

Design of Viral Dynamic Studies for Efficiently Assessing Potency of Anti-HIV Therapies in AIDS Clinical Trials

Hulin Wu*

Frontier Science & Technology Research Foundation, Inc.,
1244 Boylston Street, Suite 303, Chestnut Hill, Massachusetts 02467, U.S.A.

and

A. Adam Ding

Department of Mathematics, Northeastern University, Boston, Massachusetts 02115,
U.S.A.

SUMMARY

The study of HIV dynamics is one of the most important developments in recent AIDS research. It greatly improves our understanding of the pathogenesis of HIV infection. Recently it has been proposed to use HIV dynamics to evaluate the efficacy of antiviral treatments. Currently a large number of AIDS clinical trials on HIV dynamics are in development worldwide. However, many design issues that arise from HIV dynamic studies have not been addressed. In this paper, we study these problems using intensive Monte Carlo simulations and analytic methods. We evaluate a finite number of feasible candidate designs, which are currently used and proposed in AIDS clinical trials from different perspectives. We compare the viral dynamic marker and classical viral load change markers in terms of power for identifying treatment difference, asymptotic relative efficiency, and sensitivity. Finally we propose some useful suggestions for practitioners based on our results.

* Corresponding author's email address: wu@sdac.harvard.edu

Key Words: AIDS; Asymptotic relative efficiency; Clinical trial simulation; Design of experiment; HIV; Optimal design; Surrogate marker; Viral dynamics.

1. Introduction

The recent development of clinical studies on HIV-1 dynamics has had a strong impact on HIV/AIDS research (Ho et al., 1995; Wei, et al., 1995; Perelson et al., 1996 and 1997; Wu et al., 1997, 1999a; Luzuriaga et al. 1999). Recently it has been proposed to use viral dynamics to evaluate antiviral therapies (Essunger et al., 1997; Ho, 1998; Ding and Wu, 1999, 2001). In particular, Ding and Wu (1999) established a formal relationship between viral decay rates and treatment potency (defined by the reduction rate of virus production due to antiviral treatment), and statistical methods were also proposed for assessing the potency of antiviral therapies using viral decay rates (Ding and Wu, 2001). Viral dynamics can be used to evaluate the virological response of antiviral therapies more efficiently and in a more timely manner. This will accelerate the development of new anti-HIV drugs and speed up the life-saving process in fighting this fatal disease. Although a great number of HIV dynamic studies have been developed recently and the number is still increasing, the design of these studies has not been investigated carefully. Thus, it is badly needed to study the important design issues. In this paper we intend to address these issues based on practical considerations and provide some guidelines for practitioners.

An HIV dynamic model is a system of differential equations. The solution to the observed variable, the concentration of HIV RNA copies (viral load) in the differential equation model, can be simplified into a bi-exponential model (Wu and Ding 1999):

$$V(t) = P_1 e^{-d_1 t} + P_2 e^{-d_2 t}, \quad t \geq t_c, \quad (1)$$

where P_1 and P_2 are macroparameters which are functions of the coefficients of the “underlying” differential equations; d_1 and d_2 are called the first phase and second phase viral decay rates; t_c is the time (usually 1 to 3 days) required for disappearance of the ‘shoulder’ due to pharmacokinetic and intracellular delay (Perelson et al., 1996; Herz et al., 1996; Mittler et al., 1998; Ding and Wu, 1999). It has been shown that the viral decay

rates d_1 and d_2 can be approximated by (Ding and Wu, 1999),

$$d_1 = [1 - R_1(1 - e_1)]\delta, \quad (2)$$

$$d_2 = \left(1 - R_2 \frac{1 - e_2}{e_1}\right)\mu, \quad (3)$$

where R_1 and R_2 are the baseline reproduction/clearance ratios of virus from the two infected cell compartments (productively and long-lived/latently infected cells); e_1 and e_2 are treatment effects in these two compartments, respectively; and δ and μ are death rates of productively and long-lived/latently infected cells.

The potency e_1 and e_2 can not be determined directly from d_1 and d_2 because the parameters R_1 , R_2 , δ and μ are unknown. However, in a randomized clinical trial, we expect that these factors are homogeneously distributed in the treatment groups. Hence these monotone relationships between viral decay rates and treatment effects allow us to evaluate antiviral potency by comparing the viral decay rates, d_1 and d_2 . Notice that the viral decay rates, d_1 and d_2 , are potentially good markers for the potency (efficacy) of antiviral therapies which can be evaluated during the early stage of treatment (within several weeks after treatment initiation).

Other virological endpoints used in AIDS clinical trials, such as the success (or failure) rate of viral load suppression (below detectable levels) or time to viral load suppression, are good surrogate markers for long-term effectiveness of antiviral therapies which cannot be replaced by d_1 and d_2 , although the viral decay rates may be correlated with the long-term virological responses (Mueller et al., 1998).

To fit the bi-exponential model equation (1), two statistical methods, an individual nonlinear regression method and a nonlinear mixed-effects model method, have been proposed (Wu, Ding and DeGruttola, 1998; Wu and Ding, 1999; Ding and Wu, 2001). The nonlinear mixed-effects model method was shown to be favorable in most cases (Wu et al, 1998). Thus, in this paper we focus on the design issues raised under the setting of nonlinear mixed-effects models, which are briefly specified as follows.

Stage 1. within-subject variation in viral load measurement:

$$y_{ij} = \log_{10}(e^{p_{1i}-d_{1i}t_{ij}} + e^{p_{2i}-d_{2i}t_{ij}}) + \varepsilon_{ij}, \quad \varepsilon_i | \boldsymbol{\beta}_i \sim (\mathbf{0}, \mathbf{R}_i(\boldsymbol{\beta}_i, \boldsymbol{\xi})), \quad (4)$$

where y_{ij} is the log-transformation of the measurement of total viral load for the i th subject and at the j th time point t_{ij} , $i = 1, \dots, n; j = 1, \dots, m_i$. The viral dynamic parameters for the i th subject are denoted by $\boldsymbol{\beta}_i = [p_{1i}, d_{1i}, p_{2i}, d_{2i}]'$. The within-subject error $\varepsilon_i = [\varepsilon_{i1}, \dots, \varepsilon_{im_i}]'$ is assumed to have mean zero and variance-covariance matrix $\mathbf{R}_i(\boldsymbol{\beta}_i, \boldsymbol{\xi})$. Here the log-transformations of the viral load and parameters $p_{1i} = \ln(P_{1i})$, and $p_{2i} = \ln(P_{2i})$ are used to stabilize the variance and estimation algorithms.

Stage 2, between-subject variation:

$$p_{1i} = p_1 + b_{1i}, \quad p_{2i} = p_2 + b_{2i}, \quad (5)$$

$$d_{1i} = d_1 + b_{3i}, \quad d_{2i} = d_2 + b_{4i}, \quad (6)$$

where the population parameters are denoted by $\boldsymbol{\beta} = [p_1, d_1, p_2, d_2]'$. Random effect is $\mathbf{b}_i = [b_{1i}, \dots, b_{4i}]' \sim (\mathbf{0}, \mathbf{D})$.

In a clinical trial design, we need to determine the number of subjects (sample size n) and the number of measurements for each subject (including sampling schedules). A large number of samples from each subject would result in better estimates for some end-points (such as viral decay rates) which need multiple measurements. This will increase the power to detect the treatment differences in one way. However, when we increase the number of subjects (sample size n), it also definitely increases the power for hypothesis tests in another way. For fixed total cost, it is important to study the trade-off between the number of subjects and the number of measurements per subject, which is one of the important tasks of this paper. For convenience of clinical implementation, we assume that the sampling schedules are the same for all subjects. In the next section we introduce feasible sampling schedules and possible markers for antiviral potency. In Section 3, we

compare different sampling schedules and markers via intensive Monte Carlo simulations. In Section 4 we further study the two important markers identified from the simulation studies in Section 3. In particular, we find the optimal time to measure the viral load change as a marker for treatment potency. We also conduct a sensitivity analysis in Section 4 to investigate the robustness of our conclusions against parameter values in the viral dynamic models. We conclude our paper with conclusions, practical suggestions and some discussions in Section 5.

2. Sampling Schedules and Derived Test Variables

In a general nonlinear regression problem, an optimal design of sampling schedules can be obtained by optimizing a criterion which measures the goodness of the design. These criteria are usually based on the Fisher information matrix, which includes D-optimal, A-optimal, E-optimal, C-optimal and L-optimal criteria (Landaw, 1980; Atkinson and Donev, 1992). Among these criteria, D-optimality (determinant of the information matrix) is the most popular one and is widely used in practice (Mentre et al, 1997, Atkinson et al., 1993). The optimal design points are derived under the assumed parameter values. Therefore, they may not be optimal if the parameters are different from the assumed values. To solve this problem, other design strategies such as sequential design, Bayesian design and robust design have been developed in the past few years (Pronzato and Walter, 1988; DiStefano, 1981; Atkinson and Donev, 1992; Chaloner and Larntz, 1989; Atkinson, Chaloner, Herzberg and Juritz, 1993).

