STATISTICS IN MEDICINE
Statist. Med. **17**, 2435–2462 (1998)

STATISTICAL ISSUES FOR HIV SURROGATE ENDPOINTS: POINT/COUNTERPOINT

AN *NIAID* WORKSHOP

JEFFREY M. ALBERT[1]*, JOHN P.A. IOANNIDIS[1], PATRICIA REICHELDERFER[1], BRIAN CONWAY[2], ROBERT W. COOMBS[3], LAWRENCE CRANE[4], RALPH DEMASI[5], DENNIS O. DIXON[1], PHILLIPE FLANDRE[6], MICHAEL D. HUGHES[7], LESLIE A. KALISH[8], KINLEY LARNTZ[9], DANYU LIN[3], IAN C. MARSCHNER[10], ALVARO MUÑOZ[11], JEFF MURRAY[12], JIM NEATON[9], CARLA PETTINELLI[1], WASIMA RIDA[1], JEREMY M.G. TAYLOR[13] AND SETH L. WELLES[9] FOR THE WORKSHOP PARTICIPANTS

Division of AIDS, National Institute of Allergy and Infectious Diseases, National Institutes of Health[1], *University of British Columbia*[2], *University of Washington*[3], *Wayne State University Detroit Medical Center*[4], *Glaxo Wellcome*[5], *Institut National de la Sante et de la Recherche Medicale*[6], *London School of Hygiene and Tropical Medicine*[7], *New England Research Institutes*[8], *University of Minnesota*[9], *Harvard School of Public Health*[10], *Johns Hopkins School of Public Health*[11], *Food and Drug Administration*[12], *University of California, Los Angeles*[13].

Other workshop participants were: Heather Beacon, Harvard School of Public Health; James Bethel, Westat Incorporated; Donald Brambilla, New England Research Institutes; Brad Carlin, University of Minnesota; George C. Chao, DuPont Merck; Jay Chmiel, Abbott Laboratories; Joan Chmiel, Northwestern University Medical School; Ellen Cooper, Office of AIDS Research; Deb Dawson, Glaxo Wellcome; Victor DeGruttola, Harvard School of Public Health; Al Getson, Merck Research Laboratories; Peter Gilbert, Harvard School of Public Health; Peter Gilbert, National Institute of Allergy and Infectious Diseases; David Hall, Boehringer Ingelheim; Andrew Hill, Glaxo Wellcome; Mark Knowles, Agouron Pharmaceuticals; Andrew Kuhn, Medical College of Virginia; Ira Longini, Emory University; Dana Nickens, Pharmacia and Upjohn; Bisser Roussanov, University of California, Los Angeles; Mark Schluchter, Cleveland Clinic Foundation; David Schoenfeld, Harvard School of Public Health; Lesley Struthers, Hoffmann La Roche; Wai-Yuan Tan, University of Memphis; Melanie Thompson, AIDS Research Consortium of Atlanta; Paige Williams, Harvard School of Public Health; George Yu, Agouron Pharmaceuticals; Fan Zhang, University of California, Los Angeles.

SUMMARY

This paper summarizes the proceedings of an NIAID-sponsored workshop on statistical issues for HIV surrogate endpoints. The workshop brought together statisticians and clinicians in an attempt to shed light on some unresolved issues in the use of HIV laboratory markers (such as HIV RNA and CD4+ cell counts) in the design and analysis of clinical studies and in patient management. Utilizing a debate format, the workshop explored a series of specific questions dealing with the relationship between markers and clinical endpoints, and the choice of endpoints and methods of analysis in clinical studies. This paper provides the position statements from the two debaters on each issue. Consensus conclusions, based on the presentations and discussion, are outlined. While not providing final answers, we hope that these discussions have helped clarify a number of issues, and will stimulate further consideration of some of the highlighted problems. These issues will be critical in the proper assessment and use of future therapies for HIV disease. © 1998 John Wiley & Sons, Ltd.

* Correspondence to: Jeffrey M. Albert, Biostatistics Research Branch, Division of AIDS, National Institute of Allergy and Infectious Diseases, 6003 Executive Boulevard, Bethesda, MD 20892-7620, U.S.A. E-mail: ja24o@nih.gov

Contract/grant sponsor: National Institutes of Health
Contract/grant numbers: AI27664, AI27757, AI30731, DK49477, AI35161, AI95030, AI35043, AI29196

CCC 0277–6715/98/212435–28$17.50
Received July 1997
Accepted March 1998

INTRODUCTION

The use of surrogate markers as endpoints in clinical efficacy studies has been an area of active research and controversy. Clinical research of human immunodeficiency virus (HIV) disease is one of the disciplines where surrogates have gained increasing popularity. However, there are still many, as yet unresolved, issues regarding their use in patient management, and analysis and design of controlled clinical studies. To facilitate further understanding of these issues, the U.S. National Institute of Allergy and Infectious Diseases conducted a workshop in Memphis, Tennessee on 25–26 March 1997. A goal of the workshop was to bring together individuals representing both the statistical and clinical point of view to discuss these issues. The discussion was in the format of a debate during which individuals were asked to take either the pro or con viewpoint for a series of nine questions. Following the initial debate, other members of the workshop were invited to contribute to the discussion. What follows is a synopsis of each of the debate questions from both points of view. Unlike a conventional debate, the goal was not to draw sharp lines between sides and determine a 'winner', but rather to clarify concepts and reveal areas of agreement as well as dispute.

THE CHOICES OF MARKER METRIC AND ASSAY FOR A SURROGATE ENDPOINT HAVE NO EFFECT ON CONCLUSIONS REGARDING TREATMENT EFFICACY (MICHAEL HUGHES AND PHILIPPE FLANDRE)

Pro

In considering the impact of marker metrics, it is, first of all, important not to label intrinsically different measures as being simply different metrics. For example, absolute CD4 cell count and CD4 percentage are not simply different metrics; the former is derived from the latter, using in addition the total lymphocyte count.[1, 2] Similar comments apply to endpoints involving measurements at different time points. Such outcomes involve different information (potentially, differential variation or missingness) and cannot be expected to yield similar results.

However, using alternative *metrics* (for example, different transformations or net versus per cent change from baseline) that measure the same underlying quantity and generally have the same pattern of missingness should not impact conclusions regarding treatment efficacy. This is particularly evident if the goal is hypothesis testing. A null treatment effect on an outcome based on one metric will imply a null effect using any alternative metric. Note also that the validity of a marker as a surrogate endpoint should not be affected by choice of metric. This follows from the fact that Prentice's criteria for a surrogate endpoint[3] requires that a null effect of treatment with respect to the marker implies (and is implied by) a null effect of treatment on the 'true' endpoint. Similar comments apply to the choice of assay for measuring HIV RNA; if the assays measure the same underlying quantity, then the choice of assay should be immaterial.[4] This is true provided that the study design is appropriate for handling any differences in measurement variability, and that the statistical analysis is appropriate for handling any differences in other measurement-specific issues such as quantitation limits.[5]

Of course, in any given trial, it is almost certain that treatment comparisons made using different metrics, given their different distributional properties, will give rise to different *p*-values. If there is consistency between the results using different metrics, then the conclusion from the trial concerning differential treatment activity is unambiguous. However, it may occur that the results do not appear consistent, for example, if one is statistically significant and the other is not. This situation might arise because the study was not powered adequately to address treatment comparisons for alternative metrics, or inconsistency in the results may point to an inappropriate method of analysis for one or both of the metrics (for example, violation of assumptions in a statistical test). Setting aside possible flaws in analysis, one should be cautious about over-interpreting differences; in particular, *p*-values for different metrics that closely straddle a significance threshold are not discordant.

Con

Attempts to use laboratory markers, especially CD4 count and HIV RNA, as surrogate endpoints can be divided into two parts: those that employ fixed variables and those that use time-dependent variables. For both, the number of options for defining metrics can be enormous.

For a fixed variable, one can have interest either in changes from baseline values to a given time point, or in the level of the marker considered at a specific time. Numerous metrics have been proposed,[6] although absolute CD4 count, square root of CD4 count, area under the curve for CD4 count, log HIV RNA and area under the curve for log HIV RNA are the most popular. A shortcoming with using fixed variables is the possibility of missing data that could introduce substantial bias in the analysis.[7]

In time-dependent approaches one could either use the current value of the marker or define a time-dependent event from the marker using a previously agreed upon cut-off (for example, a 50 per cent decline in CD4 count). The former possibility is not very convenient for clinical investigators while the latter is probably an oversimplification and allows ample room for subjective and variable determination of cut-off values.

In addition to metric and type of covariate (fixed or time-dependent), the choice of time from baseline (for example, week 8, 16 …) has a substantial effect on the conclusion regarding how well the potential surrogate marker captures treatment efficacy. A good surrogate for short-term effects may fail for long-term effects.

As an example, by considering the proportion of treatment effect explained (PTE) by a surrogate marker as a validation criterion, results can differ substantially according to the metric and method used. Choi *et al.*[8] analysed data from ACTG 019 to evaluate CD4 as a potential surrogate for time to AIDS. Using CD4 count and CD4 per cent as time-dependent covariates in a Cox regression model, the PTE was estimated as -1 per cent and 37 per cent, respectively. To help delineate the treatment effect, AIDS after week 16 was also examined. For this endpoint, the estimated PTEs of CD4 count and CD4 per cent (here used as fixed covariates at week 16) were 46 per cent and 74 per cent, respectively.

