Supplemental Materials
Evaluating the Effect of Early Versus Late ARV Regimen Change if Failure on an Initial Regimen: Results from the AIDS Clinical Trials Group Study A5095

Li Li, Joseph J. Eron, Heather Ribaudo, Roy M. Gulick, Brent A. Johnson*

1 Area Under the Curve (AUC) as Endpoint

Although both treatment strategy and endpoint may depend on viral load levels, the nature of the relationship is not known and we exemplify this point via illustration. To illustrate the that switch to second-line regimen and endpoint are not positively correlated "by definition," two hypothetical viral load trajectories are displayed in Figure 1. Patient I and patient II have the same viral load trajectory over time before week six after confirmed failure. In the left panel, patient I switched to the second-line regimen within eight weeks after confirmed failure, and then viral load dropped below 200 copies/mL quickly. Patient II switched to the second-line regimen at week ten and then viral load dropped below 200 copies/mL immediately. We assume they have the same follow-up length. Therefore, cumulative viral load for patient I (AUC of purple line with squares Y_1) is less than cumulative viral load for patient II (AUC of blue line with dots Y_2) in the left panel. On the other hand, $Y_1 > Y_2$ in the right panel because viral load for patient I does not drop significantly after switching to second-line regimen. A similar phenomenon occurs for the rate of time of suppression endpoint.

[Figure 1 about here.]

2 Simulation Studies

We conducted simulation studies to examine the operating characteristics of several estimators. We consider a special case where all the patients experienced virologic failure and switched to

the next line treatment either early or late. We simulated data similar to the Monte Carlo studies in Cao, Tsiatis and Davidian (2009). For each i, $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^{\mathrm{T}}$ was generated as standard multivariate normal, and the elements of $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^{\mathrm{T}}$ were defined as $X_{i1} = \exp(Z_{i1}/2), X_{i2} = Z_{i2}/(1 + \exp(Z_{i1})) + 10, X_{i3} = (Z_{i1}Z_{i3}/25 + 0.6)^3$ and $X_{i4} = (Z_{i1} + Z_{i2})^2$. Thus, $\mathbf{X}_i$ is a nonlinear function of $\mathbf{Z}_i$. The true and posited propensity score model is $\pi_0(\mathbf{Z}_i) = \mathrm{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.1Z_{i4})$. We consider four simulation scenarios:

1. true OR model is $Y_i = \boldsymbol{\xi}^{\mathrm{T}}\mathbf{X}_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$; the posited OR model is correctly specified and $\mathbf{X}_i$ is observed.

2. same as 1. except that posited OR model uses $\mathbf{X}_i$ as well as 10 noise covariates;

3. same as 1. except that posited OR model uses $\mathbf{Z}_i$ not $\mathbf{X}_i$;

4. true OR model is $Y_i \sim \mathrm{Exp}(\lambda_i)$, with $\lambda_i = \boldsymbol{\xi}^{\mathrm{T}}\mathbf{X}_i$, while posited OR model is linear model.

For each scenario of $n = 1000$ and $n = 100$, 1000 Monte Carlo datasets were generated. Results for all simulation scenarios are presented in Table 1.

When the sample size is large, all estimators all showed small Monte Carlo bias. In large samples, AIPW, REG, RRZ, and CTD estimators exhibited similar variance with modest differences. For example, in OR model 3, RRZ and AIPW had smaller variance than both REG and CTD. The CTD estimator had the smallest variance in OR model 2 but had bias twice as large as REG. For $n = 100$, the REG estimator was the best overall performer and RRZ was the worst, excluding the IPW estimator. In OR model 1, CTD showed somewhat smaller bias than REG but its variance was much larger. Interestingly, the RRZ estimator had smallest variance in OR model 3 and similar bias to REG. In general, the sandwich estimator for the variance of the estimator was too small in small samples but improved as the sample size increased. Thus, in many scenarios, the REG estimator was competitive with other estimators in the scenarios we considered. However, based on the simulation results, we see that even the REG estimator can perform poorly in small samples, even when the PS model is correctly specified.

