DRU-1365-BMS

September 1995

RAND

The Medicine Education and Research Foundation, San Diego

Measuring Patient-Centered Outcomes in ACTG 116/117: Instruments, Methods, Application, and Results

Prepared for Bristol-Myers/Squibb by

Samuel A. Bozzette
Sandra H. Berry
Naihua Duan
David E. Kanouse

PREFACE

This report summarizes results of work sponsored by Bristol-Myers/Squibb and aimed at capturing the overall outcomes of didanosine treatment in person with advanced HIV disease. While this report concentrates on ACTG trials 116/117, it also reviews selected results from an entire program of research to develop instruments and analytic methods for capturing health status and other patient-centered outcomes in AIDS clinical trials. This is because the sponsorship of Bristol-Myers/Squibb clearly aided the overall development of the program, which was in its infancy at the time this work began. With respect to ACTG 116/117, it is unfortunate that participation at the sites was disappointing and that the goal of timely publication for the trial results was not met due to difficulties in obtaining necessary clinical information and to long editorial delays at journals that ultimately rejected the manuscript detailing results. However, such a manuscript is scheduled for publication in the in the inaugural issue of *Antiviral Chemotherapy* in early 1996, and this research has contributed greatly to the development of instruments and procedures that form the basis for current standard approaches to capturing these outcomes in HIV trials, including further trials conduced by Bristol-Myers/Squibb.

SUMMARY

Clinical trials of therapies for HIV-related illness have focused largely upon survival or freedom from opportunistic infections. Yet clinical decisions take into account expected quality as well as length of survival. Assessment within clinical trials of patients' health status and quality of life therefore add practical information beyond that captured in traditional endpoints. In this study, health status and quality of life were systematically measured as part of ACTG 116 and 117, clinical trials comparing the benefits of didanosine to zidovudine in persons with varying prior experience with zidovudine (efforts to capture similar outcomes from subjects in ACTG 118 and private companion studies failed because of non-participation). A self-administered patient questionnaire was developed that included several scales from the Medical Outcomes Study (MOS). To summarize patient experiences across these dimensions, a Perceived Health Index was developed using weights derived from regression analyses of a reference variable on these scales. The resulting index was shown to be highly reliable.

For this study, 356 participants in ACTG 116 and 117, all of whom had HIV infection and either a CD4+ cell count < 200, or a CD4+ cell count < 300 plus symptoms of HIV disease, were randomized equally within strata defined by duration of prior zidovudine therapy to receive didanosine sachets at a dose of 500 mg daily (334 mg in subjects weighing < 60 kg) or 750 mg daily (500 mg in subjects weighting <60 kg) plus inactive capsules resembling zidovudine, or zidovudine capsules at a dose of 600 mg daily plus inactive sachets resembling didanosine. Measures of self-reported health-related quality of life, health care utilization, disability, work, and symptom impact showed few differences between the treatment groups regardless of the duration of prior zidovudine use. Specifically, there were no differences in reported symptom impact or health care utilization, and most measures of disability were similar. However, in the group with more than 8 weeks of prior zidovudine, several of the health status scale scores for ongoing participants were significantly better for didanosine recipients. Use of several different approaches to combining health status and survival showed no differences in the overall quality-time experiences between the treatment groups. Persons on zidovudine, low dose didanosine, and high dose didanosine experienced 33, 34, and 35 weeks in at least the typical health state if they had less than 8 weeks of previous zidovudine and 23, 23, and 26 weeks, respectively, if they had more than 8 weeks. Results did not differ when data were analyzed within strata of persons who had any versus no prior exposure to zidovudine or AIDS versus non-AIDS status. We therefore conclude that functional status and health-related quality of life were substantially similar among persons receiving either zidovudine or didanosine, regardless of the duration of prior zidovudine treatment.

CHAPTER I: OVERVIEW

INTRODUCTION

Acquired immunodeficiency syndrome (AIDS) is a disease characterized by profound immuno-suppression that results in increased susceptibility to a variety of opportunistic infections and the occurrence of unusual forms of neoplasms. It is currently estimated that about 1 million persons in the United States have been infected with HIV, and it is expected that a large proportion of these individuals will develop AIDS. The vast majority of clinical research studies on HIV infection and AIDS have focused on clinical endpoints, such as time to opportunistic infection or death, to determine therapeutic efficacy of investigational agents.

Because of the emphasis on improving survival or preventing opportunistic infection, less attention has been paid to measuring the therapeutic impact of new agents on general health status or on health-related quality of life. In clinical practice, however, the choice between alternative therapies often rests not on their prophylactic efficacy or effects on survival, which may be quite similar, but on differences in symptom relief and side effects--that is, on which agent can be expected to yield the best level of functioning and quality of life for the patient over the course of treatment. Even where agents have different effects on survival, health status and quality of life are of concern to physicians and patients; some patients may prefer to live (or remain free of opportunistic infection) a shorter time with higher quality of life rather than have survival (or freedom from opportunistic infection) prolonged by a therapy that produces life-interfering side effects and adverse reactions.

For these reasons, assessing health status and quality-of-life may add valuable information to clinical trials that focus largely upon survival status or freedom from opportunistic infections as final endpoints, potentially broadening the basis on which new agents are evaluated for approval and informing subsequent treatment decisions by physicians and patients. To allow for such evaluation, it is important to develop reliable and valid measures that are sensitive to changes in the various domains that health status and quality of life encompass--e.g., physical functioning, distress from physical symptoms, emotional well-being, cognitive functioning, social functioning, role functioning, and general health perceptions. In addition to being reliable and valid, such measures need to be capable of being administered to a clinical population in the field settings represented in a clinical trial.

With this perspective in mind, staff at University of California San Diego and RAND undertook a systematic assessment of health status and quality of life outcomes as part of ACTG 081, a clinical trial of prophylaxis against HIV-related Pneumocystosis and invasive mycoses, and ACTG 981, a nested substudy of preventive therapy for HIV-related fungal disease. Soon after that trial began, parallel

studies were initiated for ACTG 114, a study of comparing zalcytibine to zidovudine, and 116/117, clinical trials of didanosine at two different doses versus zidovudine in patients with advanced HIV disease and varying prior experience with zidovudine. This report summarizes the developmental work performed for these trials, the field experience administering the measures, and the findings that emerged from analyses of health status and quality-of-life outcomes for ACTG 116/117.

MEASUREMENT AND ANALYTIC APPROACH

Drug therapy for HIV-related illness is prescribed to improve a broad range of patient health outcomes, including survival and various medical components of health status. The health outcomes of concern in prescribing a particular drug are often multidimensional, so that arriving at an overall assessment of the effects of treatment requires integrating information on various aspects of health status. This means weighing in some fashion, whether implicitly or explicitly, all relevant aspects of health status potentially affected by the drug, including both toxic and salutary effects. Quality of life and health status measures offer an attractive way to do this, in that they permit the effects of disparate health states to be compared in a common metric based on how they affect the patients' experience of their current health and satisfaction with it. In principle, this facilitates the process of making rational choices that are rooted in patient preferences among therapies that present qualitatively different risk/benefit profiles.

Although interest in patient-reported health status measures as outcomes in clinical trials has been growing, such measures have rarely been used as primary or secondary endpoints in actual trials. Among the reasons for this limited acceptance is the failure of common analytic procedures to meet key requirements: (1) results of clinical trials should be expressed in terms that clinicians, regulators, administrators, and patients find useful; (2) reasonable means of summarizing outcomes should be available *a priori*; (3) scale units should have real-world meaning; (4) unwarranted assumptions regarding scale properties should be avoided; and (5) adequate methods for handling attrition from death or drop-out must be found.

Much of the work performed by our group in connection with this and other ACTG trials was to develop approaches for dealing with these problems. To address requirements (1) and (2), we adapted several scales from the Medical Outcomes Study (MOS) to provide the detail necessary to characterize the varied dimensions of health-related quality of life, but we created an overall summary score (the Perceived Health Index) to use in determining the best overall treatment across all dimensions. We used the highly reliable 5-item MOS Current Health Perceptions scale as a reference variable (1), deriving weights that represent the best linear combination of domain-specific MOS health status scales for predicting Current Health Perceptions.

To address the need for scale units with real world meaning (requirement 3), we extend the language and logic of survival analysis (and related concepts such as "disease-free survival") to describe health-related quality of life more generally. The effects of alternative treatments are compared in terms of the time that patients who receive them experience health states of at least a given quality. This approach shifts the comparison from units that lack an intuitive meaning (such as average scores on an index of health-related quality of life) to units that are better understood (time). This approach also avoids the problem of unwarranted assumptions about scale quality (requirement 4): Because we do not need to average scores on health-related quality of life, our only requirement is that each scale value can be ranked relative to the others (ordinality assumption). It is much easier to order health states (and their corresponding scale values) than it is to assign an absolute value to each.

Once we have measured health states in a way that summarizes information across several domains and produces a score on an ordinal scale describing patients' overall health state, we can employ a set of analytic techniques known as multistate survival analysis to draw proper inferences from repeated outcome measures in clinical trials. Multistate survival analysis, a generalization of survival analysis developed for use in this and other ACTG clinical trials, allows for the combination of data on survival and health status and provides reasonable ways to handle attrition, thereby addressing requirement (5) above. Multistate survival analysis is described in more detail below. First, however, we summarize the psychometric results for the Perceived Health Index.

Psychometric Results

The psychometric properties of the health status scales were assessed using multitrait scaling and test-retest stability. Weights for the index were derived from regressions of Current Health Perceptions on the domain-specific health status scales. The effect of participant characteristics on weights was tested with additional regressions and sensitivity analysis. Finally, the reliability and known-clinical-groups validity of the index were assessed.