Further illustration of metric differences appears in data from the European Delta Trial involving 1280 patients (and a total of 9402 blood samples). Marker responses were considered as fixed and time-dependent covariates, change from baseline to week 8, 16, and 32, and area under the curve for HIV-RNA (considered on a log scale). In this analysis, as in Choi *et al.*'s analysis, estimated PTEs were sometimes found to lie outside the range of 0 to 1. For example, the proportion of treatment effect explained by RNA levels up to week 16 on time to AIDS/death was estimated as 183 per cent (95 per cent CI: 76–290 per cent) and 249 per cent (95 per cent CI: 83–416 per cent).[9] This occurrence, as well as the considerable variability of estimated PTE, obviously weaken its usefulness in comparing different surrogate endpoint or metrics. We may need to consider robust alternative measures.[10]

It may be argued that, for example, CD4 per cent versus total count, or fixed versus time-dependent endpoints represent different underlying parameters, and thus should not be compared. However, the scientific distinctions among such endpoints (which may come under the same conceptual heading of immune response) are often not well understood. Thus it is important to clarify and demonstrate the substantial statistical differences between such alternative measures.

THERE IS A 'SET-POINT' VALUE OF RNA ACHIEVED AFTER SEROCONVERSION THAT PROVIDES MOST OF THE INFORMATION IN THE TRAJECTORY OF RNA WITH REGARD TO DISEASE PROGRESSION (LES KALISH AND JEREMY TAYLOR)

Pro

If early 'set-point' RNA levels did not provide most of the information in the trajectory of RNA with regard to disease progression, then we would expect that the estimated association between most recent viral load and progression should be much larger than the association between baseline viral load and progression. That is, consideration of viral load as a time-varying covariate provides much more prognostic information than a single baseline measurement.

The most widely cited paper in support of viral load monitoring is based on data from the Pittsburgh portion of the Multicenter AIDS Cohort Study.[11] The concluding paragraph of this paper states '… HIV-1 RNA is the best available surrogate marker of HIV-1 disease progression.' Yet the results of the paper suggest that baseline (set-point) viral load carries about the same prognostic information as most recent value. With 10·6 years median follow-up among patients who did not develop AIDS and 5·6 years median follow-up among those who did develop AIDS, the relative hazard (RH) of death associated with a tenfold (1 log) difference in baseline viral load was 2·51 (95 per cent CI: 1·85, 3·43), adjusting for CD4. In another proportional hazards model, using time-varying covariates in place of baseline levels, the RH was 2·57 (CI: 1·85, 3·51). The similarity of these estimates suggests that ongoing monitoring of viral load provides little more information than the set-point level.

Table I summarizes results of the above study, and two additional published papers [12,13] which considered both baseline and time-varying viral load in proportional hazards models. Also shown is a meta-analysis of the three studies. The summary RH per tenfold difference in baseline levels was 2·39 (CI: 1·84, 3·11) and using time-varying covariates the RH was 2·62 (CI: 2·02, 3·41), suggesting that baseline (set-point) viral load carries nearly as much prognostic information as ongoing (most recent) monitoring. This conclusion may need to be confirmed with analyses of characteristics of the RNA trajectory other than the most recent value,[14] but it is likely in practice that subtle differences in trajectories may be masked by assay variability.[15]

Con

The RNA set-point is a frequently used term,[16] however, very little has been written to define it formally and to evaluate whether the concept is valid and useful. In the work of Mellors *et al.* on 62 seroconverters in the Multicenter AIDS Cohort Study (MACS),[17] the RNA set-point level (the viral load value attained a few months after HIV infection) was shown to have good prognostic significance for disease progression. Mathematical modelling research[18] has also assumed that there is a long stable or equilibrium period, corresponding approximately to the asymptomatic period in infected subjects, during which viral load and CD4 count are stable. Based on the above,

Table I. Meta-analysis of baseline plasma HIV-1 RNA and time-varying viral load in proportional hazards models

Author	Endpoint	Median follow-up	Viral load predictor	RH per ten-fold difference†	(95 per cent CI)
Mellors *et al.*[10]	Death	5·6–10·6 years*	Baseline	2·51	(1·85, 3·43)
			Most recent	2·57	(1·85, 3·51)
Henrard *et al.*[11]	AIDS	7·5 years	Baseline	1·77	(1·01, 3·13)
			Most recent	2·05	(1·21, 3·47)
Phillips *et al.*[12]	AIDS	<1 year	Baseline	3·90	(1·36, 11·20)
			Most recent	6·90	(2·66, 17·86)
Meta-analysis	–	–	Baseline	2·39	(1·84, 3·11)
			Most recent	2.62	(2·02, 3·41)

* 5·6 years among those who developed AIDS; 10·6 years among those who did not
† RH = relative hazard

we define the 'RNA set-point model' to have two components, the first is the level attained by an individual's RNA value soon after HIV infection, and the second the level that is maintained for an extended period of time. According to the model, the pattern of RNA values, when plotted against time for each infected asymptomatic individual, should look like parallel, horizontal lines.

We have examined the pattern of RNA change in individuals in three data sets. The first was a small trial of the effect of influenza vaccination on RNA values. There were 4 or fewer measurements per subject ($n = 45$) and the total follow-up time was 12 weeks. Over this short time period, we saw no significant changes of RNA for different patients (graph not shown).

The second data set was from the VA Cooperative group trial 298 of early versus late AZT.[19,20] Here we considered the subset of patients from the late treatment arm, who did not go on open label AZT and did not show clinical disease progression. These patients could be considered to be in the asymptomatic untreated phase. There was up to three years follow-up on these 70 patients. For this study, a pattern of parallel horizontal lines for RNA values was clearly not evident. Figure 1 shows the data from a subset of 12 of the patients chosen so that their ranking of baseline RNA values is evenly spread across the range. There is substantial fluctuation in RNA values, clearly not consistent with the set-point model. However, we note that the within-subject variation seen in these data is larger than that seen in other studies,[21] suggesting that not all study populations and/or not all methods of measuring RNA are equivalent.

The third data set consisted of a subset of Los Angeles MACS participants who had an RNA measurement on a 1985 sample and a second one on a 1992 sample. All these men were infected at least one year prior to their 1985 measurement and we excluded all men who developed clinical AIDS in 1996 or earlier. The remaining 108 subjects could be considered in the asymptomatic phase. Figure 2 shows the pattern of RNA change for a sample of 27 of these subjects; although some subjects showed little change in RNA values over the 7 year period, others showed considerable changes. Changes in CD4 number and CD4 per cent tended to show a more consistent pattern (Figure 2). Furthermore, there also was a clear decrease in CD4 per cent, illustrating that this was not a perfectly stable equilibrium period, but rather that there was slow disease progression during this time.

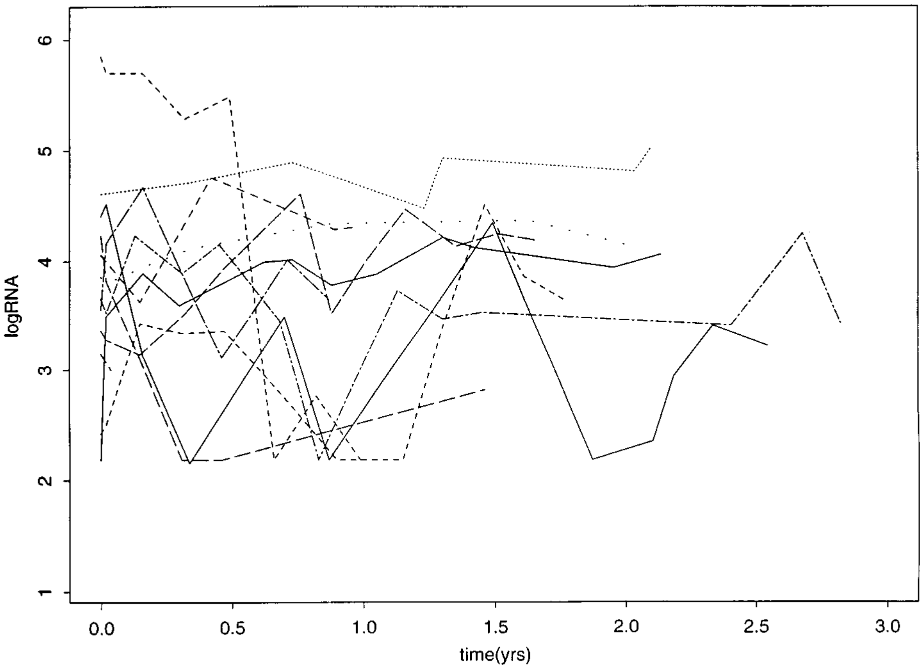

Figure 1. Log (base 10) RNA values for participants who did not progress to AIDS from the late treatment group of VA Cooperative Group trial 298 of early versus late AZT

Since the individual-level set-point definition given above implies that individual ranks on early and late viral load measurement are maintained (in other words, that there is a high level of 'tracking'), it follows that the early (set-point) level is highly correlated with later measurements. We calculated correlation coefficients between the 1985 and the 1992 measurements in the MACS data set (Table II). The two RNA measures have a moderate correlation of 0·47 which is lower than either the correlation of 0·59 between the two CD4 cell counts or the correlation of 0·63 between the two CD4 per centages. This suggests that the pattern of CD4 changes during the asymptomatic period was more predictable than the pattern of RNA changes, and that CD4 values showed more tracking than did RNA values.

