Published in final edited form as:
Stat Med. 2012 August 30; 31(19): 2086–2097. doi:10.1002/sim.5318.

Efficient use of longitudinal CD4 counts and viral load measures in survival analysis

S.E. Holte[1], **T.W. Randolph**[1], **J. Ding**[2], **J. Tien**[3], **R.S. McClelland**[4], **J.M. Baeten**[5], and **J. Overbaugh**[6]

[1]Fred Hutchinson Cancer Research Center - Biostatistics and Biomathematics, Seattle, Washington

[2]Washington University - Mathematics and Statistics, St Louis, Missouri

[3]Ohio State University – Mathematics, Columbus, Ohio

[4]University of Washington – Medicine, Seattle, Washington

[5]University of Washington - Allergy and Infectious Disease, Seattle, Washington

[6]Fred Hutchinson Cancer Research Center - Human Biology, Seattle, Washington

Abstract

CD4 counts and viral loads are dynamic quantities that change with time in HIV-infected persons. Commonly used single summary measures, such as viral load set point or early CD4 count do not explicitly account for changes in viral load or CD4 counts or other features of the overall time course of these measures. However, the efficient use of all repeated measurements within each subject is often a challenge made more difficult by sparse and irregular sampling over time. Here we illustrate how functional principal component (FPC) analysis provides an effective statistical approach for exploiting the patterns in CD4 count and viral load data over time. The method is demonstrated using data from Kenyan women who acquired HIV-1 during follow-up in a high risk cohort and were subsequently followed prospectively from early infection. The FPC scores for each woman obtained by this method serve as informative summary statistics for the CD4-count and viral-load trajectories. Similar to baseline CD4 count or viral set point, the first FPC score can be interpreted as a single-value summary measure of an individual's overall CD4 count or viral load. However, unlike most single-value summaries of CD4-count or viral-load trajectories, the first FPC score summarizes the dynamics of these quantities and is seen to reveal specific features of the trajectories associated with mortality in this cohort. Moreover, FPC scores are shown to be a more powerful prognostic factor than other common summaries when used in survival analysis.

Keywords

longitudinal data; functional principal components; CD4 counts; viral loads

Introduction

Repeated measurements of the HIV viral load in plasma and CD4 T cell counts in an HIV-1 infected individual over time represent a sampling of continuous trajectories for these dynamic quantities. A variety of summaries for these trajectories have been used in HIV

Correspondence to: Sarah Elizabeth Holte, Fred Hutchinson Cancer Research Center, Biostatistics and Biomathematics, 1100 Fairview Avenue North, M2-B500, Seattle, Washington 98109.

research, such as the first-observed CD4 count or viral load, the estimated viral set point, and the estimated slope of CD4 count trajectory. Numerous studies among untreated men and women in North America and Europe have shown that these summary measures of viral load and CD4 trajectories predict survival to some degree [1–5]. Studies in African women demonstrated similar associations between survival and summary measures including baseline viral load [6, 7] or baseline CD4 percentage [6]. Although these single value summaries are useful, most fail to account for the dynamic patterns in longitudinal data which are appropriately viewed as curves or functions over time. The analysis of viral load and CD4 count trajectories is often complicated by the sparsity and irregularity of the measurements. Indeed, estimating the full trajectory from a small number of measurements from each individual is problematic and comparison across individuals is made difficult by the fact that time periods represented by measurements from two individuals may not even overlap.

Methods for obtaining summaries of longitudinal data of this type include random-effects models using linear or piecewise linear mixed effects [2, 8–10] to obtain individual viral load or CD4 count slopes. These models assume a certain trajectory shape and they require choices about portions of the trajectories where the assumed parametric shape is appropriate. Other methods for capturing patterns in repeated measures of CD4 and viral load trajectories after treatment with anti-viral medications include mechanistic models using differential equations [11, 12]. This approach produces one or more parameters that characterize the trajectories, but imposes restrictive structure on the shape of the trajectories and fitting these models requires a substantial number of measurements from each subject. Subsequent statistical inference is difficult to interpret in these models since the estimation is based on subjective modeling assumptions. Furthermore, the models can be computationally extremely complex, especially when model parameters are treated as random effects in aggregate population analysis.

An alternative, flexible statistical framework that is well-suited to obtaining summaries for sparsely and irregularly sampled longitudinal data is provided by recent functional principal component analysis (FPC analysis) [13–15] which extends the methods for densely sampled functional data [16]. Sparsity and irregularity of the data are dealt with using an implementation described by Yao et al. [14] in which the functional principal components are estimated through conditional expectation (termed PACE). Importantly, this approach allows for a smooth estimate of covariance among the composite of all subjects' trajectories even when individuals have relatively few measurements. Moreover, no priori assumptions or constraints are placed on the shape of trajectories since the estimated trajectories arise naturally from the PACE-derived principal component functions and scores. Finally, the FPC approach is designed to capture the most subject-to-subject variability among all possible summary statistics. The FPC scores obtained in this type of analysis can be used as a summary measure in place of the more common measures such as a single CD4 measurement or estimated set point for viral load. Yao et al. [14] analyzed a dataset of CD4 count trajectories from treatment-naïve participants in the Multicenter AIDS Cohort Study [17, 18] and characterized the overall and subject specific trends in CD4 counts over time, although CD4 count profiles were not related to any specific outcome, such as survival. In a related use of FPCs, Yao [15] explored joint modeling of survival and longitudinal CD4 counts and evaluated the combined role of CD4 trajectories and treatment arm using data from a clinical trial of zalcitabine (ddC) compared to didanosine (ddI) treatment [19]. Lower CD4 counts were shown to be significantly associated with an increased risk of mortality and patients in the didanosine arm had a CD4 count adjusted increased risk of mortality compared to patients in the zalcitabine arm.

The primary focus of this work is to use a two-stage approach to examine the prognostic power of the first FPC score (for both viral load and CD4 count data) on survival in the Cox proportional hazards models and to compare these results to other two-stage analysis which use common prognostic indicators of survival such as the early CD4 count, estimated viral set point, or estimated viral load or CD4 slope. Previous analyses of data from a cohort of female Kenyan sex workers acutely infected with HIV-1 found that plasma HIV-1 viral set point predicted subsequent survival but failed to show a significant association between early CD4 count (defined as the first CD4 measurement occurring 4 to 24 months after the estimated date of HIV acquisition) with survival [7], even though it is commonly accepted that higher CD4 counts are associated with increased survival rates. We illustrate how a two-stage FPC and survival analysis of these data can be used to extract enough information to show that a significant association between CD4 count profile and survival does indeed exist. In contrast to the approach described in Yao [15], we consider specific features of the "shape" of CD4 count trajectories rather than the actual CD4 counts as prognostic factors of mortality in this paper. Our model, in which FPC scores are prognostic factors, reflects the clinical belief that survival may be affected by the change and other patterns of CD4 counts and viral loads over time.

Our approach first uses FPC analysis with data from the entire cohort to extract the primary structure in individual trajectories and provides a concise summary of each trajectory–the FPC score–without presuming a parametric model for its shape. In the second stage, this FPC score is then used as a covariate in a survival model to evaluate the relationship between CD4 or viral load trajectories and risk of mortality. We also show how an FPC analysis of these data can be used to reveal clinical features of the CD4 and viral load trajectories that are predictors of mortality risk in these data.

Methods

Study Population

Women in this analysis were identified from a prospective study of female sex workers in Mombasa Kenya. Of 1579 women enrolled in the cohort between February 1993 and March 2004 when antiretroviral therapy became available, 265 became infected with HIV-1 and 218 of these had a date of HIV-1 acquisition which could be estimated with reliable precision as defined by a combination of serology and HIV RNA testing [7]. Of these, 216 had at least one observed viral load or CD4 count after HIV-1 acquisition. These 216 women are the focus of the analysis in this work. Additional details and demographics of the cohort are described in [7] and the results section.

