Supplement for: CD4 cell dynamics in untreated HIV-1 infection: overall rates, and effects of age, viral load, gender and calendar time.

Anne Cori*[1], Michael Pickles*[1], Ard van Sighem[2], Luuk Gras[2], Daniela Bezemer[2], Peter Reiss[2], Christophe Fraser[1]
(*equal contibution)

[1] Department of Infectious Disease Epidemiology, School of Public Health,
Imperial College London, London W2 1PG, UK.
[2] Stichting HIV Monitoring, Academic Medical Centre, University of Amsterdam,
Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands.

Contents

1 Overview

In this supplement, we present the data in more detail as well as a range of results and sensitivity analyses aiming to further explore the factors potentially influencing CD4 dynamics in HIV infected individuals as well as the possible biases in the analyses presented in the main text.

For all analyses, we assessed the variability in the estimated CD4 dynamics (either across different subgroups e.g. with different set-point viral load (SPVL), or using different methods, e.g. for the smoothing) by examining whether the 95% confidence intervals (95%CI) for parameter estimates were overlapping.

For some of the analyses presented here, there were not enough (≤ 1) deaths observed amongst individuals with smoothed CD4<200. For these, estimates of q_4 could not be obtained and are shown as NA in tables, and not shown in Figures.

1.1 Sensitivity analyses

We explored three smoothing methods (monotonic spline, unconstrained spline, and linear regression). All of them lead to consistent estimates (i.e. the 95%CI were overlapping for all parameters), both for the SPVL-stratified and unstratified analyses. However, using no smoothing led to different results, with fewer individuals estimated to have high CD4 counts after seroconversion, and faster progression through the 2 first CD4 categories.

We also varied the minimal number of CD4 counts required for inclusion of individuals in the analyses, and found no significant impact on the estimated CD4 dynamics. Similarly, we varied the definition of the time period during which these CD4 counts could be measured, namely before the initiation of triple, dual or monotherapy, and again found no significant impact on the estimated CD4 dynamics. Finally, we assessed the potential bias associated with the partly retrospective nature of the ATHENA cohort, by comparing results obtained with the whole dataset and discarding individuals diagnosed before 1996. There were no significant differences.

1.2 Factors influencing CD4 dynamics

In the main text, we showed that age at seroconversion was not a predictor of CD4 dynamics, but SPVL was a very strong predictor. Here, we examined the potential influence of the following additional factors on the CD4 dynamics: gender, time of seroconversion, and transmission route.

We found no effect of gender. Time of seroconversion had a significant effect in the unstratified analysis, but the effect disappeared in the SPVL-stratified analysis. There was no systematic effect of transmission route, but a small number of the estimated parameters in the SPVL-stratified analysis were significantly different in MSM/bisexuals compared to the other individuals. In particular the MSM/bisexuals with SPVL $\geq$ 4.5 log10 copies/ml appeared to progress faster from CD4 200 to death.

We further explored the influence of SPVL on CD4 dynamics through an analysis stratified by narrower SPVL categories, which confirmed the dramatic and systematic impact of SPVL on CD4 dynamics. Finally, we used a Cox proportional hazard model to better quantify this impact. We found that the hazard of progressing from each CD4 category to the next was roughly doubled for each log10 increase in SPVL, with a stronger effect in high CD4 categories (the relative hazard was 2.17 (95%CI 1.89-2.49), 1.88 (95%CI 1.61-2.19), 1.96 (95%CI 1.41-2.73) and 1.63 (95%CI 0.77-3.44), for each of the 4 CD4 categories).

2 Data

2.1 Data selection

The ATHENA cohort comprised $21,999$ individuals in total, of who $2,858$ had a seroconversion window (defined by a negative HIV test and a positive HIV test) no longer than a year. Date of seroconversion was estimated as the mid-point between these tests. We further restricted the analysis to the $2,359$ individuals who had a known date of HAART initiation, where we defined a patient on HAART as one taking either three or more drugs from at least two distinct drug classes, or at least three nucleoside reverse transcriptase inhibitors including abacavir or tenofovir. CD4 counts in patients who had initiated non-HAART therapy were however included.

For these individuals, we considered the viral load measurements taken before the initiation of HAART, at least six months after the first positive HIV test but within two years of that test, in order to focus on viral load in untreated patients in the chronic phase. For the $1,202$ patients who had at least one such viral load measurement, we defined the set-point viral load (SPVL) as the geometric mean of all available viral load measurements. For the primary analysis we further selected those who had at least 6 CD4 counts for analysis, leading to a final sample of 873 patients (Figure S1). 1,039 individuals, who had at least 6 CD4 counts but without necessarily having any set-point viral load measurements, were included in the unstratified analysis.

Sensitivity analyses were performed for patients who had between 3 and 7 CD4 measurements (n=$1,571$ and n=903 respectively). Of these, n=$1,154$ and n=757 had one or more SPVL measurements respectively. The results are shown in the next section.

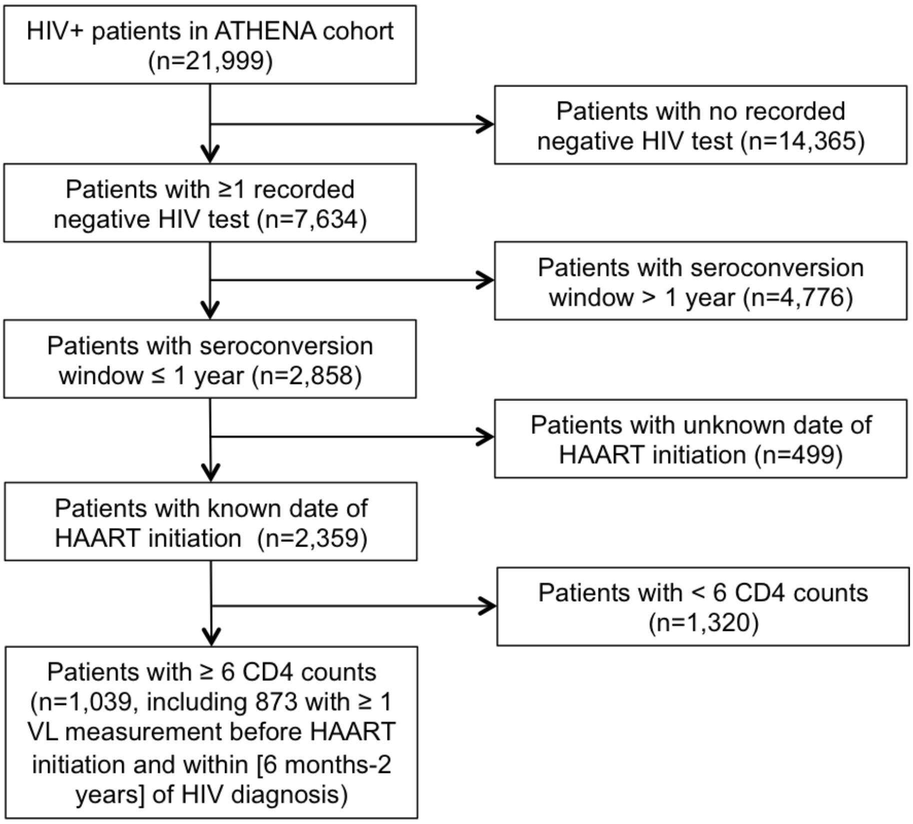
HIV+ patients in ATHENA cohort (n=21,999)
Patients with no recorded negative HIV test (n=14,365)
Patients with ≥1 recorded negative HIV test (n=7,634)
Patients with seroconversion window > 1 year (n=4,776)
Patients with seroconversion window ≤ 1 year (n=2,858)
Patients with unknown date of HAART initiation (n=499)
Patients with known date of HAART initiation (n=2,359)
Patients with < 6 CD4 counts (n=1,320)
Patients with ≥ 6 CD4 counts (n=1,039, including 873 with ≥ 1 VL measurement before HAART initiation and within [6 months-2 years] of HIV diagnosis)

Figure S1: Data selection. VL = viral load.

2.2 Estimation of set-point viral load

For each patient, we defined the set point viral load (SPVL) as the geometric mean of all viral load measurements taken during the set point window and prior to HAART initiation. Viral load units are per ml of peripheral blood plasma, and are measured using a wide range of assays. In some cases, viral loads are below or above the detection limits of the particular assay, and in this case we used the detection limit as a proxy for the viral load. We performed a more sophisticated maximum likelihood estimation of SPVL that allows for variable detection limits, but encountered difficulties with some outliers at either end of the SPVL distribution, and did not use the estimates as a result (not shown). Because our main analysis groups patients by SPVL category, this approach is unlikely to have affected our results.

Across the cohort, SPVL varied over time, as illustrated in Figure S2.