We also addressed the prognostic value of current and earlier RNA and CD4 values on a group of 183 Los Angeles MACS participants who had both 1985 and 1992 RNA and CD4 values, without excluding those who went on to develop AIDS. We used the regression model developed by Shi *et al.*[22] to model the cube root of residual time to AIDS from 1992 as a linear combination of the covariates with additive normal error. The 1992 measures of RNA and CD4 were both significantly prognostic for the development of AIDS; the coefficients in these models were $-3{\cdot}1$ for log(RNA) and 5·9 for log(CD4). We also noted that a different marker, the median number of CD38 molecules on CD8 positive cells,[23] a measure of immune activation, was of considerably more prognostic significance than either of RNA or CD4 for this data set, suggesting that RNA and/or CD4 were not necessarily optimal markers for prognosis. The 1985 measures of RNA and CD4 differed in their prognostic significance; the coefficients for RNA was $-2{\cdot}2$ and was

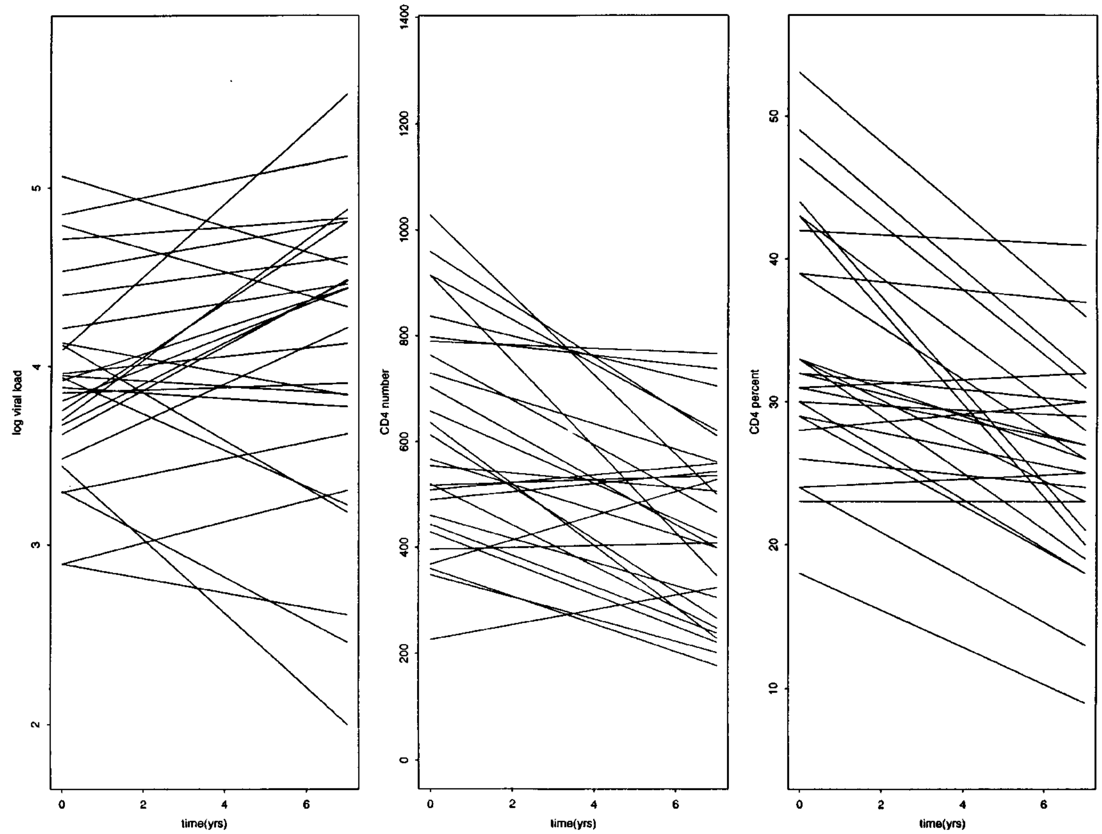

Figure 2. Log RNA, CD4 count and CD4% values for measurements taken at two time points in 1985 (time = 0) and 1992 (time = 7) for a subset of participants in the Los Angeles Multicenter AIDS Cohort Study, who did not go on to develop clinical AIDS by 1996

Table II. Correlations between 1985 and 1992 measures of RNA, CD4 count and CD4 per cent

	RNA(85)	RNA(92)	CDcount(85)	CD4count(92)	CD4%(85)	CD4%(92)
RNA(85)	1·00	0·47	−0·36	−0·28	−0·38	−0·37
RNA(92)		1·00	−0·21	−0·36	−0·27	−0·41
CD4count(85)			1·00	0·59	0·69	0·52
CD4count(92)				1·00	0·48	0·74
CD4%(85)					1·00	0·63
CD4%(92)						1·00

statistically significant while the coefficient of CD4 was 3·6 and was not statistically significant. When the 1992 and 1985 measures were considered in the same model, the earlier RNA still retained some marginal statistical significance whereas the earlier CD4 value clearly did not. With both markers at both time points considered jointly in the model, arguably the most appropriate analysis, then the 1985 values showed small non-significant coefficients, while both

1992 measurements were highly significant. Previous work with CD4 has shown that earlier values of CD4 add essentially no information about future disease progression if the current value is known[24, 25] even if values correlate in the short term.[26] For RNA, earlier values may add some information, but the current value is far more important. This observation invalidates the RNA set-point model.

If the RNA set-point model were correct, one would not need to measure viral load repeatedly in untreated asymptomatic subjects. However, such reduced monitoring is unlikely to be considered in clinical studies due to safety considerations. Also stratification or inclusion criteria for a clinical trial in asymptomatic subjects would be based on old RNA measurements; again it is doubtful this would be seriously considered. The RNA set-point model is an interesting hypothesis, but its apparent lack of validity and the fact that the set-point value is generally unknown argue against its usefulness in describing individual RNA trajectories or predicting clinical outcomes.

THERE IS A THRESHOLD IN THE DISEASE PROGRESSION RISK FOR BASELINE RNA AND CD4 AND CHANGES DURING THERAPY (JOHN IOANNIDIS AND IAN MARSCHNER)

Pro

The presence of a threshold for a biological marker (either its baseline or its change induced from an intervention), means that there is a value above or below which the observed risk changes to become zero, negligible or clinically insignificant. Although there is accumulating evidence on the biological continuum of HIV disease which should translate into a continuous risk function, there may exist strong threshold phenomena for the combination of CD4 and RNA measurements when we view disease risk from a clinical or decision making perspective. A linearly changing relative risk without a threshold translates to an exponentially changing absolute risk which strongly suggests a threshold situation. Absolute risks are more important for physicians and patients and the combination of CD4 and RNA measurements can define patient populations that have practically non-existent risks of disease progression for substantial periods of time even in the absence of therapy. Data from the Multicenter AIDS cohort study[11] suggest that for patients with CD4 $> 500/\text{mm}^3$ who have RNA <500 or even $<10{,}000$ copies/ml, the risk of AIDS in the subsequent 2–4 years is practically non-existent. In a mathematical modelling approach,[27] it has been estimated that a patient with CD4 above $500/\text{mm}^3$ and RNA of 10,000 copies/ml has at least 2·8 years before progressing to an AIDS equivalent stage and may have up to 19 years to enter the risk zone. By comparison, a patient with CD4 above $500/\text{mm}^3$, but RNA of 100,000 copies/ml is probably already within the risk zone of progressing at any moment or will enter the risk zone within 3 years at the most (Figure 3). This is a very abruptly increasing function that moves from a clinically low risk to clear and present danger with a relatively small change of the biological marker. Moreover, due to the decreasing course of the epidemic in developed countries, especially among patient subgroups that are likely to seek responsible and regular medical care, the overrepresentation of long-term non-progressors and slow progressors in these subpopulations is increasing.[28] This means that the proportion of patients who are below the threshold of meaningful short-term risk of natural disease progression is becoming larger, while at the same time they are more widely offered increasingly aggressive therapy. As a result, in the absence of robust, controlled data, we may be treating highly successfully a small segment of

