Appendix

Contents

1 Statistical analyses

1.1 Description of distributions

We assumed all time-to-event data to follow one of these three distributions: exponential, Weibull or double Weibull. A Weibull distribution has two parameters, which are conventionally noted *k* (shape parameter, *k* > 0) and *λ* (scale parameter, *λ* > 0). The probability distribution function (pdf) *f* of a Weibull distribution is of the form

$$f(t) = \left(\frac{k}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{k-1} e^{-\left(\frac{t}{\lambda}\right)^{k}},$$

where *t* is the time since a chosen origin, measured in years. From this equation, we can derive the cumulative distribution function (cdf) *F*

$$F(t) = 1 - e^{-\left(\frac{t}{\lambda}\right)^{k}}$$

and the hazard rate *h*

$$h(t) = \left(\frac{k}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{k-1}$$

The shape parameter *k* defines the behaviour of the hazard: with values of 0 < *k* < 1, the hazard rate decreases over time and with values *k* > 1 it increases. In the special case *k* = 1 the hazard rate remains constant over time, and the distribution reduces to an exponential distribution. Given a fixed shape parameter, the scale parameter is linearly related to the median survival time. In the case of the exponential distribution, the scale parameter reduces to the mean survival time, which is the inverse of the hazard rate. To avoid confusion, we use the notation where *λ* is the mean survival time, so that the parameter definitions for general

Weibull and exponential distributions would be equal. It should be noted that other parameterizations of both exponential and in general Weibull distributions are also widely used in literature.

Weibull distribution models monotone hazards only, i.e. the hazard is either always increasing or always decreasing over time. However, for events like mortality on ART, the hazard can be expected to be high in the initial months, decrease as the patient's situation improves, and eventually grow again. To take into account this non-monotonicity, a double Weibull distribution, which is a mixture of two Weibull distributions, was used. A double Weibull distribution has five parameters: the shape and scale parameters for both Weibull distributions and a weight parameter w, $0 \leq w \leq 1$, which is applied to the first component. A value of either 0 or 1 for the weight would reduce the double Weibull to an ordinary Weibull distribution. In our sensitivity analysis where we wanted to exclude high early mortality to represent the case where patients start with high CD4 counts, we therefore set w to zero and kept the other parameters unchanged.

1.2 Methods

We analysed data from Gugulethu and Khayelitsha ART programmes to estimate the hazard of virological failure, immunological failure and death.

1.2.1 Virological and immunological failure

We created continuous trajectories of HIV1-RNA viral load and CD4 count for each patient and used them to define the underlying time points of virological and immunological failure. These trajectories were interpolated linearly between the observed values of log10 viral load and CD4 count. Virological failure was defined as the time point when the viral load first crossed the threshold value of 1,000 copies/ml. Immunological failure was defined as the time point when the first of the three WHO criteria [1] were met for the first time, meaning that either (i) CD4 count crossed the value of 100 cells/μl, (ii) CD4 count crossed the baseline value, or (iii) CD4 decreased to less than 50% of the on-treatment maximum. For conditions (ii) and (iii), it was also required that the CD4 count was below 200 cells/μl. In addition, to meet a virological or immunological failure, it was required that the same criterion was observed in at least two consecutive measurements within one year. Those patients who met the criteria in their last appointment were censored from the analyses at the time of the previous appointment. Immunological failures were further divided into two categories, depending on if they occurred before (or without) a virological failure, or after a virological failure.

We used a Weibull model to estimate time to virological failure, and time to immunological failure before virological failure. We assumed that patients became at risk at 3 months after ART start, in line with the usual definitions requiring at least 6 months on first-line ART and at least 3 months from the first failure-defining appointment before switching. Time from virological failure to following immunological failure was estimated using an exponential model.

1.2.2 Mortality

The calculation of mortality consisted of several stages. First, we conducted a competing risk analysis to estimate time from ART start to death and loss to follow-up (LTFU). In this analysis, patients were censored at the time of virological failure or when switching to second-line therapy to model mortality and LTFU on successful first-line ART. Death was considered a competing event for LTFU and vice versa.