When there are k parameters to be estimated in the model ($k = 4$ in our model equation (1)), the D-optimal design usually results in replications of k distinct design points (Chapter 3, Bates and Watts, 1988). Generally these points serve as a reference in the practical design process. The actual designs, however, usually have more than k points since there are other needs than simple parameter estimation (e.g., to detect lack of fit and validate the nonlinear models). Also these optimal experimental points

may not be feasible to implement in practice, since the experimental conditions may be difficult to control in some studies, particularly in clinical trials. In this case, what the experimenters usually do is to specify a number of practically feasible candidate designs, and then evaluate them based on study objectives under various model assumptions and different parameter values. The one that can achieve study goals with less cost usually is selected for the actual experiment. We follow a similar approach in this paper.

There exist some limitations in the design of viral dynamic studies. One limitation is that the total volume of blood draws (number of samples) for viral load measurements for each individual is limited within a time period (e.g., not more than about 20 blood draws during the first month). Secondly, patients usually find it convenient to visit a clinic on the same day of the week. Thus, we cannot design too many points within a week (since some of these days may fall on weekends, when clinics may not be open). Due to these limitations and other practical considerations, we propose and consider the following sampling schedules:

- Schedule 1: Days 0, 1, 2, 3, 5, 7, 10, 14, 21, 28, 42, 56.
- Schedule 2: Days 0, 1, 3, 7, 10, 14, 21, 28, 56.
- Schedule 3: Days 0, 7, 14, 21, 28, 42, 56.
- Schedule 4: Days 0, 7, 14, 28, 56.

Schedule 1 is an ideal schedule which is frequent enough to accurately estimate both viral decay rates (d_1 and d_2). But this schedule requires 5 within-week clinic visits (Days 1, 2, 3, 5, and 10) and 4 extra weekly visits (Days 7, 14, 21, 42) beyond the standard sampling schedule currently used in AIDS clinical trials (Weeks 4, 8, etc.). Schedule 2 is designed to reduce the workload from Schedule 1, but the viral decay rates can still be reasonably estimated. However, this schedule still has 3 within-week clinic visits and 3 extra weekly visits. Schedule 3 drops all within-week samples, but has 4 extra weekly visits (Days 7,

14, 21, and 42). Schedule 4 further reduces the extra weekly visits to 2 (Days 7 and 14). Notice that, by comparing Schedule 2 and 4, we can evaluate whether the additional samples in the earlier stage (Days 1, 3, 10) can improve the estimates of viral decay rates and the power for treatment comparisons. Similarly, by comparing Schedule 3 and 4, we can evaluate whether the additional samples in the later stage (Days 21 and 42) can help. Another reason we choose these schedules for evaluation is that these schedules are proposed and currently used in some AIDS clinical trials. For these given candidate designs (schedules), both population and individual viral decay rates (d_1 and d_2) can be estimated via a nonlinear mixed-effect model approach (Wu, Ding and DeGruttola, 1998; Wu and Ding, 1999; Ding and Wu, 2001). More details of the nonlinear mixed-effect model and its applications can be found in Davidian and Giltinan (1995) and Vonesh and Chinchilli (1996).

We are interested in designing clinical trials for treatment comparisons. Hence, we will focus on evaluating the power of designs for treatment comparisons, i.e. for comparing viral decay rates, d_1 and d_2 . Since we can obtain the empirical Bayesian estimates of viral decay rates for individuals using the nonlinear mixed-effect model approach (Davidian and Giltinan, 1995), we may apply the two-sample tests such as t-test and Wilcoxon rank test to the individual estimates of viral decay rates to compare the antiviral potency between the treatments. Ding and Wu (2001) have shown that these two-sample tests (applied to the individual empirical Bayesian estimates) are valid and robust for our situations.

Besides the viral decay rates (d_1 and d_2), there are several other quantities (markers) which have been used to evaluate the efficacy of anti-HIV therapies (Weinberg and Lagakos, 2000). These quantities (markers) include viral load change (VLC) from baseline to a prespecified time (say, Day 7) in a \log_{10} scale and the area under the curve (AUC) of viral load trajectory. In Diggle, Liang and Zeger (1994), these markers, including viral decay rates, are called “derived variables”. In practice, we may use the \log_{10} AUC adjusted for baseline, i.e. it is calculated using the trapezoidal rule on the curve

of $\log_{10}[V(t)/V(0)]$, where $V(0)$ is the baseline viral load and $V(t)$ is the viral load at time t . We evaluate these markers and viral decay rates, as well as the trade-off between the number of measurements for each subject and the number of subjects, via extensive Monte Carlo simulations in the next section.

3. Design Evaluation via Monte Carlo Simulations

3.1 Simulation methods

Based on preliminary results from AIDS Clinical Trials Group Protocol 315 (Wu et al., 1999a), we design simulation experiments to evaluate the aforementioned sampling schedules and the markers (derived variables). The bi-exponential model (1) or (4) is an approximation model to a complete viral dynamic model, see Wu and Ding (1999). To consider the model approximation error in our simulation studies, we generate the simulation data from the following tri-exponential model, which is more accurate since it includes the “shoulder” effect (Perelson et al., 1996; Herz et al., 1996; Mittler et al., 1998; Ding and Wu, 1999).

$$y_{ij} = \log(e^{p_{0i}-d_{0i}t_{ij}} + e^{p_{1i}-d_{1i}t_{ij}} + e^{p_{2i}-d_{2i}t_{ij}}) + \varepsilon_{ij}. \quad (7)$$

The between-subject variations of parameters are similarly defined as in equations (5) and (6). However, we still fitted the simulated data to the more practical bi-exponential model (4).

In almost all cases, there are not enough data on the initial “shoulder” (the first one to three days of treatment) of the viral load trajectory to analyze the “shoulder” effect. Ding and Wu (2000) discussed four different methods to deal with the ‘shoulder’ effect, that is, to estimate parameters in model (4) from data generated from model (7). Among the four methods, the so-called SIMPLE method directly fits the bi-exponential model (4) with all data. That is, we just pretend that the “shoulder” effect does not exist in the data. Compare with other methods, the SIMPLE method produces the estimates with a larger

bias but a smaller variance (Ding and Wu, 2000). However, for the decay rate comparison between two treatment arms considered in this paper, since the estimation bias appeared in both treatment arms can be cancelled out, the bias does not affect the validity of the tests. In fact, the SIMPLE method gives a slightly higher power compared to other three methods (data not shown), presumably due to its smaller estimation variance. When we applied the other three methods to our cases, the results are quite similar to the SIMPLE method. Thus, we only report the results from the SIMPLE method in the following.

We chose model parameters based on several studies reported in literature, including Wu et al. (1999a), Luzuriaga et al. (1999), and Perelson et al. (1996). Here are the parameters that we used in our simulation experiments:

- The measurement error variance $\sigma^2 = 0.04$.
- The population parameters are $p_0 = 11, p_1 = 12, p_2 = 7.5, d_0 = 3, d_1 = 0.5, d_2 = 0.03$. Note that the unit for $d_k, k = 0, 1, 2$ is day^{-1} , and $p_k = \ln P_k, k = 0, 1, 2$, where the unit for P_k is the number of virions per ml plasma. From clinical studies (Ho et al., 1995; Wei, et al., 1995; Perelson et al., 1996 and 1997; Wu et al., 1999a), the estimated parameter d_1 ranges from 0.2 to 1.2, d_2 ranges from 0.001 to 0.2.
- The between-subject variance-covariance matrix for the parameters,

	p_0	d_0	p_1	d_1	p_2	d_2
p_0	2.3	-0.2	1.8	-0.05	1.4	0.01
d_0	-0.2	0.5	-0.2	-0.01	-0.1	-0.0005
p_1	1.8	-0.2	1.8	-0.01	1.6	0.01
d_1	-0.05	-0.01	-0.01	0.014	0	0.0005
p_2	1.4	-0.1	1.6	0	1.8	0.01
d_2	0.01	-0.0005	0.01	0.0005	0.01	0.0001

Our objective is to compare two treatments, say A and B. In our simulation experiments, without loss of generality we assumed that the population viral decay rates (d_1 and

d_2) in the second group (Treatment B) was higher than the first group (Treatment A). We simulated the cases that both d_1 and d_2 in Treatment B were simultaneously higher by 0%, 10%, 30%, 50% respectively compared to Treatment A.

