Published in final edited form as:
Biometrics. 2014 June ; 70(2): 441–448. doi:10.1111/biom.12140.

Viral Suppression in HIV Studies: Combining Times to Suppression and Rebound

Natalia A. Gouskova[1,*], **Stephen R. Cole**[2], **Joseph J. Eron**[3], and **Jason P. Fine**[1]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill Chapel Hill, North Carolina 27599, U.S.A

[2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

[3]Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

Summary

In HIV-1 clinical trials the interest is often to compare how well treatments suppress the HIV-1 RNA viral load. The current practice in statistical analysis of such trials is to define a single ad hoc composite event which combines information about both the viral load suppression and the subsequent viral rebound, and then analyze the data using standard univariate survival analysis techniques. The main weakness of this approach is that the results of the analysis can be easily influenced by minor details in the definition of the composite event. We propose a straightforward alternative endpoint based on the probability of being suppressed over time, and suggest that treatment differences be summarized using the restricted mean time a patient spends in the state of viral suppression. A nonparametric analysis is based on methods for multiple endpoint studies. We demonstrate the utility of our analytic strategy using a recent therapeutic trial, in which the protocol specified a primary analysis using a composite endpoint approach.

Keywords

AIDS; Clinical trial endpoint; Counting processes; Multistate models; Survival analysis

1. Introduction

A well-defined outcome is fundamental to the analysis of time to event data. However, in some settings a clear definition of the event of interest is a challenge. An example of such settings include clinical trials evaluating the difference between treatments which are intended to suppress the level of HIV-1 RNA viral load (henceforth viral load) in people infected with HIV (DeGruttola et al., 1998; Gilbert et al., 2000; Ribaudo et al., 2006).

[*]gouskova@unc.edu.

Supplementary Materials
A SAS macro implementing the proposed method and an example dataset are available with this paper at the *Biometrics* website on Wiley Online Library.

Infection with HIV is monitored by the number of copies of viral load present in circulating plasma (Mellors et al., 1996). HIV research relies heavily on viral load levels for evaluating the comparative efficacy and effectiveness of competing therapy regimens, and estimating the prognosis of HIV-infected individuals (Egger et al., 2002; Cole et al., 2007; Riddler et al., 2008). It is a desirable quality that a treatment regimen would suppress the viral load below a clinically relevant level (200 copies/ml is often used in practice) and keep the viral load suppressed. One way to quantify viral load suppression is to assess the average time between two events, the viral load suppression below a threshold level, and the viral load rebound above the threshold.

Traditionally, to analyze such data HIV researchers have created a single composite time-to-event endpoint, often called “virologic failure.” The time of virologic failure is defined as the time of rebound, given that a patient’s viral load has suppressed by some clinically relevant cut-off time point, for example, 16 weeks since the beginning of treatment. If the patient’s viral load did not suppress by 16 weeks, then the time of virologic failure is set equal to 16 weeks. Examples of such an endoint can be found in ACTG A5095 trial (Gulick et al., 2004, 2006). Numerous variations of this definition are used in practice, with varying values of the cut-off time point (Robbins et al., 2003; Fischl et al., 2003; Riddler et al., 2008). More complicated definitions of virologic failure may employ multiple criteria for viral load suppression and rebound, such as in ACTG A5142 and A5202 trials (Riddler et al., 2008; Sax et al., 2009). A related endpoint is the Federal Drug Administration’s time to loss of virologic response (TLOVR) (Guidance for Industry, 2002), where patients not suppressing by the cut-off time point are assigned zero as their time of virologic failure. Although such definitions may differ substantially, often being tailored to particular trials, they share two common features. Firstly, the information contained in two distinct events, viral suppression and viral rebound, is collapsed into a single composite event. Secondly, for patients who do not suppress by some chosen cut-off time, the time of virologic failure is assigned to a prespecified time, for example, the cut-off time. The differences between the above definitions involve different choices of cut-off timepoints and different criteria for suppression and rebound.

While such composite endpoints facilitate the application of standard methodology for right censored time to event data in an intent to treat analysis, there are practical concerns which arise from such endpoint definition. The definitions are complicated, and while understandable to most clinicians, may not generalize readily across trials and populations. For the patients who did not suppress their viral load by the cut-off time, the event time is redefined, which can have a notable impact on results, as evidenced in the simulation studies in Section 3. Finally, the early dynamics of suppression may be obscured using such composite endpoints.

To avoid the above mentioned problems, we suggest using a different endpoint and different analysis methods based on multistate models (Pepe, 1991). These methods explicitly acknowledge the fact that we have two distinct events, viral load suppression and rebound, with corresponding survival functions $S^S(t)$ and $S^R(t)$, respectively. Our proposed endpoint is based on the probability of being in suppression $G(t)$, which is simply $S^R(t) - S^S(t)$. We suggest an intuitive summary of treatment efficacy based on a weighted integral of this difference over a specified time interval of interest, say 1 year. With equal weights over time, this measure reduces to the restricted mean time suppressed over the time interval. One may tailor the weights to emphasize the timepoints of scientific interest, enabling a rigorous exploration of either early or late suppression dynamics. This endpoint is well-defined and has a clear and simple interpretation which may permit comparisons across trials and populations. The proposed analysis accounts for the fact that a proportion of patients will never suppress their viral load and allows investigators to simultaneously assess differences in both time to viral suppression and time to viral rebound, emphasizing those timepoints relevant to treatment evaluation. A simulation study assessing performance of the proposed endpoint in comparison to endpoints based on composite events is discussed in Section 3. The practical utility of the analysis is illustrated in a reanalysis of ACTG A5142 in Section 4. A discussion concludes in Section 5.

