Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2015 February 1; 64(2): 359–375. doi:10.1111/rssc.12080.

Predicting time to threshold for initiating antiretroviral treatment to evaluate cost of treatment as HIV prevention

Miranda L. Lynch and

University of Connecticut Health Center, Connecticut, USA.

Victor DeGruttola

Harvard School of Public Health, Boston, USA.

Miranda L. Lynch: mlynch@uchc.edu

Summary

The goal of this paper is to predict the additional amount of antiretroviral treatment that would be required to implement a policy of treating all HIV-infected people at time of detection of infection rather than at the time that their CD4 T-lymphocyte counts are observed to be below a threshold—the current standard of care. We describe a sampling-based inverse prediction method for predicting time from HIV infection to attainment of the CD4 threshold and apply it to a set of treatment-naïve HIV-infected subjects in a village in Botswana who participated in a household survey that collected cross-sectional CD4 counts. The inferential target of interest is the population-level mean time to reaching CD4-based treatment threshold in this group of subjects. To address the challenges arising from the fact that these subject's dates of HIV infection are unknown, we make use of data from an auxiliary cohort study of subjects enrolled shortly after HIV infection in which CD4 counts were measured over time. We use a multiple imputation framework to combine across the different sources of data, and discuss how the methods compensate for the length-biased sampling inherent in cross-sectional screening procedures, such as household surveys. We comment on how the results bear upon analyses of costs of implementation of treatment-for-prevention use of antiretroviral drugs in HIV prevention interventions.

Keywords

CD4 threshold; HIV; inverse prediction; longitudinal; multiple imputation

1. Introduction

Treating HIV-infected patients with antiretroviral therapy (ART) reduces the risk of sexual transmission to an uninfected partner (see Cohen *et al.* (2011); Donnell *et al.* (2010)), but the extent to which such treatment is cost-effective as a general modality for HIV prevention remains to be determined. Evaluating cost-effectiveness requires estimation of the additional person-years of treatment needed to provide all HIV-infected subjects in a community with

antiretroviral therapy at the time of detection of HIV infection, rather than at the time at which their CD4 T-lymphocyte counts (hereafter CD4 counts) reach a predetermined threshold to commence treatment. The goal of our research is to assess the cost of implementing a universal testing policy over the next five years among those who are diagnosed in household surveys, using information from a current household study in Mochudi, Botswana. In our analysis, the inferential target of interest is population mean additional time on ART treatment that would be incurred in the population if all HIV-diagnosed residents were offered immediate treatment with a goal of preventing spread of HIV. Estimation of this quantity is essential for predicting costs associated with implementing a treatment-for-prevention HIV intervention strategy. If this approach to prevention were effective, it could be expected to reduce the incidence and therefore the need for treatment in the future. Were household surveys repeated and all infected patients treated, disease incidence would be expected to decline over time and distributions of CD4 counts would shift. We do not attempt to evaluate such phenomena, but discuss only the predicted additional amount of treatment required to implement treatment-as-prevention as part of a cross-sectional household intervention. Below we describe the statistical challenges that arise in such analyses, and an approach to meeting them.

A complicating factor in the present analysis is that measurements of CD4 count in the cross-sectional household survey lack a time reference because dates of HIV infection are unknown. To address this issue, the dates of infection are treated as missing covariate information; the proposed methods make use of the time referencing from an auxiliary cohort study of subjects with known times of seroconversion (development of detectable antibodies). A multiple imputation (MI) framework allows us to account for the uncertainty that arises from imputing the missing time information, as well as the uncertainty in the model for CD4 decline over time. Our sample of cross-sectional data on CD4 count from the population of interest is length-biased; longer durations of time from infection to onset of treatment or death provide a greater opportunity for an infected individual to be included. Issues related to length-biasing in such cross-sectional investigations have received considerable attention (Zelen and Feinleib (1969), Brookmeyer *et al.* (1987)); below we discuss how our approach addresses this bias.

2. Cohort data on disease progression and cross-sectional survey data

Two datasets provide the information needed to address our question of interest. The first is from a population of HIV-1 Subtype C-infected patients in Mochudi, Botswana, currently under investigation in a pilot study intended to determine the feasibility of testing for HIV infection in a household setting and linking infected subjects to care. In addition to the goal of identifying undiagnosed individuals, this study also developed information regarding distributions of CD4 count and viral load among subjects who had, and had not, been diagnosed at the time of the household visit. These distributions provide information regarding the additional antiretroviral treatment required under different expanded treatment guidelines intended to reduce spread of HIV. From this cross-sectional household survey, CD4 level is available for each HIV-positive participant. For the prediction analysis using this cross-sectional household data, we considered two CD4 treatment thresholds ($CD4 = 250$ and 350 copies/mm³), which resulted in sample sizes in the cross-sectional data of 176

and 120, respectively. The most recent available guidelines indicate that treatable threshold in Botswana is 350 copies/mm³ (Botswana Ministry of Health (2012)) but the lower level was in place at the time the data were collected and so will be the focus of analysis. Summary information for the cross-sectional CD4 values for subjects in the prediction set from Mochudi, Botswana appears in Table 1.