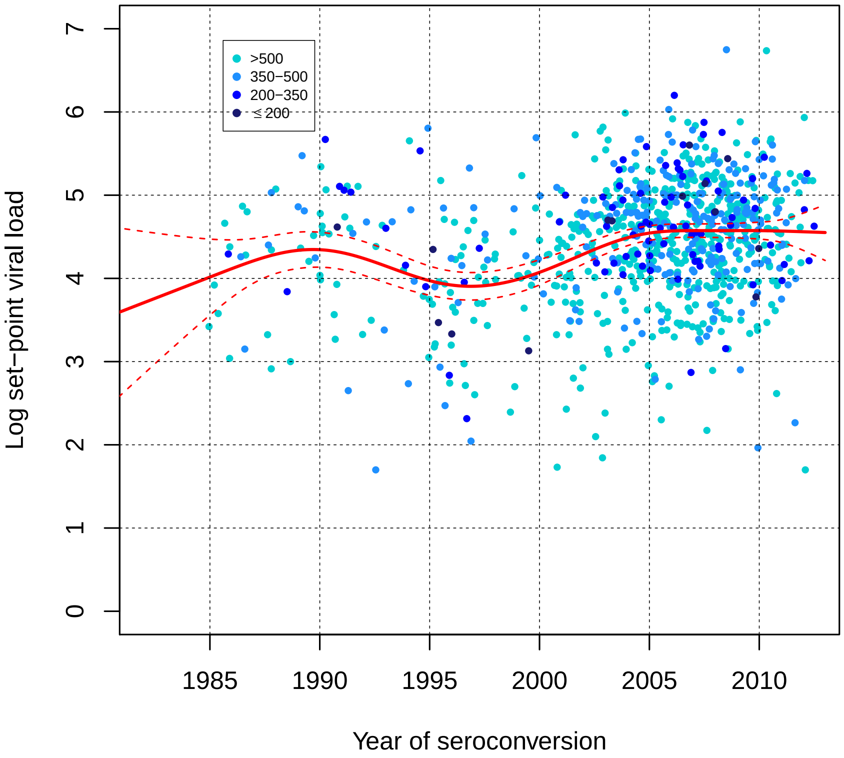

Figure S2: Set-point viral load across individuals as a function of seroconversion date, stratified by first CD4 measurement category ($>$500, 350-500, 200-350, $\leq$200 cells/mm^3).

Figure S3 shows the distribution of all individual viral load measurements throughout the set-point window, stratified by SPVL and CD4 count. Most viral load measurements of a patient lie in the range of their assigned SPVL category, even when the patient progresses to lower CD4 categories, showing that SPVL is a good predictor of viral load throughout untreated chronic infection.

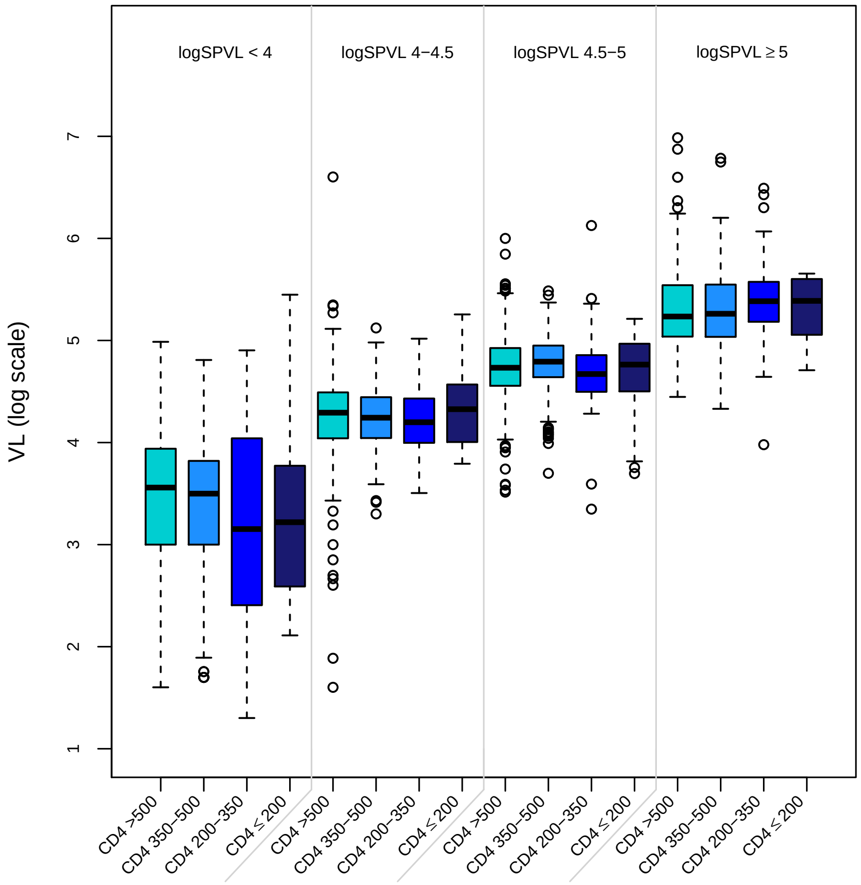

Figure S3: Distribution of all individual log10 viral load measurements, taken during the set-point window and prior to initiating HAART, within each of the 4 SPVL categories (< 4, $4-4.5$, $4.5-5$, ≥ 5 log10 copies/ml) for each CD4 category (> 500, $350 - 500$, $200 - 350$, $\leq$200 cells/mm^3)

3 Sensitivity analyses

3.1 Smoothing method

Here we compare results obtained with different smoothing methods to describe the decline of CD4 for each individual as a function of time since infection. The monotonic decreasing cubic smoothing spline was the method used in the main text. For this method, the dimension of the basis defining the spline was set to a default value of 4 and incremented to 5 or 6 if any of the predicted CD4 differed from the observed ones by more than 20%. We also considered an unconstrained cubic smoothing spline, and a linear function constrained to be decreasing. The monotonicity condition was imposed as the CD4 count would be expected to decline over time in the absence of effective treatment. For the unconstrained spline, we defined the time of transition to a given CD4 category as the earliest time when the CD4 trajectory fell under the upper bound defining that CD4 category. We also examined whether using the times at which the CD4 measurements first fell below the threshold of each CD4 category could be used - in other words a model with no smoothing. The fits of the three smoothing models to 19 randomly sampled individuals to the CD4 measurements is shown in Figure S4 (Note that individuals shown in panels A, B, G, H and M are those shown in the main Figure 1B). Parameter estimates from all models are given in Table S1, and comparison between observed and smoothed CD4 counts with the different methods is shown in Table S2.

3.2 Minimum number of CD4 measurements per patient

We next examine whether the requirement that individuals included in the analysis have 6 or more CD4 measurements alters our results. In Table S3 we show results from the monotonic cubic spline smoothing model for individuals with ≥ 5, ≥ 6 and ≥ 7 CD4 measurements respectively. We also give results from the linear smoothing model fitted using individuals with a minimum of 3-7 CD4 measurements, to examine whether omitting individuals with 3 or 4 CD4 measurements could lead to biased estimates of the rate of CD4 progression, as such individuals could potentially be faster-progressors who would start highly active antiretroviral therapy (HAART) more quickly (see Table S4).

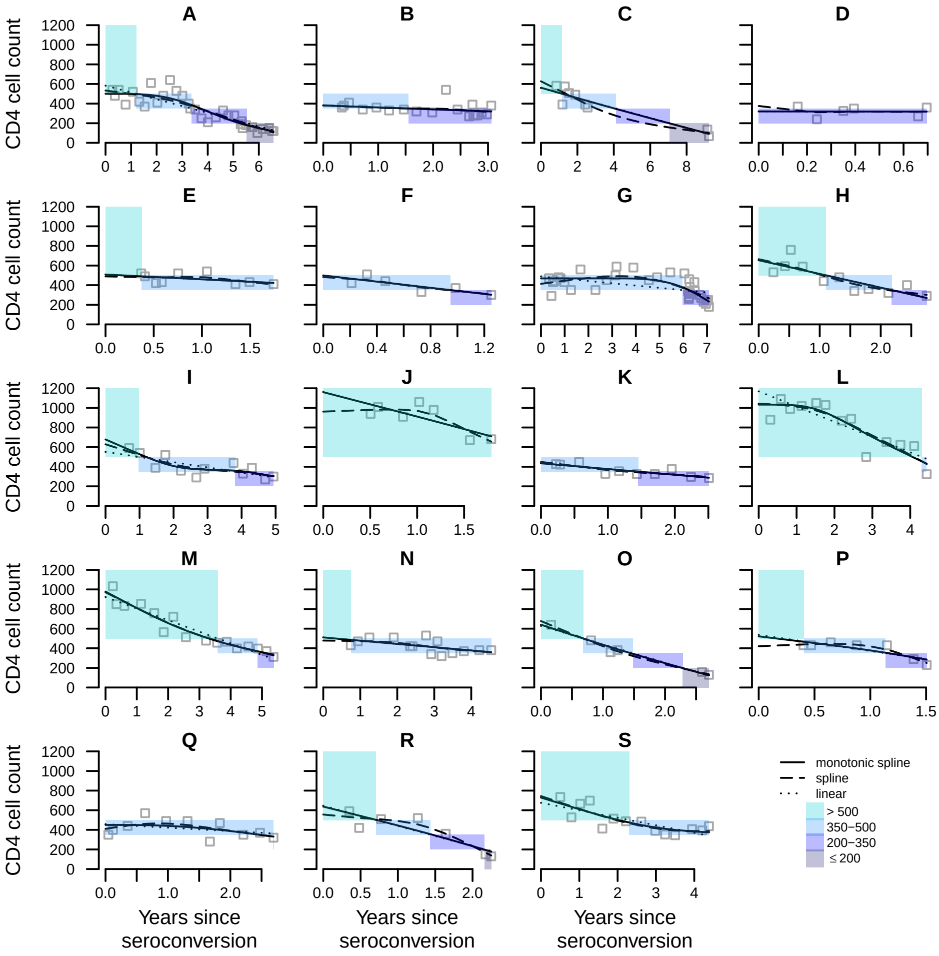

Figure S4: Fits of three smoothing models (monotonic cubic spline model, cubic spline, and linear) for a random sample of 19 individuals (including 5 patients also shown in the main text, here shown in panels A, B, G, H and M). CD4 cell count is in cells/mm^3. The coloured rectangles show the classification into CD4 categories according to the monotonic cubic spline model.