2. Methods

For patient *i*, let R_i be the treatment regimen assignment at time of randomization, with the focus being an intent to treat analysis of treatment efficacy. The potential time at which patient *i* has their viral load suppressed is denoted by T_i^S and the potential time at which patient i has their viral load rebound is denoted by T_i^R. Let C_i denote the potential censoring time for patient *i*, with the binary indicators δ_i^S and δ_i^R equallying 1 when T_i^S and T_i^R are smaller than C_i, respectively, and 0 otherwise. Furthermore, define $X_i^S = \min(T_i^S, C_i)$ and $X_i^R = \min(T_i^R, C_i)$. In general, $\delta_i^S \geq \delta_i^R$, because the time to rebound of viral load may only occur subsequent to viral load suppression. For patient *i*, the observed data consists of $(X_i^S, \delta_i^S, X_i^R, \delta_i^R, R_i)$. The main difficulty in conducting a time-to-event intent to treat analysis using this data structure is that there is not an obvious single “time to event” on which to base the analysis.

Suppose for simplicity that there are two treatment groups, r = 1 and 2, and let S_r^S and S_r^R denote the survival functions for T_i^S and T_i^R, respectively, in group r = 1, 2. The endpoint we propose for viral suppression studies is the probability of being suppressed at time *t*, $G_r(t) = S_r^R(t) - S_r^S(t)$, r = 1, 2. This endpoint is defined without conditioning on information observed post randomization and may be analyzed using intent to treat methods. However, because the event probability is the difference difference of two survival functions and is not itself a survival function for a single time to event, the Kaplan–Meier estimator and logrank test are not applicable. Inferential methods for multistate data must be used in the development of non-parametric estimators and tests for treatment differences.

Following Pepe (1991), we employ the Kaplan–Meier estimates $\hat{S}_r^S(t), \hat{S}_r^R(t)$, r = 1, 2, of survival functions for time to viral suppression and time to viral rebound respectively. Note that time to viral rebound defined as above is measured from randomization. The survival function for time to viral rebound will be the marginal survival function, not conditional on being suppressed. We can estimate the probability for a patient to be in the state of suppression, within each treatment group separately, as

$$\hat{G}_r(t)=\hat{S}_r^R(t)-\hat{S}_r^S(t)\text{for}r=1,2.$$

The variance estimator for $\hat{G}_r(t)$, r = 1, 2, is given by

$$\hat{Var}(\hat{G}_r(t))=\frac{1}{n_r^2}\sum_{i:R_i=r}\left[\hat{X}_{G_r}^i(t)\right]^2,$$

where n_r is the number of subjects in group r,

$$\hat{X}_{G_r}^i(t)=n_r\hat{S}_r^S(t)\left\{\int_0^t\frac{1}{Y_S(u)}\mathrm{d}N_S^i-\int_0^t\frac{Y_S^i(u)}{(Y_S(u))^2}\mathrm{d}N_S(u)\right\}-n_r\hat{S}_r^R(t)\left\{\int_0^t\frac{1}{Y_R(u)}\mathrm{d}N_R^i-\int_0^t\frac{Y_R^i(u)}{(Y_R(u))^2}\mathrm{d}N_R(u)\right\},$$

$N_S^i(u), N_R^i(u)$, are the counting processes for the events of suppression and rebound, respectively for a patient i, $Y_S^i(u)$ and $Y_R^i(u)$ are the at risk processes for suppression and rebound, respectively for a patient i, and

$$Y_\varepsilon(u)=\sum_{i:R_i=r}Y_\varepsilon^i(u)\text{and}N_\varepsilon(u)=\sum_{i:R_i=r}N_\varepsilon^i(u)\text{for}\varepsilon\in\{S,R\}.$$

The probability of being in suppression $G_r(t)$ for group r = 1, 2 varies over time, similarly to a survival function, albeit not a monotically decreasing function of t. As with standard time to event analyses, simple summary measures are needed for quantifying differences among treatment regimens. One should recognize that $G_r(t)$ does not have a corresponding hazard function and treatment differences cannot be summarized using hazard ratios, as they might in separate analyses of S_r^R and S_r^S. We suggest summarizing using the weighted restricted mean time a patient from group r will spend in suppression in the time interval [0, t_0], which is $\int_0^{t_0}\hat{W}(u)G_r(u)\mathrm{d}u$, where $\hat{W}(u)$ is an estimate of some appropriately chosen weight function $W(u)$ discussed below.

The analysis may be tailored to capture the information of greatest importance with a careful choice of the weight function. When $W(u) \equiv 1$, the weighted integral estimates the restricted mean time spent in viral suppression. For those interested in short term outcomes, larger weights may be applied at early time points, and vice versa for long term outcomes. For example, for those interested primarily in long term maintenance, zero weights may be employed at time points before some predetermined cut-off for suppression, for example, 24 weeks. On the other hand, for those interested in population health where individuals with circulating virus present a transmission risk, nonzero weights at early time points would be an important consideration.

Following Pepe (1991), for the purpose of hypothesis testing one may compute a simple Z type test statistic as the difference of the weighted averages in the two treatment arms. The test statistic is:

$$WG=\sqrt{\frac{n_1n_2}{n_1+n_2}}\int_0^{t_0}\hat{W}(u)\{\hat{G}_1(u)-\hat{G}_2(u)\}\mathrm{d}u.$$

Under the null hypothesis, the test statistic is asymptotically normal with zero mean and its asymptotic variance can be estimated by

$$\hat{Var}(WG)=\frac{n_1n_2}{n_1+n_2}(\hat{V}_1+\hat{V}_2,)$$

where

$$\hat{V}_r=\frac{1}{n_1^2}\sum_{i:R_j=r}\left[\int_0^{t_0}\hat{W}(u)\hat{X}^j_{G_r}\mathrm{d}u\right]^2,r=1,2.$$

Wald type confidence intervals for the weighted average time in suppression may be calculated using the asymptotic normality of the estimator $\int_0^{t_0}\hat{W}(u)\hat{G}_r(u)\mathrm{d}u$ and its variance estimator $\hat{V_r}$, r = 1, 2.