Data were obtained from 1,862 participants in various randomized controlled clinical trials of chronic therapies for advanced HIV disease conducted by the AIDS Clinical Trials Group (ACTG) who provided a total of 7,352 observations (2). Over 50% of participants were from sites enrolling at least 96% of candidates. The mean CD4 count was 131. The internal consistency reliability (Cronbach's alpha) of the multi-item scales ranged from 0.86 to 0.90, and items demonstrated excellent discrimination across scales. The domain-specific scales explained 59% of the variation in the Current Health Perceptions scale ($P < 0.00001$). The resulting Perceived Health Index was equal to 0.20*Physical Functioning + 0.15 Pain +0.41*Energy/Fatigue + 0.10*Emotional Well-Being + 0.05*Social Functioning +0.09*Role Functioning. A strong positive bivariate relationship between the Cognitive

Function/Distress scale and the Current Health Perceptions scale was subsumed by the combination of the other domain-specific scales in multiple regressions, so it does not appear independently in the index. The proportional weights used in the index were insensitive to variations in demographics. The reliability of the index was conservatively estimated to be 0.94. Patients with index scores in the lowest quartile had a 2- to 11-fold higher probability than those in the highest quartile of reporting various specific clinical events, and the index correlated significantly more highly with the number of such events than did the current health perceptions scale.

The modified health status scales included in the HIV-PARSE are reliable and valid in patients with advanced HIV disease. The perceived health index provides a reliable and valid means of summarizing self-reported current health, correlates strongly with clinical indicators, and should be useful as an outcome measure in patients enrolling into clinical trials of therapies for advanced HIV disease.

Multistate Survival Analysis

Multistate survival analysis has two components: a description of the time or survival above threshold state (TATS or SATS) and a significance test based on transitions from state. The central notion is that any observed scale score can be used as a threshold dividing participants into two groups, with those having higher than threshold scale scores being considered to have better health. For any threshold score, one calculates the total amount of time during the study that a subject has better than threshold health, or time above threshold state (TATS). Just as the usual survival curve is the complement of the survival time distribution, one can take the complement of the TATS distribution to obtain the survival above threshold state (SATS) curve.

In the calculation of the TATS and SATS, attrition is handled by extending the standard Kaplan-Meier assumption that attrition is uninformative regarding unobserved states.

The SATS curve, a generalization of the usual survival curve, gives the proportion of subjects whose TATS for a given threshold is greater than each given total duration time. If the threshold score being considered is below the lowest observed score, all surviving subjects have better than threshold scores; in this situation, the TATS is equivalent to the usual survival time, and the SATS curve coincides with the usual survival curve. Thus, incorporating mortality into the analysis requires only that death be considered worse than all recorded scores.[1]

If the threshold score is higher than the lowest observed score, the SATS score is analogous to a "survival curve" for time above a clinically defined threshold, such as the proportion of patients not

[1] Allowing live states worse than death in an intervention study is inappropriate because this could allow a fatal side effect of treatment to improve overall health status.

experiencing an opportunistic infection. However, the two curves are generally not identical, because the standard survival curve depicts the chronological time until a one-time-only event, whereas the SATS curve depicts the total cumulative time free of the event (or above the threshold). Thus, the SATS curve captures an unlimited number of deteriorations and improvements over the course of the study. This contrasts with standard survival analysis, which can be used to describe the time until the first opportunistic infection or the first drop below a health status threshold but cannot address the total duration of time free from opportunistic infections or with above-threshold health status.

The TATS and SATS can be estimated for a large number of thresholds covering the entire range of observed health status scores. Averaging the TATS across both patients and thresholds gives a summary of the overall time-quality experience of the cohort; namely, the typical time that a typical patient in the cohort spends above a typical threshold. Plotting the SATS for a given duration of time against thresholds covering the range of observed scores yields a graphical summary of that experience, known as a SATS map. (For further detail, see Bozzette, Duan, and Kanouse (a) in the Appendix.) (3)

The second component of multistate survival analysis allows for inferences to be made on the combination of survival and a continuous outcome measure, such as a health status summary score. Just as multistate survival analysis summarizes time-quality using a generalization of the standard survival curve, the significance of differences between treatments on this combined outcome can be assessed using a generalization of the standard Mantel-Haenszel test for differences in survival. In the standard Mantel-Haenszel test, the cumulative number of excess deaths on the experimental treatment (relative to the standard treatment) is calculated and an estimate of the probability of observing at least this number is made. In the generalized Mantel-Haenszel tests for multistate survival analysis, both the excess deteriorations to unfavorable states (e.g., development of an opportunistic infection, drop in health status score) and excess improvements to more favorable states (e.g., resolution of bacteremia, improvement in health status score) are accumulated over time. The overall difference between two treatments is then assessed using the net excess transitions, which is equal to the excess deteriorations minus the excess improvements. The significance of deviations from the expected number of net transitions is then tested using the permutation test. The permutation test is essentially a simulation in which subjects are repeatedly "reassigned" to the treatment groups. The amount of variation that results from random reassignment reflects the amount that would be seen if treatment did not matter, which is the null hypothesis. Experimental results that are large relative to that amount of variation are unlikely to have occurred by chance. (For further detail, see Bozzette, Duan, Kanouse (b) in the Appendix.) (4)

REVIEW OF ACCOMPLISHMENTS

The following is a brief summary of the work carried out in the research program that was supported in part by a grant from Bristol-Myers Squibb to the Medical Education and Research Foundation, with references to later chapters of this report or to appended publications that provide more detail.

Instrument Development

We developed the HIV-PARSE Instrument specifically to measure patient-reported global health status and functioning, symptom impact, disability, work, and health care service utilization of patients with HIV disease. Chapter 2 describes the instrument, how it was developed, modifications made for ACTG 116/117, and how it has been changed as a result of experience in this and other ACTG trials. We also developed a brief version of the instrument, as described below.

Administration of the HIV-PARSE Instrument

The HIV-PARSE instrument was administered to 644 men and women enrolled in the ACTG 116/117 clinical trial across 38 treatment sites. Questionnaires were administered in clinic settings, at the time of clinic visits; patients generally filled them out in the waiting area. We sought to obtain completed instruments at baseline and every 12 weeks thereafter. In addition, many patients filled out questionnaires at other clinic visits in "off weeks."

Across all clinic visits, we received a total of 1,598 questionnaires, or 2.5 per enrolled patient. Of 644 patients enrolled in the HIV-PARSE substudy, 465 (72%) completed at least two surveys. Across all "expected study weeks" we received 1,598 questionnaires, 25% of the 6,380 that would be expected if all enrolled patients completed a form at each scheduled visit. For further information on our field experience, see Chapter 2.

The questionnaires were distributed and collected by personnel at the ACTG study sites. RAND developed a protocol for administering the HIV-PARSE that was compatible with Study Protocol 116/117, and RAND staff conducted a group training session at an ACTG meeting to explain the use of the HIV-PARSE form and to answer questions from representatives of the participating sites. Study staff at RAND maintained regular contact with ACTG site staff via electronic mail and telephone and coordinated the mailing and return of completed instruments.

Data collection at each site was tracked using a database management system, which was also used to generate periodic reports on the completeness of data acquired. Data received at RAND were entered

into machine readable form and checked for consistency and accuracy. Various analyses were performed to assess the reliability and validity of the data, as described elsewhere in this report.

Scale Development

As described above, we developed a Perceived Health Index that combines measures of specific health domains considered relevant to treatment into an overall summary of health status. This work is described in Bozzette, Hays, Berry, and Kanouse (1994) (see Appendix). (2)

Short Form

We found that the health status measures in the HIV-PARSE instrument have high reliability; this suggested that several of the multi-item scales could be shortened substantially while still maintaining adequate reliability. Because acceptance of health status measures has been limited in part by concerns over item redundancy and investigator and patient burden, shorter scales are likely to be more useful in clinical trials. Accordingly, we used data from the ACTG 116/117 and other trials (n = 1,934 participants) to select items for shorter scales based on static and dynamic relationships to longer scales and to indicators of clinical and functional status. We developed a set of scales covering disability, work, utilization, and health status, and containing a total of 21 items, as compared to 38 in the longer instrument. The analytic work in developing this short form was partially supported by the grant for quality of life outcomes in this trial, and is described in Bozzette et al., (1995) (included in the Appendix). (5)

Analytic Methods

Some of the problems encountered in attempting to include quality of life outcomes in clinical trials and our conceptual approach to dealing with these problems are described in Bozzette, Duan, Berry, and Kanouse (6). The analytic methods are described in more detail in the following papers that have been submitted for publication:

> Bozzette SA, Duan N, Kanouse DE. Multistate survival analysis I: Time above threshold state. (Submitted.)
>
> Bozzette SA, Duan N, Kanouse DE. Multistate survival analysis II: Generalized Mantel-Haenszel tests of transitions from states. (Submitted.)

Copies of all these papers may be found in the Appendix.

SUMMARY OF ANALYTIC RESULTS FOR ACTG 116/117

The methods described above were applied to analyze health status and quality of life outcomes in a study comparing the effects of zidovudine and didanosine in persons with advanced HIV infection (7). ACTG study 116 and 117 participants had HIV infection and either a CD4+ cell count < 200 or a CD4+ cell count < 300 plus symptoms of HIV disease. All were randomized equally within strata defined by duration of prior zidovudine therapy to receive didanosine sachets at a dose of 500 mg daily (334 mg in subjects weighing < 60 mg) or 750 mg daily (500 mg in subjects weighing <60 kg) plus inactive capsules resembling zidovudine, or zidovudine capsules at a dose of 600 mg daily plus inactive sachets resembling didanosine. Three hundred fifty six participants participated in this substudy.

The HIV-PARSE instrument was administered to measure health-related quality of life, health care utilization, disability, work, and symptom impact. There were no differences in reported symptom impact or health care utilization, and most measures of disability were similar. In the group with more than eight weeks of zidovudine experience, several of the health status scale scores for ongoing participants were significantly better for didanosine recipients, but average differences were relatively small for these and other comparisons. Multistate survival analysis, which was reworked for this study to emphasize the mean time above threshold (MTATS) as a more comprehensive and easily understood outcome than survival above threshold states (SATS), showed no significant differences in health state transitions or the overall quality-time experiences between treatments. Persons on zidovudine, low dose didanosine, and high dose didanosine had 33, 34, and 35 weeks of 64 weeks, respectively, in at least the typical health state if they had fewer than eight weeks of previous zidovudine and 23, 23, and 26 weeks, respectively, if they had more than eight weeks. Results did not differ when data were analyzed within strata of persons who had any versus no prior exposure to zidovudine or AIDS versus non-AIDS status.