	$^1/q_1$	$^1/q_2$	$^1/q_3$	$^1/q_4$	f_1	f_2	f_3	f_4
Unstratified								
Monotonic spline	3.32 (3.07-3.58)	2.70 (2.48-2.94)	5.50 (4.69-6.54)	5.06 (3.61-7.29)	0.76 (0.73-0.78)	0.19 (0.17-0.22)	0.05 (0.03-0.06)	0.00 (0.00-0.01)
Spline	3.30 (3.05-3.56)	2.95 (2.69-3.27)	6.61 (5.51-8.05)	7.05 (5.25-9.57)	0.70 (0.67-0.73)	0.21 (0.18-0.23)	0.08 (0.06-0.09)	0.02 (0.01-0.02)
Linear	3.39 (3.15-3.66)	2.86 (2.63-3.12)	5.60 (4.74-6.72)	4.78 (3.33-7.14)	0.76 (0.73-0.78)	0.19 (0.17-0.22)	0.04 (0.03-0.06)	0.00 (0.00-0.01)
None	2.61 (2.41-2.83)	2.68 (2.48-2.90)	6.79 (5.83-7.89)	8.40 (6.33-11.81)	0.61 (0.58-0.64)	0.26 (0.23-0.29)	0.11 (0.09-0.13)	0.02 (0.01-0.03)
Stratified by SPVL								
Monotonic spline								
logSPVL < 4.0	5.35 (4.56-6.37)	3.66 (2.98-4.53)	7.62 (5.04-13.69)	6.59 (3.28-12.87)	0.86 (0.81-0.91)	0.11 (0.07-0.16)	0.02 (0.00-0.04)	0.00 (0.00-0.02)
logSPVL 4.0-4.5	3.12 (2.68-3.64)	3.09 (2.65-3.64)	8.39 (5.46-15.55)	3.26 (1.43- 6.09)	0.78 (0.72-0.83)	0.19 (0.14-0.25)	0.03 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.35 (2.08-2.64)	2.32 (1.98-2.72)	6.57 (4.73-10.22)	9.71 (4.41-23.64)	0.74 (0.69-0.79)	0.21 (0.16-0.25)	0.05 (0.03-0.08)	0.00 (0.00-0.01)
logSPVL ≥ 5.0	1.51 (1.28-1.76)	1.44 (1.22-1.69)	2.93 (2.12- 4.19)	2.14 (1.32- 3.59)	0.71 (0.64-0.77)	0.25 (0.19-0.31)	0.04 (0.02-0.08)	0.00 (0.00-0.00)
Spline								
logSPVL < 4.0	4.92 (4.17-5.76)	4.06 (3.32-5.04)	11.04 (6.96-21.35)	8.85 (4.06-21.20)	0.81 (0.75-0.86)	0.14 (0.10-0.19)	0.04 (0.01-0.07)	0.01 (0.00-0.03)
logSPVL 4.0-4.5	3.03 (2.60-3.50)	3.29 (2.79-3.88)	13.27 (8.13-26.97)	7.91 (2.70-16.19)	0.73 (0.67-0.79)	0.21 (0.15-0.27)	0.05 (0.02-0.08)	0.00 (0.00-0.02)
logSPVL 4.5-5.0	2.39 (2.11-2.69)	2.41 (2.04-2.81)	9.18 (6.17-15.10)	10.01 (5.48-20.39)	0.69 (0.63-0.74)	0.23 (0.18-0.28)	0.07 (0.05-0.10)	0.01 (0.00-0.02)
logSPVL ≥ 5.0	1.55 (1.30-1.81)	1.67 (1.35-2.00)	2.66 (2.00- 3.55)	3.30 (2.10- 5.96)	0.63 (0.56-0.70)	0.23 (0.17-0.30)	0.13 (0.08-0.18)	0.01 (0.00-0.02)
Linear								
logSPVL < 4.0	5.77 (4.86-6.90)	3.89 (3.20-4.75)	7.74 (5.03-13.16)	5.00 (2.29-11.03)	0.84 (0.79-0.89)	0.13 (0.08-0.17)	0.02 (0.00-0.05)	0.00 (0.00-0.02)
logSPVL 4.0-4.5	3.10 (2.68-3.61)	3.51 (2.99-4.16)	9.03 (5.72-17.86)	4.31 (2.26- 6.57)	0.78 (0.72-0.83)	0.20 (0.15-0.25)	0.02 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.40 (2.11-2.73)	2.40 (2.06-2.82)	6.06 (4.45- 8.77)	7.54 (3.11-19.42)	0.74 (0.69-0.79)	0.20 (0.16-0.25)	0.05 (0.03-0.08)	0.00 (0.00-0.01)
logSPVL ≥ 5.0	1.49 (1.29-1.72)	1.43 (1.22-1.67)	3.21 (2.27- 4.77)	1.75 (1.13- 2.97)	0.72 (0.65-0.79)	0.25 (0.18-0.32)	0.03 (0.01-0.06)	0.00 (0.00-0.00)
None								
logSPVL < 4.0	3.63 (3.15-4.15)	3.24 (2.77-3.85)	11.33 (7.69-17.38)	13.27 (5.40-31.38)	0.73 (0.67-0.78)	0.21 (0.15-0.26)	0.05 (0.02-0.08)	0.02 (0.00-0.04)
logSPVL 4.0-4.5	2.35 (1.97-2.79)	3.13 (2.68-3.67)	9.63 (7.01-14.37)	7.23 (3.19-16.87)	0.64 (0.57-0.70)	0.24 (0.19-0.30)	0.11 (0.07-0.15)	0.01 (0.00-0.02)
logSPVL 4.5-5.0	1.93 (1.69-2.17)	2.29 (1.98-2.63)	8.19 (6.08-12.06)	15.74 (6.49-85.84)	0.59 (0.53-0.65)	0.30 (0.25-0.35)	0.10 (0.06-0.13)	0.01 (0.00-0.03)
logSPVL ≥ 5.0	1.39 (1.20-1.58)	1.56 (1.31-1.84)	2.73 (2.09- 3.56)	4.00 (2.70- 6.50)	0.53 (0.47-0.60)	0.32 (0.25-0.38)	0.14 (0.09-0.19)	0.02 (0.00-0.04)

Table S1: Average time to progress to next CD4 stage (in years), and proportion of individuals initially in each CD4 stage, unstratified and stratified by set-point viral load, according to the different smoothing methods: mean estimate (95% confidence interval).

	Smoothed CD4>500	**Smoothed 350-500**	**Smoothed CD4 200-350**	**Smoothed CD4<200**
Monotonic spline smoothing				
Observed CD4>500	4853 (83.98%)	910 (18.69%)	24 (0.85%)	1 (0.16%)
Observed CD4 350-500	855 (14.79%)	3122 (64.12%)	566 (20.11%)	8 (1.29%)
Observed CD4 200-350	65 (1.12%)	815 (16.74%)	1961 (69.66%)	66 (10.61%)
Observed CD4<200	6 (0.10%)	22 (0.45%)	264 (9.38%)	547 (87.94%)
Total	5779 (100%)	4869 (100%)	2815 (100%)	622 (100%)
Unconstrained spline smoothing				
Observed CD4>500	4415 (81.58%)	1116 (23.87%)	208 (6.65%)	49 (5.63%)
Observed CD4 350-500	848 (15.67%)	2870 (61.39%)	784 (25.06%)	49 (5.63%)
Observed CD4 200-350	135 (2.49%)	655 (14.01%)	1925 (61.54%)	192 (22.07%)
Observed CD4<200	14 (0.26%)	34 (0.73%)	211 (6.75%)	580 (66.67%)
Total	5412 (100%)	4675 (100%)	3128 (100%)	870 (100%)
Linear regression smoothing				
Observed CD4>500	4816 (82.65%)	946 (19.31%)	25 (0.91%)	1 (0.16%)
Observed CD4 350-500	919 (15.77%)	3043 (62.10%)	584 (21.31%)	5 (0.81%)
Observed CD4 200-350	83 (1.42%)	880 (17.96%)	1852 (67.59%)	92 (14.89%)
Observed CD4<200	9 (0.15%)	31 (0.63%)	279 (10.18%)	520 (84.14%)
Total	5827 (100%)	4900 (100%)	2740 (100%)	618 (100%)

Table S2: Observed versus Predicted CD4 category according to different methods. Each line in the table shows how observed CD4 in a certain category were predicted using a given method.