We simulated the data for the four schedules introduced in Section 2 respectively, and 400 replication runs are used for each schedule (the simulation sample size of 400 runs is selected by considering both simulation error and computational limitation). The empirical powers of the Wilcoxon test (one-sided test at significance level 0.05) using markers d_1 for different schedules and using the viral load change (VLC) at Week 1 (i.e., the VLC at Day 7 or $VLC(7)$) are obtained and plotted in Figure 1 (The results from the corresponding two-sided tests are similar, which are not reported due to space limitation).

The power curve in Figure 1 are classified by the total number of subjects n (each group has $n/2$ subjects). For a fixed value of n , of course a more frequent schedule gives more power to detect the difference between treatment groups since more data is collected. Therefore, in order to compare the efficiency of different sampling schedules, in Figure 2 we plot the power function curves of different schedules versus the total number of measurements required (the product of the number of measurements for each subject and the number of subjects).

3.2 Simulation results

The empirical type I errors in the simulations range from 0.0275 to 0.0675, showing no significant deviation from the nominal level $\alpha = 0.05$ considering the simulation variation. This confirms the validity of the tests.

Place of Figure 1

From Figure 1 we can see that, when the number of subjects and the difference in

viral decay rates (d_1) between groups are large enough, all schedules and markers can detect the treatment difference with enough power, and the power function curves tend to converge together toward 1. If the treatment difference is too small, all the schedules and markers give no power. In the intermediate range of treatment difference (between 10% and 30%), different schedules and markers will make a significant difference in the power. Not surprisingly, a more frequent schedule results in a higher power for a fixed n number of subjects. However, Schedule 3 is not necessarily more powerful than Schedule 4 since the extra observations in Schedule 3 are located during the second phase decay period which may not help much to improve estimation of d_1 . But schedule 3 is more powerful at detecting the difference of d_2 between treatments (data not shown).

In terms of total measurements (cost), however, the most frequent schedule is the least efficient design as shown by Figure 2. In fact, the sparsest schedule (Schedule 4) is most efficient. Also notice that Schedule 2 is better than Schedule 3 in efficiency. This indicates that the early time points on Day 1 and 3 are very important.

Place of Figure 2

Notice that the marker d_1 can consistently dominate the marker $VLC(7)$ in power only when the most frequent schedule (Schedule 1) is used. So the marker $VLC(7)$ is very powerful and efficient. It requires only two measurements on each subject and can provide a conclusion within a week. But the power of $VLC(7)$ depends heavily on the underlying true parameters (to be discussed later). On the other hand, using the marker d_1 under Schedule 1 can improve the power for detecting the difference in antiviral potency between treatment groups by more than 20% in some cases. Our simulation results also show that the marker d_2 and AUC are generally not powerful (data not shown). Enough power can be achieved only when there is a large between-group difference ($> 50\%$) and

a large sample size ($n > 100$). The comparison results in marker d_2 between schedules are similar to those in d_1 .

4. Marker Comparisons and Sensitivity Analysis

Since productively infected cells are the major resource for virus production (almost 99% of virions are produced from this compartment according to Perelson et al., 1997), it is most important that a new antiviral therapy should be potent in this virus compartment. We will focus the following discussion on evaluating the treatment potency in this compartment (e_1 in Equation (2)). Ding and Wu (1999) established d_1 as a good marker for e_1 . However, in the last section, we observed that other markers such as $VLC(7)$ can also be very effective for evaluating treatment potency. The VLC marker only requires two measurements on each subject. It is worth further studying these two different kinds of markers from different perspectives.

To evaluate different markers, several factors need to be considered: (1) whether the marker is monotonically related to the treatment potency, and a linear relationship is the ideal; (2) whether the marker can be measured or estimated accurately, and thus can be used to identify treatment differences with a high power; (3) whether the power of the marker is robust against the values of model parameters. We address these issues in this section.

4.1 Comparison between d_1 and $VLC(t)$ as Markers of Treatment Potency.

As shown in Ding and Wu (1999), the marker d_1 is almost linearly related to treatment potency e_1 , except for extremely weak treatments. It is difficult to obtain a similar analytic relationship between $VLC(t)$ and treatment potency (say, e_1 in model (2)). We will investigate the relationship using a numerical solution. Using the original differential equation model of viral dynamics in Ding and Wu (1999) with parameter values of $\delta = 1.0$, $\mu = 0.03$, $R_1 = 0.99$ and $R_2 = 0.01$, we obtain the relationships between d_1 , $VLC(7)$, $VLC(14)$ and $VLC(28)$ versus treatment potency e_1 numerically, which are shown in

Figure 3. We scaled the $VLC(t)$ markers in the plot by time t in order to make them comparable.

Place of Figure 3

From Figure 3, we can see that both d_1 and $VLC(7)$ are almost linearly related to e_1 , and thus are good markers for e_1 . However, $VLC(14)$ and $VLC(28)$ exhibit a nonlinearity in the relationship with e_1 due to the effect of the second compartment of long-lived/latently infected cells. Thus, some information will be lost if $VLC(14)$ or $VLC(28)$ is used as a marker of e_1 .

Since both d_1 and $VLC(7)$ are good markers for potency e_1 , we would like to know which one can better distinguish treatment potencies in practice. Weinberg and Lagakos (2000) have proposed an asymptotic relative efficiency (ARE) of rank tests between different surrogate markers. The ARE can be used to assess the asymptotic performance of different markers. Under model equations (4), (5) and (6) with a normality assumption, we can numerically calculate the $ARE(d_1 : VLC(t))$ with the parameter values given in Section 3. See details in Appendix A.1. Figure 4 plotted the calculated AREs at time $t = 1, 2, \dots, 14$ days, assuming perfect measurements on both markers d_1 and VLC . We can see that the ARE of d_1 is lower ($ARE < 1$) than the $VLC(t)$ when time t is between Day 2 and Day 7. On Day 3, the VLC reaches its highest efficiency [$ARE(d_1:VLC(3))=0.84$]. The ARE of d_1 is higher ($ARE > 1$) before Day 2 due to the “shoulder” effect. After Day 8, the phase two viral decay starts to affect the efficiency of VLC , which results in a lower efficiency in $VLC(t)$. On Week 2 (Day 14), the asymptotic efficiency of d_1 can be 4.5 times higher than $VLC(t)$.

Place of Figure 4

The above ARE results, however, are based on the assumption of perfect measurements (the measurement error is 0) on both markers. With the presence of measurement errors in practice, both d_1 and the $VLC(7)$ markers have to be estimated from data. A frequent sampling schedule allows more accurate estimation for d_1 but does not affect the accuracy of $VLC(7)$ estimation. Hence, as we have shown in section 3, if the repeated measurements for each subject are frequent enough, marker d_1 is superior to the $VLC(7)$ marker for finite sample size.

As shown above, the highest ARE of VLC can be achieved at Day 3 assuming perfect measurements, indicating the existence of an optimal day on which to measure viral load change. However, in practice, measurement error is not avoidable. We would like to find the optimal VLC time with measurement errors. If we assume the measurement error to be the same as in our simulation example in Section 3 ($\sigma^2 = 0.04$), we can obtain the optimal VLC time to be $t = 6.87$ days by minimizing $ARE(d_1 : \widehat{VLC}(t))$ in time t . This optimal time is about 4 days later than the result without considering measurement errors.

The foregoing results based on ARE, however, are only valid asymptotically (when the number of subjects is large enough). To exam if the results hold for small and medium sample sizes, we carried out a simulation experiment on the design in Section 3. Under the assumptions of sample size (total number of subjects) of $n = 30$ and $n = 200$, and viral decay rate differences of 10% and 30% between two treatment groups respectively, we simulated the power of $VLC(t)$ for $t = 1, 2, \dots, 14$ days (3000 simulation runs were used). The results are given in Figure 5.

Place of Figure 5

We can see, from these results, that the highest power can be achieved between Day 6 and Day 8, similar to the theoretical optimal time from ARE calculations. Considering implementation convenience, the VLC at Day 7 (Week 1) may be the best choice based on this simulation example. However, when the parameter values in the viral dynamic model vary, this optimal VLC time as well as other conclusions about the markers may change. A sensitivity analysis will be carried out in the next subsection.

4.2. Sensitivity Studies

The results in the previous sections were obtained based on the parameter values estimated from ACTG 315 (Wu et al., 1999a; Wu and Ding, 1999). However, these parameter values may change for different studies due to different patient populations and different antiviral treatments. In this subsection, we evaluate the sensitivity of $VLC(t)$ and d_1 for different situations.

First we study the sensitivity of the optimal time of $VLC(t)$ for different parameter values. In order to avoid intensive computations in the sensitivity analysis, we need to simplify the marker $VLC(t)$. When the measurement error is considered, the power for identifying a fixed difference in treatment potency only depends on the coefficient of variation (CV) for a given sample size. Thus, to study the sensitivity of the optimal time of $VLC(t)$, we will derive the CV as a function of time t and minimize $CV(t)$, instead of maximizing the power of $VLC(t)$ directly. See detailed description of this method in Appendix A.2.