Despite the differences reported in the clinical outcomes for this trial (8, 9), functional status and health-related quality of life were substantially similar among persons receiving either zidovudine or didanosine regardless of the duration of prior zidovudine treatment. The result stands in contrast to the findings for the zalcitabine trial described above, where differences in functional status and health-related quality of life as revealed in multistate survival analysis preceded and were larger in magnitude than clinical differences.

APPLICATIONS IN OTHER CLINICAL TRIALS

The methods described above were also applied to analyze health status and quality of life outcomes in other ACTG trials such as ACTG 081/981 and 114. ACTG 081 was a trial of three different forms of prophylaxis against HIV-related Pneumocystosis while ACTG 981 was a nested trial of prophylaxis against invasive fungal infections in individuals. In ACTG 081, the goals was to use the HIV-PARSE

instrument to help synthesize the toxicity/efficacy tradeoffs between the trimethoprim/sulfamethoxazole, dapsone, and aerosolized pentamidine. In the 981 substudy, the HIV-PARSE instrument was used to illuminate the overall effects of giving prophylaxis for a relatively infrequent disease.

In ACTG 081, Clinical data indicated that differences in prophylactic efficacy between systemic and aerosolized pentamidine therapy were not large. Analysis of data from 785 patients who completed a baseline HIV-PARSE form and at least one follow-up showed that mean scores on all MOS scales were essentially identical for all scales and indices except for the Current Health Perceptions scale, which was 0.15 standard deviation higher in the aerosolized pentamidine group. Typical entrants in the trimethoprim/sulfamethoxazole, dapsone, and aerosol pentamidine arms spent 12.9, 12.5 and 13.1 months, respectively in at least a typical health state (P = 0.27 to 0.66). Disability and utilization of health care were also similar across the groups. However, when sub-groups with fewer than 100 CD4+ cells were analyzed separately, functional and health status measures became less favorable and differences between treatment groups were larger and more consistent. On essentially all these measures, dapsone recipients reported more favorable outcomes. In addition, the mean time above threshold state (MTATS) map suggests a trend toward better quality-survival among those randomized to systemic therapy.

In ACTG 981, overall outcomes differed very little overall between the two arms. However, among those entering with fewer than 50 CD4+ cells, scores on the Current Health Perceptions scale were higher in the fluconazole group while disability was more common and utilization of medical procedures and hospitals was greater in the clotrimazole group. Multistate survival analysis on the Perceived Health Index showed that a typical person entering 981 with fewer than 50 CD4 cells at baseline spent eight months with at least the typical health state if assigned to clotrimazole and 9.4 months if assigned to fluconazole (P = 0.09).

The HIV-PARSE instrument and the measures and analytic methods developed with support from this grant were also applied in ACTG 114, a study of the outcomes of prescribing zalcitabine (ddC) or zidovudine (AZT) for initial therapy of advanced HIV disease (10). Patients participating in this trial had HIV infection, fewer than 200 CD4+ cells, and either a history of *Pneumocystis carinii* or symptoms of HIV infection. In this study, differences between the treatments were striking. Zalcitabine recipients were twice as likely to undergo an invasive procedure (P = .004) or be admitted to a hospital (P = 0.12). Zalcitabine recipients reported > 40% more symptoms that interfered with their activity (P = .001) and > 50% more disability days (P<.01). They also had a 7% lower employment rate and a 35% lower monthly income. Average observed health status scores were lower in zalcitabine recipients overall, but especially in the early portion of the study. When survival and health status data were combined using multistate survival analysis, results showed that, over 76 weeks of study, a typical zidovudine recipient

spent about 4 (10%) more weeks with at least the typical health state than a typical zalcitabine recipient. These results illustrated the sensitivity of functional outcome measures and how inclusion of such measures can improve the information available from a clinical trial (10).

REFERENCES

1. Stewart AL, Hays RD, Ware JE. Health perceptions, energy/fatigue, and health distress measures. In Stewart AL, Ware JE, eds. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach.* Durham and London: Duke University Press, 1992: 143-172.

2. Bozzette SA, Hays RD, Berry S, Kanouse DE. A perceived health index for use in persons with advanced HIV disease: Derivation, reliability and validity. *Medical Care* 1994; 32:716-731.

3. Bozzette SA, Duan N, Kanouse DE. Multistate survival analysis I: Time above threshold state. (Submitted.)

4. Bozzette SA, Duan N, Kanouse DE. Multistate survival analysis II: Generalized Mantel-Haenszel tests of transitions from states. (Submitted.)

5. Bozzette SA, Hays RD, Berry SH, Kanouse DE, Wu AW. Derivation and properties of a brief health status assessment instrument for use in HIV disease. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* 1995; 8:253-265.

6. Bozzette SA, Duan N, Berry S, Kanouse DE. Analytic difficulties in applying quality of life outcomes to clinical trials of therapy for HIV disease. *Psychology and Health* 1994; 9:143-156.

7. Bozzette SA, Kanouse DE, Duan N, Berry S, Richman DD. Impact of zidovudine compared to didanosine on health status and functioning in persons with advanced HIV infection and a varying duration of prior zidovudine therapy: Results from a randomized trial. (Submitted.)

8. Kahn JO, Lagakos SW, Richman DD, et al. A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. *New Engl J Med* 1992;327:581-7.

9. Dolin R, Amato DA, Fischl MA, et al. Zidovudine compared with didanosine in patients with advanced HIV type 1 infection and little or no previous experience with zidovudine. *Archives Intern Med.* 1995;155:961-74.

10. Bozzette SA, Kanouse DE, Berry S, Duan N. Health status and function with zidovudine or zalcitabine as initial therapy for AIDS: A randomized placebo-controlled trial. *JAMA* 1995; 273:295-301.

CHAPTER II: THE HIV-PARSE INSTRUMENT: DEVELOPMENT AND FIELD EXPERIENCE

DEVELOPMENT OF THE HIV-PARSE INSTRUMENT

Interest in using patient-reported assessments as an outcome in clinical trials of chronic therapies for Human Immunodeficiency Virus (HIV) disease is increasing for several reasons (1, 2). First, many investigators believe that improvement of health, as perceived by the patient, is the principal goal of non-curative treatment. Second, most patients participating in HIV-related clinical trials are receiving multiple treatments and medications with diverse toxic and salutary effects. In a situation where patients have a progressive disease and are receiving multiple therapeutic agents, it is difficult to isolate the effect of a single new therapy solely by comparing the occurrence and the time to onset of clinical events. Further, interpretation of the relative clinical meaning of dissimilar events, such as episodes of *Candida esophagitis* versus pancreatitis, is unclear. Finally, short term prognosis for HIV disease has improved to the point that there are fewer clinical events per patient/month of followup in clinical trials, thereby reducing the statistical power of comparisons based on counts of events.

Against this background, measurement of health status and functioning can be a useful adjunct to traditional outcome measures by providing a direct measure of health in the absence of marker clinical events as well as an integrated measure of overall effect in patients who do and do not experience such events.

ACTG 081 represented the first of the NIAID HIV-related clinical trials to include health status as an official outcome of the trial. The HIV-PARSE instrument was developed to measure patient reports of health status. ACTG 116/117, which compared didanosine to zidovudine, used a version of HIV-PARSE instrument that was developed and tested for ACTG 081 which excluded symptoms but was otherwise only lightly modified

OVERVIEW OF HIV-PARSE INSTRUMENT

The HIV-PARSE instrument is designed in four main sections: demographic background, risk group and life circumstances; health status, including a quality of life measure; utilization of health services; and a checklist of symptoms and symptom impact. The purpose and source of items for each section are described below. The scoring rules for the scales are explained and the reliability of the scales, based on a HIV clinical trial population, are provided. Included as appendices are the version of the form used in the ACTG 116/117 trial and the source listing for each item in it.

Demographic Background, Risk Group, and Life Circumstances

This section includes items on work status in the past week, type of work respondent does, total personal income for the past month, type of health insurance coverage, where respondent is living, with whom respondent is living, availability of social support and extent to which care the respondent is providing for someone else interferes with the respondent's life. The main purpose of this section was to provide explanatory or control variables to use in the analysis of the effects of the drug on patient functioning and utilization of health services. In addition, certain variables were included to serve as outcome variables for some longitudinal analyses; for example, change in work status can be tracked over time.

Most items in this section were drawn from standard sources such as the Medical Outcomes Study, the US Census, and the NORC General Social Survey, often with some modifications for this survey. For example, the type of work item is drawn from the US Census, with examples associated with each category selected to include type of jobs common to the population of the study. The type of insurance item includes a category indicating whether the respondent uses his usual insurance to pay for HIV care (some HIV patients avoid doing this in order to protect their privacy). Several items were developed for this survey, including living situation, and social support provided by the respondent.

Health Status and Quality of Life

The second section is mainly drawn from items in the 24-month Patient Assessment Questionnaire, a self-administered questionnaire used in the Medical Outcomes Study and administered by mail two years after entry into the study (3, 4). Modifications were made, however, to improve the quality of measurement for an HIV population versus a general patient population.[2] Measures in this section include:

Current health perceptions,
Physical functioning
Pain
Energy/fatigue
Emotional and well-being
Psychological distress and well-being
Cognitive functioning
Social functioning
Role functioning
Quality of life
Disability days

---

[2]Main modification was to change the wording of the role functioning response categories from "yes, for more than 3 months," "yes, for less than 3 months," or "no" to "yes, all of the time," "yes, some of the time," or "none of the time."

These items are designed to serve as outcome measures that tap specific dimensions of self-reported health and quality of life. The reference period for these items is the implicit "now" for current health perceptions, physical functioning, quality of life, and will to function. For all other measures, it is the four-week period prior to administration of the questionnaire. Scale means and standard deviations are shown in Table 1.

TABLE 1

HEALTH STATUS SCALE SCORES FROM THE HIV-PARSE EXPRESSED ON 0 AND 100 SCALE WITH HIGHER NUMBERS BEING BETTER (HIGHER FUNCTIONING, LESS DISCOMFORT, ETC.)

