Supplemental Digital Content

A more detailed description of the statistical model for CD4 count trajectories and the simulation study is provided here. Specifically, for subject i with a CD4 count measurement at time t years after their first seropositive HIV test result, $CD4_i(t)$, the model is specified as:

$$\sqrt{CD4_i(t)} = \eta_{it} + \varepsilon_{it}.$$

The model describing how the latent square root CD4 count, η_{it}, changes with time is:

$$\eta_{it} = \beta_0 + \beta_1 t + b_{i0} + b_{i1} t + W_i(t).$$

In this model, β_0 is the estimated population mean square root CD4 count at time = 0 (the time of the first seropositive HIV test result) and β_1 describes the population mean rate of change over time in square root CD4 count. The estimates (standard errors) of β_0 and β_1 were 27.03 (0.29) and -2.08 (0.13), respectively. b_{i0} is patient i's deviation from the population mean square root CD4 count at time t=0, and b_{i1} is patient i's deviation from the population mean rate of change in square root CD4 count. The random effects, b_{i0} and b_{i1}, are assumed to arise from a mean zero bivariate normal distribution providing for variability in the intercept and rate of change in CD4 count among patients in the study population. The estimated variances of the b_{i0} and b_{i1} were 17.30 and 2.23, and the estimated covariance was -0.35. The term $W_i(t)$ is a stochastic process which allows for both periods of slower and periods of faster change in latent CD4 count about the subject's general linear trajectory; these changes are assumed to occur over periods of weeks or months as distinct from the very short-term variability described by ε_{it}. We found, as in other studies [12; 14], that the best-fitting stochastic process was a scaled Brownian motion process with estimated variance 1.10. The estimated variance of ε_{it} was 10.09.

The goodness of fit of the model used in the simulation study is illustrated in Figure 1 by showing that the cumulative proportions of patients starting ART for different CD4 thresholds are similar to the Kaplan-Meier estimates of these proportions obtained directly from the MACS data (i.e. without assuming a model). In producing the Kaplan-Meier estimates, if a patient had a missing CD4 measurement, follow-up was censored prior to their first missing measurement, thereby avoiding any assumption about whether it was above or below a threshold. This censoring explains why some Kaplan-Meier curves terminate before 60 months.

Having estimated the model parameters, we then used the model with these estimates to generate monthly latent square root CD4 counts, η_{it}, and monthly square root observed CD4 counts, $\sqrt{CD4_i(t)}$, for times through to 61 months after the first seropositive HIV test result. Squaring each of these two quantities gave a sequence of monthly latent CD4 counts and a sequence of monthly observed CD4 counts over a 61 month period for each of the 50,000 patients. This simulation assumes that the statistical model is appropriate for monthly measurements and hence that the term ε_{it} in the model includes very short-term biological variation over a few days while the model for η_{it} captures longer-term variation over months