We make use of data from longitudinal follow-up of patients with incident infection that provide information on rates of CD4 decline. This cohort study of HIV positive, treatment-naïve subjects (i.e. those not yet having initiated antiretroviral treatment) with known times of seroconversion was conducted in Botswana and Durban, South Africa (Novitsky *et al.* (2011)). We restrict attention to those $N_{long} = 72$ viremic subjects (i.e. the 36 people in each site with consistently detectable viral RNA measurements) in the longitudinal cohort who had at least five CD4 measurements, thereby excluding 7 subjects with very short profiles from the original cohort of viremic patients. The median number of measurements per profile was 11 (range 5 – 31), and the median duration of follow-up, 1.45 years (range 0.16 to 4.8y). All members of the longitudinal cohort were infected with HIV-1 Subtype C, as are the large majority of cases in the Southern African region (Lihana *et al.* (2012)). Women make up 79% of the Botswana cohort and all of the Durban Cohort, and 77% of the household survey members in Mochudi. Median age for the Mochudi cohort was 36, which is higher than for the longitudinal cohort, reflecting the greater age of people with prevalent rather than incident infections. While CD4 measurements vary over time within an individual, HIV infection generally results in a progressive decline over time; we use this tendency to guide our models in our prediction analysis. CD4 profiles from several representative subjects are given in Figure 1.

3. Inverse prediction models for imputing time to threshold using auxiliary longitudinal profile data

Our goal, using survey and longitudinal data on CD4 count, is to estimate the expected time for a subject identified in the survey to reach the treatment threshold. This problem is one of inverse prediction; such problems also arise in other settings including statistical calibration for laboratory assay measurements, and prediction of age from growth curve measurements. Inverse prediction refers to prediction of the regressor level t_0 given an observed measurement of a response variable y_0 and a set of observed training data $(\mathbf{t}_i, \mathbf{y}_i)$. Methods for estimating t_0 are well developed in both Bayesian (Hoadley (1970); Brown (1982)) and frequentist (reviewed in Osborne (1991)) frameworks, for both univariate and multivariate outcomes, and for in-sample and out-of-sample prediction. Situations in which measurements in the training data possess longitudinal structure have received less attention. Our proposed approach uses the longitudinal data to estimate the posterior distribution of parameters characterizing CD4 decline, which is then used to impute missing components in the cross-sectional survey data.

3.1. Notation and time referencing in unreferenced data

Our data consist of (W_{obs}, W_{mis}) , where W_{obs} has two components: 1) the $(y_{i,j}, t_{i,j})$ measurements for subject i in the longitudinal cohort, concatenated into $(\mathbf{y}, \mathbf{t})$ for the entire

longitudinal dataset, where y_{ij} is the j^{th} measurement of CD4 level at time t_{ij} on the i^{th} subject, $i = (1, \dots, N_{long})$, $j = (1, \dots, n_i)$, n_i is the number of observations on subject i , and with the times t_j referenced to the initial timepoint, $t_1 = 0$ (time of seroconversion), and 2) the two CD4 values, $(CD4_{k,M}, CD4_{k,T})'$, at monitoring and at the (fixed) CD4 threshold value, for subject k from the household survey data. We use the index i for subjects in the longitudinal cohort, and k for subjects in the cross-sectional household survey cohort. The missing data, $W_{mis,k} = (t_{k,M}, t_{k,T})'$, are the times of monitoring and of reaching CD4 threshold for subject k , respectively, both referenced to an unobserved baseline time of seroconversion. The target of estimation is the random quantity $t_A = t_T - t_M$, the time from monitoring until reaching threshold (see Figure 2). We specify a prior on the vector of missing values $(t_T, t_M)'$, and then carry out imputation by drawing the values from their posterior distribution (derived in Section 3.3), taking into account the uncertainty in the model for CD4 as a function of time. We obtain t_A by subtraction across the sampled chains.

3.2. Fitting the longitudinal profiles

CD4 decline in the longitudinal cohort is modeled using mixed effects models, where time enters as a linear term to ensure monotonicity of the profiles. Random effects models have been used to characterize repeated measures of CD4 (Lange *et al.* (1992), DeGruttola *et al.* (1991)). A general form for the mixed effects model for subject i is given in Equation 1.

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad (1)$$

where $\mathbf{Y}_i = \mathbf{y}_i$ is a vector of n_i CD4 measurements observed at times $(t_1, \dots, t_{n_i})$. CD4 values were transformed to square root scale to stabilize variance over the CD4 range; henceforth CD4 refers to the transformed value. $\mathbf{X}_i$ is the $n_i \times 2$ fixed effects design matrix $(\mathbf{1} \ t_i)$. For all models, the random effects design matrices $\mathbf{Z}_i = \mathbf{X}_i$. $\boldsymbol{\beta}$ is a 2×1 vector $(\beta_0, \beta_1)'$ of fixed effects parameters, with intercept β_0 and slope β_1 associated with time. Individual-specific random effects are given by the vector $(b_{i,0}, b_{i,1})'$. The random effects are assumed to be normally distributed $N_2(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is a 2×2 variance-covariance matrix associated with the random intercept and slope. The error terms $\boldsymbol{\varepsilon}_i$ are assumed to be $N_{n_i}(\mathbf{0}, \sigma^2 \mathbf{I})$. Hence, the $\mathbf{Y}_i$, conditional on the random effects, are distributed according to multivariate normal:

$$\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i \sim N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i, \sigma^2 \mathbf{I}). \quad (2)$$

Integrating over the random effects results in the marginal model for $\mathbf{y}$, given by:

$$\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i \sim N(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \sigma^2 \mathbf{I}). \quad (3)$$

Fitting the mixed effects model in a Bayesian framework has been addressed previously (Lange *et al.* (1992), Gelfand *et al.* (1995), Chib and Carlin (1999), Gilks *et al.* (1993), and Zeger and Karim (1991)). In our setting, the parameters of interest are given in the list $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{D}, \sigma^2)$. Estimation proceeds by sampling from full conditional posterior distributions for each parameter in a Gibbs sampling scheme. Posterior sampling was implemented in the

statistical software R (R (2012)). Note that time is the only explanatory variable entering the model for CD4 decline, as other covariate information was not available to specify the model.

