Appendix C: Modelling

Analysis model

Let $Y_{m+k,i}$ be the indicator for PCP diagnosis for subject i at the end of month m, $C_{m,i}$ (1: censored, 0 = uncensored) be the indicator for censoring at the end of month m for subject i, X_i is the regime assignment for patient i, and $L_{m,i}$ are time fixed ($m = 0$) and time varying ($m = 1, \ldots, M$) covariates at the end of $m$for subject i. In the following, over bars are used to denote histories up to and including the month defined by the subscript m.

For the analysis, we fit an inverse probability weighted, pooled logistic model to estimate the hazard ratio over all months of follow-up:

$$\begin{aligned} &\text{logit}\left[Pr(Y_{m+1,i} = 1 | Y_{m,i} = 0, C_{m,i} = 0, X_i, V_i)\right] = \\ &\qquad \beta_{0,m} + \beta_1 X + \beta_2 X t_m + \beta_3^T V, \end{aligned} \tag{1}$$

where
$\beta_{0,m}$ is the baseline hazard function which includes terms for $time$, $time^2$ and $time^3$,
β_1 is the estimated log hazard ratio for the prophylaxis regime,
β_2 is the estimate for the interaction term between time in months (t_m) and the treatment regime (to allow for non-proportional hazards), and,
β_3 is a vector of estimated log hazard ratios for baseline covariates.

To compensate for patients being involved in both arms of the trial due to duplication, we calculated robust sandwich errors to account for the intra-patient correlation.

Since we are using a parametric analysis model, we were able to estimate not only the primary endpoint, the hazard ratio, but also secondary endpoints such as the 5 year risk difference between regimes. For the latter, we used nonparametric bootstrapping with 500 samples to estimate 95% confidence intervals for the difference. We log transformed the point estimates for the 5 year survival probability on each regime prior to estimating the absolute risk difference to adjust for the asymmetric distribution of the risk estimates.

The following code implements the analysis model in R:

```
amod1 <- svyglm(EVENT~ # pcp diagnosis or death
                as.factor(rep) # regime 1 or 2
                +time
                +I(time^2)
                +I(time^3)
                # interaction between time and regime to
                # allow non-PH
                +as.factor(rep):time
                +as.factor(rep):I(time^2)
                +as.factor(rep):I(time^3)
              # time fixed covariates
                +bage
                +I(bage^2)
                +factor(gender)
                +factor(mode2) # mode of transmission
                +factor(origin3) # geographical origin
                +factor(cohort)
                +sbcd4
                +I(sbcd4^2)
                +log10brna
                +I(log10brna^2)
                +YRbase # calendar year at time0
                +as.factor(fupind) # indicator for death and drop-out
                +pc_timeoncart # % time on cART
                , family = quasibinomial()
                , design = svydesign(id = ~patient
                                     , weights = ~sw.trunc # truncated weights
                                     , data = aset))
```

Inverse probability weights

We introduced artificial censoring when patients did not follow their randomized regime, and subsequently, have to compensate for the potential selection bias this may have introduced. Making the standard assumption of no unmeasured confounding, we can eliminate this potential bias by introducing inverse probability weights (IPWs). We calculate a weight for each patient-month in the expanded (monthly) data set, which is inversely proportional to the conditional probability of the patient remaining uncensored until the end of the specific month.

We calculate the weights by fitting a logistic model with the censoring indicator as dependent variable, and independent variables which include the prophylaxis regime, along with baseline and time varying covariates. To ensure these weights provide reliable estimates, we stabilize the weights and then truncate them at the 99% point to avoid large values.

Let $C_{k,i}$ (1: censored, 0 = uncensored) be the indicator for artifificial censoring at the end of the mth month for subject i, A_i is the treatment history (on and off prophylaxis) for patient i, X_i is the assigned treatment regime for this patient ($i = 1, 2$), V_i are time fixed (baseline) covariates for subject i, and $L_{k,i}$ are time varying covariates at the end of month m for subject i. Over bars are used to denote histories up to and including month m.
The stabilised weights for all types of censoring are defined as:

$$\mathrm{SW}^{C}_{m,i} = \prod_{k=m}^{M} \frac{Pr(C_{m,i} = 0 | C_{m-1,i} = 0, Y_m = 0, X_i, V_i)}{Pr(C_{m,i} = 0 | C_{m-1,i} = 0, Y_m = 0, X_i, \bar{A}_{m-1,i}, \bar{L}_{m-1,i})},$$

The denominator is, informally, the subject's probability of remaining uncensored through period m given baseline and time varying confounders. The probability of being uncensored through month m is estimated by fitting a pooled logistic model (see example code below):

$$\text{logit}\left[Pr(C_{m,i} = 0 | C_{m-1,i} = 0, Y_m = 0, X_i, \bar{A}_{m-1,i}, \bar{L}_{m-1,i})\right] =$$

$$\psi_{0,m} + \psi_1 X + \psi_2 \bar{A}_{m-1} + \psi_3 \bar{L}_{m-1},$$

where
ψ_0 is an intercept term,
ψ_1 estimates the odds ratio comparing the regimes, and
ψ_2 is a vector of estimates for the treatment history (i.e. PcP prophylaxis) up to time $m - 1$.
ψ_3 is a vector of estimates for the covariate history up to time $m - 1$.

The numerator being defined similarly, but without including time varying covariates:

$$\text{logit}\left[Pr(C_{m,i} = 0 | C_{m-1,i} = 0, Y_m = 0, X_i, V_i)\right] = \phi_0 + \phi_1 X + \phi_2 V,$$

with estimated defined analogously to those defined as above, but this time including baseline covariates only. The numerator stabilises the weights to reduce the variance of the estimates in the final model.