	$^1/\mathbf{q_1}$	$^1/\mathbf{q_2}$	$^1/\mathbf{q_3}$	$^1/\mathbf{q_4}$	$\mathbf{f_1}$	$\mathbf{f_2}$	$\mathbf{f_3}$	$\mathbf{f_4}$
Unstratified								
$\geq$ **5 CD4**	3.17 (2.94-3.40)	2.53 (2.34-2.74)	5.51 (4.71-6.51)	4.66 (3.28-6.62)	0.73 (0.71-0.76)	0.21 (0.18-0.23)	0.05 (0.04-0.07)	0.01 (0.00-0.01)
$\geq$ **6 CD4**	3.32 (3.07-3.58)	2.70 (2.48-2.94)	5.50 (4.69-6.54)	5.06 (3.61-7.29)	0.76 (0.73-0.78)	0.19 (0.17-0.22)	0.05 (0.03-0.06)	0.00 (0.00-0.01)
$\geq$ **7 CD4**	3.54 (3.28-3.84)	2.90 (2.65-3.18)	5.46 (4.59-6.67)	5.20 (3.61-7.62)	0.77 (0.74-0.80)	0.18 (0.16-0.21)	0.04 (0.03-0.05)	0.01 (0.00-0.01)
Stratified by SPVL								
$\geq$ **5 CD4**								
logSPVL < 4.0	5.30 (4.54-6.26)	3.52 (2.90-4.38)	8.08 (5.24-15.00)	6.59 (3.13-12.53)	0.85 (0.80-0.90)	0.12 (0.08-0.17)	0.02 (0.00-0.05)	0.00 (0.00-0.01)
logSPVL 4.0-4.5	3.00 (2.58-3.46)	3.02 (2.59-3.55)	8.48 (5.36-15.39)	3.26 (1.29- 6.25)	0.78 (0.73-0.83)	0.19 (0.14-0.24)	0.03 (0.00-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.33 (2.07-2.61)	2.21 (1.90-2.52)	6.38 (4.58- 9.74)	7.93 (3.52-18.67)	0.70 (0.65-0.75)	0.24 (0.19-0.29)	0.06 (0.03-0.08)	0.00 (0.00-0.01)
logSPVL $\geq$ 5.0	1.44 (1.26-1.66)	1.38 (1.19-1.61)	3.17 (2.29- 4.45)	1.94 (1.20- 3.29)	0.69 (0.63-0.75)	0.26 (0.20-0.31)	0.05 (0.03-0.08)	0.00 (0.00-0.00)
$\geq$ **6 CD4**								
logSPVL < 4.0	5.35 (4.56-6.37)	3.66 (2.98-4.53)	7.62 (5.04-13.69)	6.59 (3.28-12.87)	0.86 (0.81-0.91)	0.11 (0.07-0.16)	0.02 (0.00-0.04)	0.00 (0.00-0.02)
logSPVL 4.0-4.5	3.12 (2.68-3.64)	3.09 (2.65-3.64)	8.39 (5.46-15.55)	3.26 (1.43- 6.09)	0.78 (0.72-0.83)	0.19 (0.14-0.25)	0.03 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.35 (2.08-2.64)	2.32 (1.98-2.72)	6.57 (4.73-10.22)	9.71 (4.41-23.64)	0.74 (0.69-0.79)	0.21 (0.16-0.25)	0.05 (0.03-0.08)	0.00 (0.00-0.01)
logSPVL $\geq$ 5.0	1.51 (1.28-1.76)	1.44 (1.22-1.69)	2.93 (2.12- 4.19)	2.14 (1.32- 3.59)	0.71 (0.64-0.77)	0.25 (0.19-0.31)	0.04 (0.02-0.08)	0.00 (0.00-0.00)
$\geq$ **7 CD4**								
logSPVL < 4.0	5.42 (4.61-6.38)	3.61 (2.91-4.53)	7.52 (5.00-13.43)	6.59 (3.14-12.53)	0.86 (0.82-0.91)	0.11 (0.07-0.16)	0.02 (0.00-0.03)	0.01 (0.00-0.02)
logSPVL 4.0-4.5	3.24 (2.77-3.79)	3.26 (2.79-3.92)	9.20 (5.70-17.13)	3.09 (1.18- 5.80)	0.79 (0.74-0.85)	0.18 (0.12-0.23)	0.03 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.53 (2.23-2.85)	2.55 (2.17-2.99)	6.52 (4.56-10.72)	9.25 (3.73-22.74)	0.73 (0.67-0.79)	0.21 (0.16-0.27)	0.05 (0.02-0.08)	0.00 (0.00-0.01)
logSPVL $\geq$ 5.0	1.66 (1.43-1.96)	1.52 (1.26-1.83)	2.81 (1.96- 4.23)	2.41 (1.57- 4.34)	0.73 (0.66-0.80)	0.22 (0.15-0.28)	0.05 (0.02-0.09)	0.00 (0.00-0.00)

Table S3: Average time to progress to next CD4 stage (in years), and proportion of individuals initially in each CD4 stage, unstratified and stratified by set-point viral load, according to the different minimum number of CD4 measurements per patient, using the monotonic spline model: mean estimate (95% confidence interval).

	$^{1}/q_{1}$	$^{1}/q_{2}$	$^{1}/q_{3}$	$^{1}/q_{4}$	f_{1}	f_{2}	f_{3}	f_{4}
Unstratified								
≥ **3 CD4**	2.96 (2.76-3.19)	2.34 (2.16-2.55)	5.02 (4.32-5.81)	4.46 (3.14-6.11)	0.70 (0.68-0.72)	0.23 (0.21-0.24)	0.07 (0.05-0.08)	0.01 (0.01-0.02)
≥ **4 CD4**	3.09 (2.87-3.32)	2.49 (2.30-2.71)	5.16 (4.40-6.05)	4.49 (3.23-6.20)	0.72 (0.69-0.74)	0.22 (0.20-0.24)	0.06 (0.05-0.08)	0.01 (0.00-0.01)
≥ **5 CD4**	3.26 (3.01-3.49)	2.68 (2.48-2.90)	5.56 (4.73-6.58)	4.48 (3.26-6.33)	0.73 (0.71-0.76)	0.21 (0.19-0.23)	0.05 (0.04-0.06)	0.01 (0.00-0.01)
≥ **6 CD4**	3.39 (3.15-3.66)	2.86 (2.63-3.12)	5.60 (4.74-6.72)	4.78 (3.33-7.14)	0.76 (0.73-0.78)	0.19 (0.17-0.22)	0.04 (0.03-0.06)	0.00 (0.00-0.01)
≥ **7 CD4**	3.63 (3.33-3.95)	3.12 (2.85-3.43)	5.60 (4.75-6.70)	4.86 (3.34-7.28)	0.77 (0.74-0.80)	0.19 (0.16-0.21)	0.04 (0.03-0.05)	0.01 (0.00-0.01)
Stratified by SPVL								
≥ **3 CD4**								
logSPVL < 4.0	5.47 (4.68-6.48)	3.54 (2.93-4.32)	8.14 (5.64-13.96)	5.00 (2.36-11.08)	0.83 (0.78-0.87)	0.15 (0.10-0.19)	0.02 (0.00-0.04)	0.00 (0.00-0.01)
logSPVL 4.0-4.5	2.89 (2.51-3.37)	3.18 (2.70-3.70)	9.04 (5.61-17.13)	4.32 (2.22- 6.16)	0.77 (0.71-0.82)	0.21 (0.16-0.26)	0.03 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.29 (2.01-2.62)	2.20 (1.94-2.56)	5.55 (4.26- 7.97)	6.53 (2.57-16.88)	0.69 (0.64-0.74)	0.24 (0.19-0.28)	0.07 (0.04-0.10)	0.00 (0.00-0.01)
logSPVL ≥ 5.0	1.32 (1.15-1.48)	1.24 (1.09-1.42)	3.07 (2.29- 4.34)	1.48 (0.96- 2.43)	0.70 (0.65-0.75)	0.25 (0.20-0.30)	0.05 (0.03-0.08)	0.00 (0.00-0.00)
≥ **4 CD4**								
logSPVL < 4.0	5.48 (4.67-6.48)	3.60 (2.96-4.40)	8.11 (5.49-14.06)	5.00 (2.38-11.47)	0.83 (0.78-0.88)	0.14 (0.10-0.19)	0.02 (0.00-0.04)	0.00 (0.00-0.01)
logSPVL 4.0-4.5	2.93 (2.53-3.37)	3.24 (2.77-3.81)	8.95 (5.70-16.66)	4.32 (2.30- 6.43)	0.77 (0.72-0.82)	0.20 (0.15-0.25)	0.03 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.33 (2.05-2.63)	2.22 (1.93-2.56)	5.50 (4.14- 7.66)	6.52 (2.78-18.60)	0.70 (0.65-0.74)	0.24 (0.19-0.29)	0.06 (0.03-0.08)	0.00 (0.00-0.01)
logSPVL ≥ 5.0	1.38 (1.20-1.57)	1.28 (1.11-1.48)	3.29 (2.48- 4.59)	1.54 (1.03- 2.52)	0.70 (0.64-0.75)	0.25 (0.20-0.30)	0.06 (0.03-0.09)	0.00 (0.00-0.00)
≥ **5 CD4**								
logSPVL < 4.0	5.71 (4.84-6.70)	3.72 (3.07-4.67)	8.01 (5.46-14.78)	5.00 (2.26-10.88)	0.83 (0.78-0.88)	0.14 (0.10-0.19)	0.02 (0.00-0.05)	0.00 (0.00-0.01)
logSPVL 4.0-4.5	2.99 (2.60-3.44)	3.41 (2.90-4.03)	9.15 (5.79-17.67)	4.31 (2.22- 6.21)	0.78 (0.73-0.83)	0.20 (0.15-0.25)	0.02 (0.00-0.04)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.40 (2.13-2.73)	2.27 (1.96-2.62)	5.90 (4.40- 8.69)	6.42 (2.81-17.15)	0.70 (0.65-0.76)	0.23 (0.19-0.28)	0.06 (0.03-0.09)	0.00 (0.00-0.01)
logSPVL ≥ 5.0	1.45 (1.26-1.67)	1.38 (1.20-1.60)	3.46 (2.50- 5.08)	1.59 (1.07- 2.66)	0.69 (0.63-0.75)	0.26 (0.20-0.32)	0.04 (0.02-0.07)	0.00 (0.00-0.00)
≥ **6 CD4**								
logSPVL < 4.0	5.77 (4.86-6.90)	3.89 (3.20-4.75)	7.74 (5.03-13.16)	5.00 (2.29-11.03)	0.84 (0.79-0.89)	0.13 (0.08-0.17)	0.02 (0.00-0.05)	0.00 (0.00-0.02)
logSPVL 4.0-4.5	3.10 (2.68-3.61)	3.51 (2.99-4.16)	9.03 (5.72-17.86)	4.31 (2.26- 6.57)	0.78 (0.72-0.83)	0.20 (0.15-0.25)	0.02 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.40 (2.11-2.73)	2.40 (2.06-2.82)	6.06 (4.45- 8.77)	7.54 (3.11-19.42)	0.74 (0.69-0.79)	0.20 (0.16-0.25)	0.05 (0.03-0.08)	0.00 (0.00-0.01)
logSPVL ≥ 5.0	1.49 (1.29-1.72)	1.43 (1.22-1.67)	3.21 (2.27- 4.77)	1.75 (1.13- 2.97)	0.72 (0.65-0.79)	0.25 (0.18-0.32)	0.03 (0.01-0.06)	0.00 (0.00-0.00)
≥ **7 CD4**								
logSPVL < 4.0	5.86 (4.96-6.89)	3.90 (3.18-4.80)	7.63 (5.23-12.57)	5.00 (2.23-10.63)	0.84 (0.79-0.89)	0.13 (0.09-0.18)	0.02 (0.01-0.04)	0.01 (0.00-0.02)
logSPVL 4.0-4.5	3.21 (2.78-3.76)	3.80 (3.23-4.60)	10.10 (6.24-20.57)	4.05 (2.19- 5.87)	0.79 (0.73-0.85)	0.19 (0.13-0.24)	0.02 (0.00-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.59 (2.25-2.99)	2.65 (2.24-3.08)	6.01 (4.39- 9.17)	7.18 (2.99-18.56)	0.74 (0.69-0.80)	0.21 (0.15-0.26)	0.05 (0.02-0.08)	0.00 (0.00-0.01)
logSPVL ≥ 5.0	1.64 (1.39-1.93)	1.53 (1.29-1.82)	3.08 (2.22- 4.63)	1.96 (1.36- 3.45)	0.74 (0.66-0.81)	0.23 (0.16-0.30)	0.03 (0.01-0.07)	0.00 (0.00-0.00)

