevant. Focus on estimating the slope will be useful in illustrating the amount of bias due to the selection mechanisms. In addition, a validation procedure of the imputed dates will be developed and discussed later, based on various slope estimates.

The proportion of patients exiting the study early due to the initiation of therapy containing two or more antiretroviral substances was 62 per cent ($n = 2627$). The number of deaths within 6 and 12 months after the last CD4 was 296 (7 per cent) and 404 (9.6 per cent). Normally, these patients should have started ART if their last CD4 counts was below 200 cells/ μL (82 per cent in both situations). Nevertheless, for the reasons elaborated in the Discussion these patients were non-informatively censored.

3.2.1. Estimation of CD4 slopes. In this subsection we present the results of the estimation of different models for the CD4 slopes to assess the sensitivity of different modelling strategies, as well as the biases due to onset confounding and informative censoring (Table I). In addition, the analyses were repeated on a sub-sample of patients who had four or more CD4 determinations to assess sensitivity to the number of individual measurements available.

The mean slope (MS) estimate, which is also the population average slope in a linear model, together with the empirical distribution of the IS was estimated using six different models:

1. unweighted mean of individually fitted least squares linear regression estimates (UWLS);
2. a linear mixed effects model of CD4 count differences (*model 2*);
3. a linear mixed effects model with time measured since first CD4 (LME);
4. a joint linear mixed effects model of CD4 count differences and drop-outs (*models 2&3*);
5. a joint linear mixed effects model of CD4 count differences, drop-outs and backward recurrence time (*models 1&2&3*);
6. a joint linear mixed effects model of CD4 count and drop-outs, which uses the imputed dates (calculated from (5)) to determine the time since infection (LMEED).

Table I. Comparison of CD4 mean slope (MS) estimates and empirical predicted distribution of individual slope (IS) estimates.

Model	MS				IS						
	Slope	Std. err.	Mean	Std. err.	Min	Max	P2.5	P50	P75	P97.5	
No. CD4 ≥ 2, N = 4217 patients											
1. UWLS*	-1.26	0.294	-1.26	0.294	-356.61	292.63	-19.86	-1.45	-0.00	25.38	
2. Model 2	-1.65	0.037	-1.64	0.015	-7.10	2.78	-3.96	-1.64	-1.18	0.13	
3. LME†	-1.70	0.040	-1.73	0.015	-6.87	2.66	-3.94	-1.73	-1.18	0.19	
4. Models 2&3	-2.20	0.049	-2.21	0.021	-7.30	2.84	-4.94	-2.09	-1.24	0.12	
5. Models 1&2&3	-2.29	0.048	-2.29	0.021	-7.71	2.82	-5.04	-2.18	-1.28	0.06	
6. LMEED‡	-2.08	0.027	-2.08	0.016	-5.76	1.26	-3.95	-2.12	-1.31	-0.23	
No. CD4 ≥ 2, N = 2507 patients											
1. UWLS	-1.72	0.078	-1.72	0.078	-23.39	90.93	-8.65	-1.38	-0.36	3.21	
2. Model 2	-1.61	0.039	-1.61	0.023	-6.93	2.77	-4.23	-1.53	-0.89	0.34	
3. LME	-1.65	0.040	-1.65	0.026	-6.84	2.63	-4.25	-1.58	-0.88	0.39	
4. Models 2&3	-1.89	0.042	-1.89	0.026	-6.95	2.81	-4.61	-1.75	-0.98	0.25	
5. Models 1&2&3	-1.95	0.041	-1.96	0.026	-7.18	2.79	-4.69	-1.82	-1.00	0.18	
6. LMEED	-1.85	0.029	-1.85	0.020	-5.04	1.21	-3.68	-1.79	-1.06	-0.05	

*UWLS, unweighted mean of individual least squares estimates.

†LME, linear mixed effects model with time measured since first CD4.

‡LMEED, joint linear mixed effects model of CD4 count and drop-outs. Estimated time since infection is used. (See definitions of models (1)–(6) given in text.)