By fitting the pooled logistic model including these per person-month weights, we create a pseudo-population in which artificial censoring has effectively been eliminated. This establishes the rationale for the analysis providing results that provide statistically valid inference, albeit with the assumption that there is no unmeasured confounding

The following code implements the IP weights in R:

```
aset[,cens_any:=any(cens_ind), by="patient"]

#-------------------------------------------------
# define the censoring weights
#-------------------------------------------------

# denominator weights

# dependent variable is probability of not being censored
# i.e. 1-indicator variable for being censored
aset$notcensor <- 1- aset$cens_ind

# denominator of IPWeights

mod <- glm(notcensor ~ as.factor(pcp_prophyl1) # A_t
           +time # time in months for baseline hazard, perhaps later with a spline
           +I(time^2)
           +I(time^3)
           # baseline covariates V
           +bage
           +I(bage^2)
           +factor(gender)
           +factor(mode2)
           +factor(origin3)
           +YRbase
           +factor(cohort)
           +sbcd4
           +I(sbcd4^2)
           +log10brna
           +I(log10brna^2)
                #+factor(cohort)
           # time varying covariates L_t
           +age
           +I(age^2)
           +scd4
           +I(scd4^2)
           +log10rna
           +I(log10rna^2)
           # the regime X=x
           +rep
           # other censoring indicators
           +fupind # added in denom as its a time varying parameter
           +pc_timeoncart
           ,family = binomial()
           ,data = aset[(cens_any==T & time!=0)]) # only those patients which have
                                                  # been censored at some point
                                                  # excluding time 0.
```

```
summary(mod)

probC.d <- predict(mod, type = 'response');length(probC.d)

aset$probC.d<-(-1)
aset$probC.d[which(aset$cens_any==T & aset$time!=0)]<-probC.d

# correct and extend weights
# those not censored have weight 1
aset$probC.d[which(aset$cens_any==F)]<-1
# those at time 0 have missing weight
aset$probC.d[which(aset$time==0)]<-NA

summary(aset$probC.d)

#------------------------------------------------------------------------------
# numerator of IPWeights

# no time dependent covariates, time, L_t and A_t

# add baseline age, cd4, rna
aset[,sbcd4:=sqrt(cd4[1]), by = "patient"]
aset$log10rna<-aset$rna
aset$log10rna[which(aset$rna==0)]<-0.1
aset[,log10brna:=log10(log10rna[1]), by = "patient"]
aset[,bage:=age[1], by="patient"]

mod <- glm(notcensor ~
           # baseline covariates V
           +factor(gender)
           +factor(mode2)
           +factor(origin3)
           +YRbase
           +factor(cohort)
           +sbcd4
           +I(sbcd4^2)
           +log10brna
           +I(log10brna^2)
           +bage
           +I(bage^2)
           +rep # X=x the regime
           +pc_timeoncart
           ,family = binomial()
           , data = aset)

summary(mod)

aset$probC.n <- predict(mod, type = 'response');length(aset$probC.n)
```

```
# correct and extend weights
# those not censored have weight 1
aset$probC.d[which(aset$cens_any==F)]<-1
# those at time 0 have missing weight
aset$probC.d[which(aset$time==0)]<-NA

summary(aset$probC.n)

# products

aset$C.numcum <- ave(aset$probC.n,aset$patient,
                     FUN=function(x) cumprod(x))
summary(aset$C.numcum)

# correct those with time 0
aset$probC.d[which(is.na(aset$probC.d))]<-1 # set 1 as its a cumulative product

aset$C.dencum <- ave(aset$probC.d,aset$patient,
                     FUN=function(x) cumprod(x))
summary(aset$C.dencum)

aset$swC <- aset$C.numcum/aset$C.dencum
summary(aset$swC);hist(aset$swC, col="lightblue", breaks=50)

#-------------------------------------------------------------------------------------------
# Truncate weights at 99% for stability

trunc.cutoff <- quantile(aset$swC,0.99,na.rm=TRUE)
aset$sw.trunc <- ifelse(aset$swC<trunc.cutoff, aset$swC,trunc.cutoff)
summary(aset$sw.trunc);hist(aset$sw.trunc, col="lightblue", breaks=50)
```

Subgroup analysis: Grace periods for the stopping regime

In the definition of the two stopping strategies, we defined a strict cut off time at which patients should stop their prophylaxis, either based on their CD4 count or confirmed viral suppression. In a further step, we allowed patients to stop prophylaxis within m months following the stopping criteria being met for the respective regime. For example, a patient with confirmed viral suppression at time point x, would still be consistent with this regime even if they were still taking prophylaxis $(x + m)$ months later; however, they would be artificially censored if they did not stop taking prophylaxis at month $(x + m + 1)$. This means that patients can have multiple periods of being on and off prophylaxis, and, at least within the m months, are allowed to be non-compliant with their regime.

We chose m=3 months for the primary analysis in the main document, and then varied the value of the non-compliance window to determine the sensitivity of the results in subgroup analyses.

With no grace period the hazard ratio increased to be marginally significant at the 5% level (0.6 [0.3, 1.0], p=0.04). With a longer 6 month grace period, the HR remained the same as in the primary analysis with 3 months, but due to the increased number of patients and PcP diagnoses include in the analysis the estimate of the hazard ratio became more precise (HR 0.8 [0.7, 1.0], p=0.05).

Using grace periods mirrors realistic clinical practice in which the decision to stop prophylaxis may be delayed, either by patient or physician. The longer 6 month grace period analysis served to confirm our initial findings, whilst the no grace period analysis, albeit less clinically realistic, highlighted a marginally lower risk from using confirmed viral suppression as stopping criteria.