We use standard conjugate prior specification and select hyperparameters that result in fairly uninformative priors. Analyses of sensitivity to these prior choices are described below. The prior on β is chosen to be $N(\mu, V_\beta)$, with $\mu = (20, 0)'$; this choice reflects the requirement that CD4 counts at time of infection be nonnegative but makes no prior assumptions about the slope of CD4 progression. The prior variance is specified as

$$V_\beta = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}.$$

This prior is dispersed over the plausible range of values; we assessed sensitivity to it by selecting a variety of additional priors of increasing variance, but observed no differences from results reported below. Diagonal structure of the variance matrix V_β implies an *a priori* assumption of independence between the fixed effects, reflecting our lack of prior beliefs regarding the structure of their relationship; we use this prior specification to allow the data to dictate the nature of the dependence. To probe the impact of this assumption, we examined a prior covariance matrix V_β which included a covariance value of 0.3 (selected to be larger than the observed posterior covariance). This prior choice affected neither the overall mean and variance of draws of the fixed effects vector β nor their covariance, so we retained the assumption of prior independence in the fixed effects.

We use a Gamma (a, b) prior on the inverse variance parameter σ^{-2} , parameterized with mean a/b and variance a/b^2 ; we select $(a, b) = (0.1, 0.1)$. In the full conditional distribution for the σ^{-2} parameter, the shape parameter a from the Gamma prior appears in the posterior mean in a sum that includes a term that is a function of sample size; consequently the hyperparameter has minimal effect on the posterior mean so long as it is fairly small in magnitude. To examine sensitivity to values of the hyperparameters, we considered hyperparameters resulting in increasing variance for the Gamma prior, $(a, b) = (0.01, 0.01)$ and $(a, b) = (0.001, 0.001)$. Increasing the prior variance had the effect of slightly reducing the posterior variance for σ^{-2} , with only minor increases in the overall posterior mean. The choice of such small values for (a, b) is common, but yields a prior that is nearly improper and substantially concentrated in the vicinity of zero, which can cause difficulties in convergence (see, for instance, Gelman (2006) for a discussion within the context of hierarchical modeling); the hyperparameters $(a, b) = (0.1, 0.1)$ that we employ reflect our goal of having a minimally informative prior that avoids these convergence difficulties.

We use a Wishart (c, V_D) prior for D^{-1} , and select hyperparameters for a reasonably uninformative prior. We set $c = 2$, and

$$V_D = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}.$$

The Wishart is parameterized to have prior expectation equal to cV_D . To examine sensitivity to this prior choice, we decreased the diagonal elements of V_D by a factor of 10, resulting in decreases in both the mean and variances of the posterior values for elements of the precision matrix; the largest effect was observed in the decrease in the random effect slope variance. This change slightly reduced the mean of the sampled values of the random intercepts and generally also decreased the mean of the random slope effects, while increasing their variance in both cases. As random effects sampled from their posterior distribution are used in the prediction analyses, this increase in variance has the potential to affect the prediction outcomes; we conducted sensitivity analysis to this Wishart specification and discuss impact on the overall prediction values in Section 5.

Sampling chains were run for 30,000 draws, with 10,000 post burn-in draws used for subsequent prediction analyses. Summary information for fixed effect parameters and random effects are given in Table 2. Posterior mode values were determined by computing kernel density estimates and locating parameter values where the density maximum occurred. For a few subjects, overall slope was nonnegative, owing to large variability in their profiles. We used graphical diagnostics to investigate our assumptions of normal specification for the within-group errors and the random effects. Standardized residuals and residuals by subject from maximum likelihood model fits support the assumption that errors are centered at zero with approximately constant variance; quantile-quantile plots of random effects show no significant departures from normality (see Figures 1 – 3 in the Supplementary Material). The residual plots show some outliers, but no major violations of underlying assumptions.

3.3. Deriving posterior distribution for imputation of missing time values

Let N_{xs} be the number of subjects in the household survey with CD4 count above the threshold and $CD4_{k,M}$ be the CD4 count for subject k in the prediction set, $k = (1, \dots, N_{xs})$, taken at the (unobserved) time since seroconversion $t_{k,M}$. For each subject, prediction of time from CD4 monitoring to attainment of threshold is based on imputation of the bivariate pair of missing values, $(t_T, t_M)'$. We develop an imputation model within an inverse-prediction framework based on a derived posterior distribution for the missing values, that makes use of draws from the observed-data posterior distribution for linear random effects model parameters $\theta = (\beta, D, \sigma^2)$, described in Section 3.2. We use Bayesian methods based on Markov Chain Monte Carlo (MCMC) to carry out draws from the derived posterior distribution.

Procedures have been described for multiple imputation of missing times of HIV seroconversion in the context of estimating the distribution of time from seroconversion to AIDS diagnosis, using time-to-event data (Taylor *et al.* (1990)). The missing times were viewed as part of the (possibly censored) outcome, and CD4 counts appeared as covariates in the models for event times. In our case, the missing values $(t_T, t_M)'$ are covariates in our model for CD4 outcomes; this approach raises challenges in the selection of the imputation model (see discussion in Kenward and Carpenter (2007)).

In their discussion of inverse prediction in a Bayesian framework, Aitchison and Dunsmore (1975) provide an estimation technique that makes use of a derived prediction density,

which they term the calibrative density. We expand this posterior density as our model for imputing the missing time components required for our analysis. Our extension accommodates bivariate outcomes, and the dependence over time in the longitudinal data. We characterize the posterior distribution for $(t_T, t_M)'$, derived as follows for a single subject among the N_{xs} subjects in the household survey. For subject k , we have the pairs

$(\begin{smallmatrix} CD4_{k,T}, t_{k,T} \\ CD4_{k,M}, t_{k,M} \end{smallmatrix})$, where $(t_{k,T}, t_{k,M})'$ are viewed as missing data. We also have data $(y_{i,j}, t_{i,j})$ for subject i in the longitudinal cohort, concatenated into $(\mathbf{y}, \mathbf{t})$ for the entire dataset.

