# Efficient Trial Designs for Studying Combination Antiretroviral Treatments in Patients with Various Resistance Profiles

**Peter Gilbert, Victor DeGruttola, and Scott Hammer**

*Department of Biostatistics, Harvard School of Public Health; Harvard Medical School, Boston, Massachusetts*

**Selecting antiretroviral therapies for human immunodeficiency virus type 1–infected persons is complicated by the availability of a vast number of potentially useful drug combinations and by extensive variation among patients in their resistance to various drugs. AIDS clinical trials have used designs in which a handful of drug regimens in a few patient classes can be compared. Here is proposed implementation of innovative designs with factorial structure that permit assessment of many treatment arms and patient classes in a single trial; when and how they can be appropriately used are discussed. These designs are efficient, permit systematic investigation of correlations between genetic mutations and in vivo drug resistance, and provide insight into important drug interactions in people that conventional designs are unable to provide. Through creative application of these designs, identification of superior drug combinations and the science of understanding in vivo joint drug dynamics and genotypic resistance will progress at an optimum pace.**

The development of powerful antiretroviral therapies has dramatically changed the expectations and capabilities of the treatment of human immunodeficiency virus type 1 (HIV-1) infection—it may be possible to suppress virus permanently in some individuals [1–4]. Because numerous antiretroviral drugs in multiple drug classes are under development or have recently been approved, there are hundreds of potential drug combinations that may be effective. In addition, there are a large number of important classifications of patients with regard to treatment history and degree of resistance to various antiretroviral drugs. Due to limited resources, only a fraction of possible drug combinations in a fraction of patient classes can be investigated in conventionally designed efficacy trials, even if the primary end point is a surrogate marker, such as virologic response. This motivates consideration of innovative trial designs that facilitate efficacy assessment of many drug combinations in many patient classes. In this paper, we argue the aptness of factorial and related structured designs.

Factorial designs have been used extensively in agricultural and industrial experiments, and to a lesser degree in clinical trials. Properly designed factorial trials answer two or more questions at the price of one conventional nonfactorial trial and provide insight into treatment interactions. An example is ISIS-4, a randomized factorial trial that simultaneously assessed three treatments for patients with suspected myocardial infarction: early oral captopril, oral mononitrate, and intravenous magnesium sulphate [5]. We propose that the current climate of combination antiretroviral trials provides many opportunities for application of these efficient designs.

When only a handful of combination HIV-1 therapies was available for efficacy testing, it was rarely appropriate to conduct a factorial trial. This is because at least 4 treatment arms must be arranged in a particular configuration, and for most sets of arms that fit this structure, at least 1 was known a priori to be inferior to the others. For a typical example, in 1996 Hoffmann-La Roche completed a 3-arm trial of saquinavir monotherapy versus dideoxycytosine (ddC) monotherapy versus saquinavir/ddC combination. A 2 × 2 factorial design would require a placebo arm. Today, with the advent of hundreds of possible drug combinations, there are many factorial configurations for which equipoise holds.

Interactions of antiretroviral drugs are commonly studied in the laboratory but not in the clinic, as nonfactorial trials do not have the capability of assessing in vivo interactions. Due to special considerations of HIV-1 disease and its management, knowledge of interactions in people might help in developing combination therapies that meet their high promise of capably suppressing virus permanently or at least for long periods [6]. The central barrier to this treatment success is HIV-1's rapid evolution within a host and the resulting selection for drug-resistant virus strains [7]. This dynamic nature of HIV-1 requires that sequences of drug combinations are judiciously tailored to patients and that patients strongly adhere to the demanding treatment regimen [8].

In this treatment context, there are several ways in which knowledge of in vivo interactions would provide valuable guidance for treatment selection and administration. For one, it is important to assess if a preefficacy trial interaction ascertained from pharmacokinetic or laboratory data extends to a clinically meaningful interaction. For instance, a 2 × 2 factorial design would have allowed the assessment of whether and to what extent in vitro zidovudine and lamivudine synergy [9] translates into positive interaction in clinical or virologic end points and

Received 25 November 1997; revised 10 March 1998.

Grant support: NIH (AI-38855 and AI-28076).

Reprints or correspondence: Dr. Peter Gilbert, Dept. of Biostatistics, Harvard School of Public Health, 677 Huntington Ave., Boston, MA 02115 (pgilbert@hsph.harvard.edu).

**The Journal of Infectious Diseases 1998;178:340–8**

0022–1899/98/7802–0007$02.00

similarly for showing that enhanced bioavailability of saquinavir by ritonavir translates into an ritonavir/saquinavir positive clinical or virologic interaction. For patient management, a physician's knowledge that two drugs work better in combination than expected from their efficacy as individual agents will direct his or her emphasis on joint administration and adherence. Further, knowledge of in vivo negative interactions would help prevent administration of antagonistic combinations.

Understanding of the relationship between mutations in the HIV-1 genome and in vivo drug resistance and cross-resistance would also greatly assist tailoring of therapies. Since genetic sequencing technology is advancing rapidly, there is a new opportunity to systematically collect this knowledge and apply it as a primary guide in patient management. As we elucidate, factorial and related designs are well suited for rapidly acquiring this understanding, because the effect of several patterns of nucleotide or amino acid mutations on in vivo efficacy of multiple drug regimens can be investigated in a single study. Moreover, factorial designs permit assessment of interactions between a mutation pattern class and a treatment class, which can potentially reveal extremely important insights for tailoring treatment assignments. For example, if a mutation in the protease amino acid residue M-46 (to I or L) and a mutation in residue V-82 (to A, F, or T) significantly negatively interact with drug regimens A and B, in that A works well for those with the M-46 mutation but poorly for those with the V-82 mutation, and vice versa for B, a very clear signal is sounded. Clinical trials designed to systematically investigate genotypic resistance patterns are keenly needed, as currently this information is accumulating via post hoc analysis of virologic failures in individual antiretroviral trials.