Statistical Analysis

Functional Principal Component Analysis—Sparse functional principal components analysis (FPCA) is performed following methods developed by Yao, Mueller, Wang and others [14, 20, 21]. We briefly summarize this work as it relates to the model used for estimating the FPC scores. Specifically, the sparsely sampled trajectories are viewed as noisy samples from a set of independent realizations of a smooth random function, X, with unknown mean $\mu(t) = EX(t)$ and covariance function $G(s,t) = \text{cov}(X(s), X(t))$, for s,t in a closed and bounded time interval T. The autocovariance operator, K, defined on the space of square-integrable functions on T, is $(Kf)(t) = \int_T G(t,s)f(s)\,ds$. When G is square-integrable on $T \times T$, then K has an orthornormal set of eigenfunctions $\{\phi_1, \phi_2, \ldots\}$ and corresponding eigenvalues $\{\lambda_1 \geq \lambda_2 \geq \ldots \geq 0\}$ (i.e., $K\phi_k = \lambda_k\phi_k$, $k = 1,2, \ldots$) such that $G(t,s) = \sum_{k\geq 1} \lambda_k\phi_k(t)\phi_k(s)$. Moreover, an individual trajectory x_i can be expressed as $x_i(t) = \mu(t) + \sum_{k\geq 1} \xi_{ik}\phi_k(t)$, where $\{\xi_{ik}\}_{k\geq 1}$ is a set of uncorrelated random coefficients having mean 0 and $\text{var}(\xi_{ik}) = \lambda_k$. We refer to ϕ_k as the k^{th} *functional principal component* (FPC). The k^{th} *FPC*

score for x is defined as the coefficient $\xi_k = \xi_k(x)$. These scores act as random effects for models using an FPCs basis representation. For the i^{th} trajectory, $x_i$the FPC scores may be calculated as

$$\xi_{ik}=\int_T [x_i(t)-\mu(t)]\phi_k(t)dt, \text{where} E[\xi_{ik}]=0, E[\xi_{ij}\xi_{ik}]=0 (j \neq k), \text{and} E[\xi_{ik}^2]=\lambda_k. \quad (1)$$

Now, for irregularly sampled data, each trajectory x_i is observed at different timepoints, $x_i(t_{ij})$, j = 1, …, N_i. The number of measurements, N_i, observed for the i^{th} subject is assumed to be random, iid, and independent of other random variables in the model. To allow for additive measurement error, Yao et al. [14, 20, 21] incorporate uncorrelated errors with mean 0 and constant variance: let ε_{ij} be mean zero iid measurement errors, with $\text{var}(\varepsilon_{ij}) = \sigma^2$, and assumed to be independent of the random scores ξ_{ik}, for i = 1, …, n; j = 1, …, N_i; k = 1, 2, …. Letting y_{ij} denote the observation of $x_i(t_{ij})$ of the i^{th} trajectory at the j^{th} timepoint made with error ε_{ij} (with $E\varepsilon_{ij} = 0$, $\text{var}(\varepsilon_{ij}) = \sigma^2$), the model is

$$y_{ij}=x_i(t_{ij})+\varepsilon_{ij}=\mu(t_{ij})+\Sigma_{k\geq 1}\xi_{ik}\phi_k(t_{ij})+\varepsilon_{ij}, t_{ij} \text{in} T.$$

Using only a single principal component, the model for the i^{th} trajectory takes the form

$$y_{ij}=\mu(t_{ij})+\xi_{i1}\phi_1(t_{ij})+\varepsilon_{ij}, t_{ij} \text{in} T \quad (2)$$

The first principal component function ϕ_1 describes the direction of most dispersion while a subject's first FPC score ξ_{i1} quantifies the extent to which their trajectory correlates with ϕ_1. Using more than one principal component function gives a more detailed fit. This is done, however, at the cost of a more complicated model, larger variation and less power as each additional principal component requires an additional term in the representation of each trajectory (2) and additional FPC scores must be estimated.

The estimation of FPC scores in (1) is complicated by the sparse and irregular sampling of each trajectory. This sparsity confounds the estimation of the population mean μ and FPC ϕ_1. The latter, being an eigenfunction of K, further depends on the estimation of $G(s,t)$. Both μ and G are assumed to be smooth and their estimates easily obtained by locally weighted least squares. In the case of μ, an estimate is obtained by smoothing the aggregate data (t_{ij}, y_{ij}) (or, without additional noise, (t_{ij}, x_i (t_{ij})). Given an estimate of the mean, $\hat{\mu}$, one may set $G_{ijk} = (y_{ij} - \hat{\mu}(t_{ij}))(y_{ik} - \hat{\mu}(t_{ik}))$ and get an estimate of the surface $G(s,t)$ from the aggregate data ((t_{ij},t_{ik}), G_{ijk}). Once μ and G are obtained, they are discretized on a finite grid; we continue to denote the discretized forms simply by μ and G. The first FPC is obtained as the eigenvector, ϕ_1, that corresponds to the largest eigenvalue, λ_1, of G.

Now, the first FPC score could be obtained as a discrete approximation to the integral in (1), but this clearly fails when the trajectories are extremely sparse and noisy. Instead, this integral is replaced by a classic multivariate formula for conditional expectation [22]. For this, let $\mu_i = (\mu(t_{i1}),.., \mu(t_{iN_i}))^T$, $\phi_{i1} = (\phi_1(t_{i1}),.., \phi_1(t_{iN_i}))^T$ and $y_i = (y_{i1},..., y_{iN_i})^T$ denote, respectively, the population mean, the first FPC, and the i^{th} trajectory, as sampled with respect to the i^{th} subject's observed timepoints $t_{i1},.., t_{iN_i}$. Also, let $\Sigma_{y_i} = \text{cov}(y_i., y_i)$ whose (j,k) entry is $G(t_{ij}, t_{ik}) + \sigma^2\delta_{jk}$ (where $\delta_{jk} = 1$ for j=k and 0 otherwise). Then define an estimate of the first FPC score as

$$\tilde{\xi}_{i1}=E(\xi_{i1}|y_i)=\lambda_1\phi_{i1}^T\Sigma_{y_i}^{-1}(y_i-\mu_i).$$

When ξ_{i1} and ε_{ij} are jointly Gaussian, and when each of μ_i, $\lambda_{,1}$, ϕ_{i1}, Σ_{y_i} are substituted by their estimates from the data ensemble, this is the best linear predictor of ξ_{i1} given information on the i^{th} subject.

In summary, sparse FPCA [14] provides a parsimonious and consistent method for estimating the FPC scores $\{\xi_i : i=1,\ldots,n\}$ in settings for which a set of n trajectories are sparsely observed at irregularly–sampled time points. In our analysis, μ, and G are estimated using local linear smoothing with Gaussian kernel on the pooled data using PACE [22]. For the viral load trajectories, a bandwidth of 350 days is used to estimate μ and G is estimated from measurements taken up to 2920 days post-infection, using a bandwidth of 300 days. For CD4 trajectories, a bandwidth of 350 days is used to estimate the mean, and the covariance surface estimated from measurements taken up to 2000 days post-infection, using a bandwidth of 300 days. Due to extreme sparsity in the entire population of CD4 measurements observed less than 120 days or more than 2100 days post date of HIV infection relative to sampling between 120 and 2100 days post infection, only CD4 counts observed between 4 and 70 months post infection are used.