The MS and IS estimates obtained using the different models differed substantially according to the time scale used and whether accounting for 'informative' drop-outs (joint modelling). Modelling the differences of CD4 counts allows estimation of the slope without requiring knowledge of the date of infection. The impact of the use of an inappropriate time scale is illustrated by the results obtained for model 2 and LME. However, when censoring is informative (with respect to slope) these estimators are biased and models 2&3 or 1&2&3 are preferable. Model LMEED also provides an unbiased estimate of the slopes. However, the standard error is underestimated because the uncertainty in the imputed dates of infection is not accounted for. Table I illustrates well the upward bias due to frailty selection. The amount of bias is particularly important in studies similar to ours where the level of censoring is high.

The individual ordinary least squares estimates are, in principle, robust to the drop-out process and the unweighted mean (UWLS) should provide a consistent estimate of the population MS. However, as expected because of the very unbalanced data set, the individual estimates showed considerable variability with very extreme values for patients with few determinations. The situation improves when considering patients with four or more CD4 determinations, although at the cost of a large reduction in sample size.

3.2.2. Estimation of the unknown date of infection. The mean width of the censoring interval was 14.1 years (IQR 11.7–17.4) and the mean number of CD4 determinations per patient 6 (IQR 3–8) with a median value of the first available CD4 measurement of 370 cells (IQR 202–568). Estimation of the final joint model (models 1&2&3) may be difficult and requires good starting values to achieve convergence. Firstly, we estimated models 2 and 3 separately. Then, the joint model 2&3 was estimated using as starting values the parameters obtained in the first step. Using the empirical Bayes estimates of the slopes from model 2&3 as input, model 1 was in turn estimated. Finally, the joint model 1&2&3 was estimated using as starting values the fixed and covariance parameters obtained in the previous steps. Table II shows the estimated coefficients from the joint model 1&2&3 with notations matching those of equations (1)–(3). The positive signs of γ_2 and α_2 indicate that a steeper CD4 slope is associated with both shorter backward recurrence time and earlier study exit. It is, unfortunately, not possible to infer the true relationship between β_{0i} and β_{1i} (i.e. from the structural equation between CD4 and time), since in model 1 we considered a simpler linear formulation. Therefore, the very low covariance between β_{0i} and β_{1i} does not imply that there exists no relationship between the CD4 set point and the slope. Note the large standard deviation of the residuals (on the CD4 square root scale), which illustrates the high volatility of the CD4 marker.

Using mean and median conditional imputation, we obtained an estimate of the unknown date of infection for each CD4 determination of every individual. The results were very similar using mean or median conditional imputation (Figure 4). However, the spread of the different date estimates within an individual was on average 1.7 years (median 1.4, IQ range 0.6–2.5, range [0; 8.3] years) but varied up to 8 years. This result illustrates the instability of a backward calculation performed using only one CD4 determination per patient. A single estimate of the unknown date of infection for each patient was calculated by the mean as well as the median of the imputed dates, though a weighted mean might have been considered as well (see Discussion). We found the mean age at infection to be 30.5 years (IQR 23.5–35, min=10, max=74) and the mean time since infection at the first CD4 determination for prevalent cases to be 4.8 years (IQR 3.2–5.7). A 95 per cent prediction interval was calculated for each CD4 determination (Figure 5). The mean width of the intervals was 7.3 years (IQR 5.5–8.9), which represented a reduction in the uncertainty by

Table II. Parameter estimates from the joint model of CD4 counts differences, drop-outs and backward recurrence time (*models 1&2&3*).