The choice of the weight function may also be directed towards improving the power of the test statistic to detect treatment differences in the probability of suppression over time. As suggested by Pepe and Fleming (1989), one may downweight at time points where $\hat{G}_1 - \hat{G}_2$ is highly variable using the weight function:

$$\hat{W}_{\mathrm{se}}(u)=\frac{1}{S\hat{E}}[\hat{G}_1(u)-\hat{G}_2(u)],$$

where $S\hat{E}[\hat{G}_1(u)-\hat{G}_2(u)]=\sqrt{\hat{Var}(\hat{G}_1(t))+\hat{Var}(\hat{G}_2(t))}$. This may also be accomplished using some function of the censoring distributions in the two groups Pepe and Fleming (1989), with weight:

$$\hat{W}_{\mathrm{cens}}(u)=\frac{[\hat{S}_1^C(t)\times\hat{S}_2^C(t)]}{[p_1\hat{S}_1^C(t)+p_2\hat{S}_2^C(t)]},$$

where $\hat{S}_r^C(t)$ is the Kaplan–Meier estimator of the survival function of C_i, $S_r^C(t)$, in group r = 1, 2 and p_r is the proportion of patients allocated to group r = 1, 2. The unity weight assigns equal weight to all time points, while the second and third weights tend to assign higher weight to earlier time points, where the estimation is typically less variable, potentially resulting in increased power. In applications with focused scientific objectives, the choice of the weight should be driven by those objectives and not by unguided power considerations.

If we have several strata *j* = 1, …, *J* and wish to conduct a stratified analysis, we can compute the above *WG* statistics separately within each stratum *j* and then let

$$SWG=\frac{\sum_{j=1}^{J}\omega_j WG_j}{\sqrt{\sum_{j=1}^{J}\omega_j^2\hat{Var}(WG_j)}}$$

where WG_j and $\hat{Var}(WG_j)$ are the test statistic and its estimated variance within stratum *j* = 1, …, *J*. The scalar ω_j determines the relative weight given to stratum *j* = 1, …, *J*. Under the null hypothesis of no difference between the groups, the *SWG* statistic is asymptticaly normal *N*(0, 1).

To perform power and sample size calculations for studies using the proposed endpoint, one can use standard formulas for continuous normally distributed outcomes. The standard deviation of the test statistic necessary for such computations can be obtained by re-analysis of previously available similar data or via simulations. For example, for the test statistic based on the unity weight, for a trial with two arms of equal size and assuming equal variances in both arms, we can take the desired effect size Δ to be a clinically relevant difference in average time spent in suppression between the treatment and control arms (e.g., 4 weeks, if weeks is the chosen time scale). If the data from an earlier similar trial is available, we can compute the WG_{prior} statistic for the prior trial data and estimate its standard error. Due to the scaling of *WG* by $\sqrt{\frac{n_1n_2}{n_1+n_2}}$, the standard error $\hat{SE}(WG_{\text{prior}})$ is an estimate of the true standard deviation of the time spent in suppression. Hence we can use Δ and $\hat{SE}(WG_{\text{prior}})$ as the effect size and the standard deviation in the standard sample size formulas. The results of a small simulation study verifying this approach are provided at the end of Section 3.

3. Simulation Results

We conducted a simulation study to compare performance of the proposed endpoint with the virologic failure endpoint used in the A5142 trial and TLOVR. For each simulated patient we generated a treatment group assignment and then, conditionally on the treatment assignment, times to suppression (T_i^S), rebound (T_i^R), and censoring (C_i). The time was on the weeks scale, and the length of the observation period was chosen to be 80 weeks. We first generated time from randomization to suppression, then time from suppression to rebound, and then computed the time from randomization to rebound as the sum of the two above times.

We employ three different simulation scenarios shown in Figure 1. In scenario 1 the treatment group was the same as control in terms of suppression and had much later rebound, thus maintaining suppression much longer than the control group. Under scenario 2, the treatment group had faster suppression but also faster rebound. On average, in scenario 2, the treatment group was suppressed longer. In scenario 3, the treatment group suppressed later than in the control group, but maintained suppression longer. Thus, under scenario 3, the treatment group had reduced probability of suppression in the beginning of the observation period which reversed at later times.

We used the Weibull distribution for all time variables in the simulations, due to its flexible shape, with the CDF function $F(t)=1-\exp\{-\left(\frac{t}{\beta}\right)^{\alpha}\}$ and the distribution parameters α and β as follows. Scenario 1: the treatment group—$\alpha^S = 0.2$, $\beta^S = 4000$, $\alpha^R_{\text{cond}}=4, \beta^R_{\text{cond}}=120$, the control group— $\alpha^S = 0.2$, $\beta^S = 4000$, $\alpha^R_{\text{cond}}=1.35, \beta^R_{\text{cond}}=64$. Scenario 2: the treatment group—$\alpha^S = 0.4$, $\beta^S = 800$, $\alpha^R_{\text{cond}}=1, \beta^R_{\text{cond}}=120$, the control group —$\alpha^S = 0.8$, $\beta^S = 320$, $\alpha^R_{\text{cond}}=1, \beta^R_{\text{cond}}=120$. Scenario 3: the treatment group—$\alpha^S = 1$, $\beta^S = 8$, $\alpha^R_{\text{cond}}=2, \beta^R_{\text{cond}}=240$, the control group—$\alpha^S = 0.1$, $\beta^S = 0.0008$, $\alpha^R_{\text{cond}}=1, \beta^R_{\text{cond}}=200$. The censoring distribution was the same in both treatment groups and across all scenarios, with $\alpha^{\text{cens}} = 1.5$, $\beta^{\text{cens}} = 400$. Treatment assignment was generated as a Bernoulli random variable with success probability 0.5. We assessed several sample sizes between 250 and 2000 patients. All simulations were conducted using 1000 samples. For the proposed method, we defined observed data as $X_i^S=\min(T_i^S, C_i), X_i^R=\min(T_i^R, C_i), \delta_i^S=I(X_i^S=T_i^S)$, and $\delta_i^R=I(X_i^R=T_i^R)$. We computed the test statistic *WG*, with each of the three weight functions described in Section 2.