For both viral load and CD4 measurements, principal component functions are estimated from the smoothed covariance surface. FPC scores for each woman are computed by conditioning on the observed data for the woman in question, and assuming normally distributed FPC scores and measurement errors. Individual trajectories (viral load and CD4 count) are estimated as the mean curve plus a weighted sum of the principal component, with weights corresponding to the individual's FPC score.

For the remainder of the paper, equation (2) is used to approximate individual viral load and CD4 trajectories. The term "first FPC" will be used to refer to the (estimated) first functional principal component, ϕ_1, and "FPC score" will refer to the (estimated) first FPC score or coefficient, $\xi_{i1}$for the i^{th} subject.

Linear mixed effects estimates—A linear mixed effects (LME) model with random slope and intercept was used to fit longitudinal trajectories, and then the obtained slope and intercept for each woman were subsequently used as predictors in a survival model. To fit the LME model, linearity of the trajectories must be assumed and so to approximate this, only measurements taken at least four months post-infection are used. Furthermore, the slopes and intercepts are only estimated for any woman having at least two measurements (viral load or CD4 count).

Survival analysis—A Cox proportional hazards survival model is used to examine prognostic power of FPC scores from both the CD4 and viral load trajectories. The results are contrasted with the Cox models using the summaries of early CD4 count and viral set point (defined as the first viral load measurement during 4–24 months of follow-up) for each trajectory. In addition, we evaluate the Cox models using the estimated intercepts and slopes from a LME as covariates.

Statistical analyses were performed with SAS version 9.2 and PACE was implemented using Matlab 7.9 and the PACE toolbox [http://anson.ucdavis.edu/~ntyang/PACE/]. Both viral load and CD4 data were $\log_{10}$ transformed prior to both the FPC and LME analyses.

Results

Among the 216 women for whom there is a reliable date of HIV-1 acquisition and at least one viral load or CD4 count measurement, the median duration of follow-up after HIV acquisition was 4.6 years and 40 deaths were observed. CD4 counts taken between 4 and 24 months post HIV infection ('early CD4 count') are available for 83 women and 168 women

have a viral load measurement observed during that time period ('viral setpoint'). The FPC analysis of CD4 count trajectories includes 132 women having at least one CD4 measurement; among these, a median of 4 (and a maximum of 16) CD4 measurements per subject were used. All 216 women with at least one viral load measurement are included in the FPC analysis of viral load trajectories; among these, a median of 6 (and a maximum of 27) measurements per subject are used.

Analysis of CD4 count data

For the CD4 count cohort, the estimated mean CD4 count trajectory and the first FPC by time since infection are shown in Figure 1. On average, CD4 counts are decreasing up to approximately 2000 days post infection. Also illustrated in Figure 1 are curves corresponding to the mean trajectory, μ, plus ξ_1 times the first FPC, for $\xi_1 = +1$ (upper dotted curve) and $\xi_1 = -1$ (lower). Figure 1 suggests that variability in CD4 trajectories increases up to approximately 1500 days post infection and levels off or decreases slightly to 2000 days post infection.

The first FPC captures 85.5% of variation of the CD4 count trajectories. With this large percentage, each woman's first FPC score, ξ_{i1}serves as an informative summary statistic for the entire CD4 trajectory. As the first FPC of the CD4 trajectories does not change sign, it serves as a surrogate measure of overall CD4 counts. That is, a positive FPC score corresponds to a higher than typical CD4 trajectory for that individual, whereas a negative score corresponds to a lower than typical CD4 trajectory. The FPC score also provides a useful interpretation about the rate of change of CD4 counts. Since the mean CD4 trajectory is monotone decreasing, and the first FPC is increasing up to about 1500 days post infection, a positive FPC score indicates that the CD4 trajectory is decreasing more slowly than average (or possibly increasing) while a negative score indicates the CD4 trajectory is decreasing more rapidly than average, up to 1500 days post infection.

The summary statistic, $\xi_1$is illustrated further in Figure 2: observed CD4 counts (points connected by dotted lines) and the corresponding FPC-estimated CD4 trajectories (solid lines) are shown for five individual women; a separate color is used for each subject and the population mean CD4 trajectory is shown in black (solid line). Note that larger positive FPC scores (red and teal) correspond to women with higher CD4 counts and more slowly decreasing (or possibly increasing) trajectories, while negative FPC scores (green, blue, and magenta) correspond to women with lower than average CD4 counts and more rapidly decreasing trajectories. This suggests that the FPC score summarizes information about the CD4 level *and* its change over time.

To evaluate the relationship between CD4 FPC scores and survival, the FPC scores are rescaled so that a one unit change in the score corresponds to approximately a 100 unit change in early CD4 counts. Table 1 provides a summary of Cox proportional hazards modeling of time-to-death using various summaries of individuals' CD4 trajectories as covariates. When the FPC score is used as a single covariate, an increase in FPC score corresponds to a significant decrease in mortality; a 0.82 hazard ratio for one unit increase in FPC score, with p-value < 0.001; Table 1, row 1. Recall that a one unit increase in FPC score captures information on both magnitude of CD4 counts and rate of change of the CD4 count trajectory over the entire period of observation so that the hazard ratio captures risk associated with both level and change of individual CD4 count profiles. Note that 132 subjects have a CD4 FPC score, whereas only 83 women having an observed early CD4 count. When the early CD4 count is used as the covariate in this survival model, a hazard ratio of 0.71 for 100 unit increase in early CD4 count is estimated but is not significantly associated with survival (p-value = 0.089, Table 1, row 2).

In [3] the authors showed that CD4 count slope explained very little of the variation in time to AIDS or time to death in a cohort of untreated men. To evaluate the relationship between CD4 trajectory slope and survival in this cohort we fit an LME model to $\log_{10}$ CD4 counts measured more than four months post infection in women having at least two measurements and used the resulting estimated individuals' slopes and intercepts as covariates in a single survival model. Among the 120 subjects with LME-estimated slope and intercept for $\log_{10}$ CD4 count, the Cox model shows that the slope is significantly associated with improved survival (a 0.14 hazard ratio for each unit increase in slope, with $p<0.001$; Table 1, row 3). However, the LME-obtained intercept is not significantly associated with survival outcomes ($p = 0.23$; Table 1, row 4).

Since the number of subjects differs in each of the three analyses shown in rows 1–4 of Table 1 (recall, for example, the early CD4 summary requires a measurement taken between 4 and 24 months), we repeat the survival analysis using FPC scores restricted to a common set of subjects. Row 5 of Table 1 shows the result of fitting the survival model with the FPC score covariate for the 83 women with an early CD4 count. The results are similar to those obtained using the larger data set (row 1). In particular, even in this much smaller set of subjects (more than 40% smaller) the first FPC score is significantly associated with improved survival (a 0.83 hazard ratio, with $p=0.003$). Similarly, rows 6–8 of Table 1 show the results from restricting the Cox models to the 113 subjects for whom both FPC scores and LME-estimated slopes and intercepts are available. The results are similar to those obtained with these same covariates in the larger populations.

Figure 3 gives a revealing look at how the CD4 trajectories are associated with survival. Figure 3A displays the trajectories for all women in the cohort, color coded by quartile of their estimated CD4 FPC score: the blue curves represent subjects having an FPC score in the lower quartile, where FPC scores are less than $Q1 = -3.58$; the cyan curves represent subjects with an FPC score larger than Q1 but smaller than the median, $Q2 = 0.80$; the black curves represent subjects with an FPC score are larger than Q2 but less than $Q3 = 5.19$; and the green curves are subjects with FPC scores larger than Q3. The Kaplan-Meier curves of these four groups are displayed in Figure 3B using the same color scheme.