Suppressing the subscript k , we let $f(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta})$ be the parametric model for $\mathbf{y}$ from the mixed effects model. Parametric representations for $\mathbf{y}$ represent either the marginal, population-averaged profile integrated across the i subjects in the longitudinal cohort, or an individual profile with mean response at the combined fixed and random effects. These allow accommodation of the correlation structure in the repeated measurements on the subjects in the longitudinal cohort.

Assuming prior independence of the vector of mixed effects model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{D}, \sigma^2)$ and the vector of missing time values $(t_T, t_M)'$, we have:

$$p\left(\begin{pmatrix} t_T \\ t_M \end{pmatrix}, \boldsymbol{\theta}\right) = p\left(\begin{pmatrix} t_T \\ t_M \end{pmatrix}\right) p(\boldsymbol{\theta})$$

The unnormalized joint posterior for all unknowns, conditional on observed quantities, is given by:

$$\begin{aligned} p\left(\begin{pmatrix} t_T \\ t_M \end{pmatrix}, \boldsymbol{\theta} \middle| \mathbf{y}, \mathbf{t}, \begin{pmatrix} CD4_T \\ CD4_M \end{pmatrix}\right) &\propto p\left(\begin{pmatrix} t_T \\ t_M \end{pmatrix}, \boldsymbol{\theta}\right) p\left(\mathbf{y}, \begin{pmatrix} CD4_T \\ CD4_M \end{pmatrix} \middle| \begin{pmatrix} t_T \\ t_M \end{pmatrix}, \mathbf{t}, \boldsymbol{\theta}\right) \\ &\propto p\left(\begin{pmatrix} t_T \\ t_M \end{pmatrix}\right) p(\boldsymbol{\theta}) p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta}) p\left(\begin{pmatrix} CD4_T \\ CD4_M \end{pmatrix} \middle| \begin{pmatrix} t_T \\ t_M \end{pmatrix}, \boldsymbol{\theta}\right) \end{aligned}$$

where $p(\boldsymbol{\theta})p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta})$ is the observed-data posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t})$ of $\boldsymbol{\theta}$ given the training set. Integrating the joint posterior of $((t_T, t_M)', \boldsymbol{\theta})$ over $\boldsymbol{\theta}$ gives the desired posterior of the bivariate times, the conditional distribution of W_{mis} given the observed data:

$$p\left(\begin{pmatrix} t_T \\ t_M \end{pmatrix} \middle| \mathbf{y}, \mathbf{t}, \begin{pmatrix} CD4_T \\ CD4_M \end{pmatrix}\right) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}) p\left(\begin{pmatrix} CD4_T \\ CD4_M \end{pmatrix} \middle| \begin{pmatrix} t_T \\ t_M \end{pmatrix}, \boldsymbol{\theta}\right) p\left(\begin{pmatrix} t_T \\ t_M \end{pmatrix}\right) d\boldsymbol{\theta}. \quad (4)$$

The expression above represents the posterior predictive distribution of the missing data given the observed data, $p(W_{mis}|W_{obs})$.

We effect the integration in Equation 4 by using posterior draws of the mixed effects model parameters $\boldsymbol{\theta}$ (obtained as described in Section 3.2) across their range of posterior sampled

values. We draw samples of $(t_T, t_M)'$ from their posterior predictive distribution following prior specification for t_T and t_M . Sampling across parameter values has been described in a similar setting (Lange *et al.* (1992)), but not in this predictive posterior framework; previous methods did not allow incorporation of prior knowledge about the times flexible of interest in the prior specification. As the sampling scheme allows flexible choices for incorporation of prior information, we are not limited to closed form, computationally tractable posteriors. Correlation is induced in the sampled times via the sampling procedures, as well as via prior specification. Note that we assume the model fitted to the data from the longitudinal cohort applies to the survey population.

3.4. Prior specification for missing times

The choice of prior specification for the pair of missing values (t_T, t_M) reflects knowledge that $t_T - t_M > 0$ and that their correlation is nonzero. An inverse Gaussian (IG) distribution, which arises as the first passage time distribution of Brownian motion (see Folks and Chhikara (1978)), is therefore an appropriate choice of prior. Coupled with an IG prior on t_T (assumed independent), there is sufficient information to induce a prior on the bivariate pair (t_T, t_M) via a transformation of variables (see Seaman *et al.* (2012)); this prior induces a correlation structure between t_T and t_M , and *a priori* specifies the constraint that $t_T > t_M$. The

IG is parameterized in (μ, λ) , with mean μ and variance $\frac{\mu^3}{\lambda}$. We chose hyperparameter values of $(\mu_2 = 10, \lambda_2 = 11)$ for the prior on t_T , and $(\mu_1 = 5, \lambda_1 = 16)$ for the independent prior on the difference $t_T - t_M$. These priors are centered around plausible values for the quantities under consideration, but with large enough variance to be reasonably dispersed over a wide range of values. In particular, we considered recent reports on rates of disease progression in HIV-1 Subtype C infected individuals in choosing the prior on t_T ; recent work examining the endpoint of reaching <350 cells/mm³ CD4 have indicated median time-to-event of approximately 5 years (Amornkul *et al.* (2013)); we chose hyperparameter values for the prior on t_T with a larger median value of around 7 to reflect the lower threshold <250 cells/mm³ used in this study and the estimated rates of decline. The independent prior on $t_T - t_M$ was selected to have smaller median than the prior on t_T , to reflect the nonnegativity constraint.