Parameter*	Estimate	Std. err.	<i>t</i> -Value	Pr> <i>t</i>	Lower [†]	Upper [†]
β_1	-2.286	0.048	-48.0	<0.0001	-2.379	-2.192
α_0	-0.931	0.064	-14.4	<0.0001	-1.057	-0.804
α_1	0.087	0.003	26.6	<0.0001	0.081	0.093
α_2	0.600	0.019	31.1	<0.0001	0.562	0.637
γ_0	-0.070	0.021	-3.3	0.0009	-0.111	-0.029
γ_1	-0.015	0.002	-8.7	<0.0001	-0.018	-0.012
γ_2	0.011	0.001	15.5	<0.0001	0.010	0.013
σ_ε	2.484	0.013	187.2	<0.001	2.458	2.510
σ_ω	0.312	0.003	118.1	<0.0001	0.307	0.317
σ_u	1.091	0.032	33.8	<0.0001	1.028	1.154
σ_{β_1}	1.705	0.045	38.2	<0.0001	1.618	1.793
σ_{β_0}	0.604	0.013	47.0	<0.0001	0.579	0.629
$\sigma_{\beta_0\beta_1}$	0.083	0.045	1.8	0.067	-0.006	0.170

*The notation matches those of equations (1)–(3).

[†]The 95 per cent boundaries were considered.

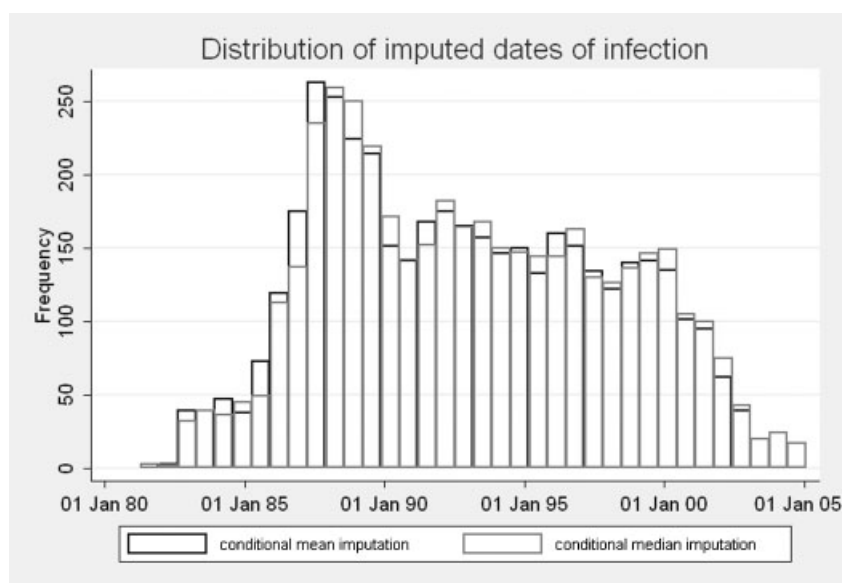


Figure 4. Comparison of mean and median imputation of dates.

half when compared with the 14 years mean width of the seroconversion window. When this distribution was limited to the first CD4 it was 6.1 years (IQR 4.5–7.7), whereas it was 9.5 (IQR 7.9–11.8) for the 15th CD4 determination, thereby illustrating that the more remote the CD4 the less precise the information or the greater the variance of the estimated ‘back’ time. Considering

