Powerful HIV Disease Prognosis despite Sparse and Irregular Sampling

May 21, 2012

ME Arnegard

Individual measurements taken at single time points are commonly used as predictors of the outcome of HIV infection, for example to help doctors decide when to recommend that patients begin antiretroviral therapy. The most common of these univariate predictors are the viral load set point and early CD4 count. The set point measures the relatively stable HIV load in plasma that occurs soon after an early burst of viral replication, as infected individuals develop HIV antibodies and begin to fight their infection. Lower set points tend to be associated with longer periods of clinical latency. CD4 counts are also useful in this regard, because HIV infects helper T cells by first recognizing the CD4 receptor expressed on the cell’s surface, eventually killing this population of immune cells. Measurement of the CD4 level during early infection is thought to predict roughly how long it will take from infection to immune deficiency (i.e., AIDS) in untreated individuals. Despite their widespread application, these univariate measures are prone to error. Using all the repeated measures of viral loads or CD4 counts can greatly enhance the predictive value of these variables. This approach, however, requires that patients be monitored at regular intervals for relatively long periods of time.

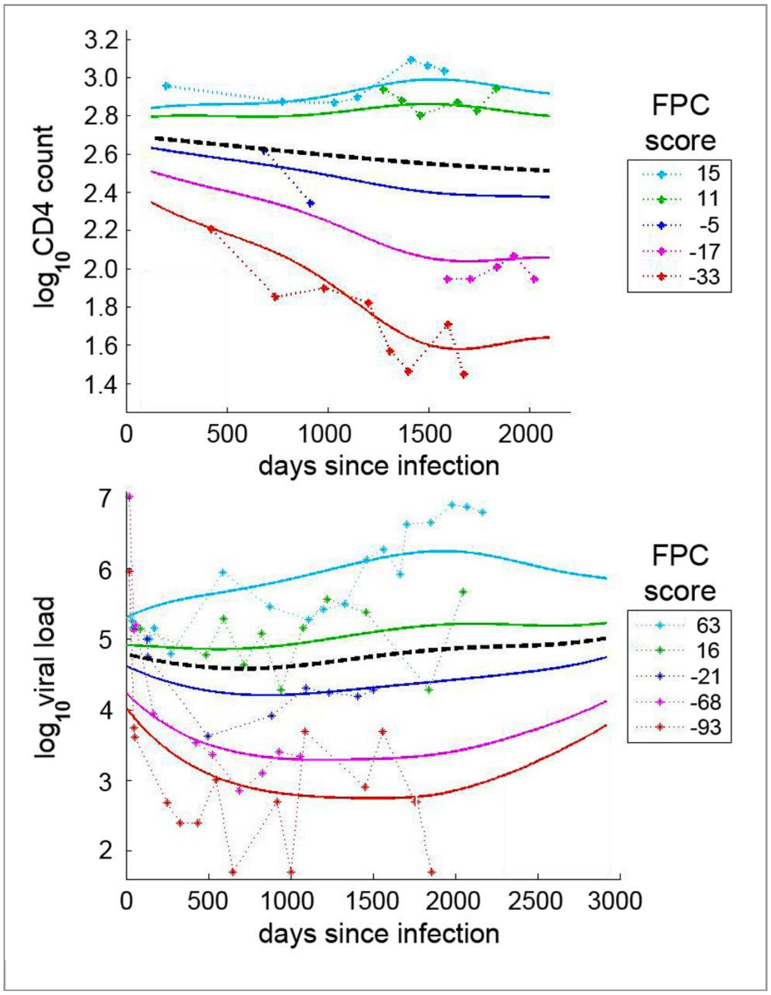
A program of regular clinical measurements can be extremely difficult to achieve where health care resources are scarce and patients lack the needed income to pay for medical expenses. To address this problem, a team of scientists led by Dr. Sarah Holte (Biostatistics and Biomathematics Program) and Dr. Julie Overbaugh (Human Biology Division) applied and refined a sophisticated new technique to predict HIV outcomes in a cohort of 216 female sex workers in Mombasa, Kenya. Dr. Timothy Randolph (Biostatistics and Biomathematics Program) and several outside collaborators also contributed to this work. The team focused on women who had become infected with HIV-1 between 1993 and 2004.

In their research, Holte *et al.* took an approach that builds on a multivariate statistical method known as ‘principal component analysis,’ or PCA. This data reduction and ordination technique converts a set of observations of possibly correlated variables into a smaller number of linearly-uncorrelated variables (or principal components) describing hidden axes of maximal variation in the original data. Scores along the ‘first principal component’ describe the axis of greatest variation in a multivariate dataset.

The authors turned to an extension of classical PCA called ‘functional principal component analysis’ (or FPC; see Yao *et al.*, 2005). Holte and co-authors applied FPC to the sparse data they had on viral loads or CD4 counts for the Kenyan cohort. It had only been possible to collect measurements from some women on as few as two or three occasions. In addition, the data series that were collected did not overlap in time between many of the women (see the distributions of points in the figure), which precluded the use of traditional repeated measures models to assist in their prognosis. The methods used by Holte *et al.* overcame these limitations and exploited the ‘shape’ of each individual’s CD4-count profile. The authors showed that these ‘shapes’ (as summarized by the first FPC score) provided a robust, single-value summary measure of the overall CD4 count profile. A similar finding was achieved for viral loads, when the scientists conducted a separate analysis using the HIV load data collected from the same cohort. Unlike the viral set point or early CD4 count, however, the first FPC scores also summarized the dynamic behaviors of these variables through time (see figure). In addition to providing greatly enhanced prognostic power in the face of sparse and irregular sampling, the FPC approach helps to reveal how the interplay between overall level and pattern of change, in either HIV load or CD4 count, has important clinical ramifications for an individual’s response to HIV infection.

Holte SE, Randolph TW, Ding J, Tien J, McClelland RS, Baeten JM, Overbaugh J. 2012. Efficient use of longitudinal CD4 counts and viral load measures in survival analysis. *Statistics in Medicine,* Epub ahead of print, doi: 10.1002/sim.5318.

Also see: Yao F, Müller H-G, Wang, J-L. 2005. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100:577-590.



Images courtesy of Dr. Timothy Randolph

The first FPC score simultaneously predicts the overall level and the pattern of change for an individual's CD4 count (above) or HIV load (below). Examples of all the data available for some women are given by the individual points, which are connected by dotted lines and color-coded to represent different individuals. The overall mean trajectory in each plot is shown by the dashed black curve, whereas the solid, colored curves represent the first FPCs scaled by their respective FPC scores (color-coded as before). The first FPC scores themselves are given in the respective legends. Note, in the top plot, that larger positive FPC scores correspond to women with higher CD4 counts and more slowly-decreasing (or even increasing) trajectories, whereas negative FPC scores correspond to women with lower average CD4 counts and more rapidly decreasing trajectories. Similarly, patterns of change in viral load vary with FPC score, as shown in the bottom plot.