STATISTICAL APPENDIX

MODELS

We present here some mathematical details of the model used in the adjusted analyses. The general exponential regression model of survival is a proportional hazards model of the hazard (the instantaneous risk of death). The general proportional hazards model is as follows:

$$\lambda_i(t) = \lambda_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

where the vector $\boldsymbol{x}$ contains the measurements of interest (in our case, gender, CD4 count and age category at ART initiation), and $\lambda_0(t)$ is the "baseline" hazard (the hazard of a person with all risk factors set at their reference values). The exponential survival model has the added simplification that $\lambda_0(t) = \lambda_0$ for all time points t (constant hazards) so that the exponential model for the hazard is

$$\lambda_i(t) = \lambda_0\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

Since the constant-hazard assumption is rather restrictive, we fit a *piecewise* exponential model where $\lambda_0(t) = \lambda_{0j}$ for t in the time interval $[\tau_{j-1}, \tau_j)$ with $\tau_0 = 0, \tau_1 = 6, \tau_2 = 12, \tau_3 = 24, \tau_4 = \infty$ and $j = 1,2,3,4$. Put more simply, we assumed that the baseline hazard of death is constant in the four time intervals (in months from ART start) $[0,6), [6,12), [12,24)$ and ≥ 24 months.

As mentioned in the methods section in the body of this document, we fit two separate models, one for the time $[0,6)$ (ART initiation to six months) and one for the three subsequent time intervals (i.e., after 6 months). This was because neither the hazard during the first six months after the start of therapy, nor the associations of the predictive factors (gender, CD4 count and age) are expected to follow the same pattern with subsequent periods (see Yiannoutsos et al., for a more mathematical treatise of change points in hazards during the period after initiation of therapy[1]). We will come back to this later on. For now, we keep considering the simpler (single) Exponential model as stated above.

Clearly, an adjustment needs to be made to account for the fact that a large proportion of the patient cohort, a subgroup including some patients with very adverse prognosis, has been lost from observation. This is because we cannot consider the subgroup of patients who remain on observation as representative of the dropouts, since patients who are lost have been observed to have much higher mortality hazard[2-4]. To do this we consider the vital status ascertained on a subset of the lost patients (which we consider a random sample of all patient dropouts) and use this information to update the hazard for all dropouts. In the case of the AMPATH cohort, the vital status of this subset was located through tracing of patients who missed scheduled visits (see Yiannoutsos et al., for a description of this program[4]). For South African cohorts, the vital status of lost patients was established through linkages to the national death registry[5]. In both cases, the subset of patients with sufficient civilian information to be successfully located was assumed to be representative (i.e., to form a random sample) of the entire lost to follow-up cohort.

Following the approach by Frangakis and Rubin[6] and An et al.[2], we generate an average of the mortality hazard weighted by the inverse proportion of the number of patients who drop out (numerator) versus the number who were located (the random sample; denominator). This has the effect of multiplying each of the located patients by a number which is larger than one, thus creating virtual "copies" of these individuals to replace those whose vital status was not ascertained (as the traced patients are considered to be representative for the individuals who are lost and not traced). We replace the remainder of the lost patients who were not located by these copies, while the data from the actual individuals (who dropped out but did not have their vital status ascertained) are weighted by zero (effectively being excluded from the analysis). Patients who remained on observation are weighted by one (i.e., they only represent themselves). Their weighted hazard is thus

$$\lambda(t) = \sum_{g=0}^{1} w_g \lambda_g(t)$$

with g representing dropouts $(g = 1)$ and non-dropouts ($g = 0$) and

$$w_g = \begin{cases} \frac{n_0}{\tilde{n}_0} & \text{if dropout and traced} \\ 1 & \text{if non} - \text{dropout} \end{cases}$$

where n_0 is the total number of dropouts and $\tilde{n}_0$ is the subset of dropouts who were traced. Transferring the same idea to our piecewise exponential regression model, we have

$$\lambda_i(t) = \sum_{g=0}^{1} w_g \lambda_{0g}(t) \exp(\boldsymbol{x}_i' \boldsymbol{\beta}) = \left\{ \sum_{g=0}^{1} \sum_{j=1}^{4} w_g \lambda_{0jg} I[t \in [\tau_{j-1}, \tau_j)] \right\} \exp(\boldsymbol{x}_i' \boldsymbol{\beta})$$

for $j = 1,2,3,4$ and $I[\cdot]$ is an indicator function.