Mathematical analysis can be used to identify important factors (parameters) for sensitivity analyses (See details in Appendix A.3). The identified major factors for the marker $VLC(t)$ include σ , Δ , $p_1 - p_0$, $p_1 - p_2$, and d_1 (Appendix A.3), where Δ is a

percentage difference in the first viral decay rates between the two treatment groups. That is, if the first group has a first decay rate of d_1 , then the first decay rate of the second group will be $(1 + \Delta)d_1$. Based on literature survey, we decided to vary d_1 from 0.2 to 1.1, $p_1 - p_2$ from 3 to 5.3 (i.e., the ratio of the viral reproduction between long-lived and productively infected cells varies from 5% to 0.5%), and $p_1 - p_0$ from 0.85 to 4.6 (i.e. the “shoulder” effect accounts for about 30% to 1% viral decay). We vary σ from 0.15 to 0.25, and Δ from 0 to 100%. The other parameters are set as the same as those in the simulation experiment in Section 3. We then vary the value of each of the parameters σ , Δ , $p_1 - p_0$, $p_1 - p_2$, and d_1 in the above range, and obtain the optimal time of $VLC(t)$ by minimizing $CV(t)$. Figure 6 shows the plot of the optimal VLC times versus the varying values of these important parameters.

Place of Figure 6

As we can see from Figure 6, the values of σ and $p_1 - p_0$ have little effect on the optimal VLC time, and Δ and $p_1 - p_2$ only have a modest effect. However, the optimal VLC time is very sensitive to the value of d_1 . As $p_1 - p_2$ increases (the domination of the productively infected cell compartment or the first phase decay increases), the optimal VLC time is delayed. As the treatment difference (Δ) is larger, the optimal VLC time moves earlier. As d_1 increases, the optimal VLC time moves earlier rapidly. Thus, the VLC measurement time needs to be adjusted based on prior information on d_1 .

Secondly, we study the sensitivity of markers d_1 and $VLC(t)$. In particular, we would like to see if the conclusions about d_1 and $VLC(7)$ reached in Section 3 changes when the parameter values vary. we carried out similar simulation experiments for marker d_1 based on Schedule 2 and 4 as well as $VLC(7)$. For a fixed $\Delta = 30\%$, the sensitivity analysis results for parameters d_1 and $p_1 - p_2$ are reported in Table 1. The powers for

identifying a 30% treatment difference in antiviral potency (d_1) are obtained for a small and a large value of d_1 and $p_1 - p_2$ for a given sample size of 60 subjects. As we increase d_1 from 0.5 to 1.1, the power only drops from 98.5% to 86.8% for marker d_1 with a frequent sampling schedule (Schedule 2). However, for d_1 with sparse measurement (Schedule 4) and $VLC(7)$, the power drops to 14.5%. As we decrease d_1 from 0.5 to 0.2, the marker d_1 with both schedule 2 and 4 still has power of above 60%, but the power of $VLC(7)$ drops to below 50%. As we have shown in Figure 6, the power increases as $p_1 - p_2$ increases. But as $p_1 - p_2$ decreases from 4.5 to 3, the marker d_1 with a frequent schedule and $VLC(7)$ have power of about 60%, the power of marker d_1 with sparse schedule drops below 50%. Thus d_1 with a frequent measurement schedule (Schedule 2) is robust to the parameter values, but d_1 with sparse measurement (Schedule 4) and VLC are sensitive.

Table 1: The empirical powers when parameter values vary ($\Delta = 30\%$, $n = 60$, $\alpha = 0.05$).

marker	Parameter values				
	$d_1 = 0.5$	$d_1 = 0.2$	$d_1 = 1.1$	$d_1 = 0.5$	$d_1 = 0.5$
	$p_1 - p_2 = 4.5$	$p_1 - p_2 = 4.5$	$p_1 - p_2 = 4.5$	$p_1 - p_2 = 3$	$p_1 - p_2 = 5.3$
d_1 (Schedule 2)	98.5%	60.5%	86.8%	62.2%	99.5%
d_1 (Schedule 4)	91.0%	64.0%	14.5%	47.3%	95.8%
$VLC(7)$	94.3%	49.0%	14.5%	59.9%	96.0%

5. Conclusions, Suggestions and Discussion

Traditionally, the design of experiments is formulated as an optimization problem. That is, to maximize (or minimize) a design criterion such as D-optimal, A-optimal, E-optimal etc. under the assumption of a proposed model. This kind of optimal design may result in infeasible designs for implementation in practice, especially for clinical trial designs. In many design problems, there only exists a finite number of feasible candidate designs. By

taking the advantage of powerful computers, we can evaluate the finite number of designs from many different perspectives, or even directly optimize our design goals, such as the power for identifying the treatment difference. We believe that the computer simulation is a very useful tool for designing clinical trials.

In this paper, we use intensive Monte Carlo simulations, combined with analytic analysis, to study the different markers and different designs for assessing the potency of anti-HIV therapies in AIDS clinical trials. We obtain the following important results: (1) the viral decay rate (d_1) in viral dynamic models is a robust marker for antiviral potency if a frequent measurement schedule of viral load is used, but it is very costly; (2) the viral decay rate, d_1 , with a sparse (weekly in the first month) measurement schedule is very cost-effective, but a large sample size (number of subjects) is needed; (3) the marker of viral load change from baseline (VLC) is an efficient marker if the optimal measurement time is captured, but the VLC is not robust; (4) In evaluating potent antiviral therapies, the viral load change at Week 1 or $VLC(7)$ is better than the VLC at later weeks, such as Week 2 and Week 4, which are currently used in AIDS clinical trials.

Although it is costly and requires frequent clinical visits by patients, marker d_1 with a frequent measurement schedule is still useful. One reason is that it can not only provide information regarding the potency of the therapy, but also can give an accurate estimate for viral/cellular dynamics such as half-life of infected cells (Perelson et al., 1996, 1997; Wu et al., 1997, 1999a; Luzuriaga et al., 1999). Secondly, the frequent measurements may help to validate the viral dynamic models used for analysis. However, this intensive study is limited to a very small sample size due to its cost and accrual difficulty.

The d_1 marker with a sparse measurement schedule and VLC are effective markers for antiviral potency. We suggest using weekly viral load measurements (or only Week 1, 2 and 4) during the first month of treatment. If only one measurement is allowed in a large study, we suggest using Week 1 or $VLC(7)$ to evaluate the potency of the therapy. A combination of a small substudy with frequent measurements and a large main study

with at least one measurement at day 7 can be used to gain benefits from both the d_1 and VLC markers.

We also want to point out that some of the results from our simulation and analytical studies are based on a selected set of parameter values. These results can not be extended to general cases. One may use the methods as proposed above to conduct similar simulation and analytical studies based on the prior information on the parameter values to determine an appropriate design for a particular study. The sensitivity results in Section 4.4 are also helpful to modify the design for different situations.

Notice that both the d_1 and VLC markers are useful for evaluating the early antiviral activities (potency) of a therapy. Other viral-load-based surrogate markers such as proportion of patients with virological failure (success) and durability of viral suppression (time to virological failure) are good to evaluate the long-term effectiveness of antiviral therapies (commonly used in Phase II and III trials). A strong correlation between these virological markers and AIDS clinical endpoints has been established from clinical studies such as Mellors et al., 1995; O'Brien et al., 1996a,b; Mellors et al., 1997. A significant correlation between the short-term responses (including d_1) and the long-term responses has been reported by Mueller et al. (1998). However, it is also very important to use markers d_1 and VLC to assess whether a new therapy is potent enough to be worthy for further evaluation. Thus, d_1 and VLC can be very efficient in Phase I/II clinical trials. Also they can be very useful in finding the therapeutic dosage for a new antiviral agent. By using d_1 and VLC with a careful design, we may accelerate the development of new antiviral drugs.

Appendix

A.1 Asymptotic Relative Efficiency (ARE) between Markers d_1 and $VLC(t)$.

Based on the results in Weinberg and Lagakos (2000), the ARE of d_1 with respect to $VLC(t)$ under the Wilcoxon rank test can be written as

$$ARE(d_1 : VLC(t)) = \left(\frac{\int_{x=-\infty}^{\infty} F_{d_1}(x) dF_{d_1}(x)}{\int_{x=-\infty}^{\infty} F_{VLC(t)}(x) dF_{VLC(t)}(x)} \right)^2, \quad (8)$$

where $F_{d_1}(\cdot)$ and $F_{VLC(t)}(\cdot)$ denote the cumulative distribution functions (CDFs) of d_1 and $VLC(t)$, respectively. If we assume that d_{1i} follows the normal distribution $N(0.5, 0.014)$ as in our simulation example in Section 3, then, $\int_{x=-\infty}^{\infty} F_{d_1}(x) dF_{d_1}(x) = -\frac{1}{2\sqrt{\pi}0.014}$. However, the integral in the denominator of $ARE(d_1 : VLC(t))$ is intractable analytically. We evaluate this integral using a Monte Carlo sampling method (of size 10,000) using parameter values given in Section 3.