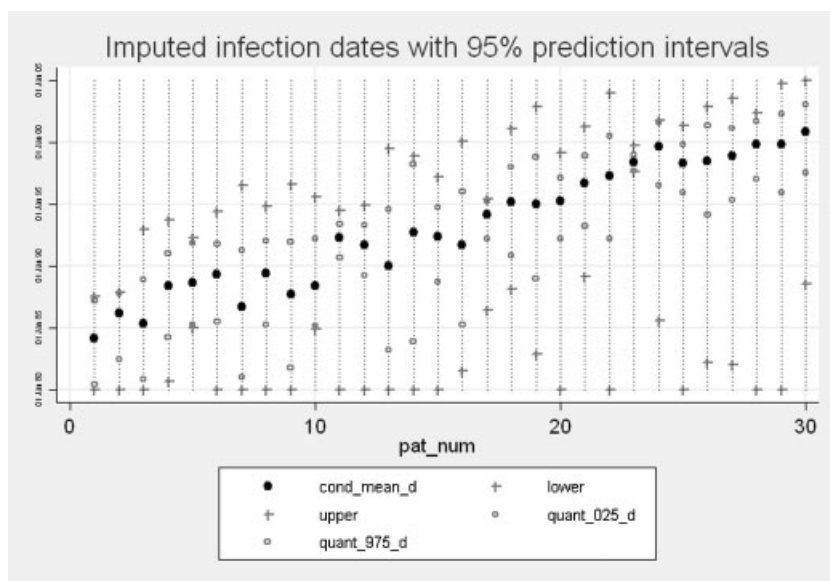


Figure 5. Imputed date along with 95 per cent prediction interval, lower and upper boundaries of the seroconversion window for a sample of 30 patients.

the interval provided by the lowest and highest boundaries the mean width was 7.8 (IQR 5.7–9.7), approximately 2 years larger than at the first CD4. An asymptotic 95 per cent confidence interval for the conditional mean parameter can be constructed using the multivariate delta method, since the parameters were estimated by maximum likelihood.

Finally, we performed a cross-validation by (interval) censoring the date of primary-infection in the subset of patients who were incident cases. Only for these patients is the date of infection really known with a good precision. The lower boundary of the window was set by drawing a random number from a uniform distribution on [0–2000] and [0–1000] for the upper boundary. The CD4 comprised within this window were erased, and as a result the sample of pseudo-interval censored cases having two or more CD4 determinations dropped to 110. The mean width of the pseudo-seroconversion window was 3.5 years (IQR 2.2–4.7). The only incident cases considered here were patients with a seroconversion window less than a year. Therefore, the estimations of the parameters were not dramatically changed and the same starting values could be used as for the original sample. We found the distance between the true and imputed date to be on average less than a year (0.9, IQR 0.4–1.2) and the mean width of the prediction intervals was 2 years (IQR 1.2–2.7). Considering the interval provided by the lowest and highest boundaries, the mean width was 3 (IQR 1.9–3.8), thereby illustrating that this approach can be quite conservative and provide very large prediction intervals. Only 72 per cent of the prediction intervals covered the true date, whereas it was 93 per cent when the lowest and highest boundaries were used to define the interval. We did not expect a good performance of the model in this simulation since incident cases are well known to be different from prevalent, as they usually progress more rapidly to AIDS. However, results of the simulation section were reassuring and showed very good performance of the model and appropriate coverage of the prediction intervals. For this reason, we adopted

a different methodology to assess the goodness of fit of our model, based on the comparison of different slope estimates (see Discussion).

4. SIMULATIONS

We evaluated the performance of our models by simulating a sample of 500 patients. The random intercepts and slopes were drawn from a bivariate Normal distribution with parameters $\mu_{\beta_0} = 0$, $\mu_{\beta_1} = 0$, $\sigma_{\beta_0} = 3$, $\sigma_{\beta_1} = 1.2$, $\sigma_{\beta_0\beta_1} = -1.6$ and the data were generated according to the model $\sqrt{\text{CD4}}_{ij} = (30 + \beta_{0i}) + (-2.1 + \beta_{1i}) * t_{ij} + \varepsilon_{ij}$, ε_{ij} iid $\cong N(0, 6)$. The value of the parameters was based on the coefficients obtained from the analysis of the incident cases. A follow-up of 20 years was generated with appointments quarterly. Staggered entry was introduced by drawing the number of years since infection in a Normal distribution with mean $3.1 + 0.5 * \text{slope}$ and variance 1.75. Values below or equal to 0 defined incident cases (10 per cent) and a maximum of 8 years was allowed. We subtracted six months from the time of first CD4 for each person to obtain the upper boundary of the seroconversion window. The lower boundary was defined by subtracting from the time of infection values drawn in a Gamma $\gamma_{(3,1)}$ distribution. The mean width of the window was 5.4 years (IQR 3.8–6.9) with a maximum of 13 years. We generated an informative right-censoring mechanism by selecting a random threshold for the start of treatment, for each individual, in a log-normal distribution, $x \cong \log N(0, 1)$, according to the mechanism $25 - 100 * \text{slope} - 10 * x$. Patients with a steeper slope had a higher threshold. Values were limited between 50 and 350. CD4 determinations were censored after the first value below the threshold.