Now, returning to the fact that a separate model was fit for the time interval $[\tau_{j-1}, \tau_j) = [0,6)$ and the last three time intervals $[\tau_{j-1}, \tau_j)$ for $j = 2,3,4$ i.e., $[6,12)$, $[12, 24)$ and $[24, \infty)$, we obtain the final model used for the analyses presented in this paper

$$\lambda_i(t) = \sum_{g=0}^{1} \left\{ w_g \lambda_{01g} I[t \in [\tau_0, \tau_1)] \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_1) + \sum_{j=2}^{4} w_g \lambda_{0jg} I[t \in [\tau_{j-1}, \tau_j)] \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_2) \right\}$$

Note that two different regression coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are calculated in these two models, corresponding to different association between the risk factors before and after six months from ART initiation respectively.

The two-part piece-wise exponential model described above was fit by the equivalent Poisson log-linear model on the constant hazards[7,8]. The analysis was implemented by the STATA software version 11.1 (StataCorp, College Station, TX, USA). The program code is given in the following Appendix.

STATA ANALYSIS

Description of the database

The following code produced the analysis presented in this white paper. It assumes a data set where the following minimum number of variables are present:

Variable name	Type	Format	Description
`ptidno`	long	8.0	Patient ID number
`dob`	date		Date of birth
`dod`	date		Date of death
`lastvisitdt`	date		Date of last visit
`death`	byte	8.0	Death indicator
`ageatarvstart`	float	8.0	Age at the start of ART
`cd4`	Integer	8.0	CD4 count (cells/µL)
`ltfu`	byte	8.0	Lost to follow-up indicator
`arvstartdt`	date		Date of ART initiation
`oraftervis`	byte	8.0	Patient outreach indicator

Here are some comments before we proceed to the code:

1. `oraftervis` is equal to 1 (“yes”) if the dropout patient was located by outreach. If not, then `oraftervis`=0 and we assume that no attempt was made (i.e., we equate patients who were not outreached with those who were outreached but not found). This is an important limitation of this study. A similar set of assumptions is made with respect to patients without sufficient information for linkage with a vital registry (they are considered as not having been found).
2. The last visit date `lastvisitdt` includes, for patients who were located via outreach, the last date known to be alive (so the last “visit” for these patients is their last contact with site outreach staff). In linkages with the vital registry, this date is the most recent date of inquiry.
3. Different lost-to-follow-up definitions produce slightly different results, as different numbers of patients are declared lost. This is not expected to be of major concern.
4. This study also assumed that all death dates (`dod`) were exact. However, some death dates were estimated (mainly due to vital status information obtained by proxy). It is tacitly assumed

that these estimates do not include a systematic error (i.e., they have random variations around the unknown true death date)

5. The indicator `death` includes all deaths, both those ascertained through passive/routine means and those ascertained through active means (i.e., patient outreach).