The second stage was to transform the obtained crude mortality and LTFU rates into a corrected estimation of HIV-related mortality. To do this, we used a function based on a review of different tracing studies [2], where mortality among patients lost to follow-up M_{lost} is calculated from the following formula:

$$M_{lost}(r) = \frac{e^{a+br}}{1+e^{a+br}},$$

where r is the proportion of patients lost and a = 0.57287 and b = -4.04409 are constants [3]. Mortality and the proportion of patients lost (r) must refer to the same time period: we used this definition continuously over time. Background mortality was estimated from the ASSA2008 estimates for Africans in Western Cape in 2007 [4] by fitting the age- and gender-specific mortality rates to a Weibull model for both genders separately. We calculated hazard functions for observed mortality (h_{obs}), mortality among LTFU (h_{lost}) and background mortality (h_{bg}). HIV related total mortality was then calculated as:

$$M_{HIV}(t) = 1 - e^{-\int_0^t h_{HIV}(\tau)d\tau}$$

with

$$h_{HIV}(t) = \big(1 - p(t)\big)h_{obs}(t) + p(t)h_{lost}(t) - h_{bg}(t),$$

where p is the proportion of patients lost to follow-up.

The HIV-related mortality function was then inserted into the Matlab Curve Fitting Tool. We specified an equation of the cumulative distribution function of a double Weibull distribution and fitted the data to gain the estimates for the parameters. To estimate the uncertainty of the parameters, we ran the entire process

10,000 times with crude mortality and LTFU rates sampled from a normal distribution around the observed rates.

The last part of the mortality estimation process was to estimate the hazard ratios related to having immunological or virological failure. We conducted a separate Cox regression analysis, where patients were also kept in the analysis after failure, and virological and immunological failures were included as time-dependent covariates. Patients were censored when they switched to second-line therapy. The results from these analyses produced hazard ratios for mortality associated with virological and immunological failure, which remained by definition constant over the time the patient spent on a failing first-line regimen. Since it could be expected that the hazard of mortality would increase over time, we examined estimates of this increase obtained from the literature [5]. We defined the hazard ratio for each three months on a failing regimen in a way that the cumulative hazard corresponds approximately to the constant hazard ratio of the Cox regression.

1.3 Patient characteristics

A total of 9,888 patients from Gugulethu and Khayelitsha were included in the analyses (Table S1). The median CD4 cell count at start of ART was 93 cells/µL, and median viral load 100,000 copies/ml. The two most common first-line regimens, used by 82.6% of patients, were lamivudine and stavudine combined with efavirenz or nevirapine (3TC/d4T/EFV or 3TC/d4T/NVP). These baseline characteristics correspond to those assumed for the hypothetical cohort of the mathematical model. A total of 7,564 patients had at least one viral load and one CD4 count measured after the start of ART and were therefore eligible for the analysis of virological and immunological treatment failure. During 15,885 person-years of follow-up, 304 (4.0%)

patients experienced immunological failure and 592 (7.8%) experienced virological failure. A total of 351 (4.6%) patients switched to second-line ART, of whom 272 (77.5%) had confirmed virological failure and 51 (14.5%) had immunological failure. The patients who switched without virological failure switched because of other reasons (e.g. after one elevated viral load value or after two detectable values below the official failure threshold of 1,000 copies/ml) or the reasons were not documented. Among patients with confirmed virological failure, the median time to switching was 21.8 months from the estimated time of failure and 13.9 months from the second (confirmative) viral load measurement. The most common second-line regimen was zidovudine, didanosine and ritonavir-boosted lopinavir (ZDV/ddI/LPV), which was used by 243 (69.2%) patients.

1.4 Viral load before and after starting antiretroviral therapy

In order to define the distribution of the viral load before ART (Figure S1a), we analyzed data from 5,556 patients in the cohorts who had at least one viral load measurement before ART. We calculated mean viral loads separately for 20 periods between 300 days before and 90 days after ART (Figure S1b). Based on these results, we assumed that viral load remains constant over one year before ART initiation. After starting ART we assumed a linear decrease in the log10 viral load value to 1 (10 copies/ml), except for patients with virological failure within the first six months of treatment, for whom the viral load was assumed to remain constant. For patients failing therapy later, we assumed that viral load started to increase approximately 3 months before the time of unobserved failure to a mean level of 10,000 copies/ml (Figure S1c). After switching to second-line therapy, viral load returned to undetectable levels within two months (Figure S1c).