The mortality rate increases from the upper to lower quartile (from green to blue), as expected, and suggests that the first CD4 FPC score is an appropriate linear covariate in a Cox proportional hazards model. In this analysis, subjects in the two upper FPC score quartiles have significantly improved survival compared to the lowest quartile. Specifically, the hazard ratio for women with scores in the highest (green) quartile is 0.20 ($p = 0.034$) while the hazard ratio for women with scores in the second highest (black) quartile is 0.16 ($p=0.019$). When the two upper quartiles are combined and compared to the two lowest quartiles, the estimated hazard ratio is 0.24 with a p-value of 0.011.

Analysis of viral load data

Figure 4 shows the estimated mean curve (black) of the viral load measurements by time since infection. There is an initial decrease in the mean curve reaching a minimum at approximately 660 days post infection.

The first FPC (shown in gray) captures 73% of the variation in the set of all viral load trajectories. As noted for the CD4 count data, due to the shape of the first FPC curve, the FPC score summarizes both viral load level and change in viral load over time. Like the first FPC for CD4 counts, the viral load first FPC does not change sign so the FPC score may be viewed as a surrogate measure of overall viral load. A positive FPC score corresponds to a viral load trajectory that is consistently greater than the population mean while a negative FPC score corresponds to a viral load trajectory that is consistently less than the population

mean. Since the first FPC is not monotone increasing (there appears to be slight dip to around 2000 days post infection), the information that is captured about rate of change of viral load trajectories is somewhat more complicated than for CD4 counts. However, up to approximately 2000 days post infection, monotonicity of the first FPC indicates that a positive FPC score corresponds to a viral load trajectory that is increasing more rapidly than average while a negative FPC score corresponds to a viral load trajectory that is decreasing more rapidly than average. This is demonstrated in Figure 5 with observed viral load measurements (points connected by dotted lines) and the corresponding FPC-estimated trajectories (solid lines) for five women. A separate color is used for each subject and the estimated population mean trajectory is shown in black (solid line).

In a previous analysis of these data, Lavreys et al. [7] used the viral set point (available for 168/216 subjects) and showed that it is strongly associated with mortality with a hazard ratio of 2.21 (p=0.0014) for each unit increase in $\log_{10}$ viral set point (see also, Table 2, row 2). An FPC score is available for all 216 women in this cohort and Table 2, row 1, shows that this summary is also strongly associated with mortality with a hazard ratio of 1.57 ($p < 0.001$) for each unit increase in viral load FPC score. For this analysis, viral load FPC scores are rescaled so that a one unit change corresponds to approximately a $\log_{10}$ change in viral set point. To compare the predictive nature of viral set point with that of the viral load FPC score, the survival analysis is repeated using only FPC scores from the 168 individuals for whom a set point is defined. The p-values and hazard ratios from this restricted analysis are similar to the analysis using the data from all 216 women (Table 2, row 5).

Another common measure of viral load trajectories is the viral load slope [9]. We fit the LME model for $\log_{10}$ viral loads measured more than four months post infection in women having at least two measurements and used the resulting estimated individuals' slopes and intercepts as covariates in a single survival model. Both the estimated slopes and intercepts are strongly associated with mortality; a hazard ratio of 1.80 (p=0.007) was estimated for each unit change in estimated slope for $\log_{10}$ viral load and a hazard ratio of 4.51 ($p < 0.001$) for each unit change in the estimated intercept for $\log_{10}$ viral load (Table 2, rows 3–4). For comparison purposes, the survival analysis using the FPC score was repeated; restricted to those 159 individuals having LME-estimated slopes and intercepts (Table 2, row 6). The results are nearly identical to the analysis conducted on the entire cohort.

Viral loads and CD4 counts are generally thought to be strongly correlated. In [3] the authors concluded that a single viral load measurement was the strongest predictor among a variety of summary measures of viral load and CD4 count trajectories of time to AIDS or death. In this cohort, longitudinal viral loads and CD4 counts are highly correlated. We conducted a linear repeated measures regression of CD4 count regressed on $\log_{10}$ viral load. The estimated coefficient of the viral load covariate was significant at level $p < 0.001$. The estimated coefficient of the viral load covariate indicates that on average, for each $\log_{10}$ increase in viral load, CD4 counts drop by 50 units. To evaluate the combined role of CD4 count and viral load trajectory we considered a survival analysis using *both* the CD4 count and the viral load FPC scores in the model. The estimated hazard ratio for the CD4 count FPC score is 0.86 ($p = 0.010$) and the estimated hazard ratio for the viral load FPC score is 1.36 (p=0.055). Both scores are borderline significant, although this analysis suggests the CD4 count is the primary indicator of survival in this cohort. This analysis can be thought of as an evaluation of the relationship between CD4 count trajectory and survival *adjusted for* viral load trajectory or, vice versa, as an evaluation of the relationship between viral load trajectory and survival *adjusted for* CD4 count trajectory.

Discussion

The first FPC score is shown here to be an effective and efficient summary of sparsely-sampled longitudinal measurements of CD4 counts and viral load. A clear benefit is that it exploits information from the population and hence can be stably calculated for individuals in a cohort having any number of measurements taken at any time during the study period. Using FPC analysis, we are able to demonstrate a significant relationship between CD4 profiles and survival outcomes when previous analyses of this data failed to find a significant association between a summary measure of CD4 counts and survival [7].

This approach is successful for a variety of reasons. Longitudinal measurements of CD4 counts and viral load are trajectories with time-dependent structure. However the classical summaries such as the estimated viral set point or early CD4 count (and others such as CD4 nadir) are single-value quantities which ignore a majority of the data from each subject. FPCA, in contrast, is well-suited for the purpose of summarizing an entire trajectory and can do so with a single quantity: the first FPC score. As seen in the analysis of this cohort, the FPC score summarizes both the overall measurement level and the measurement changes over time. Furthermore, any principle components analysis is designed to capture the most variation among subjects so that any variation in viral load or CD4 count trajectories associated with survival will be captured by the FPC scores. For this reason, the FPC scores can be used in a preliminary analysis to determine if there is *any* feature of the longitudinal trajectories associated with survival. Once established, subsequent analysis and evaluation of the characteristics of the functional principle components is needed to identify which clinical features are associated with survival.

The classical measures, viral set point or early CD4 count, are only available for a fraction of the cohort–in this case, individuals for whom measurements were taken during early HIV infection–and so the analysis must ignore a substantial portion of the cohort. Indeed, in the viral load cohort, only 168 women had data available for estimation of viral set point whereas FPC scores were estimated for all 216 women. In the CD4 cohort, only 83 out of 132 women had data available for estimation of early CD4 count. Thus, the FPC analysis has the potential to increase sample size and improve power.

An attractive feature of this analysis is that no priori assumptions are made on the trajectory shapes. Viral load trajectories are sometimes estimated using linear or piece-wise linear models [2, 9, 10] with different rates of change during and after the acute infection phase. While the acute infection stage is typically reported as the first 8–10 weeks of infection [23], Lyles et al. [10] used piece-wise linear models and found different rates of change for viral load in an Italian cohort before and after 18 months post-infection. Using a linear or piece-wise linear model, the duration of the acute infection phase must somehow be estimated–via statistical inference, from biological knowledge of the population, or by some other means. FPC obviates the need for this. Different rates of change at different times post-infection arise naturally through the estimated mean curve and principal components. Different individuals may have different trajectory shapes, depending upon their FPC scores. In the dataset analyzed in this work, the estimated mean curve and first principal component shape suggest that some individuals may have continued rates of viral load decay lasting as long as approximately 18 months post infection, consistent with the work of Lyles et al. [10] (see Figure 4).