Because the prior specification for (t_T, t_M) is informative, we performed extensive sensitivity analyses to this choice. They demonstrate robustness of our estimation procedures to a broad range of prior specifications. IG priors with the same mean hyperparameters, but with λ hyperparameters reduced to values of 8 and 5 (which increased their variances) had minimal impact on the overall distribution of median values of t_A across the population, causing only a slight downward shift in the distribution for the largest variances examined. Increasing the mean parameters μ_1 and μ_2 (for example, to $\mu_1 = 8$ and $\mu_2 = 12$) resulted in no observable changes in the distribution of median t_A values.

Additional sensitivity analyses examined alternative choices for prior distributions. One was a bivariate inverse Gaussian prior using the form specified by Kocherlakota (Kocherlakota (1986)), which allowed direct specification of the correlation ρ between t_T and t_M . We considered a range of ρ from 0 to 0.90 at different levels of prior variability. The correlations

between chains of t_T and t_M did not vary by ρ , nor did the distribution of the median values of t_A . Another choice, which again had little impact, was independent skew normal priors with differing levels of variance on both t_T and t_M . These analyses provide evidence for the robustness of the procedures to choices of informative prior specification.

3.5. Population parameter of interest

From draws from the posterior predictive distribution of W_{mis} for each subject, we sample M values of $t_A = t_T - t_M$, creating M sets of values for all subjects in the prediction set. Let Q represent the target of our analysis, the mean over the population of the amount of time to reach treatable CD4 level. Information from the M datasets can be combined using repeated-imputation inference to estimate our target (see Zhang (2003); Schafer (1999)). The observed-data posterior for Q is derived from integrating the complete-data posterior of Q over the posterior predictive distribution of W_{mis} , and is given by:

$$p(Q|W_{obs}) = \int p(Q|W_{obs}, W_{mis})p(W_{mis}|W_{obs})dW_{mis}. \quad (5)$$

For each of the M imputed datasets we can compute a value of $\hat{Q}$, an estimate of the mean additional time on treatment and $\hat{U}$, an associated squared standard error for this mean, where the standard error is $\hat{\sigma} / \sqrt{N}$. We obtain moments of the observed data posterior distribution as follows:

$$E(Q|W_{obs}) = E[E(Q|W_{obs}, W_{mis})|W_{obs}], \quad (6)$$

$$V(Q|W_{obs}) = E[V(Q|W_{obs}, W_{mis})|W_{obs}] + V[E(Q|W_{obs}, W_{mis})|W_{obs}]. \quad (7)$$

These expressions can be approximated from the mean of the M values of $\hat{Q}$, and using $(\hat{Q}) + (1 + M^{-1})\text{Var}(\hat{Q})$, where $\bar{Q}$ is the average of the $\hat{Q}$ values. The mean of $\hat{Q}$, $\bar{Q}$, provides an overall estimate of Q , and has associated standard error given by:

$$SE(\bar{Q}) = \{(1 + M^{-1})B + \bar{U}\}^{1/2}, \quad (8)$$

with

$$B = \frac{1}{M-1} \sum_{m=1}^M \left(\hat{Q}^{(m)} - \bar{Q} \right)^2, \quad (9)$$

and

$$\bar{U} = \frac{1}{M} \sum_{m=1}^M \hat{U}^{(m)}, \quad (10)$$

the between- and within- imputation variances, respectively.

4. Computational implementation

Sampling from the bivariate posterior distribution derived in Section 3.3 proceeds via a Metropolis-Hastings algorithm. The algorithm requires evaluation of the target distribution $H(\cdot)$, where $H(\cdot)$ is our posterior predictive distribution of W_{mis} . Using the mixed effects model described in Section 3.2, we can take the correlation structure into account by using either the mean structure for $\mathbf{y}$ and working with the conditional distribution given the random effects (Equation 2) or the marginal model of $\mathbf{y}$; for the latter, the correlation appears in the variance structure (Equation 3). Using the marginal model, computational difficulties arose for subjects who had large monitored CD4 levels, so we instead made use of the distribution conditioned on the random effects. The model for CD4 that appears in the integrand in Equation 4 assumes:

$$p\left(\begin{pmatrix} CD4_T \\ CD4_M \end{pmatrix} \middle| \boldsymbol{\theta}, \mathbf{b}, \begin{pmatrix} t_T \\ t_M \end{pmatrix}\right) \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2 \mathbf{I}), \quad (11)$$

where

$$\mathbf{X} = \mathbf{Z} = \begin{pmatrix} 1t_T \\ 1t_M \end{pmatrix}. \quad (12)$$

Because they are associated only with subjects in the longitudinal cohort data, the resulting random effects cannot be used directly. Therefore, we generate random effects $\mathbf{b}$ from their $N(\mathbf{0}, \mathbf{D})$ distribution, and constrain them to ranges of values relevant for each subject in the prediction set. The two constraints are:

$$\beta_1 + b_1 < 0$$

and

$$y - (\beta_0 + b_0) < 0,$$

where β_1 is the fixed effects slope representing the average decline across the longitudinal population in CD4 level over time; b_1 is a random slope effect for subject k ; β_0 is the fixed effect intercept term; and b_0 represents a subject's deviation around that term. These constraints reflect knowledge that HIV infection results in progressive decline of CD4 counts, although at different rates, and that CD4 count at the time of seroconversion is higher than at the monitoring point. Details of the implementation of the Metropolis-Hastings algorithm appear in the Supplementary Material.