In Factorial Designs (see below), we describe elementary features of factorial and related structured designs and their use in combination antiretroviral trials. We argue for implementation of these designs, not only because they provide a systematic approach to evaluation of drug interactions and genotypic resistance, but also because they require substantially fewer study subjects than conventional designs for comparisons of particular drugs or mutation patterns within a class. We compare factorial and related designs to conventional designs, both by their theoretical properties and by application to completed and recently proposed combination antiretroviral trials.

## Drug Interactions in People

Drug interactions are typically studied in the laboratory, defined through concepts such as synergy. In people, an interaction is defined relative to a measure of tangible drug efficacy. For example, synergy in the laboratory may translate into synergy in the clinic [10]. Other measures are clinical (e.g., time to AIDS-defining illness or death), virologic (e.g., suppression of plasma HIV-1 RNA level), or immunologic (e.g., increase in CD4 cell count). For a given efficacy measure, interaction of two drugs is defined relative to a rule for addition of individual drug effects. To illustrate a rule, consider two antiretrovirals, A and B, and let absolute change in plasma HIV-1 RNA level from baseline to a prespecified follow-up time be the efficacy end point. Suppose the effect of A(B) is to drop virus load by $d_A(d_B)$ logs. One rule states that the treatments are additive (positively interact) if the A/B combination drops virus load by an amount equal to (more than) $d_A + d_B$ logs. A negative interaction occurs if A/B drops virus load by an amount less than $d_A + d_B$ logs. A different addition rule would give a different interaction definition, so that the concept of drug interaction in people depends entirely on the expectation of how individual drug effects on the efficacy end point add. See [11, pp 1058–9] for further discussion on the inherent model-dependency of treatment interactions.

A drug interaction in one efficacy measure does not necessarily translate into a drug interaction in a different efficacy measure. It depends on the model linking the two measures. To illustrate this, consider efficacy defined by the capability of a protease inhibitor drug to render newly produced virions noninfectious. Define respectively $\eta_0$ and $\eta^*$ as the proportion of newly produced virions that are noninfectious before and after drug administration, so that $\eta = (\eta^* - \eta_0)/(1 - \eta_0)$ measures the efficacy of the administered drug relative to a perfect drug. Since no available assay can directly measure the proportion of new virions rendered noninfectious, in order to estimate $\eta$, it is necessary to use a viral dynamics model linking this drug efficacy mechanism to an observable efficacy end point. Consider recent viral dynamics models. Perelson et al.'s [12] popular model assumes a perfect drug ($\eta = 1$), so that $\eta$ cannot be estimated. Wu et al. [13] developed a viral dynamics model similar to that of Perelson et al. [12], from which $\eta$ can be estimated if data on infectious virus load are available. Their model expresses the observable log change, *d,* in plasma HIV-1 RNA concentrations from baseline to a prespecified follow-up time, *t,* as a function of $\eta$, *t,* and the clearance rates of free virus *(c)* and productively infected T cells ($\delta$).

How does this viral dynamics model relate definitions of drug interaction on the two efficacy measures $\eta$ and *d?* Consider sensible definitions of interaction in $\eta$ and *d.* If drug A has efficacy $\eta_A = 90\%$, and drug B has efficacy $\eta_B = 50\%$, then a natural definition of additivity is that combination A/B has drug efficacy $\eta_{AB} = 95\%$. The general formula for additivity is that combination A/B has drug efficacy $\eta_{AB} = \eta_A + \eta_B - \eta_A * \eta_B$. Define a drug interaction in *d* as above. As we elaborate in the Appendix, under Wu et al.'s [13] viral dynamics model, these two definitions correspond only if the clearance rates, *c* and $\delta$, and the plasma HIV-1 RNA measurement time, *t,* take particular values. Thus, a natural and clearly interpretable definition of additivity for the observable efficacy end point, *d,* may correspond to an arbitrary definition of additivity for the unobservable efficacy parameter, $\eta$, and vice versa. This highlights the inherent model-dependency of relating interaction definitions on different efficacy end points. More specifically, it suggests that viral dynamics models can be used to select a plasma HIV-1 RNA

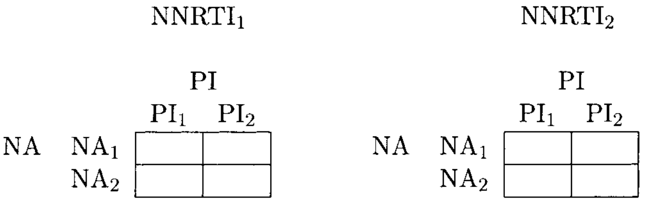

**Figure 1.** 2 × 2 × 2 factorial design with $NNRTI_1$ (nonnucleoside reverse-transcriptase inhibitor 1) present in each arm and $NNRTI_2$ present in each arm. PI, protease inhibitor; NA, nucleoside analogue.

measurement time so that an interaction in the observable measure, *d,* reflects an interaction in the biologically meaningful parameter $\eta$.

This paper aims to illustrate concepts rather than be exhaustive, so henceforth we focus on the important case in which the primary end point is log drop, *d,* in plasma HIV-1 RNA concentrations and treatment effects add on the scale for *d* defined above.

## Factorial Designs

In this section we define the factorial design, describe its elementary properties, and compare it to designs that have been extensively used in antiretroviral trials for HIV-1 infection. See [11] and [14] for further description of properties of factorial clinical trials. References [10] and [15] provide further discussion of their application specifically to antiretroviral trials.