[Table 1 about here.]

3 Details on Sensitivity and Secondary Analyses

The sensitivity our analytic results depend on many assumptions, some of which are identified and others of which are not identified by the observable data. Because these statistical assumptions play no small role in the analysis of observational data, many authors have proposed a wide range of tools for model diagnostics and sensitivity analyses (cf. Rosenbaum, 1983; Robins, 1999; Robins, Rotnizky, and Scharfstein, 1999; Rotnizky, Scharfstein, Su, and Robins, 2001). Our sensitivity analyses included, but not limited to, comparing the effect on our parameter estimates when weak *observed* confounders were removed and when all the confounders were included in the models. In addition, we will conduct analysis to look into the influence of those 50 patients who experienced virologic failure but only after a protocol-approved substitution. Investigating the sensitivity of our analytic results to nonidentifiable assumptions is beyond the scope of the current paper. Hence, our results rest on the validity of the "no unmeasured confounders" assumption. However, this assumption is ubiquitous in the literature and a well-know limitation of causal inference.

We include potential confounders only when they are significantly or mildly related to treatment switching, endpoints, or both, and used the same set of variables throughout. Different endpoints have different important covariables sets. Baseline viral load, baseline CD4 cell counts, time to viral failure on initial regimen, race and body weight are found important for cumulative viral load. Baseline viral load, HIV RNA at virologic failure before switching, baseline CD4 cell counts, baseline CD8 cell counts, time to viral failure on initial regimen, sex and race have important effects on proportion of time with non-detectable viral load. Baseline viral load, HIV RNA at virologic failure, baseline CD4 cell counts, baseline CD8 cell counts, time to viral failure on initial regimen, body weight, sex and race are significantly associated with cumulative CD4 cell counts. In Table 1, we report the point estimates of mean outcomes on treatment policies and their standard error estimates. Compared to the main analysis, we found that point estimates and standard error estimates changed little when unimportant covariates were removed from the models and the difference of mean endpoints are still significant between patients who switch to a second-line regimen within 8 weeks after failure on the first-line regimen than those patients who switch late. The findings are summarized in Table 2.

[Table 2 about here.]

[Table 3 about here.]

In our main analysis of the ACTG A5095 data, we computed our endpoints by calculating length-adjusted AUC of HIV RNA and CD4 cell counts on their log10 transformation of original scale. Alternatively, we could have computed length-adjusted AUC on original scale, taking natural logarithmic transformation on AUC. Here, we saw some deviation in the magnitude of the difference in point estimates for the cumulative HIV RNA endpoint compared to those presented Table 4 in main analysis. However, our Wald test of significance still rejected the null hypothesis at the nominal level. Thus, our conclusions remained the same as those presented in the main analysis for all three endpoints. Point estimates and standard error estimates are displayed in Table 4.

[Table 4 about here.]