Table S4: Average time to progress to next CD4 stage (in years), and proportion of individuals initially in each CD4 stage, with linear model, unstratified and stratified by set-point viral load, according to the different minimum number of CD4 measurements per patient: mean estimate (95% confidence interval)

3.3 Retrospective part of the cohort

Within the ATHENA cohort, data was collected in patients diagnosed after 1996, and patients who were diagnosed before 1996 and were still alive in 1996 had available data included retrospectively where possible. In Table S5, we compare results obtained with all data and data from individuals diagnosed after 1996 in order to evaluate the potential bias in our results due to the twofold nature of the data collection.

	$^1/q_1$	$^1/q_2$	$^1/q_3$	$^1/q_4$	f_1	f_2	f_3	f_4
Unstratified								
Full dataset	3.32 (3.07-3.58)	2.70 (2.48-2.94)	5.50 (4.69-6.54)	5.06 (3.61-7.29)	0.76 (0.73-0.78)	0.19 (0.17-0.22)	0.05 (0.03-0.06)	0.00 (0.00-0.01)
After 1996	2.99 (2.76-3.26)	2.41 (2.21-2.64)	5.99 (4.77-7.90)	7.86 (3.43-14.44)	0.77 (0.75-0.80)	0.19 (0.16-0.22)	0.04 (0.02-0.05)	0.00 (0.00-0.00)
Stratified by SPVL								
Full dataset								
logSPVL < 4.0	5.35 (4.56-6.37)	3.66 (2.98-4.53)	7.62 (5.04-13.69)	6.59 (3.28-12.87)	0.86 (0.81-0.91)	0.11 (0.07-0.16)	0.02 (0.00-0.04)	0.00 (0.00-0.02)
logSPVL 4.0-4.5	3.12 (2.68-3.64)	3.09 (2.65-3.64)	8.39 (5.46-15.55)	3.26 (1.43- 6.09)	0.78 (0.72-0.83)	0.19 (0.14-0.25)	0.03 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.35 (2.08-2.64)	2.32 (1.98-2.72)	6.57 (4.73-10.22)	9.71 (4.41-23.64)	0.74 (0.69-0.79)	0.21 (0.16-0.25)	0.05 (0.03-0.08)	0.00 (0.00-0.01)
logSPVL ≥ 5.0	1.51 (1.28-1.76)	1.44 (1.22-1.69)	2.93 (2.12- 4.19)	2.14 (1.32- 3.59)	0.71 (0.64-0.77)	0.25 (0.19-0.31)	0.04 (0.02-0.08)	0.00 (0.00-0.00)
After 1996								
logSPVL < 4.0	5.45 (4.55-6.58)	3.51 (2.75-4.42)	9.35 (4.90-24.45)	NA	0.88 (0.83-0.93)	0.10 (0.06-0.15)	0.01 (0.00-0.03)	0.01 (0.00-0.02)
logSPVL 4.0-4.5	3.04 (2.57-3.55)	2.99 (2.54-3.54)	9.77 (5.84-22.39)	NA	0.77 (0.71-0.83)	0.20 (0.15-0.26)	0.03 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.31 (2.04-2.63)	2.17 (1.89-2.57)	8.33 (5.46-16.38)	NA	0.75 (0.69-0.80)	0.20 (0.16-0.25)	0.05 (0.03-0.08)	0.00 (0.00-0.00)
logSPVL ≥ 5.0	1.37 (1.18-1.58)	1.41 (1.20-1.68)	2.65 (1.88- 3.96)	NA	0.72 (0.65-0.78)	0.25 (0.19-0.32)	0.04 (0.01-0.07)	0.00 (0.00-0.00)

Table S5: Average time to progress to next CD4 stage (in years) and proportion of individuals initially in each CD4 stage, unstratified and stratified by set-point viral load, for the full data set, and for only individuals with estimated date of seroconversion in or after 1996: mean estimate (95% confidence interval)

3.4 Criteria for initiating antiretroviral therapy

In the main text, individual CD4 measurements are included prior to the initiation of HAART (defined as taking either three or more drugs from at least two distinct drug classes, or at least three nucleoside reverse transcriptase inhibitors including abacavir or tenofovir) for the first time. This criterion is based on the assumption that progression between CD4 categories is only affected by HAART. Of course antiretroviral drugs are known to slow HIV disease progression and modify CD4 dynamics (e.g. [?]). Within the ATHENA cohort, individuals prior to 1996 were frequently put on one (mono therapy) or two (dual therapy) antiretroviral drugs when eligible, and for different clinical reasons individuals after 1996 were still sometimes initiated on one or two antiretrovirals prior to commencing HAART. Therefore, in Table S6 we show how the time spent in each CD4 category and the fraction of individuals starting in each CD4 compartment changes according to whether we include individuals prior to initiating at least one or two antiretroviral drugs for the first time, or only prior to initiating HAART for the first time as in the main text.