We have three main drug classes to investigate in combination: protease inhibitors (PIs), nucleoside analogues (NAs), and nonnucleoside reverse-transcriptase inhibitors (NNRTIs). These are factors in a factorial design. Particular drugs within a drug class are levels of the factor. A trial with factorial design is one with simultaneous randomizations to the levels of each factor. This allows investigation of multiple scientific questions at no extra cost by decomposing the analysis of treatment differences into main effects of each factor and interaction effects between factors. To illustrate, suppose we are interested in two drugs in each drug class, labeled $PI_1$, $PI_2$, $NA_1$, $NA_2$, $NNRTI_1$, and $NNRTI_2$. A 2 × 2 × 2 factorial design is formed by crossing all possible drug choices in the three drug classes (figure 1). Each subject is randomized to receive one of 8 triple combinations (i.e., simultaneously to one of the two drugs in each class). For instance, 400 subjects could be entered in random permuted blocks of 8, so that 50 individuals would receive each triple combination.

An advantage of the factorial design is symmetry—each individual drug is given to half the subjects, and each double combination to one-fourth of the subjects. Thus each subject's data contributes to many treatment comparisons. For example, there are 12 unique double combinations, and any pair of these can be compared with 100 subjects per arm. More importantly, the main effect of each of the three factors can be assessed with very good efficiency. For example, a treatment difference between the two NNRTIs is assessed by comparing the results from 200 subjects on $NNRTI_1$ with those from the 200 subjects on $NNRTI_2$. Similarly, the main effect of PIs and NAs can be assessed with 200 subjects per drug level.

The capability of the factorial trial to simultaneously assess 3 main effects has considerable advantages over the conventional nonfactorial approach, which would consist of three separate one-factor-at-a-time trials, in which the experimental factors (PIs, NAs, and NNRTIs) are varied one at a time, with the remaining factors held fixed. Each individual trial provides an estimate of the effect of a single drug class at selected fixed drug levels of the other drug classes. For this estimate to have broad clinical relevance, it is necessary to assume that the treatment effect would be the same in combinations with other drugs in the other drug classes—that, over the ranges of current interest, the drugs act on the efficacy measure additively. However, if the treatment effect is additive, the factorial design is much more efficient, and if there is nonadditivity, the factorial design, unlike the one-factor-at-a-time design, if adequately powered, can detect and estimate interactions that measure the nonadditivity [16].

To illustrate the gain in efficiency, suppose there are 200 subjects in the 2 × 2 × 2 factorial design, 25 per arm. To obtain equal precision for the estimate of PI effect, the one-factor-at-a-time trial would need to have 200 subjects, 100 for each PI level, with all observations made at some fixed levels of NAs and NNRTIs. Similarly, two further trials of 200 subjects would be required to study NAs and NNRTIs. Thus, to give estimates of the main effects of the three drug classes with the same precision as provided by the factorial design, the one-factor-at-a-time design would require 600 subjects—a 3-fold increase.

Another advantage of a factorial design is that one can test for and estimate the magnitude of interactions between drug classes. In this example, there are three possible interactions between drug classes. For instance, consider PIs and NAs. The standard measure of the strength of interaction is the difference between the average PI effect with the first NA and the average PI effect with the second NA. It is simple to detect interactions graphically as illustrated by figure 2 of Example 1. The existence of each interaction and each main effect can be tested for straightforwardly by an *F* test under standard assumptions [17]. A central advantage of testing for interactions and main effects is that it is possible to disentangle if superiority of a combination is due to main effects or positive drug interaction.

We now give an example of a completed nonfactorial trial conducted by the AIDS Clinical Trials Group (ACTG) in which a factorial design would have been efficient and illuminating.

*Example 1: hypothetical factorial trial based on ACTG 229 and ACTG 116B/117.* Clinical trial ACTG 229 compared zidovudine/saquinavir, zidovudine/ddC, and zidovudine/saquinavir/ddC [18]. Consider the available efficacy measure of absolute change in log plasma HIV-1 RNA concentrations

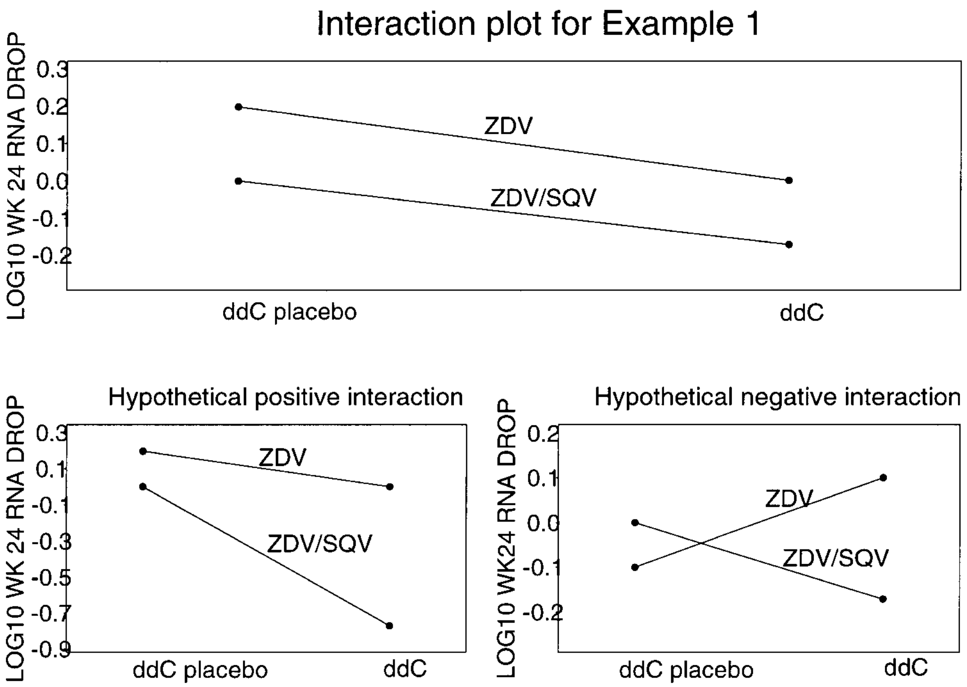

**Figure 2.** Interaction plots of mean change in plasma human immunodeficiency virus type 1 RNA concentrations from baseline to week 24, example 1. ZDV, zidovudine; SQV, saquinavir; ddC, dideoxycytosine.