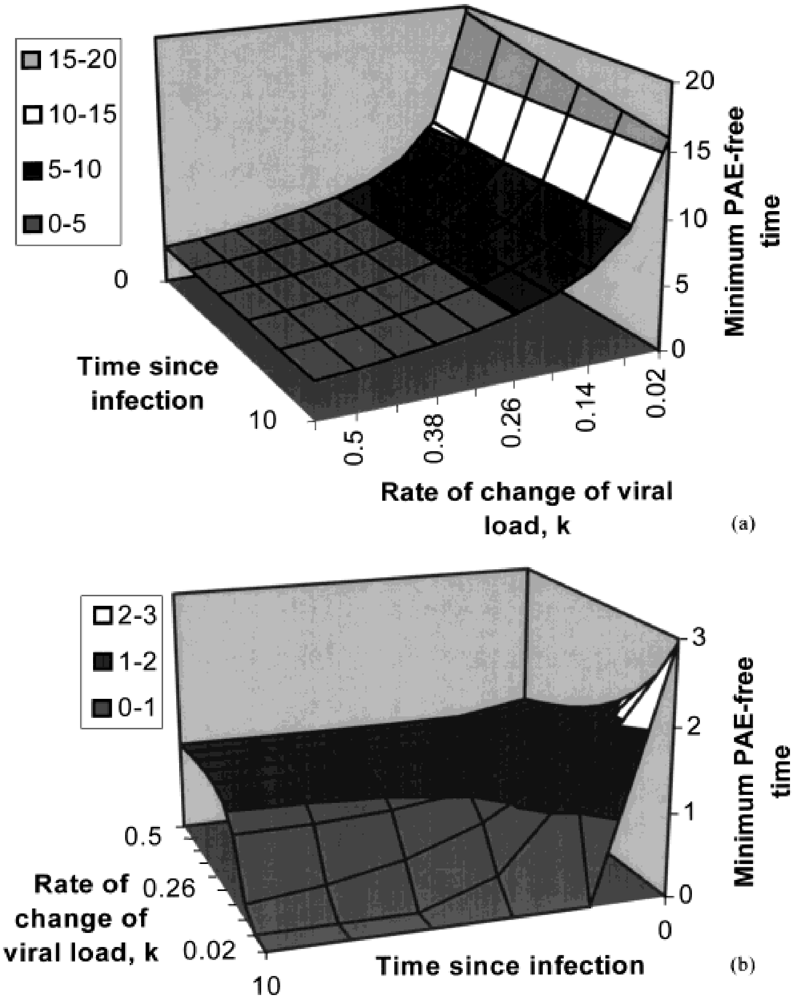

Figure 3. Estimation of the minimum time (in years) of asymptomatic, untreated HIV-infected patients with CD4 > 500 per cubic millimetre to be at risk for progression to an AIDS-equivalent stage for viral RNA load of 10,000 copies/ml (Figure 3(a)) and 100,000 copies/ml (Figure 3(b)) by NASBA or RT-PCR; probably Chiron bDNA values may be about half these numbers, that is, 5000 and 50,000 copies/ml, in a mathematical model accounting for the unknown time since infection (range 0–10 years) and the unknown rate of change of viral load over time (range 0·02–0·56 log per year). Patients with 10,000 copies/ml have at least 2·8–19 years to start being at risk for PAE, while patients with 100,000 copies/ml may develop PAE at any moment and will definitely enter the risk zone within less than 3 years. For details on the modelling see reference 26

the world pandemic that has had the most favourable disease course anyhow, while the world at large is succumbing to AIDS.

The situation is more problematic for determining thresholds of therapy induced changes, because past data pertain either to regimens of negligible potency and/or short-term effects (less than 1 year) that are clinically uninformative for a large majority of the patients. It is unknown whether intensification of regimens that have already decreased viral load to the limit of detection with current assays will achieve any meaningful reduction in the already drastically diminished clinical risk, especially when we consider absolute rates of disease progression. A decision analysis approach is required to piece together all evidence as data accumulate.

It must be acknowledged that in the absence of data, not only the presence of a threshold, but even its direction, is uncertain. For example, if complete viral eradication and cure are achievable, thresholds may exist beyond which patients with more advanced disease cannot achieve the desired optimum. Treating earlier and changing to more effective treatment earlier may indeed offer more optimal results,[29] and later stage patients beyond some yet undefined threshold may only see more modest or time-limited responses, or regimens of only moderate potency may not reach some threshold required for complete, prolonged suppression of viremia.[30]

In assessing disease risks, there are several clinically important components that unfortunately are typically ignored. In the management of a complex infection such as HIV, each of these components may have a different threshold. Examples include: the risk of toxicity, including late effects; the risk of inducing an unfavourable evolutionary spin on the virus; the cost-benefit ratio; and, most importantly, the risk of inadvertent failure with wide implementation of sophisticated interventions in real life in large patient populations rather than in controlled specialized settings.

Con

To address the issue of whether a threshold exists with respect to HIV RNA levels and clinical disease progression, it is necessary to define more precisely what is meant by the threshold. In this discussion we consider changes in HIV RNA levels in response to antiretroviral therapy, as well as absolute HIV RNA levels achieved in response to therapy, and address whether: (i) there is a maximum reduction in HIV RNA level beyond which further reduction has no additional clinical benefit; (ii) there is a minimal reduction in HIV RNA level before one achieves any clinical benefit; and (iii) there is an absolute HIV RNA level beyond which it is not clinically beneficial to attempt further reduction.

Data from the Multicenter AIDS Cohort Study[11] have indicated that there is a lower risk of progression to a new AIDS-defining event and/or death for lower levels of HIV RNA. Very similar results were obtained in a cross-study analysis of seven AIDS Clinical Trials Group (ACTG) studies,[31] where there was a progressively lower risk of clinical disease progression with lower baseline HIV RNA level. Such a relationship argues against a lower absolute HIV RNA threshold beyond which there is no further reduction in risk of clinical disease progression.

To explore further the existence of a threshold relationship with respect to therapy-induced HIV RNA levels, we consider the risk of clinical disease progression as a function of post-treatment reductions and absolute levels of HIV RNA. In ACTG trials 116A, 116B/117, 175, 197, 229, 241 and 259, there were 1000 individuals who received antiretroviral therapy, who had both baseline and follow-up measurements of HIV RNA level between 16 and 24 weeks subsequent to baseline, and who did not clinically progress (defined as a new AIDS-defining event or death) during the first 24 weeks of therapy. The reduction from baseline to the follow-up measurement was categorized into eight equal-sized groups (octiles), and a proportional hazards model was fitted adjusting for baseline level and stratifying by treatment regimen. This analysis was carried out using time from week 24 until clinical progression or censoring as the outcome variable, and a separate relative risk of clinical progression was allowed for each octile of HIV RNA reduction. A similar analysis was carried out using the absolute follow-up measurement of HIV RNA level.

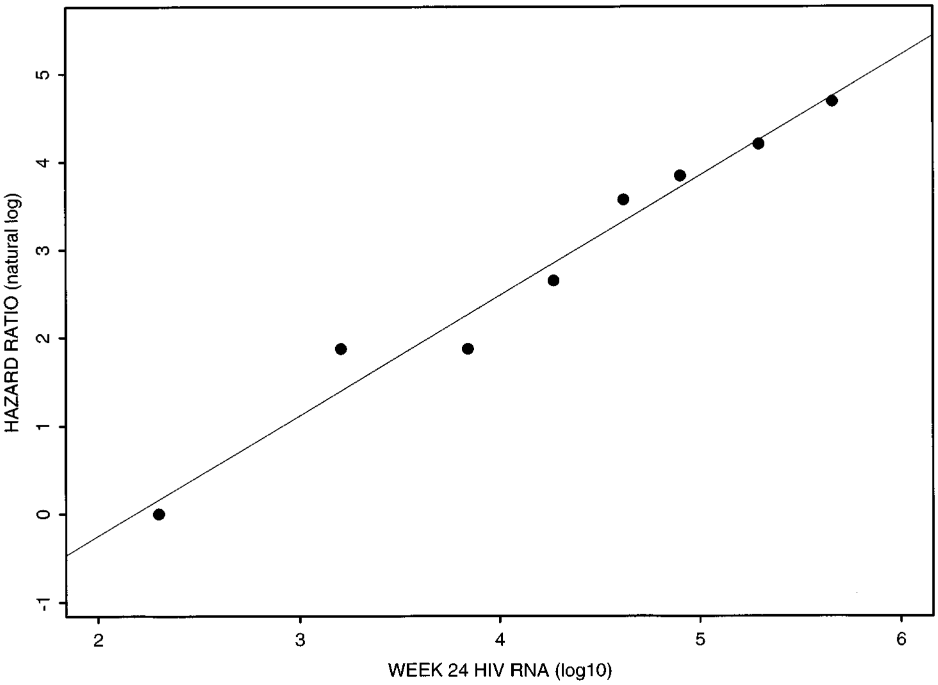

Figure 4. Relative risk of clinical progression versus octile of (log) HIV RNA level 24 weeks after treatment initiation

The results of these analyses showed a strikingly linear relationship between the (log) relative risk of progression and the HIV RNA level achieved in response to therapy, whether considered as a reduction (adjusted for baseline) or as an absolute level. Figure 4 presents the results of the latter analysis, while the results of the former analysis showed a very similar trend and are presented in Marschner *et al.*[31] This provides strong support for the linearity of HIV RNA level in the proportional hazards model. Based on these results, one would argue that the risk of clinical disease progression is related proportionally to: (i) the HIV RNA response to treatment; and (ii) the absolute HIV RNA level achieved in response to treatment. Such a relationship implies the absence of a threshold effect with respect to HIV RNA, and in particular, that larger reductions in HIV RNA are associated with greater delays in clinical disease progression, while lower absolute levels of HIV RNA are associated with a lower clinical progression risk. It also indicates that any reduction in HIV RNA (beyond the natural variation of the assay, approximately 2·5-fold)[32] is associated with a delay in clinical progression.

We need to interpret these conclusions, however, in light of the available data, which did not include a large number of individuals experiencing reductions in excess of 1·5 $\log_{10}$. As clinical endpoint data on more potent therapies become available, there is a need to investigate whether a proportional relationship persists when one achieves very large HIV RNA reductions. Furthermore, although it has been argued that reducing HIV RNA level leads to a proportional reduction in risk of clinical disease progression, untreated individuals with naturally low levels of HIV RNA (for example, <5000 copies/ml) may have such a low short to moderate term risk of clinical disease progression that further reduction in risk is not beneficial when balanced against the potential for toxicities and the considerable inconvenience of taking triple or quadruple drug regimens. However, for treated individuals who are only able to achieve low levels of HIV RNA due to therapy, further reduction (to unquantifiable levels) is likely beneficial due to protection against the development of drug resistance.[33]

CURRENT STATISTICAL METHODS FOR ADDRESSING SURROGACY ARE ADEQUATE (JEFFREY ALBERT AND DANYU LIN)

Pro

A surrogate endpoint as defined by Prentice,[3] satisfies the following: a (non-null) treatment effect on the surrogate implies, and is implied by, a (non-null) treatment effect on the 'true' endpoint. Prentice's operational criteria for surrogacy requires a marker to be correlated with the 'true' (clinical) endpoint and that the relationship between the surrogate and the clinical endpoint be the same across treatments. While the apparent stringency of this criterion is often noted, it is less frequently recognized that a marker that satisfies Prentice's criterion is not necessarily useful in a practical (predictive) sense. The latter limitation occurs because validity is specific to the set of study conditions (especially the treatments compared) for which the marker is assessed. One can imagine, for example, the above criteria being met by a treatment compliance measure in a study that compares drugs with equal biological efficacy. However, compliance is not likely valid for comparisons that involve other therapies in the likely class of interest. Thus, Prentice's criteria address what is referred to as 'statistical' surrogacy. What is desired, on the other hand, is validity over a given open-ended class of drugs – or what we could refer to as 'mechanistic' surrogacy. We cannot validate the latter entirely through statistical means and thus must involve biologic considerations, although statistical approaches can in principle invalidate a putative 'mechanistic' surrogate.