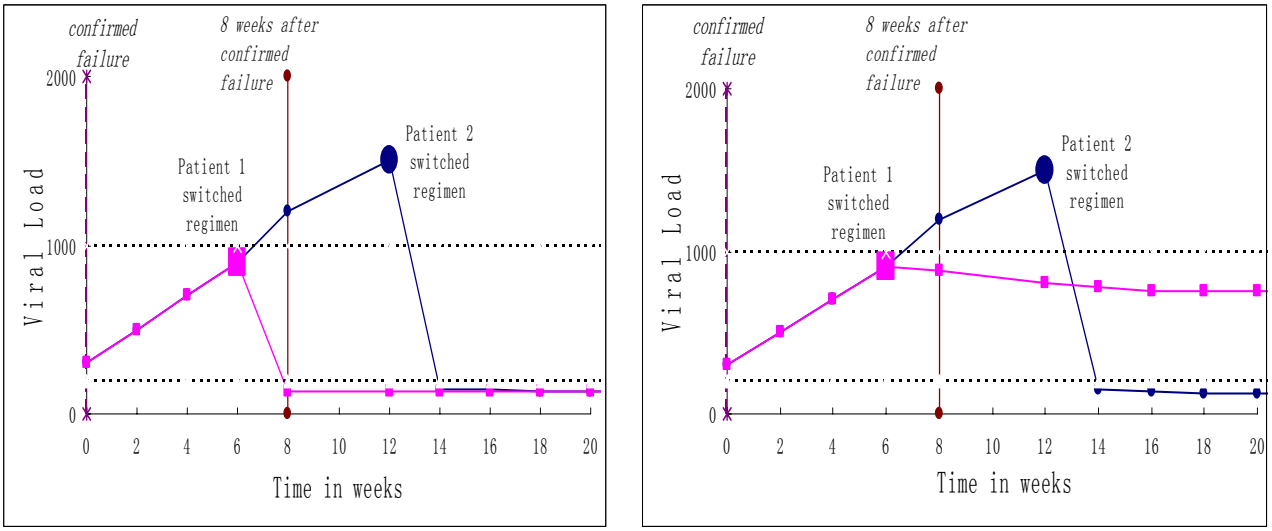

Figure 1: Two exemplary HIV trajectories: patient I has smaller AUC and longer time of viral load suppression than patient II in the left panel. Right panel shows the opposite phenomenon.

Table 1: Simulation results based on 1000 Monte Carlo data sets.

Method	n=1000				n=100			
	Bias	MCSD	AAVE	MAVE	Bias	MCSD	AAVE	MAVE
	PS correct, OR correct							
IPW	0.011	4.685	4.453	3.926	1.782	24.274	13.978	22.370
AIPW	0.029	1.149	1.147	1.147	0.127	3.658	3.605	3.596
RRZ	-0.312	1.300	1.172	1.164	5.937	9.130	5.348	4.520
REG	-0.029	1.149	1.147	1.147	0.123	3.652	3.602	3.598
CTD	-0.029	1.150	1.147	1.147	0.120	3.669	3.618	3.612
	PS correct, OR incorrect model 1							
IPW	0.040	4.869	4.488	3.975	2.660	24.154	20.533	14.165
AIPW	0.029	1.150	1.149	1.148	0.128	3.664	3.612	3.609
RRZ	-0.323	1.306	1.175	1.167	8.410	10.997	7.211	6.198
REG	0.030	1.150	1.149	1.148	0.127	3.664	3.607	3.605
CTD	0.030	1.150	1.149	1.148	0.188	3.736	3.912	3.760
	PS correct, OR incorrect model 2							
IPW	-0.497	8.324	7.685	6.757	2.676	41.472	24.535	18.256
AIPW	-0.496	2.617	2.381	2.176	0.500	7.813	6.263	5.506
RRZ	0.177	2.619	2.227	2.118	10.303	16.368	9.495	7.989
REG	-0.082	2.386	2.201	2.058	1.454	7.655	6.644	5.219
CTD	-0.177	2.218	1.973	1.918	1.798	11.279	9.920	7.022
	PS correct, OR incorrect model 3							
IPW	0.373	12.647	12.722	12.067	0.759	50.897	37.851	32.091
AIPW	0.464	12.080	12.113	11.592	3.064	42.294	35.611	32.585
RRZ	0.183	12.071	11.745	11.346	3.583	40.884	33.476	30.598
REG	0.619	12.149	11.929	11.415	3.571	45.991	42.915	33.171
CTD	0.492	13.148	11.833	11.294	3.115	77.481	61.887	44.547

NOTE: The estimators are the inverse probability weighted (IPW) estimator, augmented IPW (AIPW) estimator, a minimum variance estimator (RRZ) from Robins et al. (1995), and the proposed regression (REG) estimator. Reported statistics include bias, Monte Carlo standard error (MCSD), the mean and median of standard error estimates, AAVE and MAVE, respectively. Incorrect OR models are linear for each of models 1–2 and exponential for model 3. In model 1, the posited OR model adjusts for 10 noise covariables in addition to the important confounders. The posited OR model 2 uses nonlinear transformations $\mathbf{Z}_i$ of the original confounders $\mathbf{X}_i$.