(bDNA assay; Chiron, Emeryville, CA) from baseline to week 24. If ACTG 229 had been designed as a 2 × 2 factorial trial, including a zidovudine monotherapy arm, it would have been possible to assess the existence of an interaction between saquinavir and ddC. To mimic this, we constructed this arm by using plasma HIV-1 RNA data (bDNA assay; Chiron) from the zidovudine monotherapy arm of ACTG 116B/117 [19]. Although this example lacks randomized control, it has partial validity in that the two studies have similar patient populations (at least 4 months of prior treatment with zidovudine and CD4 cell counts in the range 50–300 copies/mL). To maximize their comparability, subjects were sampled from ACTG 116B/117 according to the distribution of baseline CD4 cell count and plasma HIV-1 RNA copy number for ACTG 229 subjects.

Figure 2 (top) displays the mean treatment response for the 4 groups, with 76, 76, 73, and 73 subjects on the arms. The results are clear: triple therapy is superior to either double therapy ($P = .0052$ vs. zidovudine/saquinavir, $P = .0042$ vs., zidovudine/ddC), the double therapies perform similarly, and zidovudine monotherapy is worse than any combination ($P = .020$ vs. zidovudine/saquinavir, $P = .019$ vs. zidovudine/ddC, and $P = .0001$ vs. zidovudine/saquinavir/ddC). The striking feature of the analysis is that the lines in figure 2 (top) are nearly perfectly parallel. This indicates the complete absence of an interaction, showing that, given a zidovudine background, saquinavir and ddC are additive (on the chosen scale) in their effect of suppressing virus load. One interpretation is that saquinavir and ddC work independently of one another without interference. This type of scientific insight was not possible with the original design. Another advantage is that, whereas in the 3-arm design three treatments can be compared pairwise, in the factorial design treatment differences can be assessed between an additional 3 pairs of arms, but the sample size has only increased by a third.

To illustrate graphically various interactions, suppose that saquinavir and ddC interacted positively, so that the triple combination zidovudine/saquinavir/ddC performed better than expected from the sum of the doubles zidovudine/saquinavir and zidovudine/ddC. Then the lines would be nonparallel and diverging, as illustrated in figure 2 (bottom left). Figure 2 (bottom right) depicts a hypothetical negative interaction (antagonism), indicated by crossing lines, in which ddC has a negative effect with zidovudine and a positive effect with zidovudine/saquinavir.

We give a second example based on recently proposed trials to evaluate novel salvage regimens in people whose treatment by PI fails.

*Example 2: protease failure salvage trial.* Although the PI era has brought with it remarkable improvements over previously available antiretroviral therapies, a substantial proportion of PI-treated subjects do not maintain adequate viral suppression. For subjects with protease failure, it is imperative that alternative regimens be developed. A number of studies, including within the ACTG, have been initiated or are in development to try to address this issue utilizing combination regimens that include approved and experimental dual PIs, NNRTIs, and the acyclic nucleoside phosphonate reverse-transcriptase inhibitor, adefovir. In this example, we discuss how factorial designs might be appropriately used for a ‘‘salvage’’ trial, which compares novel combination regimens. In con-

EFV

| | dual PI | |
|---|---|---|
| | AMP/I or N | RTV/SQV |
| ADV | A | B |
| plac | C | D |

NVP

| | dual PI | |
|---|---|---|
| | AMP/I or N | RTV/SQV |
| ADV | E | F |
| plac | G | H |

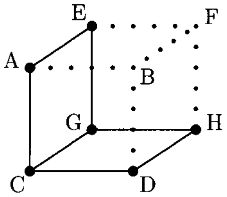

**Figure 3.** 2 × 2 × 2 factorial design with efavirenz (EFV) present in each arm (left) and nevirapine (NVP) present in each arm (middle). Right, 2 × 2 × 2 factorial design missing cells B and F. PI, protease inhibitor; AMP, amprenavir; I, indinavir; N, nelfinavir; RTV, ritonavir; SQV, saquinavir; ADV, adefovir; plac, ADV placebo.

structing treatment arms, we consider drugs proposed by groups within the ACTG. As in some current proposals, suppose all subjects receive the novel NA abacavir (also known as 1592U89) and the novel NNRTI efavirenz (DMP-266). The experimental PI amprenavir (141W94) will be included in some arms.

Two classes of therapy failures are considered: those given indinavir and those given nelfinavir. For indinavir failures, consider the following two research questions: Which dual protease, ritonavir/saquinavir or amprenavir/nelfinavir, works best? Does adding adefovir to the regimen help? These questions are also of interest for nelfinavir failures, in which case the dual protease amprenavir/nelfinavir is replaced by amprenavir/indinavir. Separate 2 × 2 factorial trials could be conducted for indinavir failures and for nelfinavir failures, each with the capability to test which dual protease works the best and whether adefovir augments therapy, and to assess an adefovir/dual protease interaction.