To define virologic failure as in the A5142 trial or for the TLOVR-like endpoint, we first chose a cut-off point γ_0, non-suppression prior to which should be considered a failure. Then, given the cut-off, we defined the observed data for the composite event in A5142 as:

$$X_i^{\text{comp}}=\begin{cases} \min(T_i^R, C_i), & 0<T_i^S \le \gamma_0, \\ \min(\gamma_0, C_i), & \gamma_0<T_i^S \end{cases}$$

and

$$\delta_i^{\text{comp}}=\begin{cases} I(X_i^{\text{comp}}=T_i^R), & 0<T_i^S \le \gamma_0, \\ I(X_i^{\text{comp}}=\gamma_0), & \gamma_0<T_i^S. \end{cases}$$

Similarly, the data for the TLOVR-like event were defined as:

$$X_i^{\text{TLOVR}}=\begin{cases} \min(T_i^R, C_i), & 0<T_i^S \le \gamma_0, \\ 0, & \gamma_0<T_i^S \end{cases}$$

and

$$\delta_i^{\text{TLOVR}}=\begin{cases} I(X_i^{\text{TLOVR}}=T_i^R), & 0<T_i^S \le \gamma_0, \\ 1, & \gamma_0<T_i^S. \end{cases}$$

Using the composite endpoint from the A5142 protocol and TLOVR independently, we separately performed two-sided logrank tests and then determined the direction of the difference by fitting the Cox model using treatment group as the sole covariate, to mimic the intent to treat analysis from Riddler et al. (2008). We looked at a range of possible cut-off points in the definition of composite events for the A5142 and TLOVR endpoints.

The observed type I error rate was close to the nominal level for all three methods, ranging from 0.041 to 0.057 for the proposed endpoint and from 0.040 to 0.060 for A5142 and TLOVR endpoints (not shown in tables). The results for power are summarized in Table 1. For the proposed method, the power to reject the null hypothesis was consistent for all scenarios, for all choices of the weight function, and increased with sample size. However, for the A5142 composite endpoint and for TLOVR, the power varied from being higher than that for the proposed method to being almost zero, depending on the scenario and the choice of the cut-off point γ_0. For scenario 1, the power for both composite endpoints was much higher than for the proposed method. For scenario 2, the power for A5142 and TLOVR endpoints was sometimes worse than for the proposed method, depending on the chosen value of the cut-off. The results for scenario 3 are the most interesting. If we look at which treatment arm was selected under scenario 3, for some values of γ_0, the A5142 and TLOVR analyses always incorrectly selected the control arm. For a large sample size (2000 patients), the null hypothesis was rejected in favor of the wrong treatment group 81% of the time using the A5142 endpoint and 92% of the time using TLOVR. Such a reversal of results happened because both composite endpoints from A5142 and TLOVR re-defined the time of event. For some values of the cut-off γ_0 (prior to 12 weeks in the scenario 3), the failures in the treatment group were forced to happen earlier than in the control group.

We also conducted a small simulation study to test the sample size computations for the proposed endpoint. We generated 1000 samples from the known distributions under scenario 3 described above, assuming a known effect size. Based on each simulated sample, we estimated standard deviations for our test statistics and computed predicted power based on the observed standard deviations and hypothesized effect size (using SAS procedure POWER). Then we compared the average predicted power with the power observed in 1000 simulations. The results summarized in Table 2 generally exhibit good agreement between the observed and predicted powers.

4. Re-analysis of the A5142 Trial

As an example, we re-analyzed the ACTG A5142 trial using the virologic failure endpoint from A5142 and the proposed method. The A5142 trial included 753 patients whose baseline viral load was at least 2000 copies/ml. Patients were randomized to one of the three treatment arms, efavirenz plus two NRTIs (efavirenz group), lopinavir–ritonavir plus two NRTIs (lopinavir–ritonavir group), or lopinavir–ritonavir plus efavirenz (NRTI-sparing arm). The median follow-up was 112 weeks, with the longest follow-up time being 157 weeks.

The definition of a virologic failure for A5142 (Riddler et al., 2008, p. 2097) was lack of confirmed viral load suppression below 200 copies/ml or by $\log_{10}$ by 8 weeks; or lack of confirmed viral suppression below 200 copies/ml by 32 weeks; or confirmed viral rebound. The definition of viral rebound also varied depending on when the rebound occurred. Early rebound (prior to 32 weeks) was defined as a viral load increase to over 1000 copies/ml for patients whose viral load had suppressed below 200 copies/ml; or viral load increase by $\log_{10}$ from the nadir value for patients whose viral load had never suppressed below 200 copies/ml. Late rebound (after 32 weeks) was defined as a viral load ≥200 copies/ml. 227 patients experienced virologic failure by the A5142 definition. The Kaplan–Meier estimators for virologic failure are shown on the top panel of Figure 2.