Table 2: Results for combined efavirenz-containing arms using reduced set of confounders in PS and OR models.

		HIV-1 RNA		Days below LOD		CD4	
Method	Switch	Est. (SE)	T	Est. (SE)	T	Est. (SE)	T
COND	Early	2.596 (0.181)	0.600	0.553 (0.052)	0.968	2.437 (0.055)	0.558
	Late	2.702 (0.071)		0.499 (0.023)		2.473 (0.055)	
IPW	Early	1.824 (0.040)	4.951	0.824 (0.023)	4.563	2.605 (0.074)	0.879
	Late	1.901 (0.033)		0.775 (0.011)		2.534 (0.013)	
IPW (No Aux)	Early	1.875 (0.051)	0.305	0.789 (0.015)	0.932	2.531(0.016)	0.355
	Late	1.901 (0.033)		0.776 (0.011)		2.540 (0.011)	
AIPW	Early	1.838 (0.044)	2.811	0.809 (0.012)	10.146	2.552 (0.013)	2.788
	Late	1.902 (0.033)		0.775 (0.011)		2.535 (0.012)	
RRZ	Early	1.826 (0.040)	4.954	0.806 (0.013)	7.867	2.543 (0.021)	0.185
	Late	1.901 (0.033)		0.775 (0.011)		2.535 (0.011)	
REG	Early	1.826 (0.039)	5.360	0.806 (0.011)	10.859	2.559 (0.015)	7.629
	Late	1.902 (0.033)		0.776 (0.011)		2.534 (0.012)	

NOTE: The table shows estimates of the mean potential outcome (standard errors) for six estimators and each of three endpoints. Each outcome (HIV-1 RNA, Days below LOD, CD4) is computed as the length-adjusted area under the curve (AUC), i.e. AUC divided by length of follow-up. For HIV-1 RNA (copies/mL), the AUC is computed after transforming HIV-1 RNA on the base-10 logarithmic scale; LOD is limit of detection; length-adjusted AUC for CD4 cell count is computed on the base-10 logarithmic scale. The estimators are conditional mean endpoint (COND) considering only those patients that failed on initial ARV, inverse probability weighted (IPW) estimator, IPW estimator using no auxiliary variables (IPW no Aux), augmented IPW (AIPW) estimator, a minimum variance estimator (RRZ) from Robins et al. (1995), and the proposed regression (REG) estimator. The Wald-type test statistic (T) tests the null hypothesis that the average causal effect between treatment strategies is zero.

Table 3: Results for combined efavirenz-containing arms excluding 45 patients who were not following initial ARV regimen at first virologic failure (i.4. $n = 744 - 45 = 699$)

		HIV-1 RNA		Days below LOD		CD4	
Method	Switch	Est. (SE)	T	Est. (SE)	T	Est. (SE)	T
COND	Early	2.596 (0.181)	0.08	0.553 (0.052)	0.62	2.437 (0.055)	0.56
	Late	2.609 (0.081)		0.517 (0.028)		2.475 (0.034)	
IPW	Early	1.805 (0.037)	0.885	0.835(0.020)	4.197	2.603 (0.060)	1.071
	Late	1.833 (0.032)		0.793 (0.011)		2.534(0.015)	
IPW (No Aux)	Early	1.829 (0.045)	0.005	0.804(0.013)	0.390	2.537 (0.014)	0.363
	Late	1.831 (0.032)		0.797 (0.011)		2.545 (0.012)	
AIPW	Early	1.813 (0.040)	0.405	0.824 (0.013)	8.106	2.558(0.012)	3.720
	Late	1.833 (0.032)		0.794 (0.011)		2.540 (0.012)	
RRZ	Early	1.804 (0.038)	0.909	0.824 (0.013)	7.609	2.548 (0.017)	0.909
	Late	1.833 (0.032)		0.793 (0.011)		2.537(0.013)	
REG	Early	1.805 (0.037)	1.094	0.820 (0.010)	10.269	2.555 (0.011)	4.152
	Late	1.834 (0.032)		0.795 (0.011)		2.540 (0.012)	