First we considered the estimation of the CD4 slopes from a sample where 85.6 per cent of the data had been informatively right-censored. The MS estimate, while not accounting for the right-censoring, was -2.04 , whereas it was -2.09 when the drop-out process was jointly estimated. At the individual level, there was a correlation of 0.94 between the joint model estimates and the true slopes. When the data were, in addition, (non-informatively) left-censored, but the true time scale was used, the slope estimate was -1.94 , whereas it was -2.22 when the drop-out process was jointly estimated. The correlation dropped to 0.87 showing the loss of precision due to left-censoring. This simulation illustrated that the impact of informative right-censoring was dependent on the amount of left-censoring. In addition, the slightly over-estimated negative slope might possibly have resulted from the loss of information.

We then used the follow-up time scale, instead of the true time measured since infection, to simulate data from a prevalent cohort. Forty-nine patients were incident (~ 10 per cent) and the mean width of the seroconversion window was 5.4 years with a minimum of 1.4 and a maximum of 13.3 years. We compared the slopes and dates of infection estimated using model 1&2, while not accounting for the informative drop-out process, and model 1&2&3, which accounted for the drop-outs. In the former the MS estimate was -1.87 with a correlation of 0.86 with the true slope, whereas it was -2.17 with a correlation of 0.87 for the latter. We calculated the distance between the imputed and true date of infection (set to be January 1 for everyone). We found the mean distance to be 0.06 years and the variance to be 0.80, with 95 per cent percentile values $\{-1.42; 1.54\}$ for the model 1&2 and mean equal to -0.12 years, variance 0.84, with 95 per cent percentile values $\{-1.58; 1.35\}$ for the model 1&2&3. In this example, the date estimates were quite similar, despite different slope estimates. The coverage percentage of the prediction intervals was very good with 95.4 per cent of true dates included in their interval.

5. DISCUSSION

In this paper we have illustrated that because of the intrinsic variability of the CD4 T cell marker, imputation of the unknown date of infection based on only one determination per patient may provide unreliable results. Therefore, we have developed a more general procedure based on all the individual CD4 measurements to obtain more robust and less biased estimates. Our methodology was based on jointly modelling the CD4 marker process, the backward recurrence time and the time to early termination of follow-up. The unknown date of infection was imputed for each CD4 determination of every patient by either mean or median conditional imputation, with results for either being quite similar. For patients with several CD4 determinations, the final estimate of the date of infection was calculated as the mean or median of the multiply imputed dates.

We have developed a procedure to estimate the individual CD4 slope and date of infection, while accounting for the likely bias due to drop-outs. These estimates are very useful from a clinical perspective, particularly concerning HIV disease progression. However, the set point of the CD4 achieved during seroconversion may also be clinically relevant. For instance, we evaluated a polymorphism on the gene CCR5-D32, well recognized to be associated with slower HIV-1 disease progression [36], using the incident cases. We found this polymorphism to be associated with a higher set point and not the slope of the CD4 marker. To confirm this result in a much larger prevalent cohort, an estimate of the individual set point is required. This is, unfortunately, not readily computable from model 1&2&3 and the mean conditional imputation procedure. However, the joint model LMEED allows one to compute empirical Bayes estimates of the IS and intercept. These estimates are not consistent, since estimated dates of infection are used. However, when comparing the mean set points between the two groups of CCR5-D32 genotypes, the bias is likely to be negligible, particularly when the measurement error variance of the imputed dates is small relative to the long follow-up periods. In a time to event setting, as well, the bias arising from using the imputed dates may be small when follow-up is long. However, some multiple imputation procedure may also be envisaged by drawing into the conditional distribution of backward time. When the focus is on the slope, assessing CCR5-D32 does not require imputation of the infection dates, models 2&3 or 1&2&3 can be used (with different slopes for the two groups), the latter in principle providing slightly more precise estimators, since additional information on the seroconversion window is used.