For the proposed approach, we defined two separate events, viral suppression and viral rebound. We defined viral suppression as viral load being reduced to <200 copies/ml for two consecutive measurements 4 weeks or less apart. We defined viral rebound as viral load being ≥200 copies/ml at two consecutive measurements 4 or less weeks apart. We had 667 patients in all treatment groups whose viral load was suppressed and 129 patients who experienced viral rebound. A plot of the estimated probability of being suppressed, over time from randomization, by treatment group, is displayed on the bottom panel of Figure 2.

Next, we computed the test statistic *WG*, using the three different weight functions introduced in Section 2. We integrated over the first 143 weeks of follow-up, capturing almost all information in the dataset. For comparison, logrank tests were also calculated based on the A5142 composite endpoint. Over the 143 week period after randomization, the analysis showed that a patient from the efavirenz group was in a state of viral suppression for 12 weeks longer on average than a patient from the lopinavir–ritonavir group (95% CI = (3, 21), p-value after Bonferroni correction p = 0.032) and for 3 weeks longer than a patient from the NRTI-sparing group (95% CI = (−3, 13), p-value p = 0.783). Moreover, a patient from the NRTI-sparing group was in the state of suppression for 5 weeks longer than a patient from the lopinavir–ritonavir group (95% CI = (−2, 16), p-value p = 0.365). The results of testing the null hypothesis of no difference between the three treatment groups using *WG* are reported in Table 3 as endpoint 1. Inferences are similar to those obtained from the original analysis.

Recognizing that there might have been clinical considerations for defining a separate early viral suppression and viral rebound, we performed additional analyses mimicking the definitions of viral suppression and viral rebound from the A5142 trial as closely as possible. We defined early viral suppression and viral rebound prior to 32 weeks as was done in the A5142 trial. Under this definition, the number of patients in all treatment groups who experienced viral suppression was 691, and who experienced viral rebound was 183. The results of this supplementary analysis are also summarized in Table 3 as endpoint 2 and are not statistically significant at 0.05 level, though the direction of the differences remained the same.

We also performed sensitivity analysis to assess how much the results of the A5142 trial depended on the choice of cut-off time of 8 weeks for early rebound and 32 weeks for late rebound. Judging by the plot of the probability to be in suppression by treatment group, we did not expect inference to change when we varied the cut-off times for early and late viral rebound. This is because the best treatment group was uniformly better than the second best treatment group both in terms of viral suppression and viral rebound, with the same ordering holding for the second and third best treatment groups. We re-defined virologic failure using cut-offs ranging from 5 to 15 weeks for early rebound and from 25 to 40 weeks for late rebound. The results confirmed our expectations: the p-values for comparison of the efavirenz and lopinavir–ritonavir groups remained significant and ranged from 0.0107 to 0.0306 (after a Bonferroni correction), all other comparisons were still not statistically significant, and all the differences between the groups were in the same direction.

In summary, certain advantages of the proposed endpoint can be clearly seen in Figure 2, where the time-specific treatment differences are cleanly summarized via the probability of being in suppression. The efavrienz group suppresses most rapidly and with higher probability and the suppression is maintained as effectively as in the NRTI-sparing arm. The NRTI-sparing arm has comparable early suppression to that in the lopinavir group, but with superior long term maintenance. Such information is not as readily gleaned from the plot of the survival curves for the A5142 composite endpoint.

5. Discussion

DeGruttola et al., (1998) were the first to discuss the use of HIV-1 RNA viral load as an outcome measure in HIV trials, both as a repeatedly-assessed continuous biomarker and as an indicator of treatment (virologic) failure. Gilbert et al., (2000) expand on the discussion of virologic failure and consider several competing definitions. Ribaudo et al. (2006) discussed design issues in HIV trials, concentrating the discussion of endpoints on further refinements in virologic failure. To the best of our knowledge, no one has previously suggested the combined endpoint we propose here.

We implemented a novel approach to defining a time to event endpoint in HIV research that combines time to viral suppression and time to viral rebound into a single measure, the probability of being suppressed over time. As demonstrated in the A5142 data analysis, this quantity precisely captures the interplay of suppression and rebound, yielding a simple graphical representation of early and late suppression dynamics which may be preferable to that for the existing composite endpoints. The integrated probability of suppression can easily be adapted by choice of the weight function to target specific time periods of interest. Employing unity weight provides a particularly attractive summary which may be interpreted as the average number of weeks suppressed over the time period of interest. As suggested by a referee, if there is scientific justification to disregard a portion of the followup period, the weights function can be set to zero for those times points.

The probability of suppression endpoint may also be useful in observational studies, albeit with the necessary caveats about confounding. Further work is needed to appropriately adjust for confounding factors in the analysis. Future research is planned into regression modeling of the probability of suppression and the associated weighted average times in suppression. However, the application of the proposed endpoint to observational studies is beyond the scope of the current manuscript which deals with intent to treat analyses in randomized HIV trials with HIV RNA measurements obtained on a specific predefined schedule.

The unifying multistate framework applied here to the HIV setting may also prove useful in other settings where endpoints are defined using biomarker threshold values, for example, hypertension as defined by elevated blood pressure. As in the HIV application, the resulting composite endpoints may be ad hoc and not easily generalizable across studies and populations. A more efficient use of the observed data in these settings might be accomplished via jointly modeling the longitudinal biomarker measurements as continuous outcomes and the event times. To adopt this strategy, rather strong modeling assumptions may be needed, the computations may be challenging, and summarizing the results from the fitted joint model may not be straightforward.

All analysis for this article has been conducted using SAS 9.3 software (SAS Institute, Cary, NC).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Dr. Gregory DiRienzo for expert advice. Dr. Cole was supported in part by NIH grant R01AI100654.