To find the optimal time for measuring VLC , we can minimize $ARE(d_1 : \widehat{VLC}(t))$ or equivalently maximize $\int_{x=-\infty}^{\infty} F_{\widehat{VLC}(t)}(x) dF_{\widehat{VLC}(t)}(x)$ in time t .

A.2 Computing the Optimal Time of $VLC(t)$ in the Sensitivity Analysis.

It is costly to compute the empirical power of $VLC(t)$ at many time points in simulations. Hence, we used an alternative method to find the optimal time for $VLC(t)$ by minimizing a coefficient of variation function $CV(t)$. A precise mathematical description of this method is given as follows.

Let $V_i(t)$ denote the viral load at treatment time t for subject i , and denote $y_i(t) = \log_{10} V_i(t) + e_i(t)$ as the observed $V_i(t)$ in \log_{10} scale. Thus, $VLC_i(t) = \log_{10} V_i(0) - \log_{10} V_i(t)$, and the observed $VLC(t)$ can be written as

$$\widehat{VLC}_i(t) = y_i(0) - y_i(t) = \log_{10} V_i(0) + e_i(0) - \log_{10} V_i(t) - e_i(t) = VLC_i(t) + e_i(0) - e_i(t).$$

For simplicity, we assume $e_i(t)$ to be *iid* with normal distribution $N(0, \sigma^2)$, and $e_i(t)$ and $VLC_i(t)$ are independent. We define the coefficient of variation (CV) of the difference in $VLC(t)$ between treatment A and B as

$$CV(t) = \frac{SD(t)}{|\mu(t)|},$$

where $\mu(t)$ is the expectation of the difference of the observed $VLC(t)$ between the two treatments and the $SD(t)$ is its standard deviation. Assuming Treatment A and B to be

independent, then, we have

$$\begin{aligned}
\mu(t) &= E[\widehat{VLC}_i^A(t) - \widehat{VLC}_i^B(t)] = E[VLC_i^A(t)] - E[VLC_i^B(t)] \\
SD(t)^2 &= Var[\widehat{VLC}_i^A(t) - \widehat{VLC}_i^B(t)] \\
&= Var[\widehat{VLC}_i^A(t)] + Var[\widehat{VLC}_i^B(t)] \\
&= Var[VLC_i^A(t)] + 2\sigma^2 + Var[VLC_i^B(t)] + 2\sigma^2 \\
&= Var[VLC_i^A(t)] + Var[VLC_i^B(t)] + 4\sigma^2,
\end{aligned}$$

where $E[VLC_i(t)]$ and $Var[VLC_i(t)]$ are the inter-subject mean and variance of $VLC(t)$. For t-test, maximizing the power for identifying a fixed difference in treatment effect is equivalent to minimizing $CV(t)$ in treatment time t . The equivalence holds for Wilcoxon rank test also if we assume that the inter-subject variance of $VLC(t)$ is the same in group A and B. The Monte Carlo method is used to evaluate $E[VLC_i(t)]$ and $Var[VLC_i(t)]$ based on equation (10) in Appendix A.3.

A.3 Important Factors in Sensitivity Analyses.

There are too many parameters in the model, which makes the sensitivity analysis difficult if all parameters are considered. However, if we can identify a few important factors that affect $VLC(t)$, we can focus our sensitivity analyses on these factors. Based on equation (7), we can write $VLC(t)$ as

$$\begin{aligned}
VLC_i(t) &= \log\left(\frac{e^{p_{0i}-d_{0i}t} + e^{p_{1i}-d_{1i}t} + e^{p_{2i}-d_{2i}t}}{e^{p_{0i}} + e^{p_{1i}} + e^{p_{2i}}}\right) \\
&= \log\left(\frac{e^{(p_{0i}-p_{1i})-(d_{0i}-d_{1i})t} + 1 + e^{(p_{2i}-p_{1i})+(d_{1i}-d_{2i})t}}{e^{(p_{0i}-p_{1i})} + 1 + e^{(p_{2i}-p_{1i})}}\right) - d_{1i}t.
\end{aligned} \tag{9}$$

From previous studies (Perelson et al., 1996; Ding and Wu 1999; Mittler, et al, 1999), we notice that $d_{0i} \gg d_{1i}$ and $d_{0i} > 3.0$. The term, $e^{(p_{0i}-p_{1i})-(d_{0i}-d_{1i})t}$ in equation (9), is negligible quickly for $t > 0$. Also, we notice that $d_{1i} \gg d_{2i}$, thus, d_{2i} is negligible. Hence

$$VLC_i(t) \approx \log\left(\frac{1 + e^{(p_{2i}-p_{1i})+d_{1i}t}}{e^{(p_{0i}-p_{1i})} + 1 + e^{(p_{2i}-p_{1i})}}\right) - d_{1i}t, \tag{10}$$

which means that the three major factors (variables) affecting the marker $VLC_i(t)$ are

$p_{0i} - p_{1i}$, $p_{2i} - p_{1i}$, and d_{1i} . Therefore, among the six parameters d_0 , p_0 , d_1 , p_1 , d_2 and p_2 , we can focus on $p_0 - p_1$, $p_2 - p_1$ and d_1 in the sensitivity analysis.

ACKNOWLEDGMENT

This work was supported by NIAID/NIH grants No. R29 AI43220, RO1 AI45356 and U01 AI38855. We would like to thank the team of AIDS Clinical Trial Group Protocol 315 for permission to use their HIV-1 RNA data, and we are grateful to the four referees for constructive and helpful comments.

REFERENCES

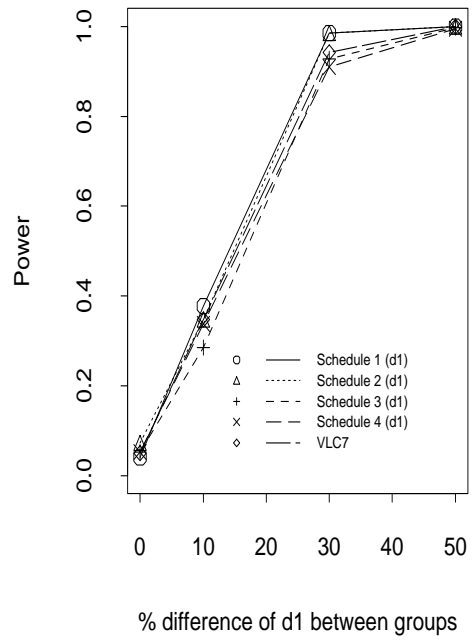
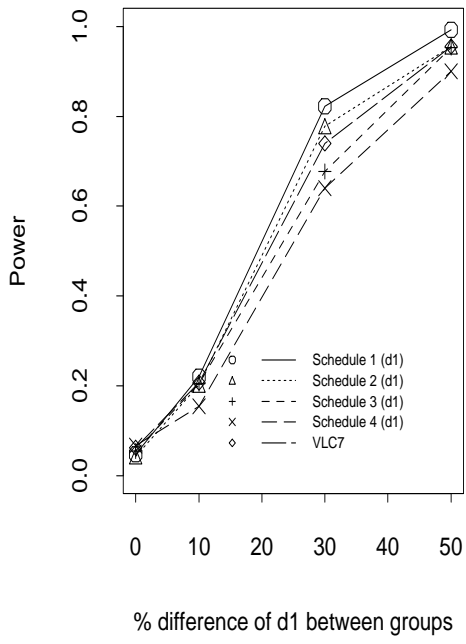
1. Atkinson, A.C., Chaloner, K., Herzberg, A.M. and Juritz, J. (1993), 'Optimum Experimental Designs for Properties of a Compartmental Model', *Biometrics*, 49 (1), 325-337.
2. Atkinson, A.C. and Donev, A.N. (1992), *Optimum Experimental Designs*, Oxford University Press Inc, New York.
3. Bates, D.M. and Watts, D.G. (1988), *Nonlinear Regression Analysis and Its Applications*, John Wiley & Sons, New York.
4. Chaloner, K. and Larntz, K. (1989), 'Optimal Bayesian design applied to logistic regression experiments', *Journal of Statistical Planning and Inference*, 21, 191-208.
5. Davidian, M. and Giltinan, D.M. (1995), *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, New York.
6. Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994), *Analysis of Longitudinal Data*, Oxford Science Publications, New York.
7. Ding, A.A. and Wu, H. (1999), 'Relationships between antiviral treatment effects and biphasic viral decay rates in modeling HIV dynamics', *Mathematical Biosciences*, 160 (1), 63-82.
8. Ding, A.A. and Wu, H. (2000), 'A comparison study of models and fitting procedures for biphasic viral dynamics in HIV-1 infected patients treated with antiviral therapies', *Biometrics*, 56 (1), 293-300.
9. Ding, A.A. and Wu, H. (2001), 'Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models', *Biostatistics*, in press.