There are a number of weaknesses in the use of the two-stage FPC/Cox proportional hazards approach for evaluating the relationship between longitudinal profiles and survival. The FPC score is not a measure that can be calculated for individuals in the absence of data from an entire cohort; in that regard it is similar, e.g., to LME estimates of slopes and intercepts. In

addition, without careful evaluation of the shape of the first FPC and its relationship to the population mean, the first (or higher order) FPC scores lack clinical interpretation and carry no explicit information about how CD4 or viral load trajectories are associated with survival. The hazard ratios estimated in our analysis do not have obvious clinical interpretations; our results indicate primarily that there is an association between both viral load and CD4 count trajectories and survival. In this study we are able to relate the FPC scores to clinically meaningful features of CD4 and viral load trajectories and provide some guidance on the interpretation of the hazard ratios, however, this is not guaranteed in general. For example, if the FPC changes sign or monotonicity over time, separate interpretations might be needed over different time intervals, possibly obscuring any relevant clinical interpretation. Thus FPC scores are not likely to be useful measures for clinical evaluation. Nonetheless, the approach is well suited to research analyses and population studies and can be used to identify if *any* features of a longitudinal trajectory are associated with an outcome of interest. When this is the case, additional analysis and careful study of the form of the functional principle components is needed to identify which clinically meaningful features of the trajectories are associated with the outcome of interest.

A further weakness, which is shared by many two-stage approaches when evaluating the relationship between longitudinal profiles and survival is that the summary measures used in the two-stage analysis (FPC score, estimated viral set point, early CD4 count, or slopes and intercepts from random effects models) are often measured with error. As such, the point estimate of the hazard ratio is known to be biased [24] towards the null with amount of bias depending on the variation in the covariate. However, as described above, the magnitude of the hazard ratio associated with the FPC score covariates lacks clinical interpretation; the strength of the approach is that it identifies the trajectories as being strongly associated with survival, leading to additional analysis to determine specific features of these trajectories which impact survival.

There are many possible applications and areas for refinement. The analysis presented here is a two-stage approach in the sense that FPC scores are obtained first and subsequently used in a survival analysis. Joint modeling of survival data together with FPC is a promising alternative approach [15, 25]. Because of the computational simplicity, our two-stage approach can serve as a good initial estimate in the joint modeling approach, which involves an Expectation Maximization (EM) algorithm to estimate parameters and may require selection of the number of FPCs within the iterations. Due to the enrollment of nonparametric longitudinal components, this estimating procedure will be more computational intensive than that in the joint models with popular parametric longitudinal submodels. Nevertheless, the proposed two-stage estimate can be used as an exploratory tool in the preliminary analysis and a starting point in the further joint models.

For simplicity and direct comparison with other commonly used summaries we have focused on only the first score. Examining finer trajectory details by including additional principal components is of interest. The number of selected principal components and smoothing parameters could be chosen objectively by automatic procedures. In this data set, adding the second principle component scores as covariates in the regression analysis did not qualitatively change the conclusions: neither the second FPC scores for CD4 count or viral load were significant predictors of survival and did not change the conclusions based on analysis with only the first FPC score as a covariate. Finally, FPC scores are used here to evaluate survival outcomes but they have application in a variety of other statistical analyses. Examples include the comparison of trajectories between two or more groups or as an adjustment covariate in regression analysis which require adjustment for viral loads or CD4 counts.

In summary, the functional principal components approach for the evaluation of longitudinal data is an attractive alternative to using single summary measures of longitudinal profiles in a research setting. It requires almost no a priori assumptions and can increase power of analysis by including all members of a cohort with observed data. Furthermore, if associations between the FPC score(s) and survival are detected, it will often be the case that additional evaluation of the profile using the functional principal components can reveal specific characteristics that are associated with the outcome of interest. As more biomarkers are identified and used to evaluate HIV survival and progression [26, 27] (such as measures of inflammation or immune activation), summarizing longitudinal trajectories of these biomarkers will become increasingly important. The FPC method presented here provides one such method that is rigorous and effective for revealing structure in these data that is associated with clinical outcomes.

Acknowledgments

The authors thank the reviewers for helpful comments which helped us to strengthen and clarify the presentation.

Financial Support: This work was supported by NIH grants AI055343, AI38518, CA126205 and CA053996.

Bibliography

1. Anastos K, Kalish LA, Hessol N, Weiser B, Melnick S, Burns D, Delapenha R, DeHovitz J, Cohen M, Meyer W, Bremer J, Kovacs A. The relative value of CD4 cell count and quantitative HIV-1 RNA in predicting survival in HIV-1-infected women: results of the women's interagency HIV study. AIDS. 1999; 13:1717–1726. [PubMed: 10509574]
2. Lyles RH, Munoz A, Yamashita TE, Bazmi H, Detels R, Rinaldo CR, Margolick JB, Phair JP, Mellors JW. Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. Multicenter AIDS Cohort Study. J Infect Dis. 2000; 181:872–880. [PubMed: 10720507]
3. Mellors JW, Margolick JB, Phair JP, Rinaldo CR, Detels R, Jacobson LP, Munoz A. Prognostic value of HIV-1 RNA, CD4 cell count, and CD4 Cell count slope for progression to AIDS and death in untreated HIV-1 infection. JAMA. 2007; 297:2349–2350. [PubMed: 17551128]
4. Mellors JW, Munoz A, Giorgi JV, Margolick JB, Tassoni CJ, Gupta P, Kingsley LA, Todd JA, Saah AJ, Detels R, Phair JP, Rinaldo CR Jr. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. Ann Intern Med. 1997; 126:946–954. [PubMed: 9182471]
5. Sterling TR, Vlahov D, Astemborski J, Hoover DR, Margolick JB, Quinn TC. Initial plasma HIV-1 RNA levels and progression to AIDS in women and men. N Engl J Med. 2001; 344:720–725. [PubMed: 11236775]
6. Hansmann A, Schim van der Loeff MF, Kaye S, Awasana AA, Sarge-Njie R, O'Donovan D, Ariyoshi K, Alabi A, Milligan P, Whittle HC. Baseline plasma viral load and CD4 cell percentage predict survival in HIV-1- and HIV-2-infected women in a community-based cohort in The Gambia. J Acquir Immune Defic Syndr. 2005; 38:335–341. [PubMed: 15735454]
7. Lavreys L, Baeten JM, Chohan V, McClelland RS, Hassan WM, Richardson BA, Mandaliya K, Ndinya-Achola JO, Overbaugh J. Higher set point plasma viral load and more-severe acute HIV type 1 (HIV-1) illness predict mortality among high-risk HIV-1-infected African women. Clin Infect Dis. 2006; 42:1333–1339. [PubMed: 16586394]
8. Boscardin WJ, Taylor JM, Law N. Longitudinal models for AIDS marker data. Stat Methods Med Res. 1998; 7:13–27. [PubMed: 9533259]
9. Lavreys L, Baeten JM, Kreiss JK, Richardson BA, Chohan BH, Hassan W, Panteleeff DD, Mandaliya K, Ndinya-Achola JO, Overbaugh J. Injectable contraceptive use and genital ulcer disease during the early phase of HIV-1 infection increase plasma virus load in women. J Infect Dis. 2004; 189:303–311. [PubMed: 14722896]
10. Lyles CM, Dorrucci M, Vlahov D, Pezzotti P, Angarano G, Sinicco A, Alberici F, Alcorn TM, Vella S, Rezza G. Longitudinal human immunodeficiency virus type 1 load in the italian

seroconversion study: correlates and temporal trends of virus load. J Infect Dis. 1999; 180:1018–1024. [PubMed: 10479126]