References

Cole SR, Hernn MA, Anastos K, Jamieson BD, Robins JM. Determining the effect of highly active antiretroviral therapy on changes in human immunodeficiency virus type 1 RNA viral load using a marginal structural left-censored mean model. American Journal of Epidemiology. 2007; 166:219–227. [PubMed: 17478436]

DeGruttola V, Hughes M, Gilbert P, Phillips A. Trial design in the era of highly effective antiviral drug combinations for HIV infection. AIDS. 1998; 12:S149–S156. [PubMed: 9632997]

Egger M, May M, Chene G, Phillips AN, Ledergerber B, Dabis F, Costagliola D, D'Arminio Monforte A, de Wolf F, Reiss P, Lundgren JD, Justice AC, Staszewski S, Leport C, Hogg RS, Sabin CA, Gill MJ, Salzberger B, Sterne JA. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: A collaborative analysis of prospective studies. The Lancet. 2002; 360:119–129.

Fischl MA, Ribaudo HJ, Collier AC, Erice A, Giuliano M, Dehlinger M, Eron JJ, Saag MS, Hammer SM, Vella S, Morse GD. Adult AIDS Clinical Trials Group 388 Study Team. A randomized trial of 2 different 4-drug antiretroviral regimens versus a 3-drug regimen, in advanced human immunodeficiency virus disease. Journal of Infectious Diseases. 2003; 188:625–634. [PubMed: 12934177]

Gilbert PB, Ribaudo HJ, Greenberg L, Yu G, Bosch RJ, Tierney C, Kuritzkes DR. Considerations in choosing a primary virological endpoint for durability in AIDS antiretroviral trials. AIDS. 2000; 14:1961–1972. [PubMed: 10997401]

Gulick RM, Ribaudo HJ, Shikuma CM, Lustgarten S, Squires KE, Meyer WA, Acosta EP, Schackman BR, Pilcher CD, Murphy RL, Maher WE, Witt MD, Reichman RC, Snyder S, Klingman KL, Kuritzkes DR. Triple-nucleoside regimens versus efavirenz-containing regimens for the initial treatment of HIV-1 infection. New England Journal of Medicine. 2004; 350:1850–1861. [PubMed: 15115831]

Gulick RM, Ribaudo HJ, Shikuma CM, Lalama C, Schackman BR, Meyer WA, Acosta EP, Schouten J, Squires KE, Pilcher CD, Murphy RL, Koletar SL, Carlson M, Reichman RC, Bastow B, Klingman KL, Kuritzkes DR. Three- vs four-drug antiretroviral regimens for the initial treatment of

HIV-1 infection: A randomized controlled trial. The Journal of the American Medical Association. 2006; 296:769–781.

Mellors JW, Rinaldo CR, Gupta P, White RM, Todd JA, Kingsley LA. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. Science. 1996; 272:1167–1170. [PubMed: 8638160]

Pepe MS, Fleming TR. Weighted Kaplan–Meier statistics—A class of distance tests for censored survival data. Biometrics. 1989; 45:497–507. [PubMed: 2765634]

Pepe MS. Inference for events with dependent risks in multiple endpoint studies. Journal of the American Statistical Association. 1991; 86:770–778.

Ribaudo HJ, Kuritzkes DR, Schackman BR, Acosta EP, Shikuma CM, Gulick RM. Design issues in initial HIV-treatment trials: Focus on ACTG A5095. Antiviral therapy. 2006; 11:751–760. [PubMed: 17310819]

Riddler SA, Haubrich R, DiRienzo AG, Peeples L, Powderly WG, Klingman KL, Garren KW, George T, Rooney JF, Barbara, Brizz, Umesh G, Lalloo MD, Robert L, Murphy MD, Swindells S, Havlir D, Mellors JW. Class-sparing regimens for initial treatment of HIV-1 infection. New England Journal of Medicine. 2008; 358:2095–2106. [PubMed: 18480202]

Robbins GK, De Gruttola V, Shafer RW, Smeaton LM, Snyder SW, Pettinelli C, Dubé MP, Fischl MA, Pollard RB, Delapenha R, Gedeon L, van der Horst C, Murphy RL, Becker MI, D’Aquila RT, Vella S, Merigan TC, Hirsch MS. Comparison of sequential three-drug regimens as initial therapy for HIV-1 infection. New England Journal of Medicine. 2003; 349:2293–2303. [PubMed: 14668455]

Sax PE, Tierney C, Collier AC, Fischl MA, Mollan K, Peeples L, Godfrey C, Jahed NC, Myers L, Katzenstein D, Farajallah A, Rooney JF, Ha B, Woodward WC, Koletar SL, Johnson VA, Geiseler PJ, Daar ES. Abacavir–Lamivudine versus Tenofovir–Emtricitabine for initial HIV-1 therapy. New England Journal of Medicine. 2009; 361:2230–2240. [PubMed: 19952143]

U. S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER). Guidance for industry: Antiretroviral drugs using plasma HIV RNA measurements—Clinical considerations for accelerated and traditional approval. 2002 Oct. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070968.pdf.