To empirically confirm that the imputation of the dates was not sensitive to the sample of patients selected, we halved the sample of patients and estimated the joint model 1&2&3 separately on the two sub-samples. The imputed dates were almost exactly the same as those obtained on the whole sample with a correlation of 0.9997.

Our simulations illustrated that in the estimation of the IS, with informatively right-censored data, the performance of the joint model that accounted for drop-outs was dependent on the amount of left-censoring, whereas the estimation of the dates was less dependent. Apparently, the model for drop-outs is most useful to correct for the estimate of the slope, but does not affect the estimate of the date much. One possible explanation is that the first frailty term in model 1 corrects for under estimated slopes. Probably, this correction works well as long as the ranking in the estimated slopes remains whatever the setting. The interrelation between the different censoring mechanisms (left, right, interval) and their impact on the performance of the joint model is a tricky issue, and extensive validation and sensitivity analysis of this modelling strategy remain to be done in future projects using a large database on incident cases and artificially censoring the date of infection.

Analysis of goodness of fit is difficult for joint models because many distributional assumptions are made and estimates are affected by frailty distributions and censoring mechanisms. Therefore, we assessed the goodness of fit of the imputed dates of infection by comparing the IS estimates obtained using model 2&3, which do not require an estimate of the date of infection, with those calculated using the joint linear mixed model with drop-outs along with the estimated time since infection (LMEED). We found the two populations of slope estimates to be in good agreement with a correlation of 0.74 ($p < 0.001$) and 0.77 ($p < 0.001$) when the analysis was limited to patients with at least four CD4 determinations. We also compared the slope estimates obtained from model 2 with those from a simple linear mixed model with estimated time since infection and no adjustment for drop-outs. Again, the two estimation procedures provided very similar results with correlation coefficients of 0.86 ($p < 0.001$), respectively, and 0.92 ($p < 0.001$).

There are some important limitations to our methodology. We included the CD4 determinations while treated with monotherapy because we believed that the long incubation period of AIDS was not significantly affected by this therapy. This hypothesis should be checked by discarding those CD4 measurements determined while receiving monotherapy. However, in the SHCS many patients have their first CD4 measurement while already treated with monotherapy and therefore omitting these patients would considerably reduce the sample size.

We focused mainly on the censoring of patients who discontinued follow-up because of the introduction of ART and considered patients who died before the introduction of ART as non-informatively censored. For this reason, we considered patients registered in the cohort since 1992. Including patients registered before that date, however, complicates the situation because before 1992 no potent (dual) ART was available and follow-up was discontinued mainly due to death or attrition. Accounting for possibly informative censoring due to death raises, however, some difficult issues. First of all, this is a competing risks situation (e.g. early exit due to either death, attrition or ART) and would require modelling the drop-out processes due to death and attrition, as well. This can be done, in principle, by extending model 3 to include the competing causes of early exit [37]. However, when the reason for early termination is death (because the CD4 value is very low or 0) one can question the relevance of the whole procedure, since in this case adjusting for informative censoring would be similar to allowing the CD4 count to be negative after death, a counterfactual situation which would be impossible to justify. In reality, if the patient were still alive, his CD4 value would probably oscillate just above 0 but at a undetectable level. In this case his CD4 curve should present a levelling off with an horizontal asymptote. The question then arises as to how long should his follow-up be accounted for? Since this is to our best knowledge still an unresolved problem, we believe it to be probably more appropriate in our setting to simply censor deaths.