2 Measures of transmission

2.1 Cohort viral load

Based on the individual viral load values, we calculated the cohort viral load (CVL) in the simulated population by adding up all individual mean viral load values over a particular year *y*:

$$CVL(y) = \sum_{i=1}^{N} \int_{t=y-1}^{y} vl_i(t)dt$$

where *N* is the total number of patients and $VL_i(t)$ the viral load of individual *I* at time *t*. CVL would be expected to be a key proximate determinant of the rate of HIV transmission from the treated population. We calculated CVL for 1,000 individuals during the first year before ART and the first five years on ART.

2.2 Number of transmissions

We used a relation of individual viral load and transmission probability introduced by Quinn *et al* [6] to estimate the number of new infections. The method calculates the probability of transmission *p* in a single sex act as

$$p(vl) = p(vl_0)k^{\log_{10}\frac{vl}{vl_0}} ,$$

where *vl* is the viral load on the absolute scale at the time of transmission, vl_0 a reference viral load for which the transmission probability is known, and *k* a constant. According to Wilson *et al* [7], we chose *k* = 2.45, vl_0 = 10 copies/ml and $p(vl_0)$ = $4.3*10^{-5}$ (male to female) or $2.2*10^{-5}$ (female to male). In our analyses, we assumed that each simulated patient has one partnership per year with an HIV negative person, with 100 unprotected sex acts. Therefore, the annual number of transmission can be written as

$$T(y) = \sum_{i=1}^{N}\left(1 - \prod_{j=0}^{99}\left(1 - p(vl_0)k^{\log_{10}\frac{vl_i\left(y-1+\frac{j}{100}\right)}{vl_0}}\right)\right)$$

where $vl_i(t)$ is the viral load of individual *i* at time *t* when $i \geq 1$, and vl_0 the reference viral load (10 copies/ml).

2.3 Additional remarks

The results concerning the new infections were based on the following assumptions: (i) all patients have 100 sexual contacts per year, (ii) all patients will change partners once a year, and (iii) the partners of all patients are HIV negative. Although these assumptions may overestimate the true level of HIV transmission in a setting where rates of sex acts, partner change and condom use differ substantially between individuals, it is clear that if we assume that the sex acts are evenly distributed over time and risk behaviour is independent of viral load, the relative difference between strategies is independent of these assumptions. Further, the absolute difference is approximately proportional to the number of sex acts as demonstrated below.

Let *p* be the prevalence of HIV in the community, $\beta_{ij}(k,S)$ the per-act transmission probability from individual *i* during the k^{th} act of the j^{th} partnership in a particular strategy *S*, n_i the number of partners of individual *i*, v_j the number of sex acts in partnership *j* and *N* the total number of individuals. Since at the origin $\boldsymbol{\beta} = \mathbf{0}$ the equations $\sum_i \beta_i = 1 - \prod_i(1-\beta_i)$ and $\frac{\partial}{\partial\beta_j}\sum_i \beta_i = \frac{\partial}{\partial\beta_j}(1 - \prod_i(1-\beta_i))$ hold for all *j*, the following approximation can be used whenever $\boldsymbol{\beta}$ is close to zero:

$$1 - \prod_{i} (1 - \beta_i) \approx \sum_{i} \beta_i$$

Then, for the absolute number of transmissions in a strategy, we can see that this can be approximated using the sum of all per-act transmission probabilities:

$$T_S = \sum_{i=1}^{N} \left(\sum_{j=1}^{n_i} (1 - p) \left(1 - \prod_{k=1}^{\nu_j} \left(1 - \beta_{ij}(k, S) \right) \right) \right) =$$

$$(1 - p) \sum_{i,j} \left(1 - \prod_{k=1}^{\nu_j} \left(1 - \beta_{ij}(k, S) \right) \right) \approx (1 - p) \sum_{i,j,k} \beta_{ij}(k, S)$$