Scale	Number of Items	Mean*	Standard Deviation*	Reliability
Current Health Perceptions	5	52	26	.88†
Physical Functioning	6	79	24	.87†
Energy / Fatigue	4	59	24	.89†
Emotional Well-Being	5	70	20	.86†
Cognitive Functioning / Distress	6	84	18	.90†
Social Functioning	1	77	27	0.74 - 0.84§
Role Functioning	2	77	29	.86†

*Based on 7352 responses in 1,856 participants.

†Cronbach's α calculated based on 1588 respondents. The maximum possible α is 1.0; scores of >0.7 are considered acceptable for group comparisons (1, 2).

§Lower estimates based on the application of the Spearman-Brown Prophesy formula to MOS data; upper estimates based on a comparison of the test-retest stability at a 4 to 8 (mean 6) week interval relative to that observed for the Current Health Perceptions scale.

Utilization of Health Services

The next section measures use of health services including whether the respondent had seen various types of providers, the number of in person contacts with providers, number of home visits from a doctor or nurse, number of telephone contacts, the number of overnight hospital stays, number of nights in the hospital, and whether the respondent was in ICU or in a hospice or nursing home. We also asked whether patients had received help with personal care at home (with bathing and dressing) and with other chores from unpaid or paid helpers. Again, the reference period in this section was the implicit "now" or the four weeks prior to administration of the questionnaire, as described above. Patients have been shown to be reliable reporters of medical care process events (5). We also asked about a list of medications and treatments they may have had, and if they had them, the number of days out of the past

4 weeks when they had them. These items were designed to allow comparison of effects of drug combinations involved in the trial through comparison of specific items (e.g., treatments, diagnostic procedures), counts of items, or dollar values of item bundles, based on a value scale for the items.[3] This scale is currently being selected and tested. Most of these items were adapted from the Medical Outcomes Study.

PRETESTING OF THE HIV-PARSE QUESTIONNAIRE

The questionnaire was pretested with approximately 40 adult patients at the U.C. San Diego AIDS Clinical Trials Group (ACTG). Each patient completed the questionnaire, recording the time started and finished. In addition, cognitive pretesting was carried out with ten patients completing the questionnaire in the presence of an interviewer who timed the administration and answered any urgent questions. After completion of the form, the interviewer debriefed each pretest respondent individually to determine whether there were any portions of the questionnaire that were unclear, confusing or offensive. This included not only asking the respondent to raise problems, but also probing for understanding of key concepts in the questionnaire and for strategies used by respondents to answer the questions. Pretest instruments were data entered and checked for scale properties (Cronbach's alpha reliability) and distributions of key variables (checking to make sure response categories captured variation among the respondents).

MODIFICATIONS TO THE HIV-PARSE

Several modifications were made to the 081 form to develop the ddI version of the instrument. The changes are noted briefly in Table 2, along with the reasons for these changes. Items deleted to reduce length were simply combined or dropped to reduce the administration time and respondent burden. In a study with these topics as a central focus, the longer version of the items could be used. In addition, the questionnaire has also been translated into Spanish. The questionnaire was translated from English to Spanish and back-translated from Spanish to English; discrepancies were resolved by discussion among the translators and the investigators. The Spanish version was administered in several locations, including Florida, New York, and southern California, and the wording did not pose problems in the field. However, the psychometric properties of the Spanish version have not been completely evaluated. The most current version of the HIV-PARSE questionnaire (English/Spanish version) is included in this report as an appendix to MR-342-NIAID for the convenience of readers.

---

[3]A value scale provides a standard dollar value associated with a particular treatment or procedure, based on an average cost or price. It is used to compare the economic value of services or treatments provided in different care systems, where cost accounting procedures may result in differences in observed price or where dollar values may not be provided (e.g., within a health maintenance organization where care is prepaid).

TABLE 2:

CHANGES MADE TO HIV-PARSE INSTRUMENT

TYPE OF CHANGE	REASON FOR CHANGE
Reduced the section on work history in the previous 12 months from four to two items.	Reduce length
Dropped the item on importance of job skills	Reduce length
Dropped items on living situation	Reduce length
Dropped items on emotional and behavioral control	Did not add unique information over and above other emotional well-being items in terms of predicting overall health perceptions
Added items on group home or hospice stays and paid in-home assistance with bathing, dressing, or household chores	New items to measure important types of utilization
Deleted symptom checklist	Reduce length; avoid dual reporting

FIELD EXPERIENCE

The HIV-PARSE instrument was scheduled to be administered to 1743 men and women over 12 years of age enrolled in the ddI clinical trials (ACTG 116/117) of therapy for advanced HIV disease across 67 treatment sites The questionnaires were planned to administered in clinic settings, at the time of clinic visits. However the PARSE component of the study was initiated at a point when over 1/3 of the patients in the trial were already enrolled, and sites never viewed it as an integrated component of the study. This reduced participation rates, compared with similar and contemporaneous trials, such as ACTG 081 (6).

Of the 67 sites eligible to enroll patients in PARSE, 38, or 71 percent did enroll patients. As shown in the first column of Table 3 the sites enrolled varying percentages of the eligible patients in the trial into the PARSE study, from as small a percentage as 0 or 1 percent to all of the eligible patients, or 100%. Column 2 shows the number of sites that enrolled each fraction of the eligible patients. Twenty-nine sites enrolled none of their eligible patients, and only one site enrolled all of them. At least half their eligible patients were enrolled into the PARSE study in 40 percent of the sites.

Table 4 shows the numbers of patients enrolled by site. Column one shows the number of patients enrolled in the site. Columns two and three show the distribution and percentages patients enrolled across sites.

TABLE 3

SITE ENROLLMENT OF ELIGIBLE PATIENTS INTO THE PARSE STUDY

Percentage of Clinical Patients Enrolled in PARSE Study by Site	Number of Sites	Percentage of Sites
0	29	43.0
2	2	3.0
3	1	1.5
5	1	1.5
9	1	1.5
10	1	1.5
19	1	1.5
30	1	1.5
36	1	1.5
40	1	1.5
47	1	1.5
53	1	1.5
54	1	1.5
55	2	3.0
57	1	1.5
59	1	1.5
60	1	1.5
62	2	3.0
63	2	3.0
64	1	1.5
66	2	3.0
67	2	3.0
68	1	1.5
69	1	1.5
70	1	1.5
72	1	1.5
73	1	1.5
77	1	1.5
81	1	1.5
83	1	1.5
88	1	1.5
89	1	1.5
100	1	1.5
TOTAL	67	100.0

SOURCE: Calculated from data supplied by Bristol-Myers Squibb Company.

TABLE 4

DISTRIBUTION OF ENROLLMENT OF PARSE STUDY PATIENTS ACROSS SITES

Number of PARSE Patients by Site	Number of Sites	Percentage of Sites
0	29	43.0
1	4	6.0
2	1	1.5
3	1	1.5
5	3	4.5
6	1	1.5
10	3	4.5
12	1	1.5
13	2	3.0
16	4	6.0
17	1	1.5
18	2	3.0
19	1	1.5
20	1	1.5
21	2	3.0
22	2	3.0
24	1	1.5
26	2	3.0
28	1	1.5
32	1	1.5
34	2	3.0
48	1	1.5
52	1	1.5
TOTAL	67	100.0

Source: Calculated from data supplied by Bristol-Myers Squibb Company.

For purposes of evaluating nonresponse, we define the response rate as the number of questionnaires received over the number of questionnaires expected from ever-baselined patients, taking into account discontinuations reported to us by the sites.

The total number of questionnaires received, including those assignable to "expected study weeks" and those received in "off-weeks," is 1598 out of a total of 6380, a total response rate of 25 percent. Because the planned analyses can make use of data from all questionnaires, this last total represents the potential sample size and response rate for those analyses.

Once enrolled as a participant in the ddI trial, patients can go off protocol for a variety of reasons, including death and dropping out of the study for other reasons. Table 5 indicates by week of protocol the number of active clinical enrollees, the number of PARSE enrollees who completed forms for that study week within two weeks before or after the scheduled administration, the percentage of active clinical enrollees who completed PARSE forms within a two weeks of scheduled administration, the number of forms for each study week that were completed more than two weeks earlier or later than the scheduled week, the percentage of forms for each study week received off schedule, and the total number of forms received. The percentage of active enrollees who completed forms as scheduled for each study week is varies from 90 percent in week 12 to 37 percent in week 108 of the study. Overall, we received 19 percent of the expected forms that were completed within a two weeks of when they should have been completed and another 6 percent of forms that were completed more than two weeks later than they should have been. Overall, we received 25 percent of the expected forms.

TABLE 5

PERCENTAGE OF PARSE FORMS RECEIVED BY STUDY WEEK

Study Week	Clinical Number Enrolled	Number of Forms Agree	Percentage of Forms Agree	Number of Forms That Do Not Agree	Percentage of Forms That Do Not Agree	Total Number of Forms Received
0	1743	451	84	89	16	539
12	1583	349	90	48	10	387
24	1399	270	79	71	21	341
36	1122	177	67	87	33	264
48	853	151	66	77	34	228
60	586	116	76	37	24	153
72	367	65	67	32	33	97
84	253	41	69	18	31	59
96	145	17	44	22	56	39
108	72	11	37	19	63	30

Source: Calculated from data supplied by Bristol-Myers Squibb Company.

SPECIAL STUDY OF PATIENT/STAFF REACTIONS TO SURVEY ADMINISTRATION

To address concerns about burden and confidentiality, we conducted a special survey of 94 patients and 20 members of the clinic staffs in March 1991 at six ACTG 081 sites, as reported in Kanouse, et al.. (6). Most patients reported that they completed the HIV-PARSE baseline form in the waiting area; about a third finished it while waiting in the examination room. Only 5 percent required assistance with the questionnaire or used audio tapes provided for people with vision or reading problems, although 5

percent said they had some difficulty reading the questionnaire and 7 percent reported difficulty with writing responses.

AIDS is a very difficult disease, and the PARSE instrument requires respondents to focus on its effects in a very systematic and personal way. Patients reported mixed views of their experience in filling out the form over the course of many visits. We asked patients to agree or disagree[4] with a series of statements about the PARSE form. The results are shown in Table 6.