10. DiStefano, J.J. (1981), 'Optimized blood sampling protocols and sequential design of kinetic experiments', *American Journal of Physiology*, 240, R259-265.
11. Essunger, P., Markowitz, M., Ho, D.D., Perelson, A.S. (1997), 'Efficacy of drug combination and dosing regimen in antiviral therapy', *The First International Workshop on HIV Drug Resistance, Treatment Strategies and Eradication*, Abstract 73.
12. Herz, A.V.M., Bonhoeffer, S., Anderson, R.M., May, R.M., and Nowak, M.A. (1996), 'Viral dynamics in vivo: Limitations on estimates of intracellular delay and virus decay', *Proc. Natl. Acad. Sci. USA*, 93, 7247-7251.
13. Ho, D.D. (1998), 'Novel approaches for the evaluation of new drugs: Approaches using viral dynamics', *The 5th Conference on Retroviruses and Opportunistic Infections*. Chicago, IL, February, 1998.
14. Ho, D.D., Neumann, A.U., Perelson, A.S., Chen, W., Leonard, J.M., and Markowitz, M. (1995), 'Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection', *Nature*, 373, 123-126.
15. Landaw, E.M. (1980), Optimal Experimental Design for Biologic Compartmental Systems with Applications to Pharmacokinetics. Ph.D. Dissertation, University of California at Los Angeles.
16. Luzuriaga, K., Wu, H., and McManus, M., Britto, P. Borkowsky, W., Burchett, S., Smith B., Mofenson, L., Sullivan, J.L., and the PACTG 356 Investigators (1999), 'Dynamics of HIV-1 replication in vertically-infected infants', *Journal of Virology*, 73, 362-367.
17. Mellors JW, Kingsley LA, Rinaldo CR, et al. (1995), 'Quantitation of HIV-1 RNA in plasma predicts outcome after seroconversion', *Ann Intern Med*, 122, 573-9.

18. Mellors JW, Munoz A, Giorgi JV, et al. (1997), 'Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection', *Ann Intern Med*, 126, 946-54.
19. Mentre, F., Mallet, A. and Baccar, D. (1997), 'Optimal Design in Random-effects Regression Models', *Biometrika*, 84 (2), 429-442.
20. Mittler, J.E., Sulzer, B., Neumann, A.U., and Perelson, A.S. (1998), 'Influence of delayed viral production on viral dynamics in HIV-1 infected patients', *Mathematical Biosciences*, 152, 143-163.
21. Mueller, B.U., Zeichner, S.L., Kuznetsov, V.A., Heath-Chiozzi, M., Pizzo, P.A., and Dimitrov, D.S. (1998), 'Individual prognoses of long-term responses to antiretroviral treatment based on virological, immunological and pharmacological parameters measured during the first week under therapy', *AIDS*, 12, F191-F196.
22. O'Brien TR, Blattner WA, Waters D, et al. (1996a) 'Serum HIV-1 RNA levels and time to development of AIDS in the Multicenter Hemophilia Cohort Study', *JAMA*, 276, 105-10.
23. O'Brien WA, Hartigan PM, Martin D, et al. (1996b) 'Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression of AIDS', *New Engl J Med*, 334, 426-31.
24. Perelson, A.S., Neumann, A.U., Markowitz, M., Leonard, J.M., and Ho, D.D. (1996), 'HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time', *Science*, 271, 1582-1586.
25. Perelson, A.S., Essunger, P., Cao, Y., Vesanen, M., Hurley, A., Saksela, K., Markowitz, M., and Ho, D.D. (1997), 'Decay characteristics of HIV-1-infected compartments during combination therapy', *Nature*, 387, 188-191.

26. Pronzato, L. and Walter, E. (1988), 'Robust experiment design via maximin optimization', *Mathematical Biosciences*, 89, 161-176.
27. Vonesh, E. F., and Chinchilli, V. M. (1996), *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, New York: Marcel Dekker, Inc.
28. Wei, X., Ghosh, S.K., Taylor, M.E., Johnson, V.A., Emini, E.A., Deutsch, P., Lifson, J.D., Bonhoeffer, S., Nowak, M.A., Hahn, B.H., Saag, M.S., and Shaw, G.M. (1995), 'Viral dynamics in human immunodeficiency virus type 1 infection', *Nature*, 373, 117-122.
29. Weinberg, J., and Lagakos, S.W. (2000), 'Linear rank tests under general alternatives, with application to summary statistics computed from repeated measures data', *Journal of Statistical Planning and Inference*, in press.
30. Wu, H., Kuritzkes, D.R., Clair, M.S., Kessler, H., Connick, E., Landay, A., Heath-Chiozzi, M., Rousseau, F., Fox, L., Spritzler, J., Leonard, J.M., McClernon, D.R., and Lederman, M.M. (1997), 'Interpatient variation of viral dynamics in HIV-1 infection: Modeling results of AIDS Clinical Trials Group Protocol 315', *The First International Workshop on HIV Drug Resistance, Treatment Strategies and Eradication*, Abstract 99,66-67.
31. Wu, H., Ding, A.A., and DeGruttola, V. (1998), 'Estimation of HIV dynamic parameters', *Statistics in Medicine*, 17, 2463-2485.
32. Wu, H. and Ding, A.A. (1999), 'Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials', *Biometrics*, 55, 410-418.
33. Wu, H., Kuritzkes, D.R., and McClernon, D.R., Kessler, H., Connick, E., Landay, A., Spear, G., Heath-Chiozzi, M., Rousseau, F., Fox, L., Spritzler, J., Leonard,

- J.M., and Lederman, M.M. (1999a), 'Characterization of viral dynamics in HIV-1-infected patients treated with combination antiretroviral therapy: relationships to host factors, cellular restoration and virological endpoints', *Journal of Infectious Diseases*, 179, 799-807.
34. Wu, H., Ruan, P., Ding, A.A., Sullivan, J.L., and Luzuriaga, K. (1999b), 'Inappropriate model-fitting methods may lead to significant underestimates of viral decay rates in HIV dynamic studies', *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 21 (5), 426-427.



(c) Total Number of Subjects = 100

(d) Total Number of Subjects = 200

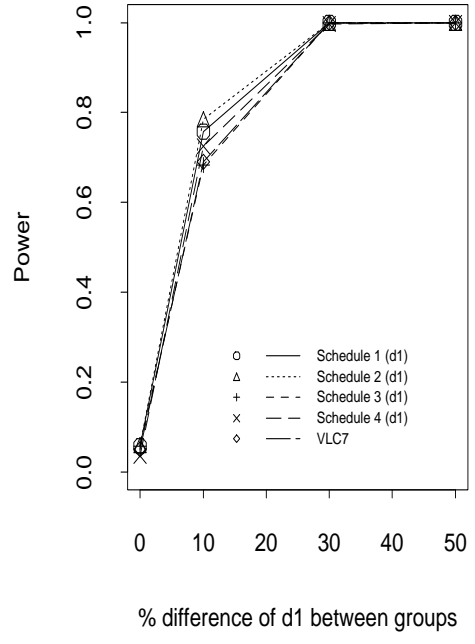
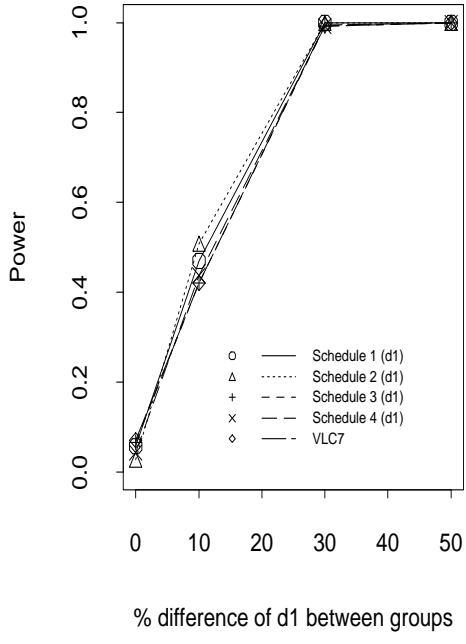
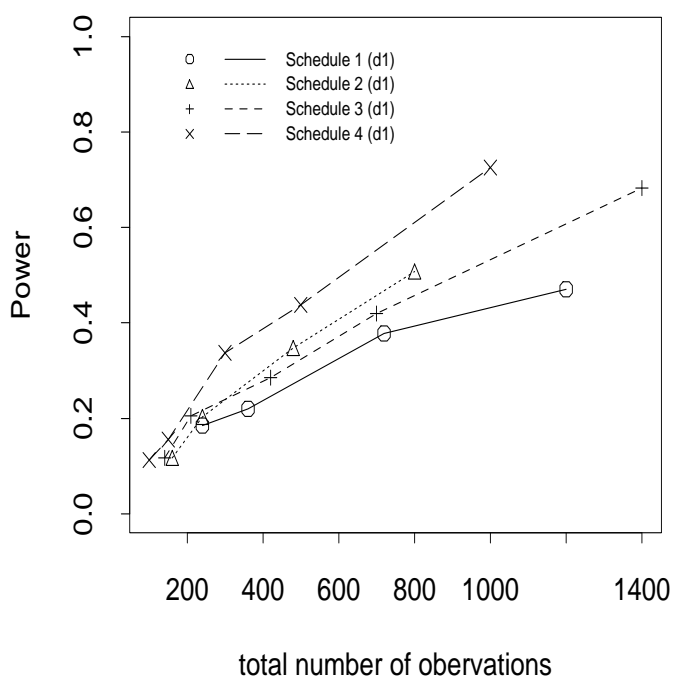