5. Application in HIV household survey study

5.1. Results from estimation procedures

For each subject in the cross-sectional household survey, the sampling procedures result in chains of posterior draws of the bivariate pair $(t_T, t_M)'$. Representative traceplots of chains of

$(t_T, t_M)'$ for two subjects with small and large CD4 levels at monitoring are given for comparison in Figure 3. As expected, for a given subject, the posterior values of t_A exhibit substantial variability (see Figure 4). 95% credible sets for the individual posterior medians tend to be wide; total interval widths range from a minimum of 5.5 to a maximum of 10.5 years around the medians. The 95% credible sets for individual chains show a fair amount of skewness for subjects with CD4 values near the threshold at monitoring, resulting from their similar predicted values of t_T and t_M . As a consequence, the generated t_A values tend to be quite small, even though fairly extreme values can be visited, resulting in skewness of their t_A distributions.

To combine the information from the multiply-imputed datasets for estimating our population-level target of inference, we took 10 draws from each subject's posterior chains of t_A . For the CD4 threshold equal to 250 copies/mm³, the population mean time to threshold is estimated to be 3.4 years, with standard error = 0.25 (computed using Eq. 8), based on $M = 10$ imputed datasets. For the CD4 threshold of 350 copies/mm³, the estimated population mean time to threshold is 3.0 years (SE 0.26). Results from our prediction analyses demonstrate the expected trend of increasing predicted additional time on treatment, t_A , with increasing measured CD4 level at time of monitoring across all subjects (see Figure 5). This result provides some validation for the way in which the modeling procedures handle CD4 counts, and demonstrates that results are not overly tracking the informative prior specification. As sensitivity analyses to prior choices showed some influence of the Wishart hyperparameters on parameter draws, we examined the impact of the alternative Wishart prior specification discussed in Section 3.2 on overall results. While these alternative hyperparameter values did have some influence on the drawn values of the precision matrix $\mathbf{D}^{-1}$, the overall prediction results were very little affected. The modified Wishart prior led to only a slight decrease in overall mean and variance of the distribution of median values of t_A and a very small decrease in population mean value of t_A compared to the results described above.

5.2. Accounting for length bias in sampling procedures

Our analytical approach reflects and accommodates the length bias of our cross-sectional sample (see Zelen and Feinleib (1969), Brookmeyer *et al.* (1987)); infected people in Mochudi are more likely to be included in the household survey if they have more recent infection or a slower decline of CD4 count. Figure 6 illustrates a hypothetical scenario in which individuals (arbitrarily observed at the same level of CD4) with less recent infection times relative to the monitoring time can only be observed in the household survey if they have relatively shallow rates of CD4 decline. Individuals with more recent infection are likely to be included in the cross-sectional sample, regardless of rate of CD4 decline. Prediction intervals resulting from the sampling methods developed above are consistent with results expected from observing a length-biased sample. Higher CD4 value at monitoring, indicative of a more recent infection, is associated with a wider range of slopes, including the steepest slopes. In our sampling-based methods, when parameter values used in evaluating the proposal are associated with a rapid CD4 decline, the probability of acceptance of a proposed $(t_T, t_M)'$ is higher when the infection is more recent (i.e. at larger CD4 values). Similarly, subjects in the cross-sectional sample with small CD4 values at the

monitoring point have higher acceptance probabilities only for proposed values associated with shallower slope parameters. Figure 7 provides representative illustrations of the ranges of slope values at which acceptance of the proposed $(t_T, t_M)'$ occurs.

5.3. Assessment of convergence

Convergence of chains was examined graphically by overlaying traceplots from multiple dispersed starting values to assess whether chains showed similar behavior and strong post burn-in overlap when initiated from different locations in $\mathcal{R}^2$. While individual profiles could be variable, resulting median values for the dispersed chains were virtually identical for all subjects in the prediction set. In addition, we examined the Gelman-Rubin convergence diagnostic metric, implemented using the CODA package (Best *et al.* (1995)) in R. This diagnostic method produces the ‘potential scale reduction factor,’ where values near one indicate parallel sequences are nearly overlapping (Gelman and Rubin (1992)). For the entire prediction set, we ran three sequences of 6000 iterations from dispersed starting values and used 5000 post burn-in values in computing the Gelman-Rubin diagnostic. Diagnostic values from subjects with small, medium and large CD4 values at monitoring are provided to illustrate convergence properties of t_A in Table 3. This observation, coupled with the observed robustness of results to our choice of prior, demonstrate that the modeling procedures using the information in the auxiliary profile data and the single CD4 value are very stably locating a consistent value of t_A .

6. Discussion and conclusion

Our analyses estimate the population-averaged additional amount of person-years of treatment required to treat all viremic patients at the time of HIV detection. The quantity t_A , the time for an individual to reach a treatable CD4 threshold from the monitoring point at which HIV and CD4 status is ascertained, serves as a proxy for the additional time that individual would spend on ART if begun immediately compared to beginning according to standard guidelines. Our estimation approach uses a missing-data framework, which allows us to incorporate errors arising from measurement (which can lead to multiple crossing times) and from model parameter estimation. The variability arising from the former is incorporated in the overall posterior distribution, which contains the appropriate variance term.

We know of no alternative methods that might be used for comparison to the proposed approach, although studies in which unobservable HIV infection times are relevant for estimating quantities of interest have been described (see, for instance, Taylor *et al.* (1990)). Our work may aid in analyses of results from such studies, in that it builds on Bayesian methods for combining different data sources, the data from which may be incomplete. Previous methods used to predict time to reaching CD4 threshold from longitudinal measurements of CD4 over time have focused on the classical calibration estimator, which inverts the relationship used to characterize CD4 as a function of time; the specified relationship has typically taken the form of a linear mixed effects model (Lange *et al.* (1992)). Such approaches can incorporate the variability in the parameters of the accompanying mixed effects model, but do not expressly allow incorporation of prior

information about the unobserved time to threshold from informative prior specification. Lange et al. (Lange *et al.* (1992)) work with something similar to the classical calibration estimator (Osborne (1991)) in a Bayesian framework for time to reaching threshold using CD4 profiles, incorporating posterior draws of the parameters associated with β_0 and β_1 , but without a directly specified (informative) prior on the unknown t_0 . Their estimator of t_0 made use of the fixed effect parameters from their linear profile fits, but did not expressly use the correlation structure assumed for the random effects. A similar approach has been examined in a Bayesian framework for predicting time to threshold in an aneurysm growth model (Sweeting and Thompson (2011)). In addition, Sweeting and others (Sweeting *et al.* (2010)) use an inverse prediction method that makes use of prior specification for an unknown time of seroconversion to estimate the distribution of the window period in recent HIV infections.