TABLE 6

PATIENTS' VIEWS OF HIV-PARSE INSTRUMENT

Percentage who indicated they "agreed strongly" or "agreed somewhat" with each statement	Statements about the HIV-PARSE instrument
29	Depressing
25	Invasion of privacy
74	Repetitive
55	Asked about important parts of their lives
58	Might help others
61	Dealt with issues that should be studied as part of clinical trials for HIV treatment

In general, staff tended to overestimate how long it took patients to complete the form, how much assistance they needed, and the extent of patients' negative reactions to the form when compared to reports by patients. About 60 percent of the staff reported that some patients had difficulty reading the questionnaire and 30 percent reported that patients had physical difficulty writing answers to the questionnaire. Nearly 60 percent of the staff members felt that the questionnaire was sad or depressing for patients and 90 percent felt that it was repetitive. However, 80 percent of the staff felt that the questionnaire asked about things that should be studied as part of clinical trials.

RELIABILITY OF THE MEASURES

In addition to concerns about the feasibility of administering these measures in a clinical practice setting, researchers are concerned about the reliability of the measures for patients with advanced HIV disease. Table 1 reports information about the reliability, central tendency, and variability of the scales scored using this method for the HIV-PARSE samples, scored as described in section 2. The reliability of the scales is very good, ranging from a low of .64 (depending on how the reliability score is calculated) to a high of .90 (7).

[4]Response categories were "agree strongly," "agree somewhat," "neutral," "disagree somewhat," or "disagree strongly."

CONCLUSIONS

Overall, our results with this form indicate that it is quite feasible to incorporate measurement of health status and quality of life into HIV clinical trials of therapies, but that it is important to explain the purpose of the form, the reliability and validity of the measures (including why questions are seemingly repetitive), and how the resulting data will be used, to both patients and clinic staff. The PARSE form is used in the context of a clinical trial that includes many tests and procedures that are not pleasant to receive or administer, but which provide valuable data for evaluating alternative therapies. Viewed in this context, the burden of completing the PARSE form is a reasonable trade-off against the value of providing reliable and valid measures of patient experience with their health during the trial.

REFERENCES

1. Wu AW, Rubin HR, Mathews WC, Ware JE, Brysk LT, Hardy WD, Bozzette SA, Spector SA, Richman DD, "A Health Status Questionnaire Using 30 Items From the Medical Outcome Study: Preliminary Validation in Persons With Early HIV Disease," *Medical Care*, Vol. 29, 1991, pp. 786-798.

2. Hays RD, Shapiro MF, "An Overview of Generic Health-Related Quality of Life Measures for HIV Research," *Quality of Life Research*, Vol. 1, 1992, pp. 91-97.

3. Stewart AL, Ware JE, *Measuring Functioning and Well-Being*, Durham and London: Duke University Press, 1992.

4. Hays RD, Sherbourne C, Mazel RM "The RAND 36-Item Health Survey 1.0," *Health Economics*, 2:217-227, 1993.

5. Brown JB, Adams ME, "Patients as Reliable Reporters of Medical Care Process-Recall of Ambulatory Encounter Events," *Medical Care*, Vol. 30, No. 5, 1992, pp. 400-411.

6. Kanouse DE, Bozzette SA, Berry SH, Duan N. *Development of Instruments and Analytic Methods for Measuring Patient-Centered Outcomes in Clinical Trials for AIDS and Application in an ACTG Prophylaxis Trial.* RAND DRU-1217-HU September 1995.

7. Bozzette SA, Hays RD, Berry S, Kanouse DE. "A Perceived Health Index For Use In Persons With Advanced HIV Disease: Derivation, Reliability And Validity." *Medical Care* 1994; 32:716-731

CHAPTER III: THE IMPACT OF ZIDOVUDINE COMPARED TO DIDANOSINE ON HEALTH STATUS AND FUNCTIONING IN PERSONS WITH ADVANCED HIV INFECTION AND A VARYING DURATION OF PRIOR ZIDOVUDINE THERAPY

INTRODUCTION

The mainstay of antiretroviral therapy for advanced HIV disease has been zidovudine (formerly azidothymidine (1). Didanosine, another antiretroviral nucleoside active against HIV, is also approved and widely accepted for therapy of advanced HIV infection (2). Comparative clinical trials of these agents conducted in persons with advanced HIV disease and varying duration of prior zidovudine treatment yielded complex results. Patients randomized to didanosine for initial therapy had a shorter survival (3). But, among patients who had at least eight weeks of prior zidovudine treatment, those randomized to didanosine had a greater period of freedom from opportunistic complications (4). Although these clinical findings are revealing, patients and doctors also want to know how well patients will function and feel before, during, and after endpoint events.

Supplementing physiologic and clinical measures with measures of functioning, health status, and utilization can capture additional information on the impact of disease and treatments on patients' well-being. Measures of health status have been shown to be sensitive to differences between clinically equivalent treatments in other settings (5). In HIV disease, several groups have shown that self-report survey instruments based on Medical Outcome Study general health and well-being measures can reliably capture health status (6, 7). However, such assessments are confounded by attrition and mortality in HIV and other fatal diseases. Although data from ongoing study participants is of interest, information obtained from supplemental health status measures is most useful when it is combined with survival data.

In one antiretroviral trial, such an integrated approach demonstrated differences between treatments long before a survival difference appeared (8). In that study, there were also large differences in simple measures of functional impairments such as bed days, hours worked, and hospital use. This study examines similar comparative information for didanosine and zidovudine. Specifically, we report the results of a substudy assessing self-reported functional outcomes within the AIDS Clinical Trials Group's (ACTG) randomized controlled trials of zidovudine versus didanosine for advanced HIV disease.

METHODS

Participants in ACTG 116/117 were persons with HIV infection and either a CD4+ cell count of fewer than 200, or a CD4+ cell count of fewer than 300 plus symptomatic HIV disease. The study

medications, procedures, and centers were similar in the two trials. Participants were randomized equally within strata defined by duration of prior zidovudine therapy (116a:0-7 weeks, 8-15 weeks; 116b/117: 16-47 weeks, ≥48 weeks) to receive didanosine at a dose of 500 mg daily (334 mg in subjects weighing < 60 kg) or 750 mg daily (500 mg in subjects weighting <60 kg) plus inactive capsules identical to zidovudine or zidovudine at a dose of 600 mg daily plus sachets identical to didanosine. For participants experiencing severe toxicity or an opportunistic complication signaling progression of disease, treatment was switched to the alternative drug in such a way that blinded administration was maintained.

Data on all patients, all subgroups, all *a priori* strata, and all persons with no prior experience with zidovudine were analyzed separately. However, this report follows the lead of the clinical reports in focusing on participants concurrently randomized to the three treatment arms. Persons enrolled into ACTG 116 after December 1990 were excluded because a decision to terminate randomization to the 750 mg didanosine arm led to the erroneous non-random assignment of all participants to zidovudine at some sites after that date (R. Dolin , University of Rochester, personal communication). Also as in the clinical reports, this report focuses on the period of therapy prior to crossover to the alternate therapy to maximize the opportunity to discover differences between the treatments. Finally, to avoid problems of small numbers in the multiple *a priori* strata and to provide data comparable to that found in the clinical reports, this report stresses analyses conducted within the strata of persons with fewer than eight weeks of prior zidovudine therapy and a group formed by aggregating strata of persons having 8 to 16 weeks, more than 16 weeks, and more than 48 weeks of prior zidovudine therapy.

At entry and monthly follow-up, participants received standard clinical and laboratory evaluations. Persons agreeing to participate in the substudy completed the HIV-PARSE survey instrument at baseline and every 12 weeks for the duration of the study. The HIV-PARSE survey instrument includes questions on utilization, disability, work, and symptom impact as well as adapted versions of the Medical Outcomes Study (MOS) scales covering Current Health Perceptions (5 items), Physical Functioning (6 items), Pain (1 item), Energy/Fatigue (4 items), Emotional Well-Being (5 items), Cognitive Functioning/ Distress (6 items), Social Functioning (1 item), and Role Functioning (2 items) (9). These versions of the MOS scales have been shown to be highly reliable and valid in this population, to yield scores approximately normally distributed, and to correlate with health care utilization as well as with clinical and physiologic status.

MOS scale scores were analyzed individually and summarized into an index. The perceived health index is a weighted average of the scale scores, with the weights reflecting the importance of the scales in predicting a summary perceived health scale score. The index, which has an estimated reliability coefficient of .94, is equal to (0.20 * Physical Function score) + (0.15 * Pain score) + (0.41 *

Energy/fatigue score) + (0.10 * Mental Health score) + (0.05 * Social Functioning score) + (0.09 * Role Functioning score). Comparisons of the scale and index scores were standardized by dividing difference scores by the standard deviation for the relevant scale at baseline, because interest lies in differences that are substantial relative to the variation in the population. A large difference in this setting has been characterized as 0.5 standard deviations. The significance of differences between treatment groups was initially assessed in a standard fashion using ANOVA to control for the effect of time on study.

This standard approach to the analysis of MOS-based scales is problematic in a trial where mortality is common because it does not incorporate a score for death (i.e., it treats death as a form of attrition or censoring). To test the effect of this potential bias on conclusions, MOS general health perception scale scores were converted to an anchored 0 to 10 scale with 0 being dead (or rated as bad as being dead) and 10 being perfect health. This was accomplished by regressing the responses of 370 persons with advanced HIV disease participating in ACTG clinical trials to a 0 to 10 categorical rating scale on their responses for the general health perceptions scale. The resulting predictive equation, which had an $R^2 = .38$ ($P < .001$), was:

$$\text{categorical rating score} = 46.0 + (.450 * \text{general health perceptions scale score}).$$

Predicted scores from ongoing patients were combined with scores of 0 assigned to patients who died and analyzed as above. In addition, to assess the effect of bias arising when sicker patients drop out (causing the average scores of remaining patients to rise), predicted rating scores were used to "weight" the survival of individual patients and standard methods of survival analysis were used to estimate quality-adjusted life years.