	1/q$_1$	**1/q$_2$**	**1/q$_3$**	**1/q$_4$**	**f$_1$**	**f$_2$**	**f$_3$**	**f$_4$**
Unstratified								
HAART	3.32 (3.07 - 3.58)	2.70 (2.48 - 2.94)	5.50 (4.69 - 6.54)	5.06 (3.61 - 7.29)	0.76 (0.73 - 0.78)	0.19 (0.17 - 0.22)	0.05 (0.03 - 0.06)	0.00 (0.00 - 0.01)
Dual therapy	3.34 (3.04 - 3.64)	2.80 (2.40 - 3.14)	4.82 (3.58 - 5.82)	4.05 (2.52 - 5.79)	0.76 (0.74 - 0.79)	0.19 (0.17 - 0.22)	0.04 (0.03 - 0.06)	0.00 (0.00 - 0.01)
Mono therapy	3.24 (2.97 - 3.52)	2.38 (2.19 - 2.60)	4.51 (3.23 - 5.94)	3.46 (1.92 - 4.54)	0.77 (0.75 - 0.80)	0.19 (0.16 - 0.21)	0.04 (0.03 - 0.05)	0.00 (0.00 - 0.00)
Stratified by SPVL								
HAART								
logSPVL < 4.0	5.35 (4.56 - 6.37)	3.66 (2.98 - 4.53)	7.62 (5.04 - 13.69)	6.59 (3.28 - 12.87)	0.86 (0.81 - 0.91)	0.11 (0.07 - 0.16)	0.02 (0.00 - 0.04)	0.00 (0.00 - 0.02)
logSPVL 4.0-4.5	3.12 (2.68 - 3.64)	3.09 (2.65 - 3.64)	8.39 (5.46 - 15.55)	3.26 (1.43 - 6.09)	0.78 (0.72 - 0.83)	0.19 (0.14 - 0.25)	0.03 (0.01 - 0.05)	0.00 (0.00 - 0.00)
logSPVL 4.5-5.0	2.35 (2.08 - 2.64)	2.32 (1.98 - 2.72)	6.57 (4.73 - 10.22)	9.71 (4.41 - 23.64)	0.74 (0.69 - 0.79)	0.21 (0.16 - 0.25)	0.05 (0.03 - 0.08)	0.00 (0.00 - 0.01)
logSPVL $\geq$ 5.0	1.51 (1.28 - 1.76)	1.44 (1.22 - 1.69)	2.93 (2.12 - 4.19)	2.14 (1.32 - 3.59)	0.71 (0.64 - 0.77)	0.25 (0.19 - 0.31)	0.04 (0.02 - 0.08)	0.00 (0.00 - 0.00)
Dual therapy								
logSPVL < 4.0	4.51 (3.90 - 4.95)	2.93 (2.50 - 3.38)	4.02 (2.58 - 5.51)	3.15 (1.23 - 3.43)	0.87 (0.82 - 0.92)	0.12 (0.07 - 0.17)	0.01 (0.00 - 0.03)	0.00 (0.00 - 0.00)
logSPVL 4.0-4.5	3.02 (2.61 - 3.39)	2.67 (2.33 - 3.05)	5.17 (3.01 - 6.41)	1.80 (0.77 - 3.44)	0.78 (0.72 - 0.83)	0.20 (0.14 - 0.25)	0.03 (0.01 - 0.05)	0.00 (0.00 - 0.00)
logSPVL 4.5-5.0	2.26 (2.01 - 2.53)	2.10 (1.80 - 2.40)	4.39 (2.56 - 5.29)	2.81 (1.11 - 3.44)	0.74 (0.69 - 0.79)	0.21 (0.16 - 0.26)	0.05 (0.02 - 0.07)	0.00 (0.00 - 0.01)
logSPVL $\geq$ 5.0	1.47 (1.28 - 1.73)	1.43 (1.18 - 1.72)	2.82 (1.77 - 3.97)	1.72 (0.55 - 2.59)	0.71 (0.64 - 0.77)	0.25 (0.19 - 0.31)	0.04 (0.02 - 0.08)	0.00 (0.00 - 0.00)
Mono therapy								
logSPVL < 4.0	4.45 (3.87 - 4.86)	2.84 (2.41 - 3.28)	4.53 (2.34 - 5.85)	NA	0.89 (0.84 - 0.93)	0.10 (0.06 - 0.15)	0.01 (0.00 - 0.03)	0.00 (0.00 - 0.00)
logSPVL 4.0-4.5	3.04 (2.58 - 3.41)	2.66 (2.31 - 3.00)	5.16 (2.80 - 6.61)	1.51 (0.72 - 1.82)	0.77 (0.72 - 0.83)	0.20 (0.14 - 0.25)	0.03 (0.01 - 0.05)	0.00 (0.00 - 0.00)
logSPVL 4.5-5.0	2.25 (1.99 - 2.52)	2.03 (1.75 - 2.30)	4.57 (2.55 - 5.69)	2.39 (1.16 - 2.52)	0.74 (0.69 - 0.80)	0.21 (0.16 - 0.26)	0.05 (0.03 - 0.08)	0.00 (0.00 - 0.00)
logSPVL $\geq$ 5.0	1.45 (1.25 - 1.67)	1.41 (1.15 - 1.69)	3.10 (1.93 - 4.33)	0.85 (0.46 - 1.80)	0.70 (0.64 - 0.77)	0.25 (0.18 - 0.31)	0.05 (0.02 - 0.08)	0.00 (0.00 - 0.00)

Table S6: Average time to progress to next CD4 stage (in years) and proportion of individuals initially in each CD4 stage, unstratified and stratified by set-point viral load, estimated using data on individuals prior to initiating HAART, dual therapy and mono therapy for the first time: mean estimates (95% confidence intervals).

4 Additional factors potentially influencing CD4 dynamics

4.1 Gender

Although the HIV epidemic in the Netherlands is mainly concentrated in MSM, and hence few women are infected, we wanted to explore whether the CD4 cell dynamics after infection was different between men and women. Results are presented in Table S7.

4.2 Time

In this paragraph, we examine whether there is any difference by calendar time (before 1995, 1995-2000, after 2000), motivated by the corresponding changes in mean population-level SPVL (Main Text Figure 3). Results are presented in Table S8.

4.3 Transmission route

We also explored whether CD4 progression was different according to HIV transmission route. We compared results within the MSM/bisexual population compared to other individuals. Results are shown in Table S9.

	$1/q_1$	$1/q_2$	$1/q_3$	$1/q_4$	f_1	f_2	f_3	f_4
Stratified by Gender								
Females	3.37 (2.54-4.48)	3.02 (2.07-4.49)	4.21 (2.87-6.76)	8.00 (3.30-21.24)	0.79 (0.69-0.88)	0.14 (0.06-0.22)	0.05 (0.00-0.10)	0.03 (0.00-0.08)
Males	3.31 (3.05-3.57)	2.67 (2.45-2.94)	5.67 (4.72-6.92)	4.38 (3.13- 6.63)	0.75 (0.73-0.78)	0.20 (0.17-0.22)	0.05 (0.03-0.06)	0.00 (0.00-0.01)

Table S7: Average time to progress to next CD4 stage (in years) and proportion of individuals initially in each CD4 stage, stratified by gender: mean estimate (95% confidence interval)

	$1/q_1$	$1/q_2$	$1/q_3$	$1/q_4$	f_1	f_2	f_3	f_4
Unstratified								
Before 1995	5.15 (4.33-6.10)	4.00 (3.21-4.96)	5.01 (3.90-6.44)	4.76 (3.31-7.03)	0.67 (0.60-0.74)	0.21 (0.15-0.28)	0.09 (0.05-0.14)	0.03 (0.01-0.05)
1995 – 2000	5.08 (3.86-6.71)	2.46 (1.98-3.00)	5.95 (3.79-11.65)	NA	0.81 (0.72-0.88)	0.14 (0.06-0.22)	0.04 (0.00-0.09)	0.01 (0.00-0.04)
After 2000	2.80 (2.59-3.07)	2.41 (2.19-2.63)	5.94 (4.61-8.16)	6.65 (2.66-12.96)	0.77 (0.74-0.80)	0.19 (0.17-0.22)	0.04 (0.02-0.05)	0.00 (0.00-0.00)
Stratified by SPVL								
Before 1995								
logSPVL < 4.0	5.02 (3.53-6.80)	4.90 (2.89-8.54)	7.29 (3.48-19.80)	4.79 (2.43- 7.58)	0.79 (0.62-0.95)	0.17 (0.04-0.33)	0.04 (0.00-0.14)	0.00 (0.00-0.00)
logSPVL 4.0-4.5	3.94 (2.44-5.63)	3.93 (2.43-6.67)	5.91 (2.46-13.01)	2.67 (0.95- 5.96)	0.92 (0.73-1.00)	0.00 (0.00-0.00)	0.08 (0.00-0.27)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.98 (1.99-4.07)	4.02 (2.22-7.21)	3.49 (1.95- 7.14)	9.14 (2.70-22.53)	0.58 (0.35-0.81)	0.32 (0.11-0.53)	0.05 (0.00-0.17)	0.05 (0.00-0.17)
logSPVL ≥ 5.0	3.69 (1.72-8.62)	1.50 (1.00-2.11)	3.81 (1.88- 8.22)	2.15 (1.39- 3.25)	0.57 (0.30-0.86)	0.29 (0.07-0.55)	0.14 (0.00-0.36)	0.00 (0.00-0.00)
1995 – 2000								
logSPVL < 4.0	8.48 (5.44-13.24)	2.60 (1.93-3.47)	NA	NA	0.80 (0.66-0.93)	0.14 (0.03-0.26)	0.03 (0.00-0.09)	0.03 (0.00-0.10)
logSPVL 4.0-4.5	2.26 (1.44- 3.34)	2.52 (1.58-3.97)	NA	NA	0.73 (0.53-0.92)	0.23 (0.05-0.41)	0.05 (0.00-0.17)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	3.00 (1.22- 5.86)	1.45 (0.78-2.80)	NA	NA	1.00 (1.00-1.00)	0.00 (0.00-0.00)	0.00 (0.00-0.00)	0.00 (0.00-0.00)
logSPVL ≥ 5.0	1.91 (0.71- 3.76)	2.34 (1.21-4.64)	NA	NA	0.75 (0.00-1.00)	0.25 (0.00-1.00)	0.00 (0.00-0.00)	0.00 (0.00-0.00)
After 2000								
logSPVL < 4.0	4.82 (4.05-5.78)	3.66 (2.80-4.85)	11.10 (5.37-26.55)	NA	0.89 (0.84-0.94)	0.10 (0.05-0.15)	0.01 (0.00-0.03)	0.00 (0.00-0.00)
logSPVL 4.0-4.5	3.15 (2.66-3.68)	3.09 (2.58-3.72)	11.73 (5.99-32.43)	NA	0.77 (0.71-0.84)	0.20 (0.15-0.26)	0.02 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.28 (2.02-2.56)	2.19 (1.88-2.55)	8.23 (5.36-14.49)	NA	0.74 (0.68-0.79)	0.21 (0.16-0.27)	0.05 (0.03-0.08)	0.00 (0.00-0.00)
logSPVL ≥ 5.0	1.38 (1.18-1.60)	1.41 (1.18-1.69)	2.67 (1.85- 4.18)	NA	0.72 (0.64-0.78)	0.25 (0.19-0.31)	0.04 (0.01-0.07)	0.00 (0.00-0.00)

Table S8: Average time to progress to next CD4 stage (in years) and proportion of individuals initially in each CD4 stage, unstratified and stratified by set-point viral load, by seroconversion date: mean estimate (95% confidence interval)