If we write this with the mean per-act probability $\bar{\beta}(S)$, this can be further written as

$$T_S \approx (1 - p) M \bar{\beta}(S)$$

where *M* is the total number of sexual contacts. Since we have assumed that (i) the patients' risk behaviour is independent of viral load (and therefore infectiousness), and (ii) sex acts are evenly distributed across time, we can assume that the mean per-act probability would remain approximately the same in all risk behaviour scenarios. Therefore, the relative reduction between the strategies depends only on the mean per-act transmission probability:

$$\frac{T_{S_1} - T_{S_0}}{T_{S_0}} \approx \frac{(1 - p) M \bar{\beta}(S_1) - (1 - p) M \bar{\beta}(S_0)}{(1 - p) M \bar{\beta}(S_0)} = \frac{\bar{\beta}(S_1) - \bar{\beta}(S_0)}{\bar{\beta}(S_0)}$$

3 The mathematical model

3.1 Overview

The model consists of a set of separate scripts, used in a nested structure. The two main components are the individual patient simulation programme and the cohort summary programme. The cohort summary produces an array of outcomes for a pre-specified number of patients, by calling the patient simulation programme separately for each patient and storing its outputs in an array.

Several additional programmes are used: whenever a random time of an event has to be generated, the patient simulation calls an external random number generator, which generates the time of an event from the desired distribution, e.g. exponential, Weibull or double Weibull distribution. The output of the cohort summary programme is a table of the individual patients' outcomes, which can then be analysed with additional programmes in a similar way to analysing observational data.

Before calling the cohort summary, we already define a set of random uniform and standard normal deviates for each patient. The set of sampled parameters and random numbers are used as input for the cohort summary and define the characteristics of each patient. If we compare several strategies, we can therefore generate the same patients for each cohort by using the same input parameters. This will guarantee that all differences in outcomes are due to differences in strategies.

3.2 Patient simulation

The patient simulation is built in the following way: First, each patient is assigned an age and gender, and based on these, a time of natural death based on HIV-free age-

and gender-specific life expectancy is generated from a Weibull distribution. To define the survival time on ART, an HIV-related time of death will later be generated based on a specific hazard function. Second, the times of unobserved first-line virological and immunological failures are generated. Unobserved virological failure is defined as the time point when the viral load first exceeds the value of 1,000 copies/mL. Unobserved immunological failure refers to the time point when the CD4 count first fulfils the WHO immunological failure criteria [1]. Two independent immunological failure times are generated; one happening as a consequence of virological failure and one independently of the viral load. Only the first of these will lead to a failure that can be observed on first line, while the second will usually remain latent because the CD4 count would already be below the threshold levels at that stage. Depending on the monitoring strategy, either CD4 counts or viral loads are monitored according to a chosen monitoring frequency. No explicit values of CD4 or viral load are modelled at this stage.

Failures are observed and confirmed in the following way:

- In the CD4 monitoring strategy, the confirmation is three months from the first appointment after immunological failure, assuming that immunological failure is not observed already during the first appointment
- In routine viral load monitoring strategy, the confirmation is three months from the first appointment after virological failure

If immunological failure is observed already during the first appointment (CD4 monitoring strategy), this is defined as a ‘false failure’, which would not lead to a switch, but which makes it impossible to use immunological criteria for failure detection, since the CD4 count will remain under the failure threshold.

After observation of the failure, the patient switches to a second-line therapy. Patients with ‘false failure’ will be assigned a random time of switching, which represents the possibility that he could be switched e.g. after a severe opportunistic infection. Timing of second-line virological failure is defined the same way as in first-line failures but may also depend on the time the patient spent on a virologically failing first-line regimen. This is taken into consideration by an exponential factor: if t'_{2ndF} is the time of second-line failure defined directly from the Weibull distribution, the true time of second-line failure t_{2ndF} will be

$$t_{\text{2ndF}} = t_{\text{switch}} + 0.25 + (t'_{\text{2ndF}} - t_{\text{switch}})e^{-(t_{\text{switch}} - t_{\text{1stF}})},$$

where t_{switch} is the time of switching and t_{1stF} the time of first-line failure. If second-line virological failure happens within six months of switching, we interpret the failures as one single treatment failure and ignore the effect of switching.