Finally, to reduce reliance on assumptions regarding the scale properties of the health status measures, additional analyses were performed using a novel approach to combining survival and health status data known as multistate survival analysis. Standard approaches to failure-time data estimate the mean or median time alive or in some clinically defined state such as free of a new opportunistic infection. In multistate survival analysis, estimates are made of the mean-time-above-threshold-state (MTATS) or mean total duration of time that persons have health status scores above a specified threshold. When the threshold chosen is below the lowest recorded score, all ongoing (living) participants have scores above threshold and the MTATS is identical to the mean survival time. When higher thresholds are used, the two times are different because only the time above threshold rather than all time alive is considered in calculating the MTATS, and because all time above the given threshold is considered rather than just the time before the first drop to below the threshold as in standard survival analysis. The estimated MTATS can be calculated for many threshold scores spanning the range of observed values. The resulting family of MTATS values can be plotted on the Y axis against the range of thresholds on the X axis to yield a

MTATS map. The MTATS map allows for visual inspection and comparisons of the time-quality experience of cohorts. In addition, the MTATS can be averaged across the range of thresholds to give the average mean time above threshold state (AMTATS), which can be interpreted as the typical time that a typical patient in the cohort spends above a typical threshold during the trial.

In multistate survival analysis, the significance of differences in the time-quality experience of two treatment cohorts is assessed using a generalization of the standard Mantel-Haenszel test for differences in survival. Rather than just considering the number of excess deaths or deteriorations, this approach simultaneously considers both excess deteriorations (i.e., transitions to a worse health status) and excess improvements (i.e., transitions to a better health status) in the intervention group compared to the standard treatment group. The statistical significance of the observed number of excess transitions under the null hypothesis is assessed using a permutation test (10, 11).

The utilization, disability, work, and income data were analyzed using a 2-part model wherein the mean probability of having any events or days during a person-month and the mean number of events or days per person-month among those having any are estimated separately, and an estimate of the average number per month is obtained by multiplying these two values (12). This was done to avoid difficulties with statistical testing that arose because, as is usually the case, the distributions of utilization and disability variables were extremely non-normal in this study. Correction for variation in baseline values was accomplished by entering the difference between individual and mean baseline scores into the regressions. t-statistics for the differences in the estimated averages between treatment groups were obtained using estimates for the standard error of the difference calculated using a Taylor expansion as previously described.

All persons participating in the main study signed written informed consent forms approved by the relevant local Institutional Review Boards, and both this project and the HIV-PARSE survey instrument were reviewed by the RAND Human Subjects Protection Committee.

RESULTS

The substudy did not open until the main studies were approximately one-third enrolled, and only a minority of sites participated. A total of 465 persons completed at least two surveys. After removal of patients randomized after December of 1990 and of data generated after patient crossovers, 1,543 responses from 336 individuals remained available for the primary analysis. This is 27% of all those enrolled into ACTG 116/117, but includes 72% of all those eligible for the substudy at participating sites. Overall, 50% of survey participants were from sites enrolling at least three-fourths of candidates and 75% were from sites enrolling at least two-thirds of candidates. Survey participants were similar to

non-participants and, among participants, persons randomized to the three arms were also similar in terms of both clinical and sociodemographic data (Table 1a and b).

TABLE 1a.

BASELINE CHARACTERISTICS OF SURVEY PARTICIPANTS AND NON-PARTICIPANTS AND, AMONG PARTICIPANTS, OF TREATMENT GROUPS. VALUES GIVEN ARE PERCENTAGES OR MEANS.

	PARSE Participation		Among PARSE Participants		
	No	Yes	Zidovudine	750 mg/day didanosine	500 mg/day didanosine
N	924	336	113	107	116
In ACTG 116	60%	61%	64%	60%	64%
< 8 week prior zidovudine	28%	36%	34%	34%	39%
Age	37	36	36	37	36
White	78%	74%	74%	73%	74%
Male	94%	94%	93%	96%	93%
Gay	69%	80%	81%	77%	89%
Weight	73	73	74	72	74
Hemoglobin	12.8	13.3	13.2	13.3	13.4
White Blood Cells (thousands)	3.8	4.0	3.9	4	4.1
Platelet count	218	219	220	219	219
CD4	118	126	138	133	138

Specific Functional Measures

There were no significant differences between the treatment groups in reported number of hospital admissions, hospital days, phone contacts, or home care visits (Table 2). There were also no significant differences between the treatment groups in the number of office visits or regular medications, but these would be expected to be less sensitive to treatment differences due to the large amount of protocol-driven care. The number of symptoms which patients reported as interfering at least moderately with functioning were also similar (Table 2).

Reported disability days were generally lowest for low dose didanosine recipients, but differences compared to zidovudine were only significant for days "feeling less well than usual" in participants with more than eight weeks of prior zidovudine (3.7 versus 5.4, P = .03), and for days of missed work among the employed in those with fewer than eight weeks of prior zidovudine (0.7 versus 1.3, P<.04)(Table 2). In the same group, days of missed work were also significantly lower for the high dose didanosine group(0.6 versus 1.3, P = .02), but hours worked were lower among employed didanosine recipients (38 and 39 versus 43 hours/week, P<.01)(Table 2). Among those with more than eight weeks of prior

TABLE 1b

FUNCTIONAL CHARACTERISTICS OF PARTICIPANTS AT BASELINE. VALUES GIVEN ARE PERCENTAGES OR MEANS.

	600 mg/day zidovudine	750 mg/day didanosine	500 mg/day didanosine
Proportion of candidates enrolled.	73%	72%	74%
High School Graduate	92%	92%	86%
College Graduate	43%	39%	39%
White collar occupation (ever)	68%	63%	63%
Physical Functioning (0-100 scale)	84	80	83
Role Functioning (0-100 scale)	78	80	79
Social Functioning (0-100 scale)	78	77	80
Mental Health/Emotional Well Being (0-100 scale)	68	69	70
Cognitive Functioning/Distress (0-100 scale)	85	82	84
Bodily Pain (0-100 scale)	74	74	76
Energy/Fatigue (0-100 scale)	60	60	60
Current Health Perceptions (0-100 scale)	53	56	56
Perceived Health Index (0-100 scale)	70	70	71
Hospital days/month	0.6	0.5	0.5
Office visits/month	3.1	3.3	3.5
Invasive or diagnostic procedures/month	1.0	1.1	1.1
Home care visits/month	0.1	0.1	0.1
Provider telephone contacts/month	1.6	1.4	1.8
Medications	2.6	2.6	2.6
Symptoms interfering moderately with functioning	3.9	3.8	4.3
Bed days/month	1.6	1.9	1.5
Days of missed work/month	1.7	1.2	1.4
Days of reduced activity	2.9	2.4	3.1
Days feeling less well	4.9	4.2	5.0
Currently Employed (full or part-time)	61%	62%	63%
Monthly earned income	$1924	$1824	$2123
Hours worked/week (including unpaid labor)	25	25	26

TABLE 2

HEALTH CARE, SYMPTOM IMPACT, DISABILITY, AND WORK. ALL FIGURES ARE MEAN VALUES PER MONTH OVER 12 MONTHS OF FOLLOW-UP

	< Eight Weeks Prior Zidovudine			> Eight Weeks Prior Zidovudine		
	600 mg/day zidovudine	500 mg/day didanosine	750 mg/day didanosine	600 mg/day zidovudine	500 mg/day didanosine	750 mg/day didanosine
Hospital Admission	0.07	0.06	0.04	0.04	0.06	0.04
Hospital Days (all participants)	0.6	0.9	0.4	0.4	0.4	0.2
Office Visits	3.0	2.9	2.2	1.5	1.7	2
Invasive Procedures	0.12	0.18	0.12	0.37	0.38	0.28
Invasive Procedures	0.9	1.2	0.8	0.9	1.1	1.1
Phone Contacts	0.1	0.2	0.1	0.4	0.7	0.3
# Regular Medications	1.4	1.4	1.2	1.2	1.3	1.1
# Moderate Symptoms	0.7	1.1	0.9	5.3	6.8	4.2
Bed Days	1.6	1.3	1.9	2.3	1.8	1.8
Days of Missed Work	1.3	0.7*	0.6*	1.2	0.7	0.8
Days of Reduced Activity	2.4	2.0	3.1	3.4	2.8	3.5
Days Feeling Less Well	3.0	2.7	4.3	5.4	3.7*	5.2
Employed	0.66	0.63	0.56	0.67	0.55**	0.55**
Work Hours	43	38*	39*	41	43	42
Income	$1,555	$1910*	$1980*	$2,488	$3,626*	$2,573

* P < .05 ** P < .001

zidovudine, the average employment rate was significantly lower for persons receiving didanosine (.55 and .55 versus .67, P<.01)(table 2). Average earned income, an integrated measure of work, was highest in the low dose didanosine group in both groups (Table 2).

Health Status Scale Scores

Among patients with fewer than eight weeks of prior zidovudine, none of the differences between ongoing patients in either didanosine group and the zidovudine group would be characterized as large (at least .5 standard deviations) and none were significant at the .05 level. Over 48 weeks on therapy, the average of scale score differences for ongoing patients in either the low dose didanosine versus zidovudine ranged from a .12 standard deviation advantage for zidovudine on the social functioning scale to .19 advantage for didanosine on the mental health scale, with the average difference in the perceived health index being only .06 standard deviations in favor of didanosine (Figure 1a). For the high dose didanosine group, the differences for ongoing patients ranged from a .18 standard deviation advantage for zidovudine on pain to a .22 advantage for didanosine on current health perceptions, but the average of differences in the perceived health index was only .02 standard deviations (Figure 1a).