Recent results demonstrating an impact of ART in reducing HIV transmission in serodiscordant couples (Cohen *et al.* (2011)) motivate our investigation, which can aid in assessing the feasibility of widescale use of ART as a HIV prevention intervention mode. Cost-effectiveness analyses have examined ‘test-and-treat’ prevention interventions in resource-limited settings (Dodd *et al.* (2010)), but our efforts are based on evidence from actual field studies in Botswana. Therefore we are able to estimate the amount of additional treatment needed to introduce a policy of treat at time of detection of HIV infection for a specific population. Our results demonstrate that for the monitored population in Mochudi, the mean additional time on treatment under the treatment guideline of 250 copies/mm³ would be 3.4 years (with standard error of 0.25); this requirement for additional treatment is quite modest, given that patients may stay on these drugs for decades. Further modeling efforts are required to establish that cost of this treatment would ultimately be less than the cost associated with treating the additional infections that would result under the current treatment policy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Dr. Francesca Dominici for helpful comments about formulating the problem, and Drs. Vladimir Novitsky and Rui Wang for assistance with the analysis of the longitudinal cohort data. We thank Dr. Max Essex, Hermann Bussmann and researchers associated with the Mochudi pilot project for providing data and discussing the context in which they were collected. We also thank the editor, associate editor, and referees for numerous helpful comments and suggestions that clarified and improved the work. We acknowledge grant support from NIH A1 R37 51164 and NIH R01 A1083036.

References

- Aitchison, J.; Dunsmore, IR. Statistical Prediction Analysis. Cambridge, UK: University Press; 1975.
- Amornkul P, Karita E, Kamali A, et al. Disease progression by infecting HIV-1 subtype in a seroconverter cohort in sub-Saharan Africa. AIDS. 2013; 27:2775–2786. [PubMed: 24113395]
- Best, N.; Cowles, M.; Vines, K. Coda: Convergence diagnosis and output analysis software for gibbs sampling output, version 0.30 Technical report, MRC Biostatistics Unit. University of Cambridge; 1995.

- Botswana Ministry of Health. [Accessed April 27, 2012] Botswana National HIV/AIDS Treatment Guidelines, 2008 Version. 2012. Available from www.moh.gov.bw/templates/moh/File/BOTSWANA_HIVAIDS_TREATMENT_GUIDELINES (November 1 2008).pdf
- Brookmeyer R, Gail MH, Polk BF. The prevalent cohort study and the acquired immunodeficiency syndrome. *American Journal of Epidemiology*. 1987; 126:14–24. [PubMed: 3647718]
- Brown PJ. Multivariate calibration. *J. R. Statist. Soc. B*. 1982; 44:287–321.
- Chib S, Carlin BP. On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing*. 1999; 9:17–26.
- Cohen MS, Chen YQ, McCauley M, et al. Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine*. 2011; 365:493–505. [PubMed: 21767103]
- DeGruttola V, Lange N, Dafni V. Modeling the progression of HIV infection. *J. Am. Statist. Ass.* 1991; 86:569–577.
- Dodd PJ, Garnett GP, Hallett TB. Examining the promise of HIV elimination by ‘test and treat’ in hyperendemic settings. *AIDS*. 2010; 24:729–735. [PubMed: 20154580]
- Donnell D, Baeten JM, Kiarie J, et al. Heterosexual HIV-1 transmission after initiation of antiretroviral therapy: A prospective cohort analysis. *Lancet*. 2010; 375:2092–2098. [PubMed: 20537376]
- Folks JL, Chhikara RS. The Inverse Gaussian Distribution and Its Statistical Application—A Review. *J. R. Statist. Soc. B*. 1978; 40:263–289.
- Gelfand AE, Sahu SK, Carlin BP. Efficient parametrisations for normal linear mixed models. *Biometrika*. 1995; 82:479–488.
- Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*. 2006; 1:515–533.
- Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. *Statistical Science*. 1992; 7:457–511.
- Gilks WR, Wang CC, Yvonnet B, Coursaget P. Random-effects models for longitudinal data using Gibbs sampling. *Biometrics*. 1993; 38:441–453. [PubMed: 8369380]
- Hoadley B. A Bayesian look at inverse linear regression. *J. Am. Statist. Ass.* 1970; 65:356–369. (1970).
- Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Statistical Methods in Medical Research*. 2007; 16:199–218. [PubMed: 17621468]
- Kocherlakota S. The bivariate inverse gaussian distribution: an introduction. *Communications in Statistics - Theory and Methods*. 1986; 15:1081–1112. 4, (1986).
- Lange N, Carlin BP, Gelfand AE. Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *J. Am. Statist. Ass.* 1992; 87:615–626.
- Lihana R, Ssemwanga D, Abimiku A, Ndambi N. Update on HIV-1 Diversity in Africa: A Decade in Review. *AIDS Rev*. 2012; 14:83–100. [PubMed: 22627605]
- National AIDS Coordinating Agency. HIV/AIDS in Botswana: Estimated Trends and Implications Based on Surveillance and Modeling. 2008 Accessed October, 2013 (Available from http://www.unaids.org/en/dataanalysis/knowyourepidemic/countryreport-sonhivestimates/20080701_botswana_nationalestimate2007_en.pdf)
- Novitsky V, Ndung'u T, Wang R, Bussmann H, Chonco F, Makhema J, DeGruttola V, Walker BD, Essex M. Extended high viremia: A substantial fraction of individuals maintain high plasma viral RNA levels after acute HIV-1 subtype C infection. *AIDS*. 2011; 25:1515–1522. [PubMed: 21505307]
- Osborne C. Statistical calibration: A review. *International Statistical Review*. 1991; 59:309–336.
- R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria: Vienna; 2012. URL <http://www.R-project.org>
- Schafer JL. Multiple imputation: A primer. *Statistical Methods in Medical Research*. 1999; 8:3–15. [PubMed: 10347857]
- Seaman JW III, Seaman JW Jr, Stamey JD. Hidden dangers of specifying noninformative priors. *The American Statistician*. 2012; 66:77–84.