11. Perelson AS, Essunger P, Cao Y, Vesanen M, Hurley A, Saksela K, Markowitz M, Ho DD. Decay characteristics of HIV-1-infected compartments during combination therapy. Nature. 1997; 387:188–191. [PubMed: 9144290]
12. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science. 1996; 271:1582–1586. [PubMed: 8599114]
13. James GM, Hastie TJ, Sugar CA. Principal component models for sparse functional data. Biometrika. 2000; 87:587–602.
14. Yao F, Muller HG, Wang JL. Functional data analysis for sparse longitudinal data. Journal of the Americal Statistical Association. 2005; 100:577–590.
15. Yao F. Functional principal component analysis for longitudinal and survival data. Statistica Sinica. 2007; 17:965–983.
16. Ramsay, JO.; Silverman, BW. Functional data analysis. 2nd edn. Springer; New York: 2005.
17. Diggle, P.; Heagerty, P.; Liang, K-Y.; Zeger, SL. Analysis of longitudinal data. Oxford New York: Oxford University Press; 2002.
18. Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR Jr. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. Am J Epidemiol. 1987; 126:310–318. [PubMed: 3300281]
19. Abrams DI, Goldman AI, Launer C, Korvick JA, Neaton JD, Crane LR, Grodesky M, Wakefield S, Muth K, Kornegay S, et al. A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. The Terry Beirn Community Programs for Clinical Research on AIDS. N Engl J Med. 1994; 330:657–662. [PubMed: 7906384]
20. James G. Generalized Linear Models with Functional Predictor Variables. Journal of thte Royal Statistical Society Series B. 2002; 64:411–432.
21. James G, Hastie T, Sugar C. Principal Component Models for Sparse Functional Data. Biometrika. 2000; 87:587–602.
22. Mardia, KV.; Kent, JT.; Bibby, JM. Multivariate analysis. Academic press; London: 1979.
23. White, DFF. Medical Virology. Academic Press; New York: 1994.
24. Raboud J, Reid N, Coattes RA, Farewell VT. Estimating risks of progressing to AIDS when covariates are measured with error. Journal Royal Statistical Society Series A. 1993; 156:396–406.
25. Ding J, Wang JL. Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. Biometrics. 2008; 64:546–556. [PubMed: 17888040]
26. Lundgren JD, Baxter J, Deeks SG, Lane HC. Biomarkers in HIV disease. Curr Opin HIV AIDS. 5:459–462. [PubMed: 20978387]
27. Neaton JD, Neuhaus J, Emery S. Soluble biomarkers and morbidity and mortality among people infected with HIV: summary of published reports from 1997 to 2010. Curr Opin HIV AIDS. 5:480–490. [PubMed: 20978391]

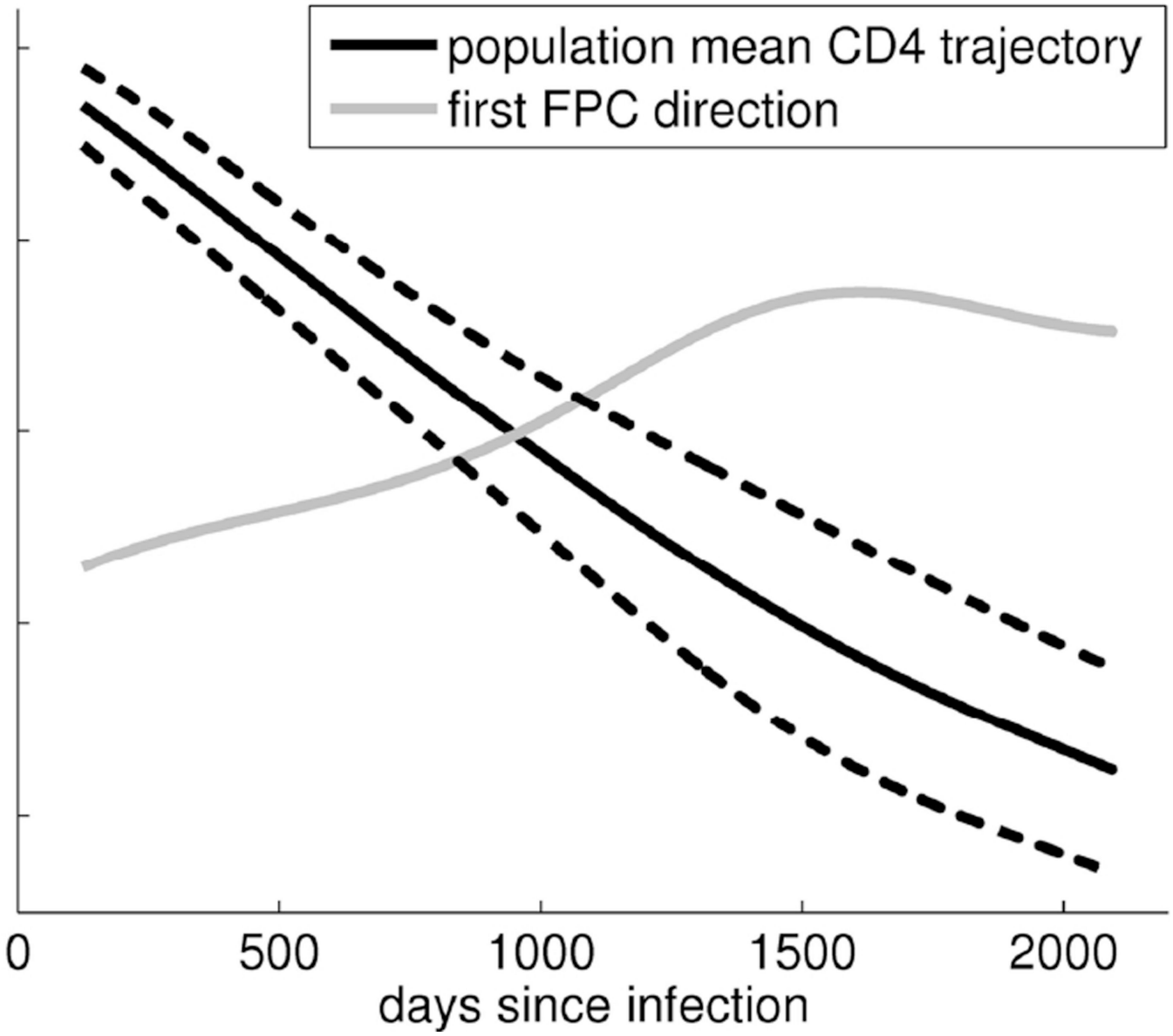

Figure 1.
Mean CD4 trajectory, μ, (black), first CD4 principal component direction, Φ_1, (gray; scaled and shifted for display), and $\mu \pm 1*\Phi_1$ (dotted).