Figure 1: The empirical power of one-sided Wilcoxon test using markers d_1 for different measurement schedules and $VLC(7)$. Type I error $\alpha = 0.05$



(b) Group Decay Rate Difference = 30%

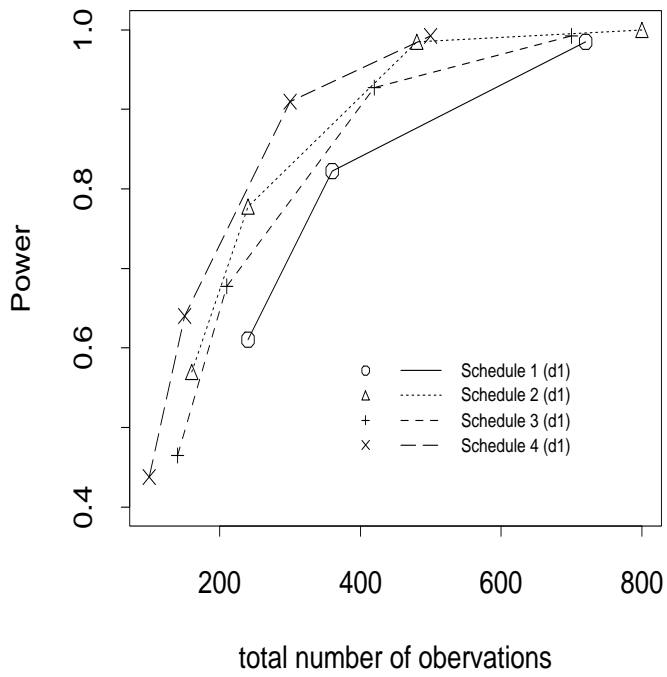


Figure 2: The empirical power of one-sided Wilcoxon test using markers d_1 versus total number of measurements. Type I error $\alpha = 0.05$

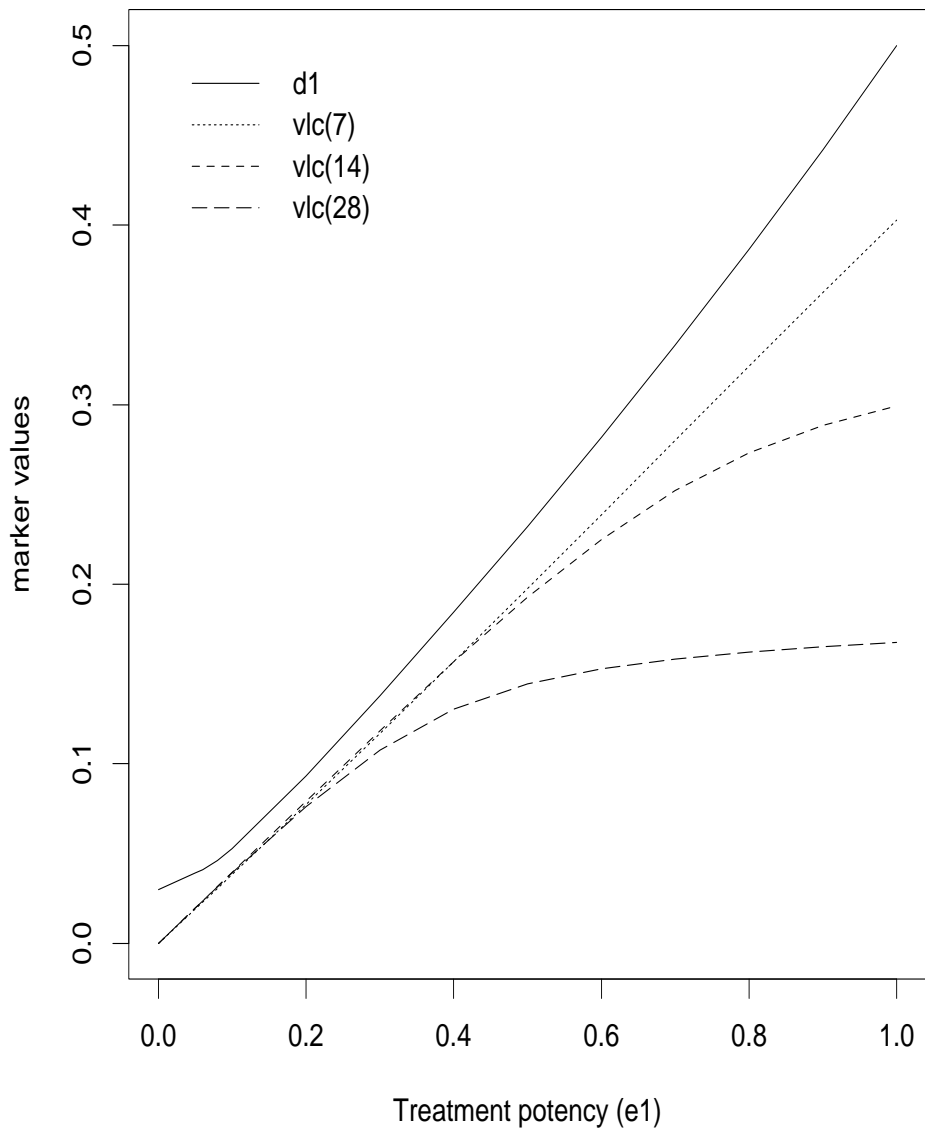


Figure 3: Numerical solutions of the relationship between the potency markers and treatment potency (e_1). The $VLC(t)$ is scaled by time t .