Now consider indinavir and nelfinavir failures in the same study. Then amprenavir in combination with indinavir or nelfinavir can be taken as a PI level, and a single 2 × 2 factorial trial stratified by indinavir or nelfinavir failure could be conducted. Next, suppose a third randomization is of interest. For instance, subjects could be randomized to an NNRTI, efavirenz or nevirapine, rather than all being assigned efavirenz. This creates a 2 × 2 × 2 factorial design with factors NNRTI (efavirenz vs. nevirapine), dual PI (amprenavir/indinavir or nelfinavir vs. ritonavir/saquinavir), and adefovir (adefovir vs. adefovir placebo) (figure 3). A full analysis of 3 main effects and 3 interactions could be carried out as described for the design illustrated in figure 1.

### Factorial Designs with Missing Cells

In practice, there are many sets of treatment combinations that can be arranged in a factorial configuration. However, because some of the combinations are not worth testing (e.g., zidovudine with stavudine, or ddI [dideoxyinosine] with ddC), there will be some empty cells. In this situation, benefits can be reaped if the design contains factorial substructure. For instance, there may be an all-cells-filled factorial substructure within the full design for which the methods of analyzing factorial designs apply. We illustrate with the 2 × 2 × 2 factorial design discussed in Example 2. Some investigators might be reluctant to add the fifth drug, adefovir, to a regimen containing ritonavir/saquinavir, in which case arms B and F are unacceptable. In this case, there are still two 2 × 2 factorial designs embedded within the remaining 6 arms, one comprising arms A, C, E, and G and the other comprising arms C, D, G, and H (figure 3, right). For each of these squares, 2 main effects and an interaction can be investigated. If another arm were unacceptable, for instance, arm D or H, then the trial would have 5 arms, of which 4 still compose a 2 × 2 factorial design. The key point is that in selecting treatment arms for inclusion in combination antiretroviral trials, it is prudent to search for groups of arms that at least partially fit a factorial structure so as much as possible can be learned at the cheapest price. We discuss sample size issues for these designs below (Sample Size Advantages of Factorial and Latin Square Designs).

If one assumes there are no drug class interactions, then for many factorial designs, the mean treatment response of the combination arms in the empty cells can be imputed, with the same precision as for the filled cells [20]. It is usually unadvisable to make this assumption, as it requires an excellent understanding of joint in vivo drug dynamics. When the assumption is valid, however, it permits ascertainment of treatment effects of the unrepresented combinations.

### Factorial and Related Structured Designs for Studying Genotypic Resistance

Investigations relating mutations in the HIV-1 genome to in vivo drug resistance and cross-resistance have been limited to exploratory analyses of virologic failures in particular clinical trials, without systematic assessment. Due to the vast number of resistance profiles and drug combination options, we need studies that facilitate systematic assessment of these correlations for many mutation patterns and drugs. A mutation pattern is any representation of change in the HIV-1 genome that is of interest, such as a single-residue mutation (e.g., T-215-Y), a combination of mutations (perhaps in some temporal sequence), or any arbitrary delineation.

We illustrate through examples how structured designs facilitate systematic and efficient assessment of genotypic resistance. In the analysis of 29 phenotypically indinavir-resistant HIV-1 isolates, Condra et al. [21] found that the relationship between protease mutations and resistance is complicated and difficult to fully elucidate. Mutations in at least 11 protease amino acid residues were associated with resistance, so apparently there are multiple resistance pathways. However, mutations at positions 46 or 82 were found in all indinavir-resistant isolates, and the number of protease mutations correlated with the magnitude of resistance, suggesting hypotheses about kinds of mutational patterns that are most predictive of loss of drug efficacy.

Factorial designs might be useful for genotypic resistance trials. For example, we might conduct a salvage trial for indinavir and nelfinavir failures with the 4 treatment arms in the left or middle of figure 3. Consider a classification into two patterns of mutation: pattern I = mutation at 82 but not at 46, and pattern II = mutations at both 46 and 82. Enrolling subjects so that half have mutation pattern I and half have mutation pattern II would permit systematic genotypic resistance assessment without increasing the sample size or compromising the original objectives of the trial. With factorial designs, interactions between mutation pattern classes and drug treatment classes can be investigated in the same way that interactions between two drug classes are investigated. Results should be interpreted more cautiously though, because mutational pattern is a classification factor rather than a randomization factor.

When more than one genetic classification factor is of interest, the simple factorial design may not be practical. A trial with ''Latin square'' design (Latin square designs are commonly used in pharmacokinetic and dosing clinical trials for many drugs, including antiretrovirals) could be conducted to test hypotheses about two classes of mutation patterns simultaneously and to compare salvage regimens. To illustrate, consider one class of three indinavir-resistance mutation patterns: mutation at 46 but not at 82, mutation at 82 but not at 46, and mutations at both 46 and 82. Define another class of three mutation patterns: $\leqslant 3$ protease mutations, 4–6 protease mutations, and $\geqslant 7$ protease mutations. A 3 × 3 Latin square design with these two mutation pattern classes as rows and columns and treatment regimens A, B, and C is depicted in figure 4 (left). Instead of randomly assigning three treatments to each stratum (defined by row and column), only one treatment is assigned, hence the number of cells is reduced from 27 to 9.

The attraction of the design depicted in figure 4 is that it permits investigation of multiple scientific questions with the same number of subjects needed to investigate just one of these questions. First: Which treatment works best overall? Second and third: Does the extent of resistance vary by mutation patterns in the rows, or in the columns? These three questions can be answered through $F$ tests of main effects under the assumption of no interactions between any of the three factors [17, pp 1091–6]. Fourth: For patients in each mutation classification (e.g., 46, 82), which treatment is the best? Fifth: How predictive of resistance is each combination of mutation patterns (e.g., $\leqslant 3$, 82)?