	$1/q_1$	$1/q_2$	$1/q_3$	$1/q_4$	f_1	f_2	f_3	f_4
Unstratified								
MSM/BI	3.25 (2.98-3.54)	2.61 (2.38-2.85)	5.64 (4.67-6.93)	4.90 (3.28-8.07)	0.76 (0.73-0.79)	0.20 (0.17-0.22)	0.04 (0.03-0.06)	0.00 (0.00-0.01)
Not MSM/BI	3.74 (3.05-4.50)	3.28 (2.49-4.25)	5.01 (3.80-6.96)	5.26 (2.98-9.71)	0.74 (0.68-0.82)	0.18 (0.12-0.24)	0.07 (0.03-0.11)	0.01 (0.00-0.04)
Stratified by SPVL								
MSM/BI								
logSPVL < 4.0	5.08 (4.20-6.03)	3.56 (2.85-4.58)	7.49 (4.62-14.56)	NA	0.87 (0.82-0.92)	0.10 (0.06-0.15)	0.02 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.0-4.5	3.23 (2.77-3.75)	3.18 (2.70-3.70)	7.92 (5.13-15.23)	NA	0.76 (0.70-0.82)	0.21 (0.16-0.27)	0.03 (0.01-0.05)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.36 (2.07-2.68)	2.23 (1.91-2.59)	6.29 (4.37-10.02)	NA	0.75 (0.70-0.81)	0.20 (0.15-0.25)	0.04 (0.02-0.07)	0.00 (0.00-0.01)
logSPVL ≥ 5.0	1.45 (1.23-1.71)	1.40 (1.18-1.65)	2.91 (2.11- 4.22)	NA	0.71 (0.64-0.77)	0.25 (0.19-0.32)	0.04 (0.01-0.07)	0.00 (0.00-0.00)
Not MSM/BI								
logSPVL < 4.0	6.87 (4.75-10.20)	4.08 (2.39-7.10)	8.14 (3.58-14.50)	NA	0.82 (0.68-0.93)	0.16 (0.05-0.28)	0.00 (0.00-0.00)	0.03 (0.00-0.08)
logSPVL 4.0-4.5	2.27 (1.47- 3.69)	2.37 (1.55-3.66)	11.22 (3.07-18.03)	NA	0.94 (0.80-1.00)	0.00 (0.00-0.00)	0.06 (0.00-0.20)	0.00 (0.00-0.00)
logSPVL 4.5-5.0	2.24 (1.60- 2.96)	3.13 (1.80-5.55)	8.39 (4.25-14.71)	NA	0.61 (0.43-0.78)	0.26 (0.11-0.41)	0.13 (0.03-0.27)	0.00 (0.00-0.00)
logSPVL ≥ 5.0	2.20 (1.41- 3.29)	1.94 (1.04-3.73)	3.03 (1.14- 9.01)	NA	0.69 (0.44-0.92)	0.25 (0.06-0.50)	0.06 (0.00-0.20)	0.00 (0.00-0.00)

Table S9: Average time to progress to next CD4 stage (in years) and proportion of individuals initially in each CD4 stage, unstratified and stratified by set-point viral load, for MSM/bisexual individuals compared to other individuals: mean estimate (95% confidence interval)

5 Further analysis of the influence of SPVL on CD4 dynamics

In our main analysis, we found that the CD4 dynamics was largely dependent on the SPVL category of individuals, with SPVL split into 4 categories roughly corresponding to the SPVL quartiles in our dataset: <4, 4-4.5, 4.5-5, ≥ 5 log10 copies/ml. In this section, we further examine the dependence of CD4 dynamics on SPVL by considering narrower SPVL intervals, and then by using a Cox proportional hazard model to directly quantify how variations in SPVL alter individual CD4 declines.

5.1 Narrower set-point viral stratification

First, to broadly assess whether the Cox model is suited to analyse the relationship between SPVL (on the log scale), and the times of transition from one CD4 category to the next, we repeated the previous analyses but with narrower equally wide SPVL categories. Figure S5 shows the rate of progression from one CD4 category to the next as well as the proportion starting in each CD4 category after seroconversion. SPVL still appears to be a strong predictor of CD4 progression rate, whereby high viral loads lead to faster progression. A linear model to explain CD4 progression rates as a function of log10 SPVL seems to be a reasonable assumption. For instance, the mean estimated rate of progressing from CD4>500 to CD4 350-500 (q_1) is well explained by a linear model of the SPVL central values of the categories (using 2.75 and 5.75 for the non extreme categories), with an adjusted R^2 of 0.91.

5.2 Cox proportional hazard model

In order to better explore the relationship between CD4 progression and SPVL, we used a Cox proportional hazard model. We denote $h_k^v(t)$ the hazard, for individual with log10 SPVL v, of moving from the k^{th} to the $(k+1)^{th}$ CD4 category. The Cox proportional hazard model assumes that $\log h_k^v(t) = \alpha_k(t) + \beta_k v$, i.e. the rate of progressing from one CD4 category to the next is assumed to be a linear function of SPVL, which seems supported by the results shown in the previous paragraph. e^{β_k} can then be interpreted as the relative hazard of progressing from CD4 category k to $k+1$ for each log10 increase in SPVL. We found this relative hazard was 2.17 (95%CI 1.89-2.49), 1.88 (95%CI 1.61-2.19), 1.96 (95%CI 1.41-2.73) and 1.63 (95%CI 0.77-3.44), for each of the 4 CD4 categories.

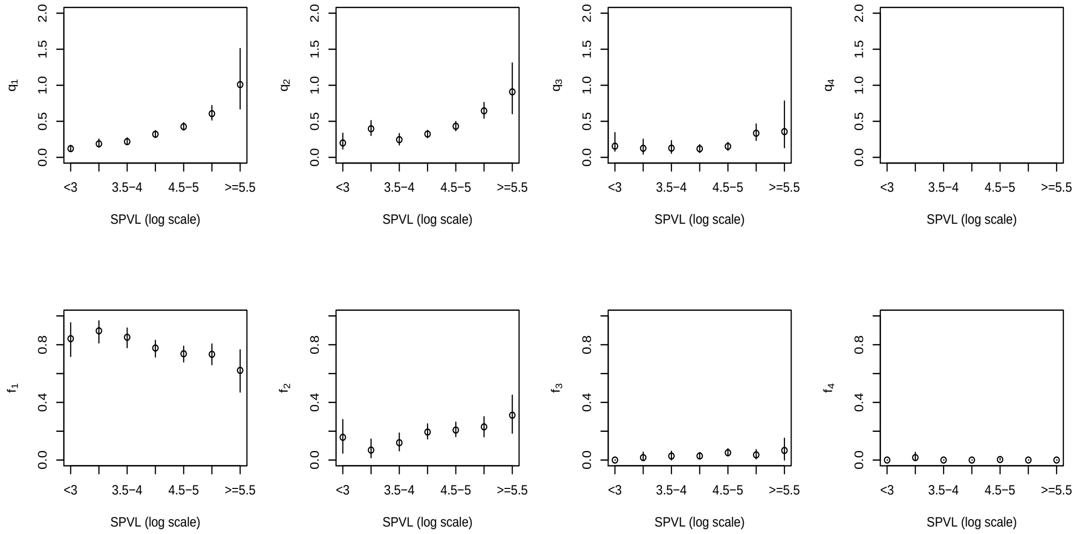

Figure S5: Average rate of progression to next CD4 stage (top row, per year) and proportion of individuals initially in each CD4 stage (bottom row), stratified by set-point viral load (<3, 3-3.5, 3.5-4, 4-4.5, 4.5-5, 5-5.5, ≥5.5).

6 Predicting future true CD4 categories given current observed CD4 category

We use our model to predict, given the first observed CD4 category of a patient, the times at which the smoothed CD4 for that patient would transition from one category to the next. Our predictive model accounts for the potential mismatch between the first observed and smoothed CD4 categories as well as the progression of smoothed CD4 over time.

Let C_t^o be the observed CD4 category at a given time t, and C_t^s be the corresponding smoothed CD4 category at time t. We denote 1 to 4 the following CD4 categories: > 500, $350 - 500$, $200 - 350$, $0 - 200$, and we denote 5 the category of deceased individuals.

Our objective is to predict, given the an observed CD4 category C_0^o (we set that time to 0 without loss of generality), what the true (or smoothed) CD4 category is at a subsequent time t. We therefore want to determine $P\left(C_t^s|C_0^o\right)$, which can be decomposed into:

$$P\left(C_t^s = i|C_0^o = j\right) = \sum_{i=1}^{4} P\left(C_t^s = i|C_0^s = k\right) P\left(C_0^s = k|C_0^o = j\right) \qquad (1)$$

The first factor in this sum describes the true CD4 progression, whilst the second describes the mismatch between the initial observed and true CD4 counts.

6.1 Misclassification model

In our estimation procedure, we obtained a misclassification matrix $M = (m)_{i,j=1,\ldots,4}$ shown in Table 1 of the main text, where $m_{i,j}$ is the number of observations for which the observed CD4 category is i and the smoothed CD4 category is j. From this matrix, we can compute: $P\left(C_0^s = k | C_0^o = j\right) = \frac{m_{j,k}}{\sum_{s=1}^4 m_{j,s}}$. Note that the probabilities shown in Table 1 are different as they represent $P\left(C_0^o | C_0^s\right)$.

6.2 True CD4 progression model

Here, we use the fact that conditionally on remaining alive and untreated, the time to progression from category k to category $k+1$ is exponentially distributed with rate q_k. We denote F_k the cumulative density function of an exponentially distributed variable with rate q_k: $F_k(x) = 1 - e^{-q_k x}$ for all $x > 0$, and $F_k(x) = 0$ for all $x \leq 0$.