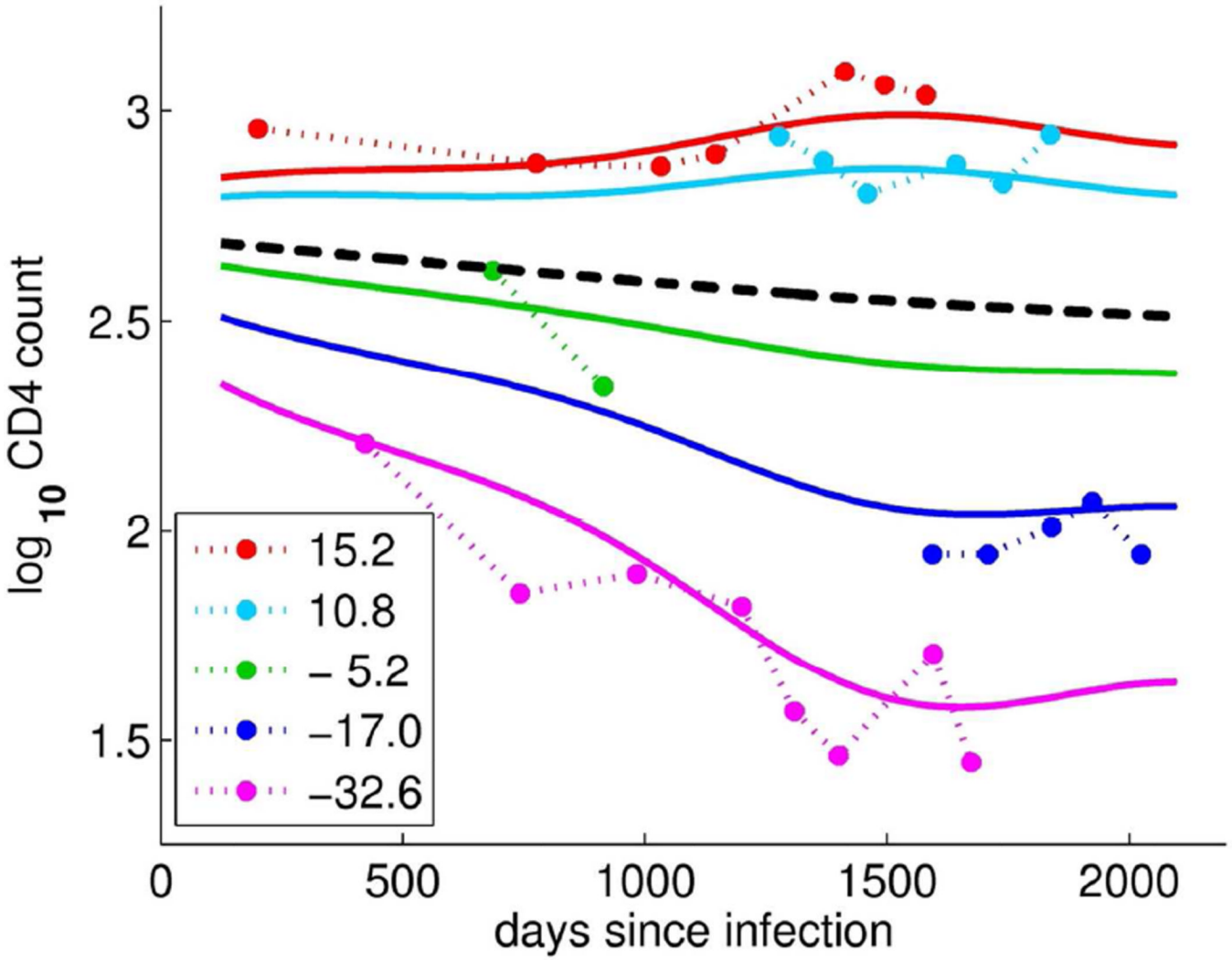

Figure 2.
Observed $\log_{10}$ CD4 count trajectories (dotted, *) for five representative women and the corresponding color-matched first FPCs scaled by their respective FPC scores (solid lines); the FPC scores are given in the legend. The overall mean $\log_{10}$ CD4 count trajectory is shown in black (dashed).

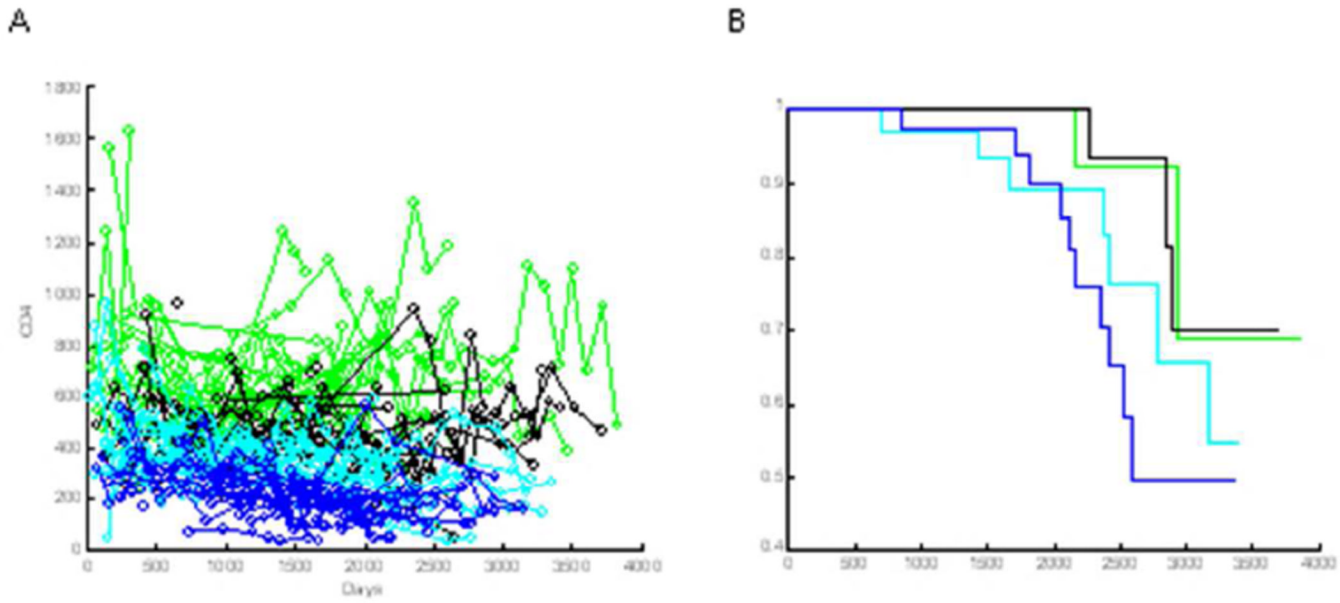

Figure 3.

A) CD4 counts by quartile of FPC score: blue=lowest, cyan=second lowest, black=second highest, green=highest quartile of FPC score. Circles connected by lines indicate sampled CD4 counts from one individual. B). Kaplan-Meier survival curves for groups defined by each of the four quartiles of CD4 FPC scores.