- Sweeting MJ, De Angelis D, Parry J, Suligoi B. Estimating the distribution of the window period for recent HIV infections: A comparison of statistical methods. *Statistics in Medicine*. 2010; 29:3194–3202. [PubMed: 21170913]
- Sweeting MJ, Thompson SG. Making predictions from complex longitudinal data with application to planning monitoring intervals in a national screening programme. *J. R. Statist. Soc. A*. 2011; 175:1–18.
- Taylor JMG, Muñoz A, Bass SM, Saah AJ, Chmiel JS, Kingsley LA. Multicenter Aids Cohort Study. Estimation the distribution of times from HIV seroconversion to AIDS using multiple imputation. *Statistics in Medicine*. 1990; 9:505–514. [PubMed: 2190287]
- Zeger SL, Karim MR. Generalized linear models with random effects: A Gibbs sampling approach. *J. Am. Statist. Ass.* 1991; 86(413):79–86.
- Zelen M, Feinleib M. On the theory of screening for chronic diseases. *Biometrika*. 1969; 56:601–614.
- Zhang P. Multiple imputation: Theory and method. *International Statistical Review*. 2003; 71:581–592.

Fig. 1.

Longitudinal measurements from 4 representative subjects from the incident cohort. CD4 values are given on the square root scale. The population-averaged profile resulting from fitting a linear mixed effects model is given for reference.

Fig. 2.

Time relationships between points of interest. I is the zero time point, the unobserved time of seroconversion ($t_1 = 0$). M is the monitoring point at which the cross-sectional measurement of CD4 is effected at time t_M . T is the treatable CD4 threshold reached at time t_T . The target of inference is t_A , the time from monitoring to reaching CD4 threshold.

Fig. 3. Representative traceplots of t_T (black) and t_M (gray) for subjects with smaller (upper panel) and larger (lower panel) CD4 cell counts at time of monitoring, shown for the first 5000 iterations.

Fig. 4.

Representative traceplot of t_A . This traceplot is associated with a subject with $\sqrt{CD4}$ value = 19. Median value show with horizontal line.

Fig. 5. Median of sampled predicted times from CD4 monitoring to reaching threshold = 250 copies/mm³ (with interquartile credible set shown in gray) versus subject's $\sqrt{CD4}$ value.

Fig. 6.

Schematic representation of length-biasing as it would be manifested in 3 hypothetical individuals who have infection times (shown as solid symbols on the horizontal line representing a CD4 level) prior to the monitoring point. For each individual, 3 slope values are shown (in different line types) to demonstrate the possible range of rates of CD4 decline following infection. Individuals whose CD4 line crosses the vertical bar representing the monitoring point above the threshold level have the potential to be observed in the sample. Point of crossing for an individual is shown by the matching symbol on the vertical monitoring time bar. Individuals with recent infection are more likely to appear in the sample.

Fig. 7.

Boxplots of the total slope parameters ($\beta_1 + b_1$) that were used in the iterations in which acceptance of proposed values of $(t_T, t_M)'$ occurred, shown for subjects with very small (left) and large (right) CD4 values at monitoring.

Table 1

Summary of characteristics of CD4 values for subjects in the prediction dataset. The total dataset included all HIV positive treatment-naïve subjects from the 2010–2011 household survey in Mochudi, Botswana before removing those who were already below treatable CD4 threshold. CD4 values are on the square root scale, except for the thresholds which are in the original units of copies/mm³.

	Total Dataset (N = 244)	Threshold > 250 (N = 176)	Threshold > 350 (N = 120)
Minimum	5.66	16.0	18.7
1st Quartile	15.5	18.3	20.3
Mean	19.1	21.6	23.6
3rd Quartile	22.6	24.1	25.9
Maximum	36.7	36.7	36.7

Table 2

Summary of features of posterior distributions associated with fixed effects parameters and the 2×2 random effects precision matrix $\mathbf{D}^{-1}$, for models fit to longitudinal cohort data. The average correlation between the random slope and random intercept is -0.204 .

Parameter	Mode	Mean	Standard Deviation	95% Credible Set
β_0	21.22	21.07	0.526	20.27, 21.96
β_1	-1.46	-1.45	0.364	-2.10, -0.762
σ^2	0.225	0.210	0.079	0.146, 0.252
$(D^{-1})_{11}$	0.072	0.072	0.012	0.050, 0.098
$(D^{-1})_{12}$	0.023	0.025	0.018	-0.010, 0.063
$(D^{-1})_{22}$	0.186	0.202	0.047	0.123, 0.310

Table 3

Gelman-Rubin diagnostics for subjects with CD4 values of 16.48, 20.35 and 30.04 at monitoring time, which is the 0.05, 0.50, and 0.95% quantiles for the CD4 data.

Point estimate	Upper limit
1.01	1.01
1.00	1.00
1.01	1.03