3 × 3 Latin square

| | 46 | 82 | 46 + 82 |
|---|---|---|---|
| $\leq 3$ | A | B | C |
| $4-6$ | B | C | A |
| $\geq 7$ | C | A | B |

2 × 2 Latin square

| | 82 | 46 + 82 |
|---|---|---|
| $\leq 5$ | A | B |
| $> 5$ | B | A |

**Figure 4.** Latin square trial designs for linking mutation patterns to drug resistance. Subjects are stratified into 9 (left) or 4 (right) mutation pattern classes. Drug regimen A, B, or C is assigned to participants in stratum indicated by row and column. For example, in design at left, persons with mutation at 46 and $\leqslant 3$ protease mutations receive regimen A.

Instead of investigating two classes of mutation patterns and one class of treatments, two drug classes (say, NAs and PIs) and one class of mutation patterns could be investigated through a Latin square design. This kind of design permits assessment of which NA works the best for each mutation pattern, which PI works the best for each mutation pattern, and which mutation pattern overall is most predictive of resistance.

### Sample Size Advantages of Factorial and Latin Square Designs

Generally, in the absence of an appreciable negative interaction, a factorial trial can detect meaningful main effect treatment differences with small sample sizes. However, a factorial design with high power to detect main effects will have substantially lower power to detect interactions (figure 5).

To illustrate these points concretely, consider the discussion in example 2 about the design of a salvage trial. Various trial designs could potentially be used that have between 5 and 8 treatment arms and various degrees of factorial structure. We compare the powers of competing designs to detect a 0.5 log HIV-1 RNA change (treatment difference) in main and interaction effects. To illustrate an observed set of treatment responses consistent with this difference, consider the 2 × 2 factorial design depicted in figure 3 (left), and suppose there is a mean response of 2.0 log for both amprenavir-containing arms (each containing amprenavir/indinavir or amprenavir/nelfinavir) and 1.5 log for both ritonavir/saquinavir-containing arms. This specifies no interaction, so that the dual PI main treatment effect is the difference between the responses for the two amprenavir cells (each with amprenavir/indinavir or amprenavir/nelfinavir) and the two ritonavir/saquinavir cells, equal to 0.5 log. If the mean response of the four treatment combinations is 1.5 log for the upper-right cell and 1.0 log for the other three cells, then the alternative is a 0.5 log positive interaction. For the power calculations, we assume the log HIV-1 RNA change is normally distributed, with the same SD ($\sigma$) = 0.7 (suggested by previous studies) in all arms.

First, suppose the 8-arm 2 × 2 × 2 factorial trial depicted in figure 3 (left and middle) was conducted, with 10 subjects per arm, a total of 80 subjects. Power to detect each of the

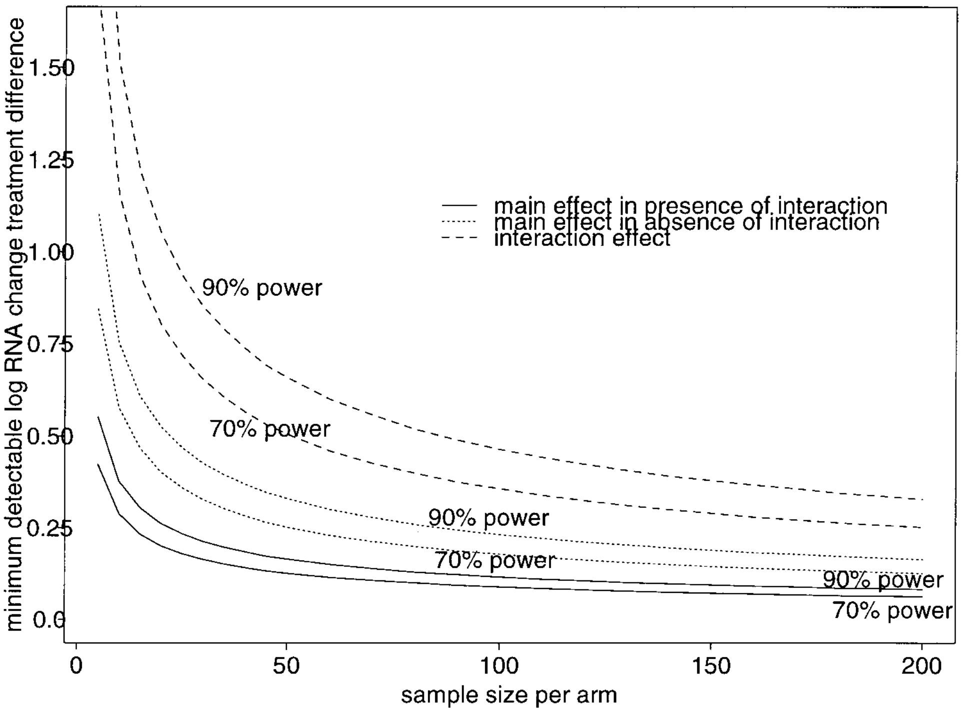

**Figure 5.** Power to detect main effect and interaction effect in 2 × 2 factorial design. $Log_{10}$ plasma RNA change is assumed to be normally distributed with 0.7 SD in each arm.

three drug class main effects (without interaction) is 81.6%. If three separate trials were run, we would require 80 patients in each study, a total of 240 patients, to have the same power to detect the main effects. If the sample size is increased to 40 per arm in a single factorial study of 320 patients, the power to detect each of three drug class main effects is >99.9%, while power to detect each of three interactions between drug classes is 81.6%. Moreover, for each of the 6 embedded 2 × 2 factorial designs, there is 99.0% power to detect each of 2 main effects and 57.5% to detect an interaction. Next, suppose that arms B and F are considered unacceptable, so that the design is reduced to 6 arms with 2 embedded 2 × 2 factorial designs (figure 3, right). The total sample size has fallen from 320 to 240 subjects, but due to the 2 missing cells, 4 rather than 12 main effects and 2 rather than 6 interactions can be investigated. For the investigatable effects, power is the same as given above, 99.0% or 57.5%. Finally, suppose arm H is also unacceptable, so that now there are 200 subjects on 5 arms. Now 2 main effects and 1 interaction can be studied, with 99.0% or 57.5% power.