FIGURE 1 (TOP)

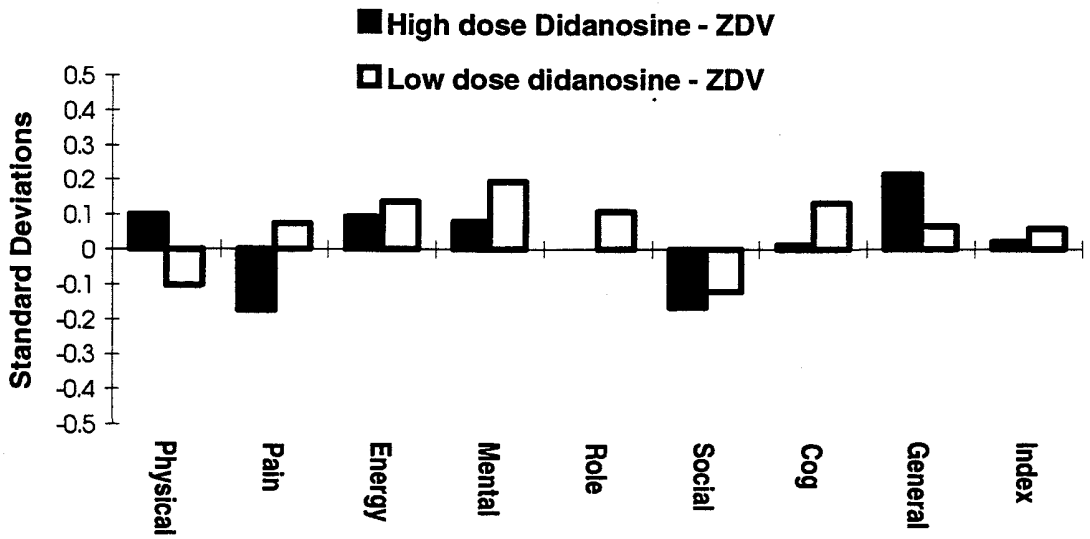
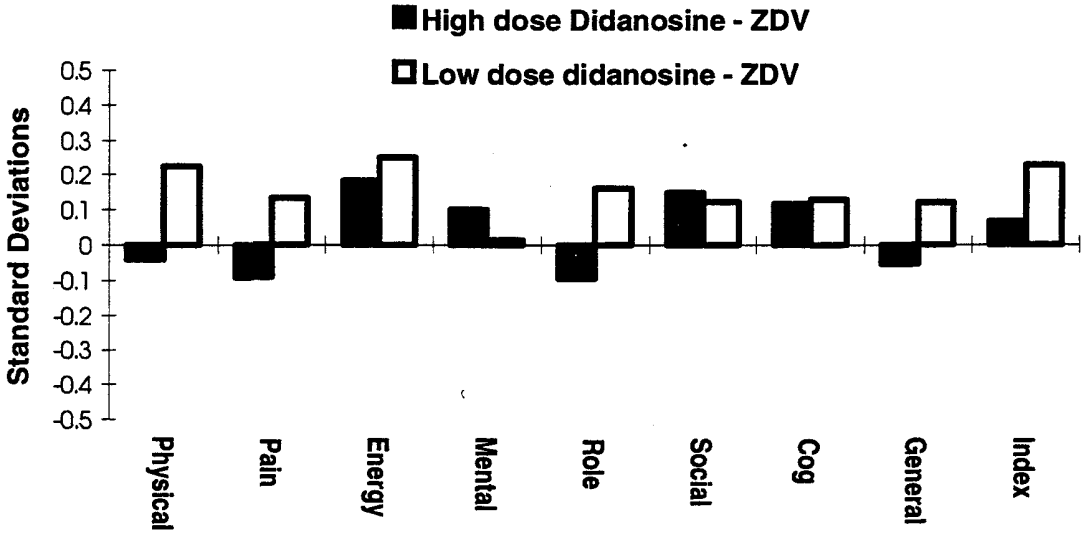


FIGURE 1 (BOTTOM)



Average of Group Differences in Health Status/Functioning Scales over 48 Weeks. Shown are the average of standardized differences in scale scores for the didanosine (ddI) groups compared to the zidovudine (ZDV) group. Scores were recorded at 12 week intervals. A difference of .5 standard deviations would be considered a large effect. Represented are scales for Physical (Functioning), Role (Functioning), Social (Functioning), Cognitive (Functioning /Distress), (Bodily) Pain, Energy/fatigue, and Current Health Perceptions as well as the Perceived Health Index, a combination of seven of the previous scales. **Top Panel**: Average of difference scores in the group with fewer than eight weeks of prior zidovudine. None of the differences were statistically significant. **Bottom Panel**: Average of difference scores in the group with more than eight weeks of prior zidovudine. Controlling for week, differences between the low-dose didanosine group and the zidovudine group were statistically significant for the energy/fatigue and physical functioning scales as well as the perceived health index, but effect sizes were small.

In the group with more than eight weeks of prior zidovudine, none of the differences between ongoing patients in either didanosine group and the zidovudine group would be generally characterized as large, but some average difference scores significantly favored the low dose didanosine group compared to the zidovudine group, including the difference in the energy/fatigue scale (average = .25 standard deviations, P = .03), the physical functioning scale (average = .23 standard deviations, P = .05), and the perceived health index (average = .23 standard deviations, P = .06)(Figure 1b). This was not true of the comparison between the high dose didanosine group and the zidovudine group, with the average difference scores for ongoing patients ranging from a .10 unit advantage for zidovudine on the role functioning scale to a .18 unit advantage for high dose didanosine on the energy/fatigue scale. The average difference in perceived health index scores was only .07 units in favor of high dose didanosine (Figure 1b).

The hypothesis that relative drug effects change over time was evaluated by testing the significance of an interaction term between drug assignment and duration of prior zidovudine exposure, controlling for time on study and main drug effect. Significance of this term would support a change in relative drug effects, but it was not significant for either low dose or high dose didanosine compared to zidovudine (F = 1.34, P = .26 and F = 0.72, P = .49 respectively).

Analyses incorporating death and compensating for attrition did not alter the conclusion that differences between the treatment groups are small. Analysis of the predicted categorical rating data (which incorporates death by giving it a score of 0) indicated that, among those with fewer than eight weeks of prior zidovudine treatment, the average differences between zidovudine group and either the low or high dose didanosine groups were only .05 standard deviations (P for overall difference = .41). For those with more than eight weeks of prior zidovudine, these differences were .20 and .27 standard deviations for the comparison with low and high dose didanosine, respectively (P for overall difference = .27). Converting these data to quality-adjusted life years also did not result in significant differences between the treatment groups in either strata.

Multistate survival analysis indicated that the number of persons in each treatment group experiencing improving or deteriorating health was similar in the zidovudine and low, high, or any didanosine dose groups within strata of prior zidovudine use (for <8 weeks prior zidovudine t = .34-1.2, P = .73-.24; for >8 weeks prior zidovudine t = .01-.32, P = .99-.75; for no prior zidovudine t = .02-1.1, P = .99-.27, for any prior zidovudine t = .11-.58, P = .91-.56). The MTATS maps did not suggest dominance by any of the treatments (Figure 2). For the strata with fewer than eight weeks of prior zidovudine use, the mean survival was highest for those receiving zidovudine, but persons receiving low dose didanosine spent more time with at least the fifth to thirtieth percentile of health status and persons on high dose didanosine spent the most time with health status above the forty-fifth percentile (Figure 2A). Overall, a

FIGURE 2 (TOP)

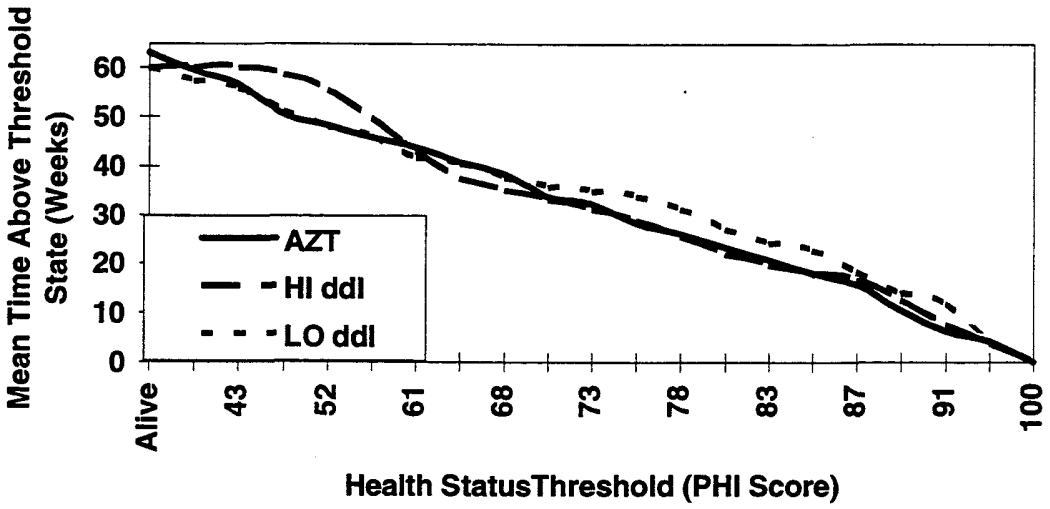
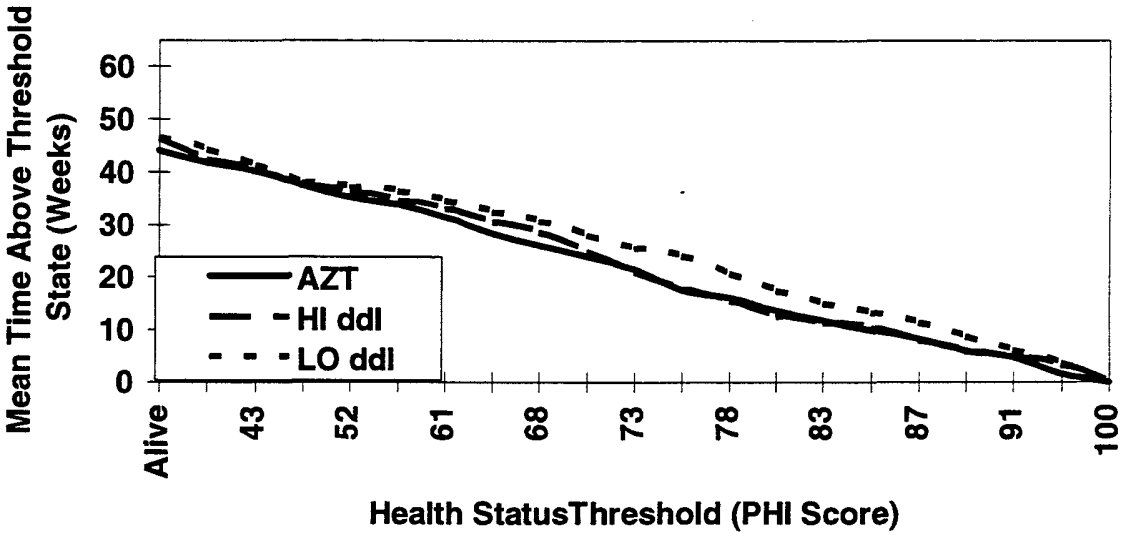


