In this issue of the Journal, results from a major collaboration of 12 cohort studies—the Antiretroviral Treatment Cohort Collaboration (ART-CC)—are presented, comparing rates of progression to the combined end point of a new AIDS-defining illness or death (AIDS/death) according to initial highly active antiretroviral therapy (HAART) regimen [1]. The results focus on the comparison between efavirenz (EFV)–based regimens and regimens that include either nevirapine (NVP), a protease inhibitor (PI), a ritonavir (RTV)–boosted PI, or abacavir (ABC). There was a significant increase, or a trend toward an increase, in risk of AIDS/death and of death from any cause for all regimens compared with EFV-based regimens, with the exception of a trend favoring ABC-based regimens for progression to AIDS/death (but not to death). The results are reassuring, because they support recommendations given in international treatment guidelines that EFV-based HAART be considered a preferred regimen for the initial treatment of HIV infection [2–4].

The randomized trial is recognized as the best study design for comparing the effects of treatments. During the last decade, however, randomized trials of initial HAART regimens have compared effects either on rates of virological failure or on a composite end point of virological failure or the discontinuation of the randomized treatment (“treatment failure”), sometimes over periods of up to 3 years but often of only 48 weeks. The limited size and duration of these trials provides minimal power to detect clinically relevant effects on disease progression and mortality. Two large randomized trials have shown similar or lower rates of virological or treatment failure for EFV- and PI-based regimens [5, 6], which is consistent with the ART-CC’s finding for disease progression. However, another large randomized trial has demonstrated the inferiority of an ABC-based regimen versus an EFV-based regimen with respect to the rate of virological failure [7, 8], whereas the ART-CC found a similar or perhaps better effect of this regimen on progression to AIDS/death, as well as an increased short-term rate of virological and treatment failure, for NVP-based regimens. Inconsistencies between the ART-CC and randomized trials such as these leave open important questions about treatment-management guidelines, particularly with respect to the merits of PI-boosted and NVP-based regimens, which are recommended alternatives to EFV-based regimens as initial treatment.

There are many reasons why results from observational studies and randomized trials may differ [10], but 2 issues are noteworthy here: the potential failing of surrogate end points and the biases that might arise from the use of different designs. The first issue concerns whether a difference in the short-term rate of virological or treatment failure between randomized treatments is a good surrogate for a difference in the longer-term rate of disease progression or death. It is clear that substantial treatment-mediated reductions in HIV-1 RNA level correlate with improved prognosis, as is evident from clinical end-point trials of early HAART regimens and the reductions in mortality seen in surveillance data after HAART became available [11]. Thus, there is good rationale for approval of HAART regimens that suppress HIV-1 RNA levels to un
detectable levels in a large majority of patients. However, the use of surrogate end points for comparing HAART regimens raises important but subtle issues. Modest short-term differences in HIV-1 RNA levels between randomized treatments are not strongly associated with corresponding differences in rates of progression to AIDS/death [12]. Thus, in equivalence or noninferiority trials, the lack of a difference in rates of virological or treatment failure may not be a reliable indicator of the equivalence or noninferiority of initial treatment with respect to AIDS/death. A particular concern here is that the course of HIV infection among subjects experiencing failure is rarely evaluated, but it is these subjects who are likely at a higher risk of experiencing clinical events. Trials of initial treatments that use surrogate end points should, therefore, compare not only rates of virological or treatment failure but also changes in HIV-1 RNA levels and CD4 cell counts after failure and, hence, the complete distribution of HIV-1 RNA levels and CD4 cell counts across all subjects according to randomized treatment.

As for the second issue, the potential for bias in comparisons of initial treatments in randomized trials is likely to be minimal, because the quality of conduct of these trials is generally very high. A potential concern, however, is the generalizability of results from trials. Lack of generalizability could arise if patient characteristics, patient decisions, or treatment management differ between the trial setting and the broader clinical setting in such a way as to modify the relative effects of treatments in the randomized trial versus the broader clinical setting. This is difficult to evaluate, but there is little evidence, for example, that treatment effects vary substantially among different population subgroups, and reversal of effects in different populations seems, therefore, unlikely.