Sample-size comparisons between conventional designs and factorial designs are qualitatively similar for binary, failure time, or other end points. There are, however, important considerations unique to each end point. We briefly discuss the comparison for a failure time end point, such as time to ‘‘virologic rebound.’’ For this end point, interaction is sensibly defined in terms of the drug efficacy measure multiplicative reduction in the hazard of virologic failure. Consider a 2 × 2 factorial design, with factors A and B, and a 2-arm trial in which the arms are the levels of A. Suppose the same number of subjects are evenly allocated in each trial, and compare power to detect a main A treatment effect. In the absence of an interaction, power may be mildly worse in the factorial than the 2-group design. This would occur solely because persons who receive a factor B treatment may have fewer events. Thus, to keep the power the same, the sample size should be raised by the ratio of the probability of an event in a nonfactorial trial with one effective treatment to the probability of an event in a factorial trial in which both treatments are effective [10, p 258]. For instance, consider example 1. Suppose the probability of failure on the zidovudine monotherapy arm is 50% and the effect of each combination zidovudine/saquinavir and zidovudine/ddC is to reduce this failure hazard by 30%. The sample size then would need to be increased by 13% [10, p 259]. This increase rises for lower probabilities of failure and for greater reductions of the hazard by the combination therapies.

Now suppose there is an interaction. If it is in the same direction as the main A effect, then the standard test (stratified log rank) for testing an A treatment effect is at least as powerful as this test in the 2-group trial [22]. On the other hand, power is less for the factorial trial if the interaction is negative. Of course, the same power to detect a main A effect in the factorial trial is available to detect a main B effect. Thus, even in the case in which power is reduced, 2 questions can be addressed at nearly the cost of 1.

See [10, pp 258–60] and [11, p 1060] for further discussion about the power of factorial designs for detecting main effects and interactions. We note that Latin square designs are optimally efficient for investigating main effects in that the fewest study subjects are needed to achieve high power [17].

## Discussion

For many sets of antiretrovirals, factorial and Latin square designs allow for more rapid, precise, and insightful evaluation than do other approaches. These designs are also efficient for learning about genotypic resistance of patients in various fixed mutation pattern classifications. This understanding may lead to identification of a simple rule, based on a genotype assay, for tailoring therapy to patients. Ultimately, we need to develop rules for assigning sequences of drug combinations to patients with changing mutation profiles over time. To build these rules, ''adaptive'' trials are needed, in which trial participants with certain new mutations emerging during the course of follow-up are randomized to a new regimen. In this paper, we have not discussed use of efficient designs for adaptive trials, because drug-switching rules predictive of long-term virologic or clinical status have not yet been established. A staging of trials is called for—first nonadaptive trials can be used to identify sound switching rules, which can then be appropriately implemented in adaptive trials.

We discuss some criteria for judging the appropriateness of inclusion of treatments in factorial and Latin square trials. First, for inferences about main effects to be interpretable, the investigators must believe that the results about any given treatment factor can be combined over the levels of the other treatment factors. To illustrate this concept, in order for the factorial design of example 1 to be appropriate, it is required that the benefits of zidovudine/ddC over zidovudine are expected in those on saquinavir and saquinavir placebo, and the benefits of zidovudine/saquinavir over zidovudine are expected in those on ddC and ddC placebo. Second, for the main effect estimates to have an unconfounded interpretation, it is necessary that all the combined drug levels can be given together without modification [11]. Third, factorial designs are inappropriate if some combinations of drug levels have contraindications, because equipoise fails and antagonism may occur, which impairs power to detect main effects.

Blinding, toxicity management, and other study characteristics can be logistically complicated in trials involving multiple arms and combination drug regimens. One implication is that highly structured trials will often need to be open-label. There are other ways of reducing the logistical complexity of factorial trials, for example by partitioning treatment arms into subsets of arms by site so that each hospital or clinic only assigns a few of the arms.

We have two conclusions. First, screening a large set of combination antiretroviral candidates and identification of the best combination therapies can proceed most rapidly and efficiently by factorial and Latin square designs. Moreover, factorial designs give insight into important drug interactions in people that conventional designs cannot provide. The potential barrier to implementation of these designs is that the set of treatment arms must fit a structure. The more rigid and complete the structure, the more efficient and informative the design. However, with the enormous number of new antiretroviral options, there now exist many sets of comparison arms of equally compelling interest that partially or wholly fit into the framework of factorial designs and are appropriate for these designs. Indeed, factorial antiretroviral trials are beginning to be implemented, as evidenced by ACTG 368, 388, and 384. The simple 2 × 2 factorial trial ACTG 368 addresses two questions: Should abacavir be added to a PI/NNRTI regimen? and Is twice-daily dosing viable, while the 3 × 2 factorial trial ACTG 388 is designed to compare three combination regimens and to assess an adherence-intervention. A proposed design of ACTG 384 specifies a 6-arm trial that fits the partial factorial structure of figure 3 (right) with A = ddI/stavudine/nelfinavir placebo/efavirenz, E = zidovudine/lamivudine/nelfinavir placebo/efavirenz, C = ddI/stavudine/nelfinavir/efavirenz, G = zidovudine/lamivudine/nelfinavir/efavirenz, D = ddI/stavudine/nelfinavir/efavirenz placebo, and H = zidovudine/lamivudine/nelfinavir/efavirenz placebo. In this era of studying combination antiretroviral therapies, a fundamental component that should be included in the process of selecting treatment regimens for clinical trials is conformity of the arms to factorial or partial factorial structure.