Some aspects of the Prentice operational criteria imply observational rather than direct causal inference. For example, comparing the 'relationship' between marker and clinical endpoint (as in the assessment of 'proportion treatment effect explained'), involves conditioning on a post-randomization variable (the marker response) which precludes casual inference without additional assumptions.[34] On the other hand, the original definition that a marker effect is equivalent to a clinical effect suggests that the proper unit of analysis is the study (or 'matched' pair). The basis for causal inference is tenuous since we cannot expect to sample studies randomly from a population of studies that represent the class of interest.

Recognizing the limitations of 'statistical' surrogacy, namely its restriction to the studies observed, and its exploratory (that is, correlative) rather than confirmatory (or causal) nature, alternative methods have been proposed. Fleming[35] presented a meta-analysis that assessed the association among studies of anti-HIV regimens of progression to AIDS/death and statistically significant CD4 effects. He found that out of eight studies with positive clinical effects, seven had significant CD4 effects. However, this apparently high sensitivity came at the price of low specificity, as six out of eight studies that showed no or negative clinical effects also had significant results for CD4. Fleming concluded in this admittedly simple approach that CD4 is an inadequate surrogate marker – a conclusion consistent with current consensus. Daniels and Hughes[36] proposed a meta-analytic method that uses a two-stage model to take into account measurement error while relating the underlying true surrogate and clinical responses. Using a Bayesian approach, they place prior distributions on the parameters in this regression relationship. This method has a number of attractive features: it allows for a test of association between marker and clinical endpoint; the model can incorporate covariates to assess a possible lack of heterogeneity among studies; one can accommodate studies that compare more than two treatments; and it allows inference for predicted clinical effects for specified marker differences. Daniels and Hughes applied this approach to a set of 16 AIDS clinical trials that involved the class of nucleoside analogue drugs. They determined that there was a significant association between difference in mean CD4 and the hazard ratio for progression to AIDS/death. However, they found that a very large CD4 effect (30–40 cell difference) was needed for 95 per cent confidence of a positive predicted clinical effect. Thus, the limitations of CD4 as a surrogate endpoint were also brought out by this approach. In the case of single-study analysis, aside from the proportion treatment effect explained approach, one could consider methods analogous to the above meta-analyses, but where subgroups (based on baseline covariates) provide the unit of analysis. Although there is much room for extension and further development, available methods provide reasonable and informative tools for the assessment of surrogacy.

Con

While a variety of metrics have been proposed for evaluating the degree to which the Prentice condition holds, such metrics are subject to misinterpretation because of the multiplicity of mechanisms by which drugs operate.[37] Without detailed understanding of these mechanisms, metrics of 'surrogacy' are not directly interpretable. Even when all of the mechanisms are understood, these metrics are associated with a high degree of uncertainty unless either treatment effects are large in moderate-size studies, or sample sizes are large in studies of moderately effective treatments.

Lin *et al.*[38] derived variance estimates and confidence intervals for the proportion of treatment effect explained by a surrogate marker (PTE) in the context of the Cox regression model. The PTE is obtained from the estimated regression coefficients of treatment with and without the marker in the model; as it involves the ratio of the two estimated parameters, it is highly variable. In addition, the statistic is in fact not a true proportion, as it is not guaranteed to lie within 0 to 1. Such inadequacies of PTE were demonstrated using data from ACTG 019 (see Volberding *et al.*[39] for the primary results). For example, the 95 per cent confidence interval for PTE using CD4 count as a surrogate for progression to AIDS after week 16 was $(-0{\cdot}14, 1{\cdot}08)$. As a result of the limitations of PTE our assessment of CD4 and RNA as surrogate endpoints is indeterminate. Existing studies do not provide strong, consistent evidence that either CD4 or RNA explains a large proportion of (aggregated) treatment effect.[8,21,40]

There has been little work on alternative statistical approaches. A meta-analysis approach seems desirable to reduce variability. Nevertheless, we need to resolve basic problems in the interpretation of measures of surrogacy such as PTE as well as questions about the biologic mechanisms of drug action.

THERE ARE NO SUBGROUP DIFFERENCES IN THE PROGNOSTIC ABILITY OF SURROGATE MARKERS (KINLEY LARNTZ AND SETH WELLES)

Pro

While, it may be conceded that there are differences in various subgroups for some studies, the problem arises when analysis of subgroups is not defined prior to study initiation. Often a clinical trial seeks to determine if A is better than B. Statistical analysis may find no evidence of difference for the main study question, but further analysis reveals some 'interesting subgroup findings'. In a recently analysed unpublished study there were 7 responses and 20 subgroup factors. Interaction tests were conducted to determine if there was a different effect of treatment (A or B) for each combination of response and subgroup factor. In all, 140 tests were done. The results were that 134 tests resulted in p-values greater than 0·05, 6 with p-values between 0·01 and 0·05, and none with p-values less than 0·01. The investigators concluded that 5 of the 6 nominally significant differences in relationships were plausible and that for several that did not attain the 0·05 level of significance it was worth reporting trends. From a statistical viewpoint, one must suspect that we are merely seeing the results of chance error.[41]

Clinical trials are powered to answer the main endpoint and are typically underpowered to address secondary endpoints.[42] They are even less suited to study variable effects of surrogate predictors in small subgroups. We should be sceptical of unanticipated differences in relationships and in most cases such differences have emerged with after-the-fact analyses.

Con

The ability of a marker to monitor treatment efficacy depends on the range of a marker in a patient group, relative to the magnitude of association of a marker with clinical outcome. For instance, how useful is CD4 count as a marker when the median level of cell count in a patient group is quite low, for example, 50 or 100 cells/mm^3? Unless very small changes of markers are related to observable large changes in clinical progression, CD4 might not be useful in a very homogenous patient group.

With respect to plasma HIV-1 RNA levels, one can see the effect of restricting the range of this marker on its prognostic capacity from the published 116A analysis[43] on 187 patients selected for virus substudy. In this study, there were two groups of patients, those with a short course of prior ZDV therapy ($\leqslant$16 weeks) and those who were ZDV drug-naive. Patients on a short course of therapy came into the study with significantly lower plasma HIV-1 RNA levels than patients who were drug naive (median levels 109,000 versus 171,000 copies, respectively; $p = 0{\cdot}03$); we would expect this due to ZDV's efficacy in lowering viral load in the absence of drug-resistance.

Changes of plasma HIV-1 RNA levels in response to antiretroviral treatment also varied by subgroup; patients with short-term prior ZDV therapy did not have decreases of viral load whether they were continued on ZDV or switched to ddI, while drug-naive patients had significant drops over 24 weeks of study, irrespective of treatment. However, should we interpret this as meaning that one or both drugs are not effective in treating previously ZDV-treated patients? In this short-term treatment subgroup, with low viral load at entry, both drugs maintained lower levels of circulating virus.

The problem with utilizing this prognostic marker in the short-term ZDV treatment subgroup is underscored by using Cox regression to analyse the prognostic value of plasma HIV-1 RNA; there was no significant association of baseline RNA level with time to new OI or death in this subgroup (RH = 1·11 for having a three-fold higher RNA level at baseline; $p = 0{\cdot}65$). Contrast this result with the significant association of this marker with progression in drug-naive patients who enrol with plasma HIV-1 levels that reflect their immune systems' abilities to suppress viral load; RH = 1·87; $p = 0{\cdot}004$. Thus, the utility of this marker to predict clinical progression varied by treatment history.

What about the use of laboratory markers to monitor treatment efficacy in diverse racial, ethnic or risk-profile subgroups, including African-Americans, women only, or male homosexuals? If these demographic factors are associated with having drug-resistant HIV or being drug-naive, then these demographic factors become indicators of treatment history, and we would expect that the utility of virologic laboratory data to predict disease progression might also vary by demographic group.