NOTE: The table shows estimates of the mean potential outcome (standard errors) for six estimators and each of three endpoints. Each outcome (HIV-1 RNA, Days below LOD, CD4) is computed as the length-adjusted area under the curve (AUC), i.e. AUC divided by length of follow-up. For HIV-1 RNA (copies/mL), the AUC is computed after transforming HIV-1 RNA on the base-10 logarithmic scale; LOD is limit of detection; length-adjusted AUC for CD4 cell count is computed on the base-10 logarithmic scale. The estimators are conditional mean endpoint (COND) considering only those patients that failed on initial ARV, inverse probability weighted (IPW) estimator, IPW estimator using no auxiliary variables (IPW no Aux), augmented IPW (AIPW) estimator, a minimum variance estimator (RRZ) from Robins et al. (1995), and the proposed regression (REG) estimator. The Wald-type test statistic (T) tests the null hypothesis that the average causal effect between treatment strategies is zero.

Table 4: Results for combined efavirenz-containing arms using alternative transformation of AUC endpoints

		HIV-1 RNA		Days below LOD		CD4	
Method	Switch	Est. (SE)	T	Est. (SE)	T	Est. (SE)	T
COND	Early	9.557 (0.321)	0.790	0.592 (0.054)	1.018	5.692 (0.056)	0.524
	Late	9.306 (0.129)		0.534 (0.024)		5.764 (0.058)	
IPW	Early	7.875 (0.105)	5.553	0.836 (0.024)	3.689	5.972(0.156)	0.500
	Late	8.083 (0.073)		0.790(0.011)		5.859(0.029)	
IPW (No Aux)	Early	8.120 (0.103)	0.541	0.806 (0.015)	1.001	5.853 (0.037)	0.311
	Late	8.059 (0.073)		0.791 (0.011)		5.870 (0.028)	
AIPW	Early	7.914 (0.079)	12.734	0.826(0.014)	8.477	5.904(0.032)	4.234
	Late	8.087 (0.073)		0.791 (0.011)		5.859(0.028)	
RRZ	Early	7.896 (0.082)	10.983	0.826 (0.013)	8.794	5.896(0.040)	1.313
	Late	8.063 (0.074)		0.790 (0.011)		5.859 (0.028)	
REG	Early	7.901 (0.080)	13.884	0.827(0.010)	16.901	5.904 (0.029)	6.700
	Late	8.087 (0.073)		0.791 (0.011)		5.859 (0.028)	

NOTE: The table shows estimates of the mean potential outcome (standard errors) for six estimators and each of three endpoints. Each outcome (HIV-1 RNA, Days below LOD, CD4) is computed as the length-adjusted area under the curve (AUC), i.e. AUC divided by length of follow-up. For HIV-1 RNA (copies/mL), the length-adjusted AUC is computed on the original scale, then we take the natural logarithm; LOD is limit of detection; length-adjusted AUC of CD4 cell count is computed on the original scale, then transformed on the natural logarithm scale. The estimators are conditional mean endpoint (COND) considering only those patients that failed on initial ARV, inverse probability weighted (IPW) estimator, IPW estimator using no auxiliary variables (IPW no Aux), augmented IPW (AIPW) estimator, a minimum variance estimator (RRZ) from Robins et al. (1995), and the proposed regression (REG) estimator. The Wald-type test statistic (T) tests the null hypothesis that the average causal effect between treatment strategies is zero.