The second conclusion is that factorial (and Latin square) designs are tailor-made for efficiently and systematically studying genotypic correlates of drug resistance. These correlations for two drug classes can be studied for the price of a conventional nonfactorial study of associations for one drug class. Moreover, only with structured designs is it possible to make the important assessment of interactions between specific mutation patterns and drug classes.

## Acknowledgments

We thank John Mellors, David Schoenfeld, and Marvin Zelen for helpful discussions, as well as the reviewers for insightful suggestions.

### References

1. Hammer SM. Advances in antiretroviral therapy and viral load monitoring. AIDS **1996**;10(suppl)3:S1–11.
2. Mathez D, De Truchis P, Gorin I, et al. Ritonavir, AZT, ddC, as a triple combination in AIDS patients (abstract 258). In: 3rd Conference on Human Retroviruses and Opportunistic Infections: program and abstracts (Washington, DC). Alexandria, VA: Infectious Diseases Society of America, **1996**.
3. Gulick R, Mellors J, Havlir D, et al. Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. N Engl J Med **1997**;337:734–9.
4. Markowitz M, Cao Y, Vesanen M, et al. Recent HIV infection treated with AZT, 3TC, and a potent protease inhibitor (abstract). In 4th Conference on Retroviruses and Opportunistic Infections: program and ab-

stracts (Washington, DC). Alexandria, VA: Infectious Diseases Society of America, **1997**.

5. ISIS-4 Collaborative Group. A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58050 patients with suspected acute myocardial infarction. Lancet **1995**;345:669–85.
6. Ho D. Time to hit HIV, early and hard. N Engl J Med **1995**;333:450–1.
7. Carpenter CC, Fischl MA, Hammer SM, et al. Antiretroviral therapy for HIV infection in 1997. Updated recommendations of the International AIDS Society—USA Panel. JAMA **1997**;277:1962–9.
8. Deeks SG, Smith M, Holodniy M, Kahn JO. HIV-1 protease inhibitors: a review for clinicians. JAMA **1997**;277:144–53.
9. Staszewski S. Zidovudine and lamivudine: results of phase III studies. J Acquir Immune Defic Syndr Hum Retrovirol **1995**;10(suppl 1):S57.
10. Schoenfeld DA. AIDS clinical trials. Chapter 17. New York: John Wiley & Sons, **1995**.
11. Byar DP, Piantadosi S. Factorial designs for randomized clinical trials. Cancer Treat Rep **1985**;69:1055–63.
12. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **1996**;271:1582–6.
13. Wu H, Ding AA, DeGruttola V. Estimation of HIV dynamic parameters. Stat Med **1998** (in press).
14. Stampfer MJ, Buring JE, Willett W, Rosner B, Eberlein K, Hennekens CH. The 2 × 2 factorial design: its applications to a randomized trial of aspirin and carotene in U.S. physicians. Stat Med **1985**;4:111–6.
15. Ellenberg SS, Finkelstein DM, Schoenfeld DA. Statistical issues arising in AIDS clinical trials. J Am Stat Assoc **1992**;87:562–9.
16. Box GEP, Hunter WG, Hunter, JS. Statistics for experimenters. New York: John Wiley & Sons, **1978**.
17. Neter J, Wasserman W, Kutner MH. Applied linear statistical models. 3rd ed. Chapters 18, 22. Homewood, IL: Irwin, **1990**.
18. Collier AC, Coombs RW, Schoenfeld DA, et al. Treatment of human immunodeficiency virus infection with saquinavir, zidovudine, and zalcitabine. N Engl J Med **1996**;334:1011–7.
19. Kahn JO, Lagakos SW, Richman DD, et al. A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. N Engl J Med **1992**;327:581–7.
20. Searle SR. Linear models for unbalanced data. Chapter 5. New York: John Wiley & Sons, **1987**.
21. Condra JH, Holder DJ, Schleif WA, et al. Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. J Virol **1996**;70:8270–6.
22. Slud EV. Analysis of factorial survival experiments. Biometrics **1994**;50:25–38.

## Appendix

From panel 15 of Wu et al. [13], for a measurement time, *t,* larger than a few days, the $\log_{10}$ change, $d(t) = log_{10}(V(t) - V(0)),$ in plasma RNA concentrations from baseline satisfies the approximation

$$d(t) = \log_{10}\left\{\frac{1}{(1-\eta)}\left[1 + \frac{\delta + c - 2c\eta}{\sqrt{(c+\delta)^2 - 4c\delta\eta}}\right]\right\}$$

$$+ \log_{10}e * \left[\frac{-(c+\delta) + \sqrt{(c+\delta)^2 - 4c\delta\eta}}{2}t\right],$$

where $c > \delta$. This follows because the exponential terms on the second line of panel 15 are negligible for $t$ not small. Thus, the viral dynamics model expresses $d_A$, $d_B$, and $d_{A,B}$ as complicated functions of $\eta_A$, $\eta_B$, and $\eta_{A,B}$. Therefore, the sensible definition of additivity, $d_{A,B} = d_A + d_B$, in the $\log_{10}$ virus load change measure prescribes a complicated definition of additivity on the $\eta$ scale, which depends on $c$, $\delta$, and $t$. Conversely, the sensible additivity formula, $\eta_{A,B} = \eta_A + \eta_B - \eta_A * \eta_B$, in the proportion reduction measure $\eta$ prescribes a complicated definition of additivity in $d$, which only approximates the sensible definition $d_{A,B} = d_A + d_B$, if the clearance rates and measurement time take particular values. Given estimates of $c$ and $\delta$, a viral dynamics model could be used to select the measurement time so that interaction in the observable measure $d$ reflects an interaction in the biologically meaningful parameter $\eta$.