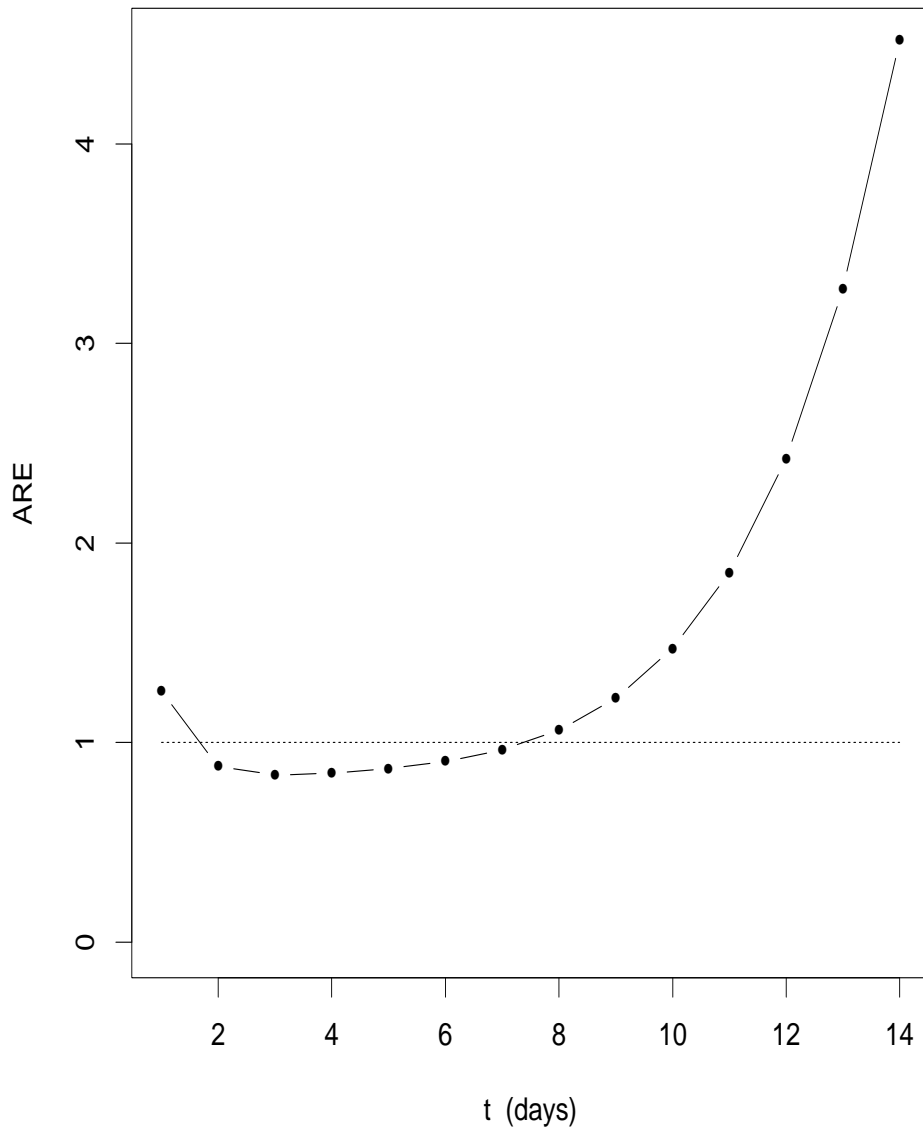


Figure 4: Numerical solutions of asymptotic relative efficiency, $ARE(d_1 : VLC(t))$.

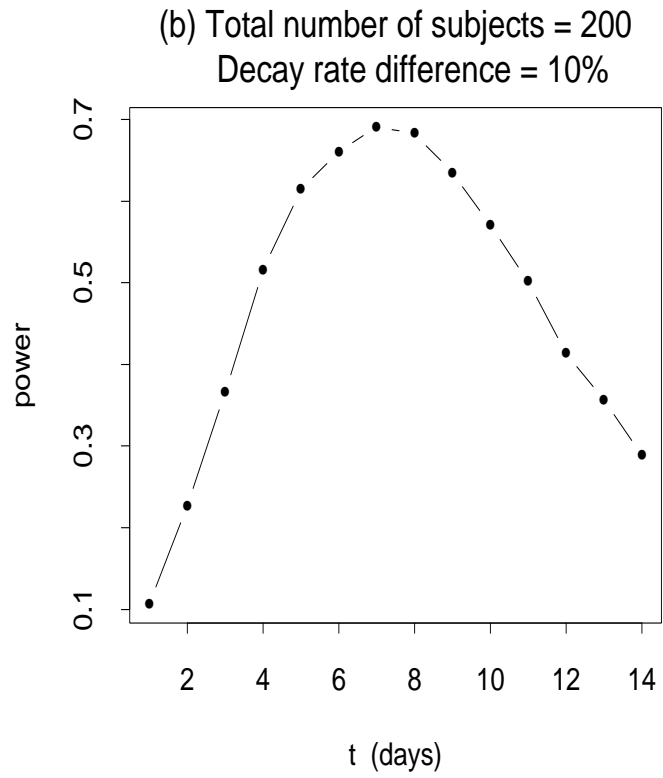
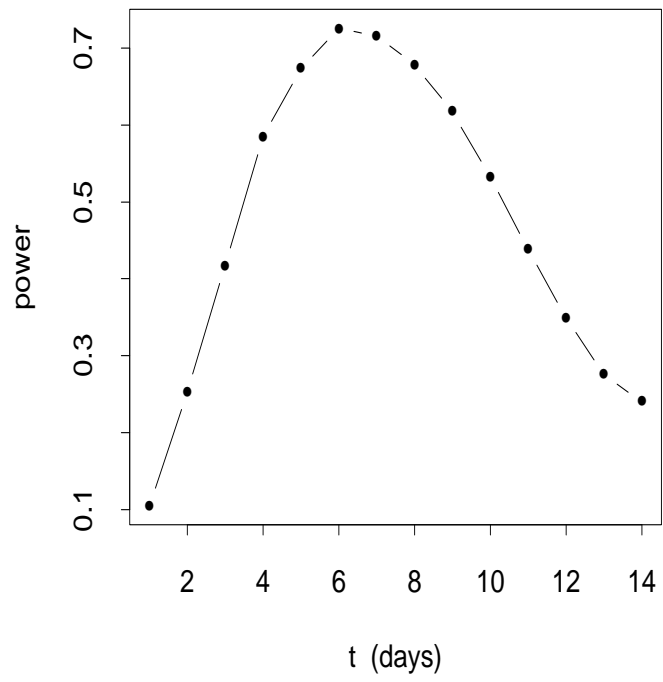


Figure 5: The empirical power of one-sided Wilcoxon test using markers $VLC(t)$: the viral load change from baseline after t days. Type I error $\alpha = 0.05$

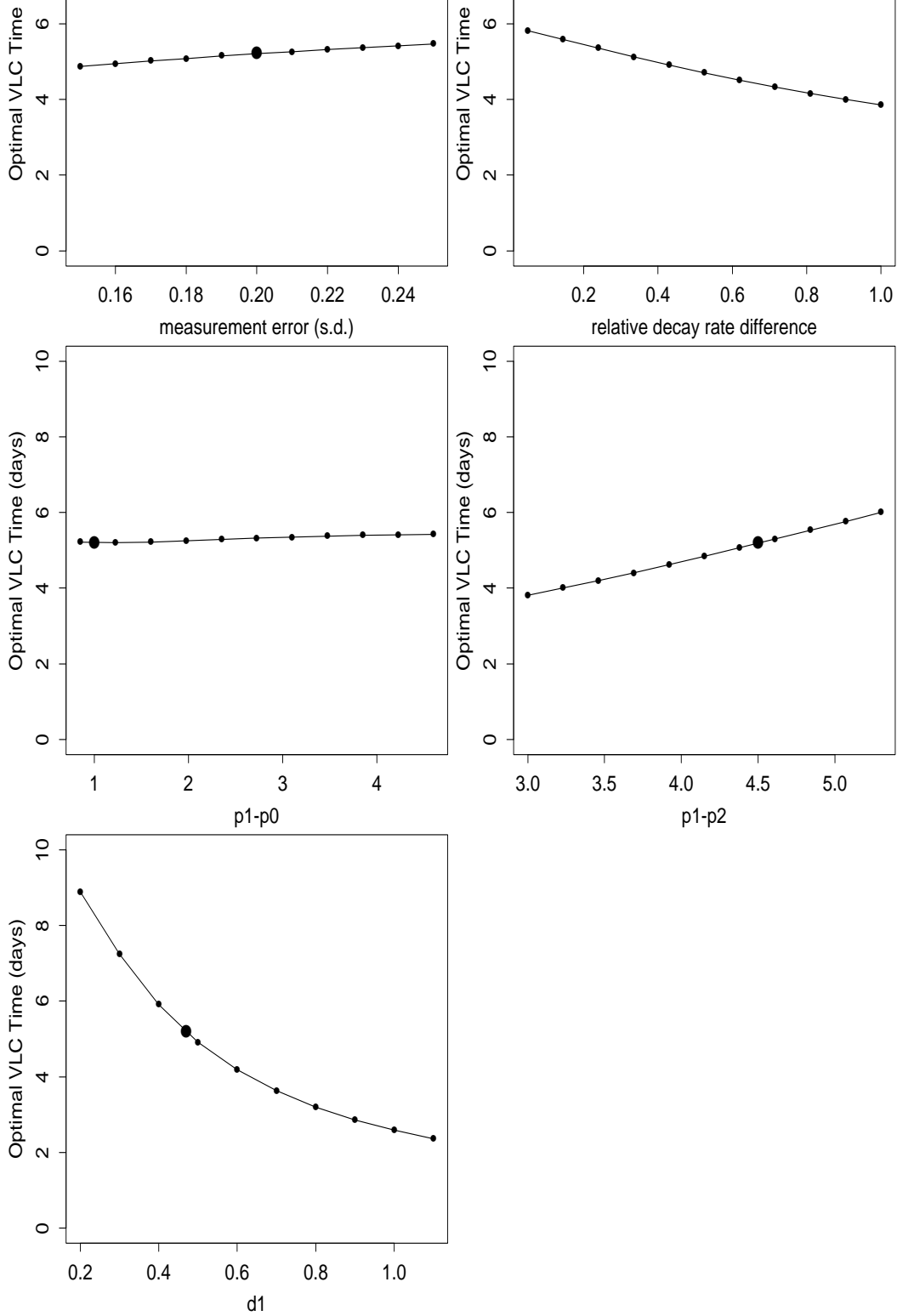


Figure 6: Sensitivity analysis: the change of optimal *VLC* time as parameter values vary. The results corresponding to the parameter values used in simulation of Section 3 is enlighten as “*” .