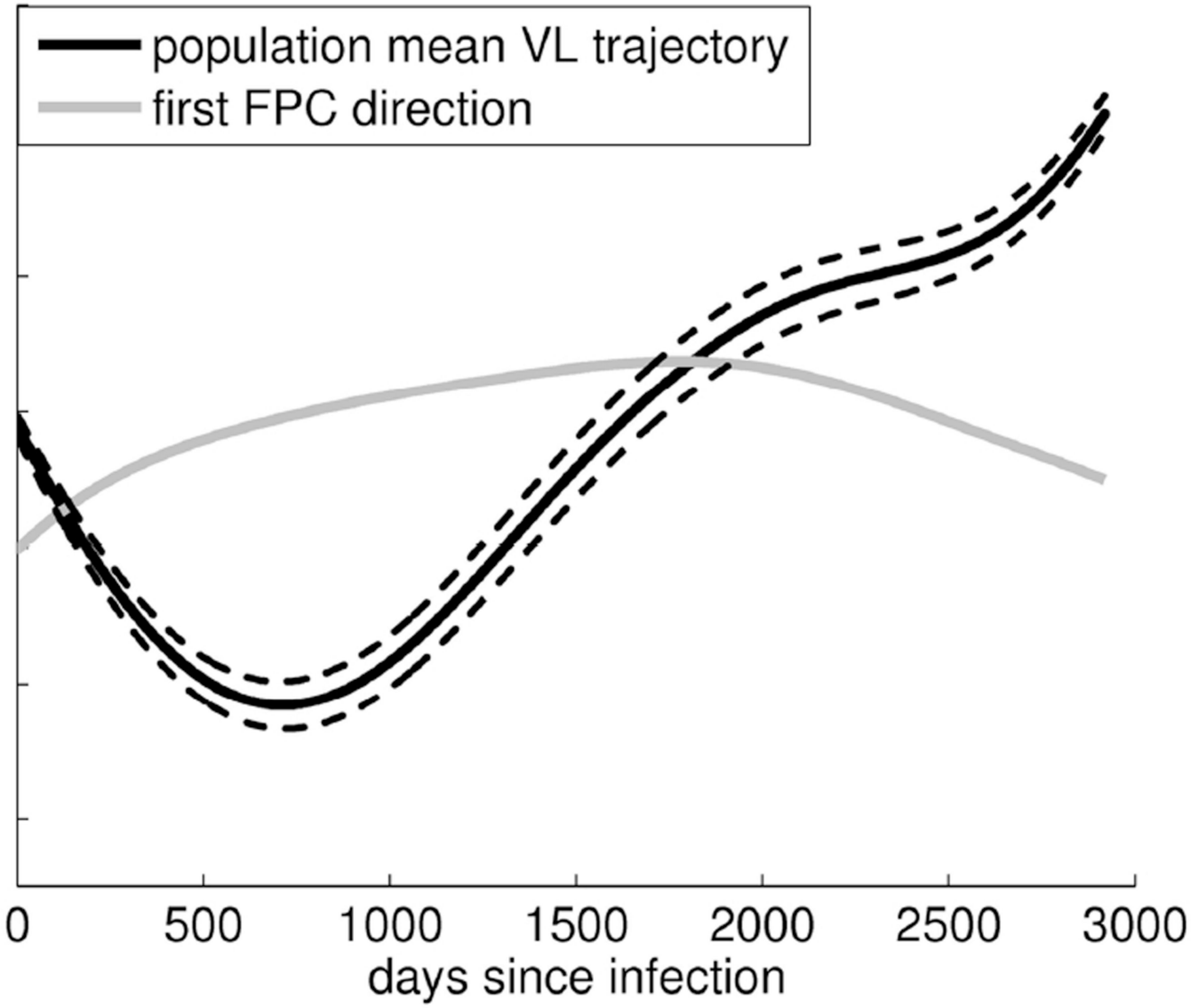

Figure 4.
Mean viral load trajectory, μ, (black), first CD4 principal component direction, Φ1 (gray; scaled and shifted for display), and $\mu \pm 1*\Phi_1$(dotted).

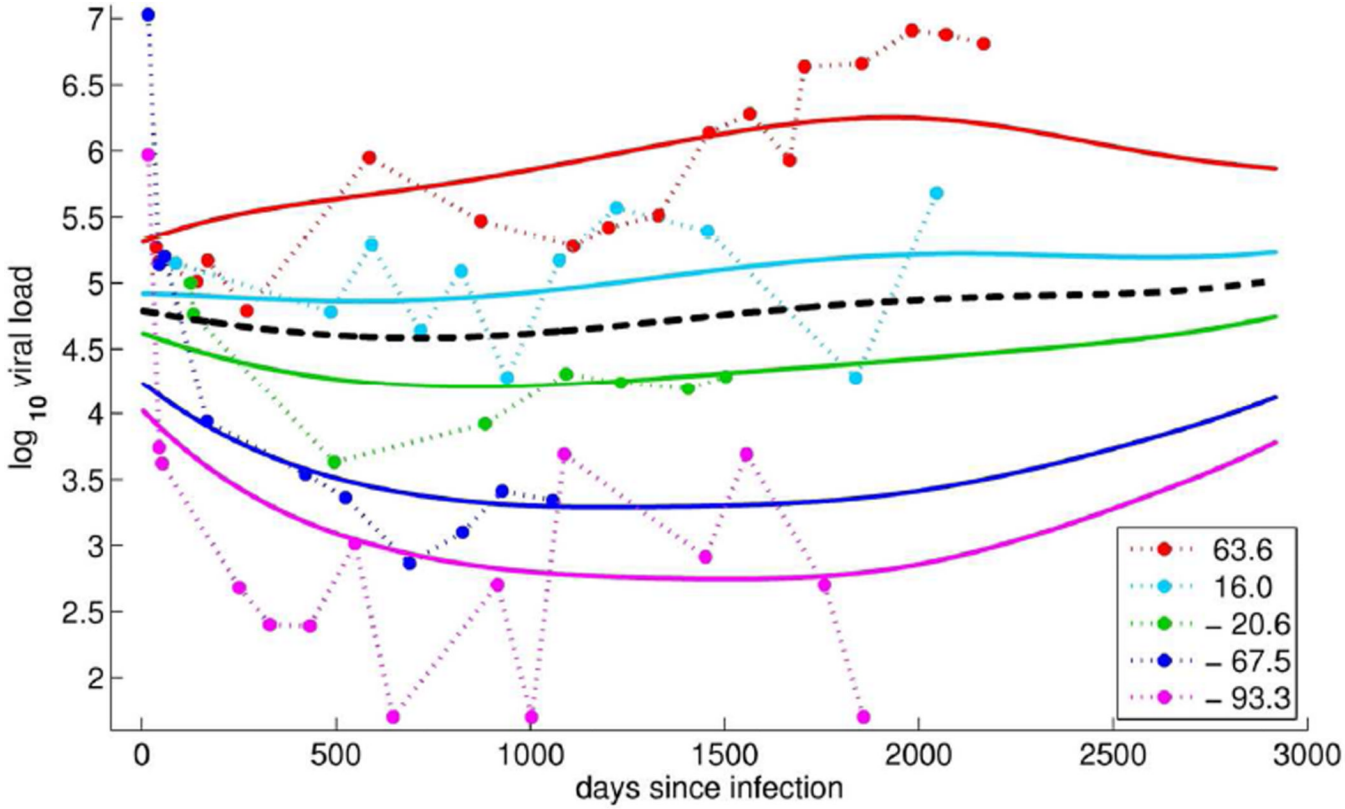

Figure 5.
Observed $\log_{10}$ viral load trajectories (dotted, *) for five representative women and the corresponding color-matched first FPCs scaled by their respective FPC scores (solid lines); the FPC scores are given in the legend. The overall mean $\log_{10}$ viral load trajectory is shown in black (dashed).

Table 1

Survival Analysis and measures of CD4 count data.

Predictor(s)	Df	N	Hazard Ratio for a 1 unit increase in predictor (95% CI)	P
1st FPC score	1	132	0.82 (0.74, 0.91)	<0.001
Early CD4 count	1	83	0.71 (0.49, 1.05)	0.089
LME slope +	2	120	0.14 (0.06, 0.34)	<0.001
intercept			0.16 (0.01, 3.16)	0.23
Subjects with both an FPC score and an early CD4 count				
1st FPC score	1	83	0.83 (0.73, 0.94)	0.003
Subjects with both an FPC score and LME-estimated slope and intercept				
1st FPC score	1	113	0.82 (0.73, 0.91)	<0.001
LME slope +	2	113	0.15 (0.06, 0.36)	<0.001
intercept			0.17 (0.01, 3.31)	0.24

Table 2

Survival Analysis and measures of viral load trajectories

Predictor(s)	Df	N	Hazard Ratio for a 1 unit increase in predictor (95% CI)	P
1st FPC score	1	216	1.57 (1.25, 1.98)	<0.001
Viral set point	1	168	2.21 (1.36, 3.60)	0.0014
LME slope +	2	159	1.80 (1.18, 2.76)	0.007
intercept			4.51 (2.17, 9.38)	<0.001
Subjects with both an FPC score and a viral set point				
1st FPC score	1	168	1.67 (1.29, 2.14)	<0.001
Subjects with both an FPC score and LME-estimated slope and intercept				
1st FPC score	1	159	1.64 (1.26, 2.13)	<0.001