We did not account for differential length sampling. However, the bias is likely to be small when patients are followed-up for a long time as in this cohort. The modelling relies on rather strong parametric assumptions, which are difficult to assess. We have principally used the log-normal distribution for time and the bivariate normal density for the frailties. Other time distributions can be investigated and different frailty distributions can be implemented using the probability integral transformation. In addition, our model for CD4 marker process is rather simple and an independent measurement error was adopted, though more sophisticated models have appeared in the literature [15] using, for instance, an Ornstein Uhlenbeck process. However, given the already great complexity of the joint model and relatively scarcity of available CD4 measurements before ART is initiated, we adopted a pragmatic approach and selected a relatively simple and computationally more tractable model.

Finally, when a single date estimation is required, the multiple estimates for each individual should be appropriately combined. A simple way is to compute the arithmetic mean. The disadvantage of this method is that the increasing variance, and loss of precision due to increasing 'back' time at each new CD4, is not accounted for. A weighted mean could be devised using some time origin, such as the lowest estimated date, to allow the weight to depend on 'back' time. However, this time origin is not well defined and the CD4 level may momentarily rise simply because of intrinsic fluctuations. One way to solve this issue could be to estimate the 'true' latent CD4 level and use it instead of the observed values. Another related question is how to combine the multiple prediction intervals. As the simulations have shown, considering the lowest and highest limits of all the prediction intervals for an individual may be too conservative and provide very large intervals. These issues should be addressed in further reports. When a patient presents with only one CD4 measurement, the MS estimate can be used along with the last negative and first positive HIV tests.

In conclusion, our aim was to develop a modelling strategy that used all of the individual repeated marker values to impute the unknown date of infection, while accounting for informative censoring. We proposed a method to assess the goodness of fit of our model based on the comparison of two models to estimate the CD4 slopes, one based on our estimated dates of infection and the other not relying on those estimates. Results were in very good agreement, suggesting that our imputation procedure worked well. However, further work is required on sensitivity and goodness-of-fit analyses, and validation on data from large seroconverter cohorts would be helpful.

ACKNOWLEDGEMENTS

This study was financed in the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation (Grant no. 3345-062041). *Conflict of Interest Statement:* The authors report no financial or other associations that may represent any conflicts of interest.

REFERENCES

1. Brookmeyer R, Gail MH. Biases in prevalent cohorts. *Biometrics* 1987; **43**:739–749.
2. CASCADE Collaboration. Effect of ignoring the time of HIV seroconversion in estimating changes in survival over calendar time in observational studies: results from CASCADE. *AIDS* 2000; **14**:1899–1906.
3. Alcabes P, Pezzotti P, Phillips AN, Rezza G, Vlahov D. Long-term perspective on the prevalent-cohort biases in studies of human immunodeficiency virus progression. *American Journal of Epidemiology* 1997; **146**:543–551.
4. Brookmeyer R, Gail MH, Polk BF. The prevalent cohort study and the acquired immunodeficiency syndrome. *American Journal of Epidemiology* 1987; **126**:14–24.
5. Brookmeyer R, Gail MH. *AIDS Epidemiology*. Oxford University Press: Oxford, 1994.
6. Brookmeyer R, Goedert JJ. Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics* 1989; **45**:325–335.
7. Bacchetti P. Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *Journal of the American Statistical Association* 1990; **85**:1002–1008.
8. Longini IM, Clark WS, Byers RH, Ward JW, Darrow WW, Lemp GF, Hethcote HW. Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine* 1989; **8**:831–843.
9. Marschner IC. Using time of first positive HIV test and other auxiliary data in back-projection of AIDS incidence. *Statistics in Medicine* 1994; **13**:1959–1974.
10. Brookmeyer R, Gail MH. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association* 1988; **83**:301–308.
11. Cui J, Becker NG. Estimating HIV incidence using dates of both HIV and AIDS diagnoses. *Statistics in Medicine* 2000; **19**:1165–1177.