FIGURE 2 (BOTTOM)



MTATS Maps. The Y axis depicts the MTATS (mean time above threshold state) and the X axis depicts the range of observed health status scores/thresholds . The left-most points (at the origin of the X axis) indicate the time with at least the lowest observed health status, which is equivalent to the mean survival time. The right-most points indicate the time with health status higher than the highest observed score and are necessarily zero. Each tick mark indicated a decile of the scores observed at baseline; e.g., 52 is the 20th percentile, 73 is the median, and 87 is the 80th percentile. Solid line indicates the zidovudine (ZDV) group, the dashed line indicates the high dose didanosine (ddI) group, and the dotted line indicates the low-dose didanosine group. **Top panel**: Strata with fewer than eight weeks of prior zidovudine treatment. **Bottom panel**: Strata with more than eight weeks of prior zidovudine treatment.

typical person in this strata randomized to zidovudine spent 33 of 64 possible weeks (52%) with at least the typical health state, while typical persons randomized to low- or high-dose didanosine spent 34 (53%) and 35 (54%) weeks, respectively, and typical persons in a stable population would spend 50% of possible time or 32 weeks. The MTATS was much lower for those receiving more than eight weeks of prior zidovudine use than for those receiving fewer than eight weeks, and it was similar for all treatments except for a slight advantage for low-dose didanosine above the twentieth percentile (Figure 2B). Overall, a typical person in this group randomized to zidovudine spent 23 of 64 weeks (36%) in a typical health state, while typical persons randomized to low- or high-dose didanosine spent 23 (36%) and 26 (40%) weeks, respectively.

For all of the above analyses, conclusions do not change when considering the pooled data from both didanosine groups, when analyzing within strata of never had or had any prior zidovudine therapy, or when considering strata of disease stages (e.g., AIDS versus non-AIDS diagnosis).

DISCUSSION

In this study, functional status and health-related quality of life were substantially similar among persons receiving either zidovudine or didanosine regardless of the duration of prior zidovudine treatment. That is, the treatment differences demonstrated in the clinical analyses of these trials did not translate into substantial differences in functional status or health-related quality of life for typical patients in the substudy, even in analyses designed to highlight potential differences. Several possible explanations must be considered, including differences in the types of outcomes and analyses, selection effects, inadequate power, and the existence of time/quality trade-offs.

First, the differing results may reflect the fact that the clinical and functional results were obtained using fundamentally different kinds of assessments. The clinical results were driven by the minority of participants who experienced provider-defined endpoint events, whereas most of the functional and all of the health-related quality of life measures capture self-report information for all participants at all times. Also, the clear interpretability of standard time-to-event analyses comes from a focus on detecting differences between the times to the first discrete event. As a result, the effect of prodromes, sequelae, and recurrences of events as well as changes in status too small to be considered events make no contribution to the measured outcome. In contrast, functional assessments fail to capture important discrete deleterious events, but do directly assess variation in the quality of time both pre- and post-event. For these reasons, time-to-event analyses and survival-in-state or other analyses that directly incorporate health status are complementary, but cannot always be expected to yield similar findings.

Second, selection effects may have acted to minimize differences between the treatments in this substudy, as only 27% of all participants in the main studies are included in this report. However, many subjects

were excluded by the randomization error in ACTG 116 and the selection that did occur from the main study into the substudy was mostly by center. Overall, 75% of candidates at sites participating in the substudy were enrolled. Further, selection prior to randomization cannot introduce bias, only a lack of generalizability. Nonetheless, exclusion of 25% of candidates at participating sites and, from one perspective, all candidates at non-participating sites could affect generalizability if the substudy population were sufficiently different from the main study population to mask treatment differences that would have appeared if all candidates were enrolled. However, this seems somewhat unlikely as participants in the substudy were physiologically similar to non-participants, survival estimates were similar in the substudy and main study, and the estimates of effect within the randomized substudy were not appreciably changed by restricting the analysis to participants from sites enrolling at least 75% and 90% of candidates. However, we recognize that the generalization of these and other clinical trials results to general HIV patient populations can be problematic, especially as the epidemic moves into populations with less access to clinical trials (13).

Third, it is possible that this study suffers from inadequate power. We consider this unlikely as standard calculations indicate that a study of this size has a greater than 95% power to detect a 0.5 standard deviation average difference in health status scale scores between the treatment groups. Additionally, earlier work by our group indicated that transition-from-state analysis can be more efficient than survival analysis in showing differences between antiretroviral treatment groups. Finally, this study may best be viewed as a "pragmatic" exercise in estimation of effect size, rather than as an exercise in seeking statistical "proof." From that perspective, it seems very unlikely that large differences in functional status paralleling the clinical findings were missed, as the vast majority of the many comparisons of both specific and psychometric measures yielded values near the null.

Finally, the summary inference drawn from the clinical data, that zidovudine is better for the zidovudine naive but didanosine is better for the zidovudine experienced, may require qualification over the time frames studied here. In the main trials, the analysis of time-to-death showed a significant advantage for zidovudine among the zidovudine naive, and the analysis of time-to-opportunistic-complication showed a significant advantage for didanosine among the zidovudine experienced. However, the converse was not shown, and direct tests of the interaction between time and relative drug effects are not available for these outcomes. Further, survival was worse among those entering the trial with HIV strains demonstrating *in vitro* zidovudine resistance compared to those entering with zidovudine sensitive strains, even if the randomized treatment assignment switched then to didanosine.

Nonetheless, the results of this substudy should not be considered inconsistent with the clinical findings in the parent trials. Rather, it should be considered as providing additional perspective on the reported clinical differences. For example, the main clinical trial indicated a survival advantage for zidovudine-

naive patients randomized to zidovudine. In this substudy, the mean time with any recorded health state (i.e., mean survival) was greater for zidovudine recipients in the strata with fewer than eight weeks of prior zidovudine, but reported health states were generally higher for didanosine recipients. This time/quality trade-off is apparent when comparing the survival with the quality-adjusted survival data. Its nature is amplified by examination of the MTATS map, which indicates the greater survival of the zidovudine group at the leftmost axis and the crossing of the zidovudine and didanosine curves at higher health states (Figure 2A). The existence of this time/quality trade-off does not mean that the survival advantage demonstrated in the parent trial is not real or exceedingly important, or that an individual valuing survival very highly would not most reasonably choose zidovudine for initial therapy. Rather, it means that a person valuing higher quality survival would not be unreasonable in choosing didanosine as initial therapy.

There was also little difference between treatment groups in reported utilization and disability, except for those related to work. Beyond their significance as an indicator of clinical status, these data may be important for administrative and societal decisionmaking. The utilization data suggest that the secondary costs associated with didanosine and zidovudine treatments are likely to be similar, so that from the payer's perspective, differences in the cost of treatment will likely relate primarily to differences in the costs of the drugs themselves. The differences in work are of interest from the societal perspective. If within the general population of persons with HIV disease, didanosine recipients are able to earn higher incomes as reported here, then these patients' productive activities would offset a larger share of the economic cost of their treatment, and the true net cost of treatment to society with didanosine will be proportionately lower.

In a previous antiretroviral trial, specific functional and health status endpoints uniformly preceded and amplified differences between the treatments, and the incorporation of these endpoints increased the sensitivity for detecting differences between the treatments over clinical assessments alone. In this study, these broader outcomes did not reveal earlier or more significant differences than those reported in the clinical results. Rather, the differences in functioning and health-related quality of life were less striking than the clinical results, in part due to the existence of an apparent time/quality trade-off which could not be captured by clinical data alone. Thus, these findings emphasize another aspect of the need for routine collection of broader outcome information within clinical trials of chronic therapies if such trials are to support informed individual decisionmaking by doctors and patients.

REFERENCES

1. Fischl MA, Richman DD, Grieco MH, et al. The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *N Engl J Med* 1987;317:185-91.

2. Yarchoan R, Mitsuya H, Thomas RV, et al. *In vivo* activity against HIV and favorable toxicity profile of 2',3'-dideoxyinosine. *Science* 1989;245:412-415.

3. Dolin R, Amato DA, Fischl MA, et al. Zidovudine compared with didanosine in patients with advanced HIV type 1 infection and little or no previous experience with zidovudine. *Archives Intern Med.* 1995;155:961-74.

4. Kahn JO, Lagakos SW, Richman DD, et al. A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. *New Engl J Med.* 1992;327:581-7.

5. Croog SH, Levine S, Testa MA, et al. The effects of antihypertensive therapy on the quality of life. *N Engl J Med.* 1986; 314:1657-64.

6. Wu AW, Rubin HR, Mathews WC, et al. A health status questionnaire using 30 items from the medical outcomes study. Preliminary validation in persons with early HIV infection. *Medical Care* 1991;29:786-798.

7. Wachtel TW, Piette J, Mor V, et al. Quality of life in persons with human immunodeficiency virus infection: measurement by the medical outcomes study instrument. *Ann Intern Med.* 1992;116:129-37.

8. Bozzette SA, Kanouse DE, Berry S, Duan N. Health Status and Function with Zidovudine or Zalcitabine as Initial Therapy for Advanced HIV Disease: A Randomized Controlled Trial. *JAMA.* 1995;273:295-301.

9. Stewart AL, Sherbourne CD, Hays RD, et al. Summary and discussion of MOS measures. In: Stewart AL and Ware JE, eds. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach.* Durham, N.C.: Duke University Press; 1992:345-371.

10. Fisher, R.A. *The Design of Experiments.* Edinburgh: Oliver & Boyd; 1935.

11. Lehmann, E.I. *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day; 1975.

12. Duan N, Manning WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. *J Business and Econ Stat.* 1983;1:115-26.

13. Cunningham WE, Bozzette SA, Hays RD, Kanouse DE, Shapiro MF. Comparison of health-related quality of life in clinical trial and non-clinical trial cohorts of persons with advanced HIV disease. *Medical Care*, 33(4):AS15-AS25, 1995