STRATIFICATION BASED ON BOTH CD4 AND RNA IS NECESSARY AND SUFFICIENT (ALVARO MUÑOZ AND JIM NEATON)

Pro

Using data from 1604 HIV positive participants of the Multicenter AIDS Cohort Study followed for over ten years with prognostic factors measured at a single time point around September of 1985, HIV RNA was the single best predictor of disease progression to clinical AIDS (that is, not simply reaching CD4 cell count below 200), followed in order of predictive strength by CD4 cell count and serum neopterin, and beta$_2$-microglobulin. To assess the relative prognostic power of each of the markers, we used a direct extension of q-q plots[44] to compare the percentile values of each marker in the group who developed AIDS with the percentile values in the group who remained AIDS-free. Specifically, if a value x of a marker was the pth percentile of the marker in the AIDS-free group, we determined the value y of the marker that was the same pth percentile of the marker in the group who developed AIDS. A graph of the natural logarithm (ln) of the ratio of the percentile values: $\ln(y/x)$ versus x, shows the difference in the distribution of the marker between the AIDS and the AIDS-free groups. If there is no difference in the distribution of the marker between the groups, then $y \approx x$ and $\ln(y/x) \approx 0$. Markers that are higher (or lower) in the AIDS group will have $\ln(y/x)$ values that are above (or below) zero; and we can use the distance from zero of the $\ln(y/x)$ values to compare the strength of the prognostic information. Namely, the further away the values of $\ln(y/x)$ are from zero, the better is the prognostic marker. The advantage of this extension of q-q plots is that we can compare markers that have different scales because the ln of the percentile value ratio is unitless and we can undertake formal testing by use of accelerated failure time distributions.[45]

Of the 1604 individuals whose viral load and CD4 cell counts were measured at a single time point (September 1985), 998 developed clinical AIDS during the follow-up to 1 July, 1995. The 25th, 50th and 75th percentile values of HIV-RNA in the AIDS-free group were 1111, 3636 and 10,340 copies/ml, respectively. The corresponding percentile values (25th, 50th and 75th) for the AIDS group were 7153, 19,145 and 52,900 copies/ml, respectively. The natural logarithms (ln) of the ratios of the percentile values in the AIDS: AIDS-free groups were $1{\cdot}86 = \ln(7153/1111)$, $1{\cdot}66 = \ln(19{,}145/3636)$, and $1{\cdot}63 = \ln(52{,}900/10{,}340)$, respectively. Part (a) of Figure 5 shows these three values with open circles marking the 25th and 75th percentiles and × marking the 50th percentile. Parts (b)–(d) in Figure 5 show the ln of the ratios of the percentile values for CD4 lymphocyte count, neopterin and beta$_2$-microglobulin, respectively. Since the distance from the zero line indicates the strength of the marker, HIV-1 RNA was the strongest predictive marker, followed by CD4 lymphocyte count, neopterin and beta$_2$-microglobulin, with CD4 lymphocyte count being only slightly more predictive than neopterin.

Multivariate analysis showed that plasma viral load discriminated risk at all categories of CD4 cell count. Conversely, CD4 cell count was significantly associated with the hazard of AIDS after

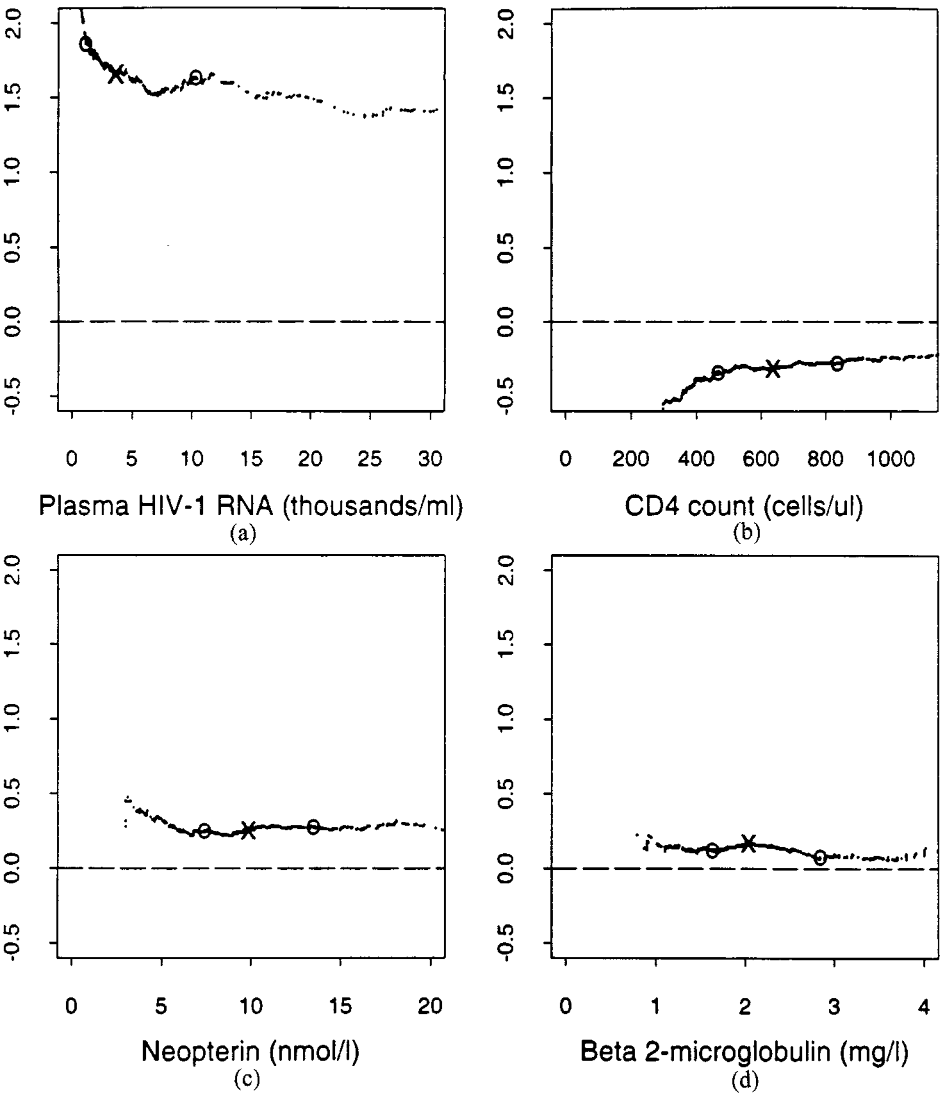

Figure 5. Natural logarithm of the ratio of the percentile values of the markers in those who developed AIDS compared to those who remained AIDS-free during follow-up. The data points show the natural log values of the ratios. The distance of the data points from 0 (dashed line) indicates the strength of the association of the marker with AIDS development. The values of the markers in the AIDS-free group are shown on the abscissa: '×' marks the median (50th percentile) value of the marker and '○' marks the 25th and 75th percentile

adjusting for HIV-1 RNA in a proportional hazards model.[46] A regression tree using recursive partitioning showed that HIV-1 RNA and CD4 lymphocyte counts jointly provided better outcome discrimination than either marker alone, defining categories of AIDS risk within six years ranging from <2 per cent to 98 per cent.[46] Although not fully sufficient, the level of risk discrimination provided by HIV-1 RNA and CD4 lymphocyte count may suffice for prediction of disease progression and for a wide range of issues regarding clinical management of HIV infected individuals.

Con

In general, the research question defines the eligibility criteria for clinical trials which in turn affects the sample size, recruitment/screening effort, and potential generalizability of the findings. Homogeneous groups of patients are desirable in some studies for reasons of efficiency, because of the endpoints considered, and patient safety concerns. In other studies, it may be either reasonable to assume that the treatment will help a diverse group of patients or important to show that it does. In those cases, one prefers or requires broad inclusion criteria.[47] These general statements apply to CD4 and HIV RNA as well as to other criteria, for example, asymptomatic/symptomatic disease, antiretroviral drug history, background treatment regimen.

CD4+ cell count and presence/absence of symptomatic HIV disease have been used to define inclusion criteria for most antiretroviral trials. Both are easily assessed and strong predictors for disease progression or death.[48] Both have a large impact on sample size. HIV RNA is another important predictor of disease progression,[46] but only a few studies have employed it to define trial eligibility. Thus, trials that appear homogeneous because of tight CD4+ criteria, may actually involve patients with a substantial range of risk of disease that one could reduce with consideration of HIV RNA.

The certainty about when to start treatment and how aggressive treatment should be argued for much broader inclusion criteria. Since even highly active HIV treatments may have time-limited efficacy, it is important to capture fully the range of risk of patients in future trials. Patients at higher risk of disease may benefit from treatment, but in the same follow-up period, the benefit for patients at lower risk of disease may either be uncertain or not present. Thus, for trials of treatment strategies, we should record but not restrict both CD4+ and HIV RNA.

CD4+ and HIV RNA are strong prognostic variables for disease progression and death. Control of both, either with inclusion criteria or pre- and/or post-stratification, reduces variability among patients and makes for more efficient conduct of some short-term pilot/early phase studies. The recording of both at baseline in larger, long-term strategy trials will provide data to assess homogeneity of treatment effects. Along the same lines, in some studies we may find it worthwhile either to freeze specimens for future analysis or to record additional information on other markers that may be important predictors of disease outcome, as the determination of such markers becomes simplified and we delineate and ascertain their predictive potential. Such markers could include SI/NSI[49] phenotype, viral resistance phenotype and genotype[50,51] immunogenetic parameters,[52] and immune activation markers,[24] among others. The levels of proof for the predictive value of each of these markers differ substantially, but it is unlikely that CD4 and viral load alone capture all the prognostic information. Limiting eligibility criteria in clinical trials according to values of these numerous markers is difficult, and probably unnecessary, but their explicit recording might provide useful information. One could more safely extrapolate such broadly inclusive trials with more clearly defined risk groups and use them to improve treatment guidelines.

SURROGATE MARKERS ARE PREFERABLE TO CLINICAL EVENTS AS PRIMARY ENDPOINTS IN EFFICACY TRIALS (JEFF MURRAY AND ROBERT COOMBS)

Pro

For the purpose of this discussion, we define efficacy trials as those studies that illustrate the efficacy of a drug treatment. We call trials that answer treatment strategy questions strategy trials to differentiate them from efficacy trials. For some strategy trials, such as defining when to start treatment, clinical events will probably remain as the endpoints of choice, however, for drug efficacy trials there are several reasons why one would prefer surrogate markers, such as HIV-RNA, over clinical endpoints. These reasons fall under two categories: (i) logistical/philosophical; and (ii) scientific.