4 Sensitivity and uncertainty analyses

4.1 Additional sensitivity analyses

In addition to the five sensitivity analyses described in the main text, we conducted two additional sensitivity analyses to examine the influence of the following assumptions on the results: the risk of second-line virological failure increases rapidly if switching is delayed (Analysis S1); and the hazard of mortality increases over time after virological failure (Analysis S2).

In analysis S1, we ran the model without the resistance penalty, i.e. with equal hazards of virological failure (after ART start) and second-line ART (after switching) virological failures. Average annual CVL was 2.58 (95% CI 2.06-3.10) million copies/ml with CD4 monitoring and 1.69 (1.41-1.98) million copies/ml with routine viral load monitoring. The average number of new infections per year was 6.26 (5.63-6.89) with CD4 monitoring and 4.29 (4.07-4.52) with routine viral load monitoring. The reduction in cohort viral load (33.6%) and in new HIV infections (31.2%) was similar to the main analysis (30.6% and 30.6%, respectively).

In analysis S2, we used a constant hazard ratio of 1.21 for mortality on virologically failing ART, instead of a hazard increasing with time. Average annual CVL was 2.58 (2.08-3.08) million copies/ml with CD4 monitoring and 1.71 (1.40-2.02) million copies/ml with routine viral load monitoring. The average number of new infections per year was 6.27 (5.62-6.92) and 4.33 (4.08-4.57) for CD4 and routine viral load monitoring, respectively. The mean reduction was again similar to the main analysis for both cohort viral load (32.8% versus 30.6% in main scenario) and new infections (30.8% versus 30.6%).

4.2 Uncertainty analyses

4.2.1 Methods

We conducted an uncertainty analysis, where the cohort was simulated 1,000 times for both strategies and the key parameters were sampled from distributions for each simulation separately using Latin Hypercube sampling. Parameters related to the following distributions were sampled: (i) 1st line virological failure (shape and scale), (ii) 2nd line virological failure (shape and scale), (iii) immunological failure related to virological failure (rate), (iv) immunological failure unrelated to virological failure (shape and scale), (v) HIV-related mortality without failures (shape and scale for both components, weight) and (vi) effect of virological and immunological failure on HIV-related mortality (hazard ratios). We assumed that all of these six distributions were independent of each other, i.e. the covariance between two parameters from two different distributions was zero. We calculated mean values of CVL and new transmissions over 5 years for both strategies as well as the relative reduction. Results were presented as mean values with 95% confidence intervals.

4.2.2 Results

In 1,000 simulations with CD4 monitoring, the average annual CVL (95% confidence interval) was 2.55 (2.02-3.08) million copies/ml and the average number of new infections per year was 6.23 (5.50-6.95). For routine viral load monitoring, the corresponding figures were 1.71 (1.40-2.02) million copies/ml and 4.32 (4.07-4.57) new infections. Both the mean values and variability of the results were consistent with the results of the analysis with fixed parameter values, presented in the main text.

Table S1. Characteristics of 9,888 patients starting antiretroviral combination therapy in Khayelitsha and Gugulethu, Cape Town, South Africa. The distribution of age, gender and baseline viral load are identical in the simulated cohort used to construct the mathematical model.

	Value
Age (years)	
Median (IQR)	33 (29-39)
Gender	
Male	3240 (32.8%)
Female	6648 (67.2%)
Cohort	
Gugulethu	2658 (26.9%)
Khayelitsha	7230 (73.1%)
CD4 cell count (cells/µl)	
Median (IQR)	93 (41-158)
HIV-1 viral load (log10 copies/ml)	
Median (IQR)	5.0 (4.5-5.5)
First-line regimens	
3TC d4T EFV	4985 (50.4%)
3TC d4T NVP	3182 (32.2%)
3TC ZDV NVP	1031 (10.4%)
3TC ZDV EFV	680 (6.9%)
d4T ddI EFV	7 (0.1%)
ZDV ddI EFV	3 (0.0%)

IQR, interquartile range; 3TC, lamivudine; d4T, stavudine; ZDV, zidovudine; ddI, didanosine; NVP, nevirapine; EFV, efavirenz.