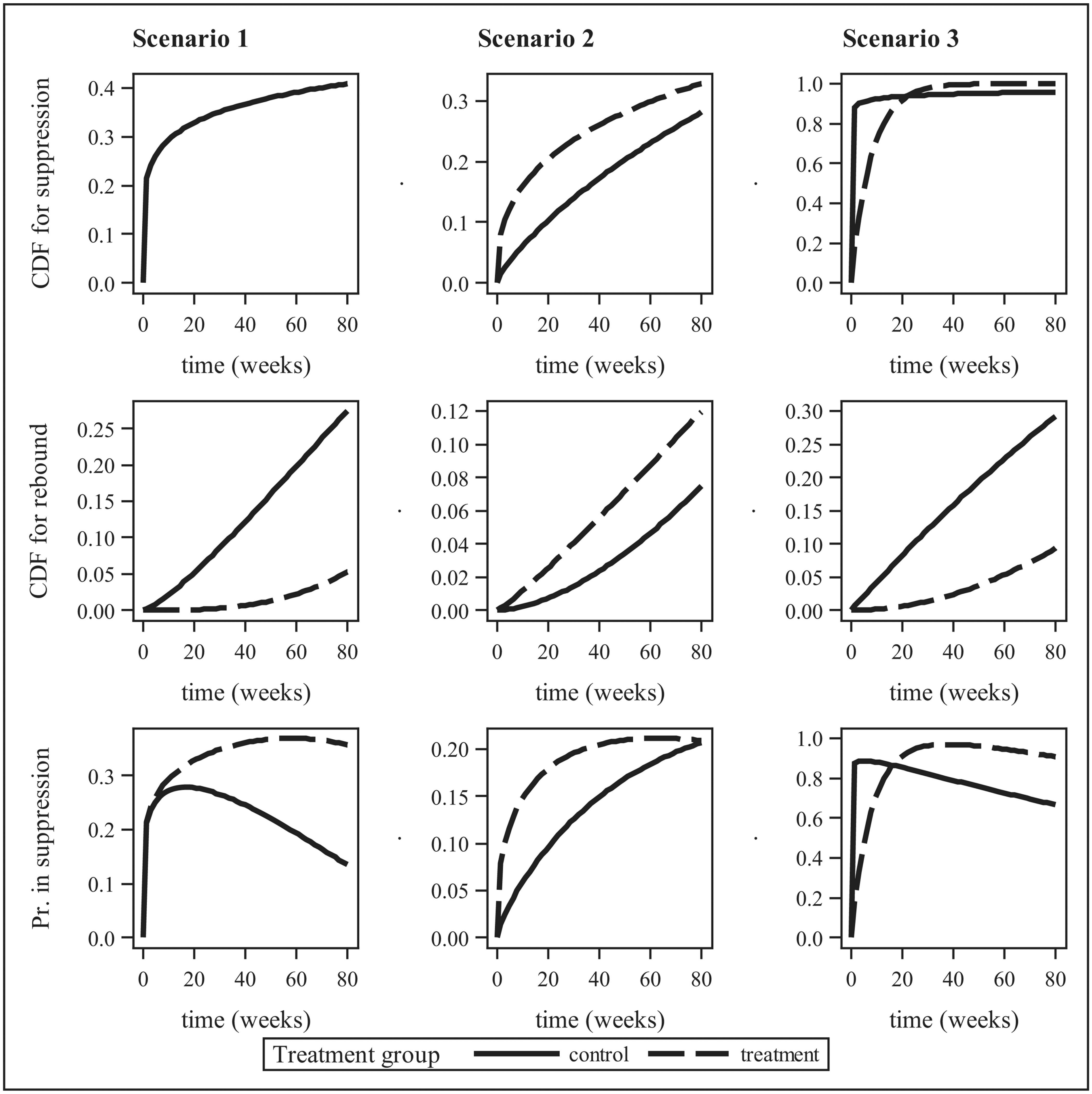

Figure 1.

Simulations scenarios: CDF for time to suppression, CDF for time to rebound, and probability to be in suppression, by treatment group.