Logistically, surrogate markers such as plasma HIV RNA are preferable from both an investigator and trial participant perspective. With surrogate markers used as study endpoints, participants are not required to remain on a fixed drug regimen until documentation of clinical deterioration. Participants who have shown virologic or immunologic deterioration may seek optimal therapy before experiencing significant progression, as they would in real life. Flexibility, consistency with routine clinical practice, and protection of the patient's well-being are reasons for preference of surrogate markers over clinical events.

For investigators, surrogate markers are logistically preferable for the same reasons as listed above but are also more attractive because surrogate marker studies are generally easier to conduct. Enrollment is easier in trials that give patients 'escape' options before clinical deterioration, and, if one measures time to virologic failure as the primary endpoint, patients can switch therapies prior to clinical failure yet still contribute to the study endpoint. With a surrogate endpoint, treatment compliance and switch-overs become less of an issue which also reduces the difficulty in evaluating the study data. Surrogate markers are easy to measure, requiring a simple blood test. In contrast, one may require a series of invasive diagnostic procedures to document clinical endpoints. Finally, one can implement surrogate marker trials, as compared to clinical event trials, more quickly in patients with earlier HIV infection in which clinical events occur infrequently over prolonged follow-up.

We cannot justify use of surrogate markers solely for convenience if there is a lack of scientific rationale to support the predictive abilities of the markers. However, there is substantial scientific support for the clinical relevance of plasma HIV RNA. Plasma HIV RNA measurements served as the basis for accelerated approval of four drugs, Epivir (lamivudine), Invirase (saquinavir), Norvir (ritonavir) and Crixivan (indinavir). All these drugs were studied in 'confirmatory' trials that used clinical events. In all cases the treatment arm that showed the greatest reduction in plasma HIV RNA also showed clinically and statistically significant reductions in disease progression and death.

Biologically, there is also scientific support for the use of plasma HIV RNA as an efficacy endpoint. Antiretroviral drugs exert their effect by halting viral replication and infectivity. The ability to reduce viral particles is the activity measure of interest. Large decreases in plasma HIV RNA levels are associated with parallel decreases in the lymphoid compartment.[53] From this aspect, one may think of HIV RNA suppression not as a surrogate endpoint but as the direct outcome of what antiretroviral drugs are designed to do, suppress virus.

Proponents of the formal statistical validation of plasma HIV RNA[3,54] leave unclear what percentage of treatment effect explained is adequate to consider a surrogate marker 'validated' for use as an endpoint. It is unlikely that any surrogate explains 100 per cent of the treatment effect. Also, not all clinical events are necessarily related to HIV disease progression. For example, the development of some infections, such as *Pneumocystis carinii* pneumonia, may sometimes relate more to lapses in prophylactics rather than HIV progression or lack of antiretroviral efficacy.

The clinical endpoint currently used in clinical event trials is a composite of death and approximately 30 infections and other conditions (modified from a CDC definition of AIDS used for epidemiology purposes). Endpoints other than death usually comprise the majority of events. This complex clinical endpoint has never been held to the standards of a formal validation as proposed for HIV RNA.

Con

Surrogate endpoints often do not predict the true clinical effects of a therapeutic intervention.[38,55] The reasons for this include either the presence of causal pathways of the disease process that are not mediated through the surrogate marker or the intervention has unintended mechanisms of action independent of the disease process, are quite unanticipated and are usually unrecognized until one conducts a clinical-endpoint-driven clinical trial. These surrogate marker considerations are particularly important with a chronic illness such as HIV infection.[37]

The validation criteria developed by Prentice[3] for phase III clinical trials require that a surrogate must be a correlate of the true clinical outcome and must fully capture the net effect of treatment on the clinical outcome.[37] Seldom, if ever, has the latter criterion been rigorously established in other areas of medicine let alone HIV-1 infection. For example, the wide acceptance of suppression of arrhythmias as a surrogate for prevention of mortality in survivors of myocardial infarction (simply because arrhythmias correlated with sudden death) led to a tragically misleading adoption of antiarrhythmic agents for such indications.[56] The acceptance of a surrogate by the FDA may be a decision made in the context of pressing practical considerations (in particular, to provide an additional treatment option to patients with few or no alternatives), but does not prove the validity of a surrogate. In fact, the unbridled exuberance shown by some clinical investigators, physicians and patients for plasma HIV-1 RNA as a surrogate for clinical outcome in HIV-1 infection is based on the misconception that if an outcome is a correlate (through observational studies) we can use it as a valid surrogate endpoint and replacement for the true clinical outcome in prospective studies. The results of ACTG protocol 076 that looked at the effect of zidovudine on plasma HIV-1 RNA level and the association of both with vertical HIV transmission is an example of the dissociation between a correlate and a surrogate.[57]

The complete validation of plasma HIV-1 RNA requires a better understanding of the causal pathways of the HIV-1 disease process that includes both virologic and immunologic parameters, as well as a therapeutic intervention's intended and unintended mechanisms of action.[37] As such, it is premature for us to use a single marker of disease, such as plasma viral RNA level, which is sufficient for short-term preliminary phase II studies, as a replacement for clinical endpoints in phase III clinical trials of HIV-1 therapies. The purpose of a phase III clinical trial is to learn about how to treat the patient, not just the patient's viral RNA!

IMMUNOLOGIC, VIROLOGIC AND CLINICAL ENDPOINTS SHOULD ALL BE INCORPORATED IN THE ASSESSMENT OF TREATMENT EFFICACY (LARRY CRANE AND RALPH DEMASI)

Pro

Measurements of virologic and immunologic endpoints are powerful tools for the appraisal of treatment efficacy in HIV disease. Unfortunately, surrogate marker treatment prediction effect is incomplete. The literature is replete with studies that demonstrate the limitations of CD4+ lymphocyte measurements.[8,35,58,59] Measurement of plasma HIV RNA response to therapy represents a significant advance. While there are encouraging preliminary results of ongoing studies of intensive or highly active antiretroviral treatment (HAART) limited to patients with early infection,[60,61] there is incomplete proof of the concept that serial measurements of plasma HIV RNA reflect virus activity in other compartments. Further, generalization of these observations to include all stages of HIV disease is, at present, inappropriate. While these trials have provided significant insights into the pathogenesis of HIV disease, it is noteworthy that we still do not fully understand the mechanisms, likely immune, behind the low RNA/high CD4+ cell counts observed in long-term non-progressors.[62] Indeed, these immune mechanisms are clearly distinct from the treatment effect of antiviral drugs. In that regard, the currently available surrogate markers may not apply to all treatment modalities, particularly immune- and gene-based therapies.

There are circumstances where we obtain important information by measuring clinical endpoints in antiretroviral trials. The opportunistic event profile in AIDS was drastically altered by prophylaxis.[63] Recently, HAART has altered the event profile even further. For the clinician who cares for persons with HIV disease, and for those with HIV disease, there are pragmatic, compelling reasons to provide information on, and to measure, clinical endpoints. Even with HAART, the immune system has a limited quantitative and qualitative capacity to replace T-cells lost to infection.[64] There are recent anecdotal reports of altered, unexpected or unusual opportunistic events occurring at the initiation of, or during HAART.[65,66] Monitoring event profiles should therefore provide important clues to clinicians and their patients for prophylaxis and treatment strategies. Furthermore, clinical endpoint trials still have great utility in the advanced patient – salvage trials. Strategy trials in less advanced patient populations are best addressed by combined or staged endpoints that utilize laboratory and clinical endpoints.

Con

Immunologic and virologic endpoints alone are necessary and sufficient for the assessment of treatment efficacy in all phases of controlled clinical trials. Moreover, powering clinical trials to detect treatment differences in clinical endpoints is no longer necessary to the extent that virologic and immunologic parameters can reliably predict the treatment effect of antiretrovirals on clinical endpoints.[46,67] In other words, confirmatory clinical endpoint studies (which necessarily have the same design as a previous trial with virologic and immunologic endpoints) are redundant since there is a high probability that observed meaningful treatment effects on virologic and immunologic endpoints (that is, larger and more sustained virologic and immunologic responses) will translate into meaningful treatment differences on the (unobserved) clinical endpoints.[21]

One may argue that there is too much unexplained variation in the estimate of the treatment effect on the clinical endpoint (that is, unexplainable by the treatment effects on virologic and immunologic endpoints alone). Measurement error associated with virologic and immunologic endpoints is often cited[59] as the primary reason that virologic and immunologic endpoints may not capture fully the treatment effect on the clinical endpoints. However, one must not dismiss the measurement error associated with the clinical endpoints in this shortcoming, since it is well-known that so-called non-differential misclassification biases[68] the estimate of the true treatment effect towards the null.