In observational studies comparing treatments, a major source of bias arises when a factor, such as patient’s disease status, determines which treatment a patient receives and when he or she receives it. In the ART-CC, there is evidence of such “confounding by indication.” For example, patients who received an NVP-based regimen as their initial treatment had less-advanced HIV infection than did those who received an RTV-boosted PI regimen: the median CD4 cell counts were 293 versus 130 cells/μL, the median HIV-1 RNA levels were 4.7 versus 5.2 log copies/mL, and the percentages with Centers for Disease Control and Prevention clinical stage C infection were 12% versus 34%. In the ART-CC, the presented results were adjusted for these measures of disease status as well as for other factors that might cause confounding. Also, the analysis was stratified by cohort and by calendar year of starting HAART. The latter are important, because they likely reflect underlying factors such as treatment-management practices, concomitant care, and available drugs. The extent of confounding by these variables can be seen by comparing the adjusted and crude hazard ratios for treatment comparisons in the ART-CC results: many change by >10%, which epidemiologists often use as a guide for identifying confounding. Treatment comparisons will still, however, be biased if other confounding variables exist that are not included in the analysis. Often these are less-tangible factors. There will be no residual bias only if there are no other factors that affect both disease progression and the decision to use one regimen over another. In essence, this would mean that, apart from the variables included in the analysis, all choices of treatment are made randomly. It is unlikely that this would happen, and, thus, there can never be certainty about the extent to which treatment comparisons in observational studies might reflect confounding rather than true differential treatment effects.

Observational studies may also not provide as much power for the evaluation of treatment effects as first appears. For example, although the ART-CC is impressive in achieving collaboration among so many cohort studies, the confidence intervals (CIs) for the differences in effects between drugs are wide. For example, for the comparison between IDV and EFV, the 95% CI for the adjusted hazard ratio for AIDS/death was 0.94–1.47, such that the study does not rule out either true equivalence or a true difference exceeding a 40% increase in the risk of disease progression. An issue here is that the method of statistical modeling used—the stratification, in particular—meant that the power to compare 2 drugs was determined not so much by the size of the study population but, more particularly, by the number of subjects in strata defined by cohort and by calendar years in which both drugs were used as initial treatment. Thus, the imprecision in the comparison between IDV and EFV reflects the fact that IDV was mainly used as initial treatment until 1999 and that EFV was mainly used as initial treatment from 1999 on. Although alternative methods of statistical analysis may improve power, these may require further model assumptions (e.g., proportional hazards). Such assumptions may be difficult to justify and, if incorrect, may also bias estimates of treatment effect.

In the future, a number of developments might be beneficial and help identify the best regimens for the initial treatment of HIV infection. Cohort studies that are as well designed as possible to deal with issues of confounding [13, 14] should continue to play an important role in exploring the clinical effects (both beneficial and adverse) of treatment-management strategies and in evaluating the reliability and limitations of surrogate markers for predicting these relative effects. The latter will be critically important as new classes of drugs, such as CCR5 antagonists, become available. As the ART-CC appropriately concluded, however, it is critical that randomized trials be pursued to address key questions about treatment management—for example, questions concerning the relative long-term clinical effects of initial regimens that include EFV versus those that include an RTV-boosted
The results of the ART-CC suggest that such trials might require the enrollment of a few thousand subjects who are followed for upwards of 3 years. Furthermore, the diversity of treatments currently used as initial therapy observed in the ART-CC suggests that randomization to different initial therapies might be feasible. For randomized trials that use surrogate end points to compare treatments, it is important to evaluate effects over periods much longer than 48 weeks and to follow all randomized patients so that outcomes in patients who experience failure with the initial randomized regimen are evaluated. Although it is unrealistic to expect all trials of initial treatment to be powered to evaluate clinical end points, a policy requiring follow-up of all randomized patients for long-term vital status (and, wherever possible, for significant disease and drug-related morbidity) would facilitate cross-trial analyses that might help to address important questions about treatment management and, by building on the randomization, to reduce the potential for confounding by unmeasured factors that is inherent in cohort studies.

References