In our model the smoothed CD4 cannot increase so that:

$$P\left(C_t^s = l | C_0^s = k\right) = 0 \text{ if } k > l. \tag{2}$$

The probability of smoothed CD4 being in category k at time t given the initial smoothed CD4 category was k is given by $P\left(C_t^s = k | C_0^s = k\right) = 1 - F_k(t)$, that is:

$$P\left(C_t^s = k | C_0^s = k\right) = e^{-q_k t}. \tag{3}$$

One can show (by induction) that the probability of smoothed CD4 being in category $k+n$ ($n \geq 1$) at time t given the initial smoothed CD4 category was k is given by:

$$P\left(C_t^s = k+n | C_0^s = k\right) = \left[\prod_{l=0}^{n-1} q_{k+l}\right] \left[\sum_{l=0}^{n} \frac{e^{-q_{k+l}t}}{\prod_{j=0, j\neq l}^{n} \left(q_{k+j} - q_{k+l}\right)}\right]. \tag{4}$$

6.3 Full predictive model

Equations (1-4) fully determine the distribution of future smoothed CD4 categories, given the current observed CD4 category. Figure 4 in the main text presents the results of this full predictive model, for all patients as well as stratified by SPVL.

7 Relevance for pre-ART monitoring

Figure 4 underlines the influence of SPVL on a patient's future clinical needs, and therefore highlights the clinical value of viral load monitoring. Currently, WHO guidelines for patient monitoring in pre-ART care recommend a CD4 test every 6 to 12 months, and do not take SPVL into account. According to the predictive model presented above, after 6 (respectively 12) months 28% (respectively 38%) of individuals overall with observed CD4$>$ 500, who may not be eligible to ART according to national guidelines in some countries, would have already progressed to a true CD4$<$ 500. However 43% (resp. 59%) of those with SPVL$>$ 5.0 log10 copies/ml would have already progressed by then.

Motivated by matching the unstratified predictions, we examined the 28th and 38th percentiles of the SPVL-stratified predicted times from a given observed CD4 $>$ 500 to the true CD4 falling below 500. We found, as expected, that these fixed percentiles corresponded to decreasing amounts of times as SPVL increased. The 28th percentile corresponded to 6 months for the overall distribution, but to 13, 6, 3 and 2 months for individuals with SPVL$<$ 4.0, 4.0-4.5, 4.5-5.0 and $>$ 5.0 log10 copies/ml respectively. Similarly the 38th percentile corresponded to 12 months overall, but to 23, 12, 8 and 4 months for each of these four SPVL categories.

Hence, our results suggest that if SPVL is known for some patients, the frequency of CD4 monitoring could be adapted to account for that, with a CD4 test every 13-23, 6-12, 3-8 and 2-4 months for individuals with SPVL$<$ 4.0, 4.0-4.5, 4.5-5.0 and $>$ 5.0 log10 copies/ml respectively, corresponding to 28% to 38% of individuals in each of those categories having truly progressed to CD4$<$ 500 before they return for a CD4 test. These proposed delays are only indicative, and should be tailored to each patient in particular given how close their observed CD4 count is to the threshold 500.

8 Validation

In order to validate our method, we assessed its ability to predict future true CD4 dynamics given the current observed CD4 category.

8.1 Data used for validation

We used as a validation set the 514 patients with seroconversion window between 1 and 2 years, and at least 6 CD4 counts prior to HAART initiation.

The median age at seroconversion of these individuals was 34.9 years (interquartile range 28.3-42.4 years), very similar to the individuals included in the main analysis. 39 (7.6%) were female. By the end of the study period 40 patients had died, while only 3 were still HAART-naïve. The median initial CD4 count was slightly lower than the population included in the main analysis, at 510 (interquartile range 393-668) cells/mm^3 and a median of 10 CD4 measurements were taken over 2.9 years. 92.5% of the 201 patients with recorded subtype were infected with subtype B. SPVL was available for 434 (84%) patients and based on 1-13 measurements. The distribution of the SPVL was similar to that in the individuals included in the main analysis, with a median of 4.5 (interquartile range 4.1-5.0) log10 copies/ml.

8.2 Comparison of 'observed' and predicted smoothed CD4 transition times

For each patient, we used monotonic spline smoothing to determine the time course of the true (or smoothed) CD4 cell counts, as was done for the main analysis. From these, we computed the 'observed' times at which the smoothed CD4 cell counts decreased from one CD4 category to the next. We then used a non parametric Kaplan-Meier estimate to describe the 'observed' time from the initial observed CD4 count to these true CD4 category transitions. This allowed to account for right-censored 'observations'. Figure S6 shows the Kaplan-Meier survival curves together with the theoretical survival curves derived from the predictive model described in section 6. The 'observed' and predicted times of transition to true CD4 < 500 are in extremely good agreement. Although there were very few observed deaths amongst the patients selected for validation, their timing is also in good agreement with the predictive model. For the intermediate CD4 categories, the predicted times of transitions match the 'observations' very well up to about 5 years, after which the predictions seem to be more pessimistic than the observations. This could be due informative-censoring by HAART initiation in these intermediate categories, whereby the tail of the distribution is informed by only a few very slow progressors, all other patients having already initiated HAART earlier.

The 'observed' and predicted survival curves shown in Figure S6 use non-SPVL stratified estimates of CD4 progression. Similar figures stratified by SPVL were produced, but are not shown here. Although these confirmed that higher SPVL lead to faster progression, the number of patients was too small to perform meaningful comparisons between predicted and observed curves.

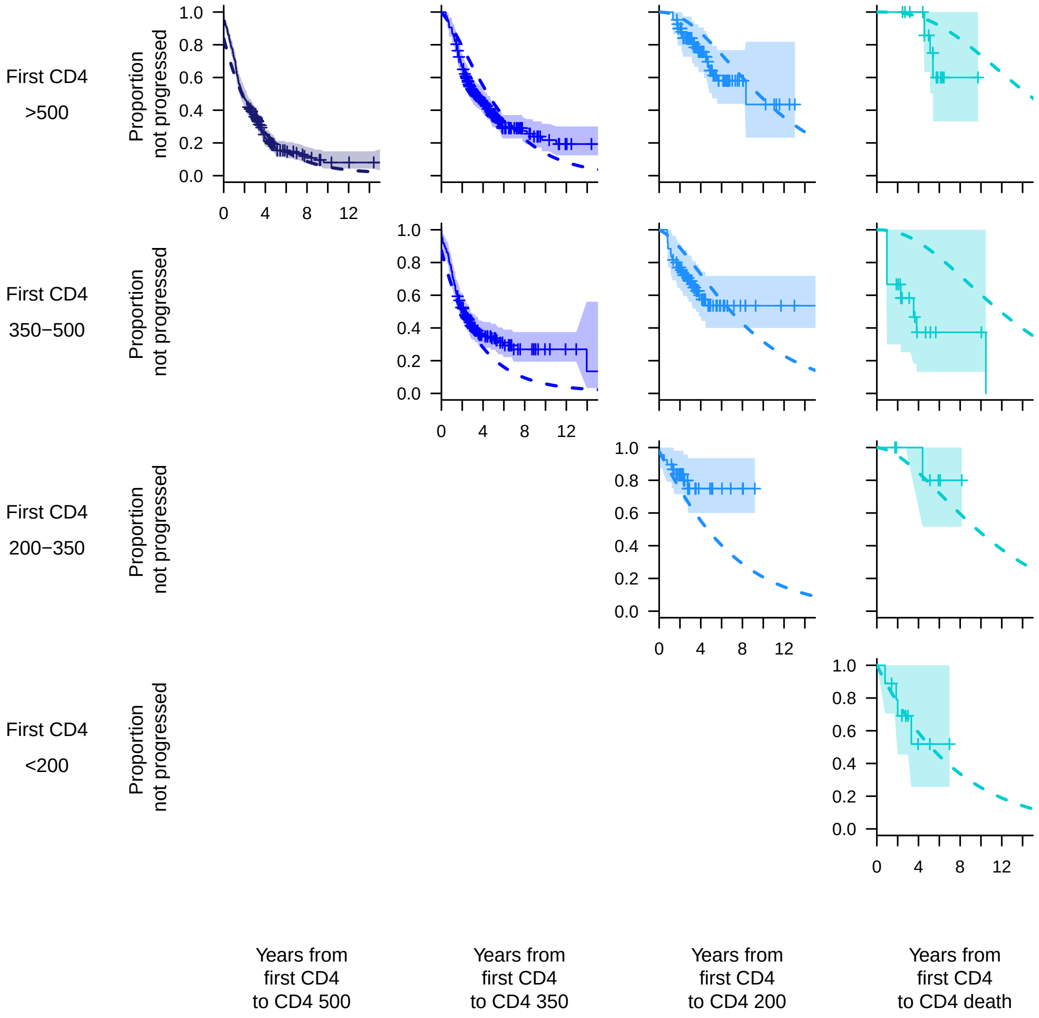

Figure S6: 'Observed' and predicted survival figures showing decline of true CD4 above the thresholds 500, 350, 200 cells/mm^3 and death (from left to right), given the first observed CD4 category (> 500, 350-500, 200-350 and ≤ 200 cells/mm^3 from top to bottom). Time is counted in years since first observed CD4 count. Crosses and shaded areas indicate non-parametric Kaplan-Meier survival estimates with 95% confidence intervals, obtained from the smoothed CD4 counts of the 514 patients included in the validation dataset. Dotted lines show the predicted survival curves according to the predicted model described in section 6, using non-SPVL stratified estimates of CD4 progression. The number of patients was too small to perform SPVL-stratified comparisons.