Figure S1. Viral load values before starting antiretroviral therapy.

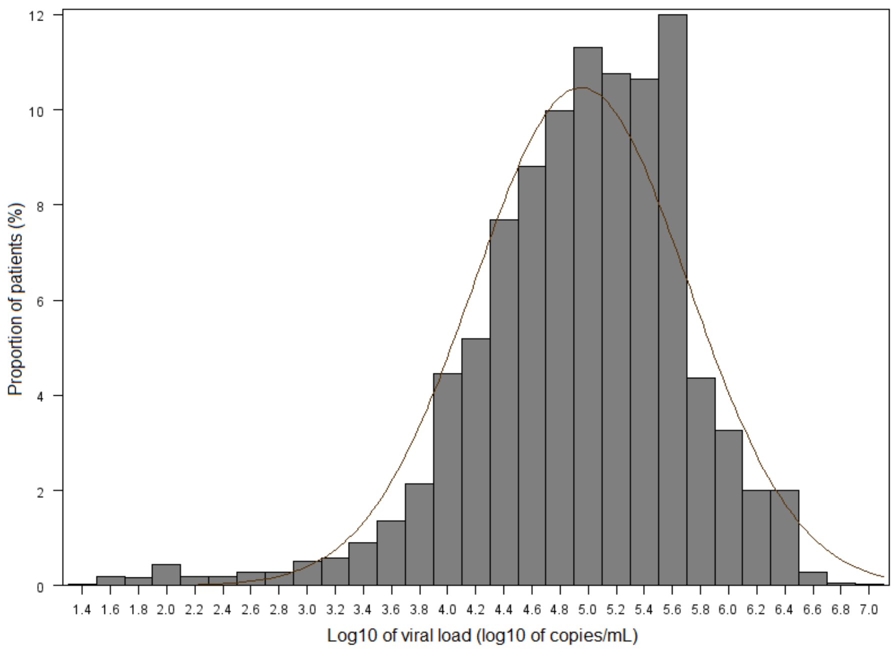
All patients from two sites in South Africa (Gugulethu, Khayelitsha) with available data were included. Viral load is shown as log10 of copies/ml.

Figure S1a: Distribution of viral load at baseline (the last available value before starting therapy), based on data from 5,556 patients. The solid line shows a normal fit.

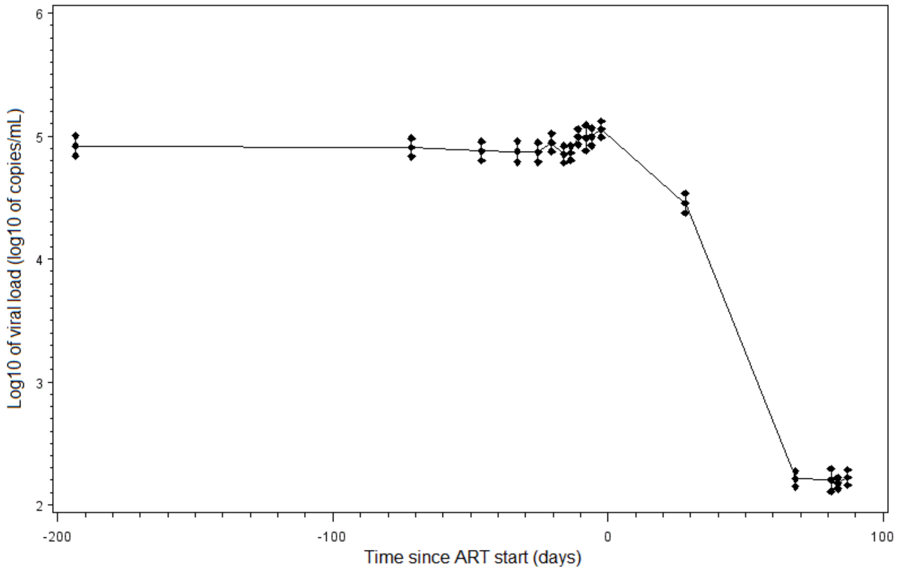
Figure S1b: Mean values of viral load of 5,813 patients over the time period from 300 days before to 90 days after ART start. The times of viral load measurements are divided into 20th quintiles, and within each of these time intervals, the mean viral load with 95% confidence intervals is shown.