STATA analysis code

```
* Ensure that last visit date is updated for date of death

gen new_lastvisitdt=max(dod,lastvisitdt) if dod ~=.
replace new_lastvisitdt=lastvisitdt if dod==. & death==0

preserve

* split for duration: origin and entry are first ART visit
stset new_lastvisitdt , fail(death==1) scale(365.25) id(ptidno) /*
*/ enter(arvstartdt) origin(arvstartdt)

stsplit durcat , at(0.5 1 2)

* split for age: origin is dob, entry is arvstartdt

stset new_lastvisitdt , fail(death==1) scale(365.25) id(ptidno) /*
*/      enter(arvstartdt) origin(dob)

stsplit new_agecat , at(15 25 35 45)

* * * * * * * * * * * * * * Weighted analysis * * * * * * * * * * * * *
* Create the weights

gen weight=1

* Count deaths among LTFU or outreached patients
count if (oraftervis ==0 & ltfu ==1 & death~=1)| (oraftervis ==1)

* Create weights for outreached (i.e., located) patients only

replace weight =r(N) if (oraftervis ==1)
count if oraftervis ==1
replace weight =weight/r(N) if oraftervis ==1

* Exclude dropouts who were not outreached from the analysis

drop if oraftervis ==0 & ltfu ==1 & death~=1


* The following code is a check that weights fill in excluded patients

sum weight
di r(mean)*r(N)

* The result of the above calculation must equal to the original N!

* reset to have origin at arvstart with pweight=weight

stset new_lastvisitdt [pweight=weight], id(ptidno) failure(death==1) /*
*/ enter(time arvstart)  origin(time arvstart) scale(365.25)

* reset to have origin at arvstartdt
```

```
stset new_lastvisitdt , fail(death==1) scale(365.25) id(ptidno) /*
*/ enter(arvstartdt) origin(arvstartdt)

* Change time scale per 100 person-years

gen pyo100=(_t-_t0)/100
```

```
**************************** Major point ***************************
* Note that, when intervals get split, STATA does not update the
* death indicator correctly.  The death indicator must be zero in
* all but the final interval and equal to the original death indicator
* in the last interval.
* Thus, we need to use the STATA internal event indicator _d in the
* calculations instead of the original death indicator because
* (correctly) _d=0 for all intervals prior to the last one and _d=death
* at the last interval.
* Thus, an analysis involving the death variable would be wrong!
**********************************************************************

* Poisson models

* Model for the first 6 months since ART initiation
xi: poisson _d i.male i.new_agecat i.cd4cat if(pyo100>0 & durcat==0), /*
*/  exposure(pyo100)

* Model for after the first six months since ART initiation
xi: poisson _d i.male i.new_agecat i.cd4cat i.durcat  /*
*/  if(pyo100>0 & durcat>0), exposure(pyo100)

* * * * * * * * * * * End of weighted analyses * * * * * * * * * * * * *
```

References

1. Yiannoutsos CT. Modeling AIDS survival after initiation of antiretroviral treatment by Weibull models with changepoints. *J Int AIDS Soc* 2009;12(1):9.
2. An MW, Frangakis CE, Musick BS, Yiannoutsos CT. The need for double-sampling designs in survival studies: an application to monitor PEPFAR. *Biometrics* 2009;65(1):301-6.
3. Egger M, Spycher BD, Sidle J, Weigel R, Geng EH, 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(1):e1000390.
4. Yiannoutsos CT, An MW, Frangakis CE, Musick BS, Braitstein P, Wools-Kaloustian K, et al. Sampling-based approaches to improve estimation of mortality among patient dropouts: experience from a large PEPFAR-funded program in Western Kenya. *PLoS One* 2008;3(12):e3843.
5. Van Cutsem G, Ford N, Hildebrand K, Goemaere E, Mathee S, Abrahams M, et al. Correcting for Mortality Among Patients Lost to Follow Up on Antiretroviral Therapy in South Africa: A Cohort Analysis. *PLoS One* 2011;6(2).
6. Frangakis CE, Rubin DB. Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. *Biometrics* 2001;57(2):333-42.
7. Holford TR. The analysis of rates and of survivorship using log-linear models. *Biometrics* 1980;36(2):299-305.
8. Laird N, Olivier D. Covariance analysis of censored survival data using log-linear analysis techniques. *J Am Stat Assoc* 1981;76:231-40.