12. De Angelis D, Gilks WR, Day NE. Bayesian projection of the acquired immune deficiency syndrome epidemic. *Applied Statistics* 1998; **47**:449–498.
13. Pawitan Y, Self S. Modeling disease marker processes in AIDS. *Journal of the American Statistical Association* 1993; **88**:719–726.
14. De Gruttola V, Lange N, Dafni U. Modeling the progression of HIV-infection. *Journal of the American Statistical Association* 1991; **86**:569–577.
15. Taylor JMG, Cumberland WG, Sy JP. A stochastic-model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* 1994; **89**:727–736.
16. De Gruttola V, Tu XM. Modelling progression of CD4 lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**:1003–1014.
17. Tsiatis AA, De Gruttola V, Wulfsohn MS. Modelling the relationship of survival to longitudinal data measured with error. Application to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* 1995; **90**:27–37.
18. Munoz A, Carey V, Taylor JM, Chmiel JS, Kingsley L, Van Raden M, Hoover DR. Estimation of time since exposure for a prevalent cohort. *Statistics in Medicine* 1992; **11**:939–952.
19. Dubin N, Berman S, Marmor M, Tindall B, Des Jarlais D, Kim M. Estimation of time since infection using longitudinal disease-marker data. *Statistics in Medicine* 1994; **13**:231–244.
20. Berman SM. A stochastic model for the distribution of HIV latency time based on T4 counts. *Biometrika* 1990; **77**:733–741.
21. Geskus RB. Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored. *Statistics in Medicine* 2001; **20**:795–812.
22. Porter K, Johnson AM, Phillips AN, Darbyshire JH. The practical significance of potential biases in estimates of the AIDS incubation period distribution in the UK register of HIV seroconverters. *AIDS* 1999; **13**:1943–1951.
23. Laird N. Missing data in longitudinal studies. *Statistics in Medicine* 1988; **7**:305–315.
24. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**:112–1121.
25. Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* 1998; **93**:1321–1339.
26. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 1999; **94**:1096–1120.
27. Hogan JW, Laird NM. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine* 1997; **16**:259–272.
28. Vonesh EF, Greene T, Schluchter MD. Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine* 2006; **25**:143–163.
29. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000; **1**:465–480.
30. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* 1992; **11**:1861–1870.
31. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Statistics in Medicine* 1999; **18**:1215–1233.
32. Boscardin WJ, Taylor JM, Law N. Longitudinal models for AIDS marker data. *Statistical Methods in Medical Research* 1998; **7**:13–27.
33. Hausman JA, Taylor W. Panel data and unobserved individual effects. *Econometrica* 1981; **49**:1377–1398.
34. Fress EW. *Longitudinal and Panel Data-Analysis and Applications in the Social Sciences*. Cambridge University Press: Cambridge, 2004.
35. Lindstrom MJ, Bates D. Nonlinear mixed effects models for repeated measures data. *Biometrics* 1990; **46**:673–687.
36. Ioannidis JPA, Rosenberg PS, Goedert JJ, Ashton LJ, Benfield TL, Buchbinder SP, Coutinho RA, Eugen-Olsen J, Gallart T, Katzenstein TL, Kostrikis LG, Kuipers H, Louie LG, Mallal SA, Margolick JB, Martinez OP, Meyer L, Michael NL, Operskalski E, Pantaleo G, Rizzardì GP, Schuitemaker H, Sheppard HW, Stewart GJ, Theodorou ID, Ullum H, Vicenzi E, Vlahov D, Wilkinson D, Workman C, Zagury JF, O'Brien TR, for the International Meta-Analysis of HIV Host Genetics. Effects of CCR5-D32, CCR2-64I, and SDF-1 3' alleles on HIV-1 disease progression: an international meta-analysis of individual-patient data. *Annals of Internal Medicine* 2001; **135**:782–795.
37. Elashoff RM, Li G, Li N. An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Medicine* 2007; **26**:2813–2835.