Figure S1c: Mean viral load of 351 patients who start second-line ART, from 90 days after ART start until last follow-up visit, censoring or 90 days before death, calculated in a similar way as in Figure S1b.

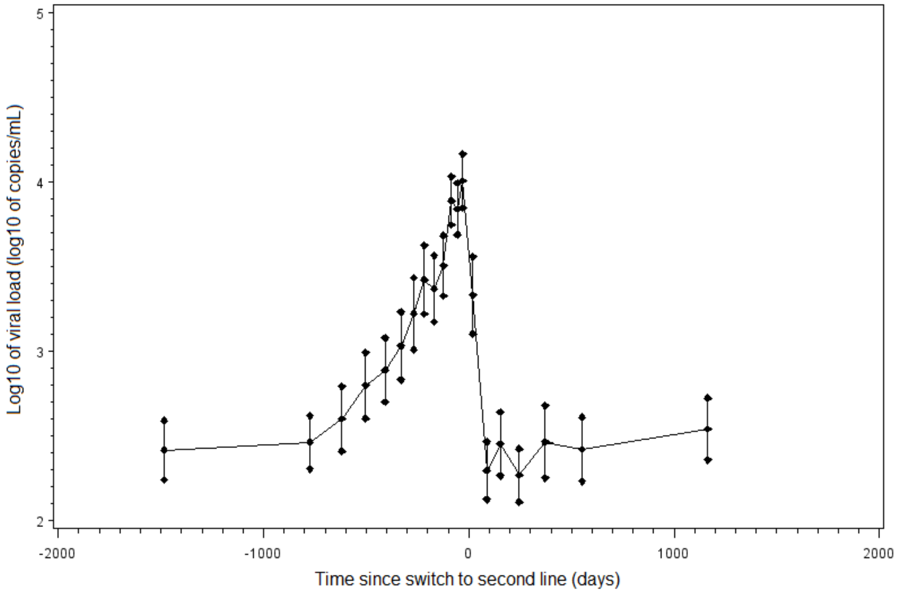
S1a



S1b


Log10 of viral load (log10 of copies/mL)
6
5
4
3
2
-200
-100
0
100
Time since ART start (days)

S1c


Log10 of viral load (log10 of copies/mL)
5
4
3
2
-2000
-1000
0
1000
2000
Time since switch to second line (days)

References

1. World Health Organization. Antiretroviral Therapy for HIV Infection in Adults and Adolescents: Recommendations for a public health approach. 2010 revision. 2010. Available at: http://whqlibdoc.who.int/publications/2010/9789241599764_eng.pdf. Accessed 17.02.2012.
2. Brinkhof MW, Pujades-Rodriguez M, Egger M. Mortality of patients lost to follow-up in antiretroviral treatment programmes in resource-limited settings: systematic review and meta-analysis. *PLoS One* 2009,**4**:e5790.
3. Egger M, Spycher BD, Sidle J, Weigel R, Geng E, Fox MP*, et al.* Correcting mortality for loss to follow up: A nomogram applied to antiretroviral treatment programmes in sub-Saharan Africa. . *PLoS Med* 2011,**8**:e1000390.
4. Actuarial society of South Africa. 2011. Available at http://www.actuarialsociety.org.za/. Accessed 24.9.2011.
5. Petersen ML, van der Laan MJ, Napravnik S, Eron JJ, Moore RD, Deeks SG. Long-term consequences of the delay between virologic failure of highly active antiretroviral therapy and regimen modification. *AIDS* 2008,**22**:2097-2106.
6. Quinn TC, Wawer MJ, Sewankambo N, Serwadda D, Li C, Wabwire-Mangen F*, et al.* Viral Load and Heterosexual Transmission of Human Immunodeficiency Virus Type 1. *N Engl J Med* 2000,**342**:921-929.
7. Wilson DP, Law MG, Grulich AE, Cooper DA, Kaldor JM. Relation between HIV viral load and infectiousness: a model-based analysis. *Lancet* 2008,**372**:314-320.