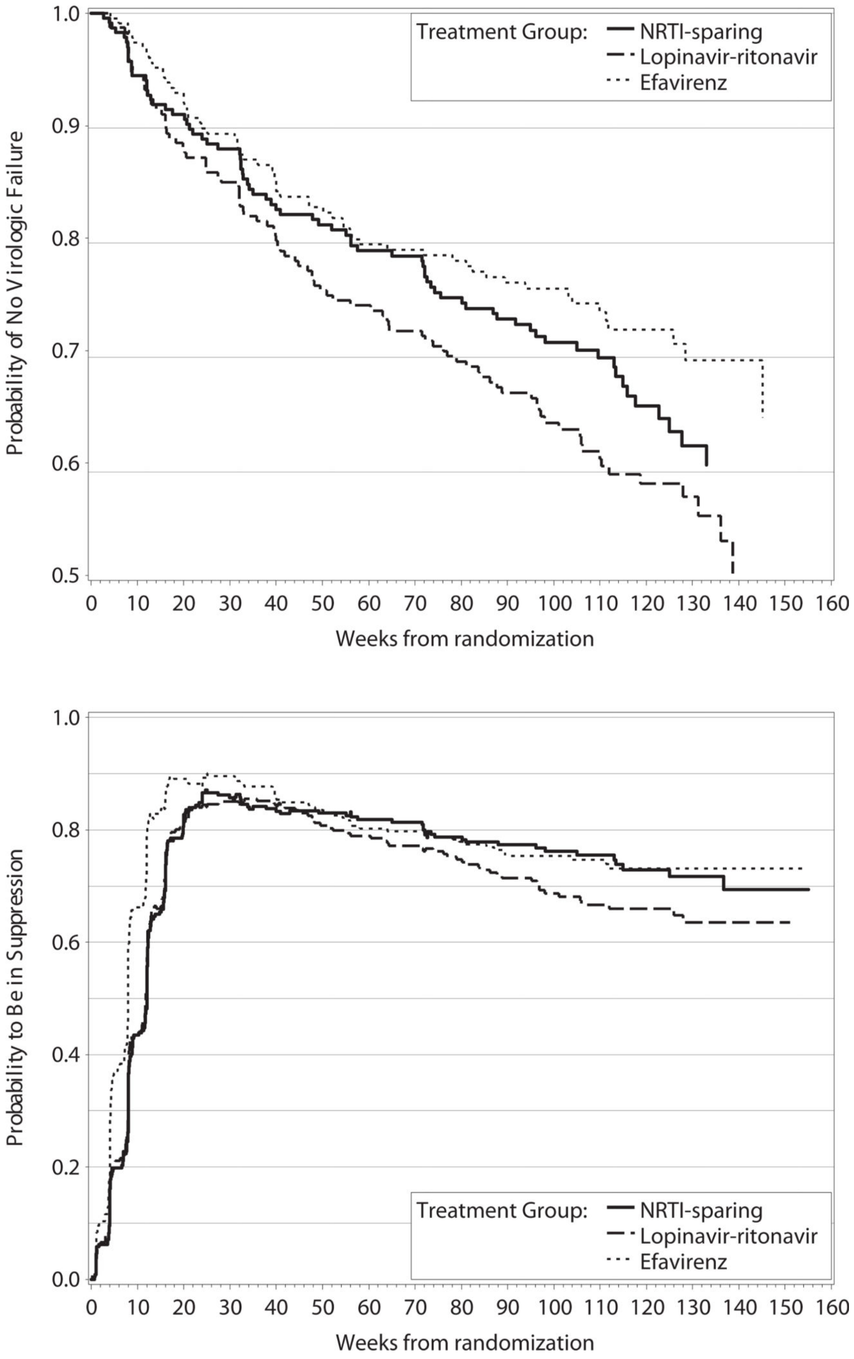

Figure 2.

Top panel: Survival functions for virologic failure as defined in A5142 trial, by treatment group. Bottom panel: Probability to be in suppression, by treatment group.

Table 1

Simulation results: Power to reject the null hypothesis, and the preferred treatment arm, by value of the cut-off time point for the A5142 and TLOVR endpoints, and by weight function for the proposed method

		Method												
Scenario 1		A5142 trial					TLOVR					Proposed		
		Cut-off (weeks)					Cut-off (weeks)					Weight		
Sample size		8	16	24	32	40	8	16	24	32	40	Unity	1/se	Cens
125	Power	0.581	0.701	0.790	0.857	0.897	0.532	0.572	0.618	0.623	0.639	0.386	0.388	0.357
250	Power	0.892	0.945	0.977	0.992	0.997	0.849	0.908	0.915	0.932	0.923	0.667	0.670	0.635
500	Power	0.997	1.000	1.000	1.000	1.000	0.991	0.998	0.999	1.000	1.000	0.919	0.916	0.888
1000	Power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990	0.990	0.985
		Method												
Scenario 2		A5142 trial					TLOVR					Proposed		
		Cut-off (weeks)					Cut-off (weeks)					Weight		
Sample size		8	16	24	32	40	8	16	24	32	40	Unity	1/se	Cens
125	Power	0.298	0.191	0.137	0.106	0.064	0.346	0.263	0.211	0.188	0.154	0.314	0.347	0.319
250	Power	0.501	0.358	0.220	0.149	0.097	0.567	0.477	0.330	0.304	0.233	0.574	0.629	0.586
500	Power	0.790	0.606	0.432	0.257	0.145	0.852	0.762	0.677	0.554	0.434	0.851	0.892	0.858
1000	Power	0.976	0.889	0.684	0.449	0.264	0.990	0.965	0.911	0.816	0.689	0.990	0.993	0.990
		Method												
Scenario 3		A5142 trial					TLOVR					Proposed		
		Cut-off (weeks)					Cut-off (weeks)					Weight		
Sample size		8	16	24	32	40	8	16	24	32	40	Unity	1/se	Cens
250	Choose treatment	0.000	0.050	0.253	0.577	0.800	0.000	0.027	0.184	0.484	0.744	0.839	0.873	0.813
	Choose control	0.154	0.011	0.000	0.000	0.000	0.210	0.020	0.000	0.000	0.000	0.000	0.000	0.000

500	Choose treatment	0.000	0.046	0.435	0.845	0.973	0.000	0.029	0.306	0.757	0.953	0.987	0.993	0.979
	Choose control	0.286	0.008	0.000	0.000	0.000	0.394	0.016	0.000	0.000	0.000	0.000	0.000	0.000
1000	Choose treatment	0.000	0.066	0.725	0.981	1.000	0.000	0.028	0.555	0.959	0.998	1.000	1.000	1.000
	Choose control	0.523	0.005	0.000	0.000	0.000	0.663	0.021	0.000	0.000	0.000	0.000	0.000	0.000
2000	Choose treatment	0.000	0.104	0.939	1.000	1.000	0.000	0.032	0.814	1.000	1.000	1.000	1.000	1.000
	Choose control	0.811	0.096	0.000	0.000	0.000	0.920	0.023	0.000	0.000	0.000	0.000	0.000	0.000

Table 2

Simulation results: Predicted versus observed power

Sample size	Weight					
	Unity		1/se		Censoring	
	Observed	**Predicted**	**Observed**	**Predicted**	**Observed**	**Predicted**
125	0.482	0.504	0.517	0.564	0.439	0.470
250	0.839	0.817	0.873	0.859	0.813	0.782
500	0.987	0.984	0.993	0.991	0.979	0.975

Table 3

p Values comparing between treatment groups in the A5142 trial, original analysis versus proposed method

	Method						
	Riddler et al.	Proposed, endpoint 1			Proposed, endpoint 2		
		Weight			Weight		
Comparison		Unity	1/se	Cens	Unity	1/se	Cens
EFAV versus LOP/RIT	0.006	0.037	0.023	0.019	0.197	0.185	0.178
EFAV versus NRTI	0.490	0.994	1.000	0.521	1.000	1.000	0.807
NRTI versus LOP/RIT	0.130	0.315	0.292	0.400	0.785	0.738	1.000

Endpoint 1: Single threshold of 200 copies/ml in the definitions of suppression and rebound. Endpoint 2: Different definitions for early and late suppression and rebound. p Values adjusted for multiple comparisons using the Bonferroni correction.