It also has been argued that the clinical effects of an anti-HIV treatment may operate through a causal pathway which excludes virologic and immunologic responses (and therefore we may miss a good treatment with respect to clinical outcomes if we rely only on virologic and immunologic endpoints); however, this argument is biologically implausible given recent advances in our understanding of HIV-1 pathogenesis.[18]

It also has been argued that an anti-HIV treatment may actually be unduly toxic and as such, we should avoid sole reliance on virologic and immunologic endpoints since we may miss the toxic effects (or they could contribute to the efficacy analysis if we consider all-cause mortality). This underlies the importance of thorough treatment comparisons of clinical adverse event rates and abnormal laboratory profiles. There are other disadvantages of using clinical endpoints for assessment of treatment efficacy in controlled clinical trials, namely: (i) with the new and improved anti-HIV regimens and the emphasis on early treatment, clinical endpoints are becoming increasingly rare, and as such, studies designed to detect clinical endpoints will require more patients with longer follow-up periods; (ii) with the increasing use of real-time viral load monitoring treatment, we can identify failures early (often within two to four weeks of treatment initiation); thus patients are unwilling to remain on such a failing regimen and non-compliance results, which blurs the true treatment difference and makes the results difficult to interpret; and (iii) the excess morbidity and mortality observed in a confirmatory clinical endpoint trial is a prerequisite for a successful trial (that is, a trial that confirms treatment efficacy), but presents ethical dilemmas if such a trial promotes sub-optimum patient care.

THERE IS A NEED FOR STANDARDIZATION IN ANALYSIS AND REPORTS WITH RESPECT TO SURROGATE ENDPOINTS (BRIAN CONWAY AND DENNIS O. DIXON)

Pro

The goal of antiretroviral therapy in 1997 is to reduce plasma viral load below the limit of quantitation of currently available assays (generally speaking, below 400 copies/ml plasma). It is clear that a wide range of drug regimens can accomplish this, particularly in treatment-naive individuals. More sensitive tests will become widely available in the near future, allowing accurate quantitation down to 50 copies/ml plasma or less. Nevertheless, we need agreement on the most appropriate endpoint to compare the efficacy of different highly effective treatment regimens. A number of specific issues need addressing: (i) Are the kinetics of plasma viral load clearance relevant in comparing the ability of different regimens to suppress HIV replication? (ii) What threshold of suppression should we utilize to compare regimens? (iii) Should we use a composite measure that integrates the magnitude and durability of virologic suppression? (iv) Should we integrate additional factors into a composite virologic efficacy parameter?

In a study[69] of 151 drug-naive patients randomized to zidovudine/didanosine (ZDV/ddI) versus zidovudine/nevirapine (ZDV/NVP) versus zidovudine/didanosine/nevirapine (ZDV/ddI/

NVP), no significant difference in virologic suppression between the ZDV/ddI and the ZDV/ddI/NVP groups was observed at 52 weeks ($-1{\cdot}1$ versus $-1{\cdot}3$ log 10 copies/ml, $p = 0{\cdot}95$) when conventional viral load test results (limit of quantitation 400 copies/ml) were considered. When the analysis was repeated using a more sensitive viral load assay (limit of quantitation 20 copies/ml),[70] a difference between the two groups was now readily evident ($-1{\cdot}5$ versus $-2{\cdot}1$ log 10 copies/ml, $p = 0{\cdot}001$). The clinical relevance of this result is supported by the fact that patients with maximal virologic suppression (<20 copies/ml) had a more rapid and durable response to therapy and were less susceptible to compliance-related loss of virologic containment than patients who were not maximally suppressed (20–400 copies/ml).

The standard analysis for comparing different antiretroviral therapy regimens should involve the reduction in viral load compared to baseline using a test capable of quantitating HIV down to the lowest practicable level. One could use secondary analysis (including the kinetics of initial virologic suppression and the durability of the response) to provide supportive evidence of therapeutic efficacy.[71]

Con

Standardization is good only when based on good analyses. At present, data are not available to guide the definition of marker-based endpoints. It is likely to be more useful to standardize the data to be collected, so that alternative approaches to analysis are possible, but even such standardization is a formidable challenge.

Choice of a particular primary endpoint should not be arbitrary. It should reflect a belief on the part of the investigators that one metric carries more meaning than other possible choices. We need to know much more about how RNA changes over time in the presence of antiretroviral treatments in order to define summaries with evident meaning.

Perhaps the most controversial issue is whether the RNA levels ultimately matter as surrogates for clinical events in the traditional way or whether they are important by themselves, that is, whether viral load reductions measure efficacy. In the latter case, resources that one would otherwise use for prolonged follow-up and collection of laboratory and clinical data might be used instead for more intensive observation over a specified, relatively short time period. Data relevant to this point are emerging, but only for relatively advanced patient populations.

Until we have a solid basis for focusing on a particular metric, it is surely better to encourage diversity of approaches so as to explore the widest possible variety.

It might be claimed that we should at least be able to examine results of different studies using the same metric. This means that we need agreement on a basic data set for collection, even if it involves some observations for which there is no immediate analysis plan. (It will be tricky to justify the extra burden of data collection at the level of the informed consent.)

Suppose two trials are both evaluating mean reduction in log (RNA) at 6 months, but one requires single measurements at baseline and at 6 months, whereas the other requires replicate measurements at baseline separated by at least 7 days and requires averaging of measurements at 5, 6 and 7 months for the 6-month value. Standardizing data requirements entails a significant additional burden on the first trial participants and investigators.

Given the severe limitations on current understanding, it is more productive to obtain information about a range of endpoint definitions than to standardize.

Table III. Description of debate questions and summary conclusions

Question	Summary
The choice of marker metric and assay for a surrogate endpoint have little effect on conclusions regarding treatment efficacy	If assays measure the same quantity then the choice of assay should be immaterial. Inconsistency of results for different metrics may point to design or analysis flaws, or inconclusive data. Estimates of proportion of treatment effect explained can be very different according to the metric used, but its typically high variability makes comparisons difficult.
The 'set-point' provides most of the information in the trajectory of RNA with regard to disease progression	The 'set-point' may be examined from both the group and individual perspective. From the group perspective, a meta-analysis of three natural history studies suggested that baseline values (considered as set-points) carried as much prognostic information as the most recent value. On an individual basis, longitudinal measurements of RNA over time show very substantial variability which contradicts the set-point concept of a long, stable RNA level and undermines its practical use in study design and analysis. The usefulness of the set-point model is also questionable in the context of effective treatment.
There is a threshold in the disease progression risk for baseline RNA and CD4 and changes during therapy	For asymptomatic patients with CD4 $>$ 500 who have RNA $<$10,000 the risk of AIDS within 2–4 years is practically non-existent. Determining thresholds for changes is problematic since the available data is primarily short term. In a cross-sectional analysis of several protocols there is a linear response, with no evidence of any threshold, however changes in RNA were modest. In the absence of data on the efficacy of potent regimens and the possibility of eradication even the direction of a threshold is uncertain. All reductions of viral load are likely to result in some benefit. However, quantitation of the benefit in absolute magnitude for individual patients and populations is also important.
Current statistical methods for addressing surrogacy are adequate	The proportion of treatment effect explained by a surrogate endpoint is subject to large variability and difficulties in interpretation. An alternative approach is to model the relationship between marker changes and clinical response, either through meta-analysis or within a single study by looking at subgroups. All methods are subject to bias due to the observational nature of the question. Extrapolation to new drugs requires a biologic understanding of drug action.
There are no subgroup differences in the prognostic ability of surrogate markers	Subgroup analysis should be defined prior to the initiation of the main analysis, and driven by biological considerations. It may be argued that the prognostic ability of HIV RNA differs in naïve and treated patients. However, a difficulty in such subgroup analyses is that the range of marker values within subgroups is often reduced.
Stratification based on both CD4 and RNA is necessary and sufficient	The combination of CD4+ cell count and RNA level provides better outcome discrimination than either marker alone and explain to a large extent the variability in time to disease progression. Homogeneity is desirable for some studies to reduce variability among patients for short-term/pilot studies. Heterogeneity is desirable in larger studies and incorporation of CD4 and RNA measurements in the trial design may allow an assessment of the generalizability of the results.

Table III. Continued

Question	Summary
Surrogate markers are preferable to clinical events as primary endpoints in efficacy trials	Use of surrogates as opposed to clinical endpoints in efficacy trials as opposed to strategic trials appears to give increased flexibility and the outcome measures in the short term correlate clinically and virologically. For a chronic illness such as HIV where the causal pathway is not fully known, it is important to consider clinical endpoints in patient management. Adequacy of surrogate endpoints such as HIV RNA may not carry over across different circumstances, for instance from treating adults to preventing maternal-infant transmission.
Immunologic, virologic, and clinical endpoints should all be incorporated in the assessment of treatment efficacy	Clinical endpoints are still needed for advanced disease patients, strategy designed clinical trials, and trials designed to address the relative contribution of the immune system versus viral load at various stages of disease. They may not be needed in drug licensure studies.
There is a need for standardization in analysis and reports with respect to surrogate endpoints	So far no particular summary of RNA copy number data has emerged as most useful for comparing effects of treatments (although a metric based on the ability to suppress viral load to below the level of quantitation of the most sensitive assays is promising). Until we have a solid basis, however, for focusing on one metric or another, it is probably better to encourage a diversity of approaches.

CLOSING REMARKS

The primary focus of this workshop has been to draw attention to key considerations in the use of surrogate markers of HIV disease. Table III provides a summary of the major points made for each of the nine questions discussed. It is clear that many questions have compelling arguments on both sides and we emphasize that the purpose of this paper has been to promote a greater understanding of these issues rather than to provide final answers. Challenges in the use of surrogate markers in HIV clinical trials are likely to increase as the use of new potent antiretroviral therapies becomes more widespread and the need escalates to contain and treat the pandemic worldwide, and not only in selected populations. Further consideration of these issues, together with many of the insights elucidated here, are crucial to the proper use of surrogate markers in future HIV clinical research and practice.

ACKNOWLEDGEMENTS

This work was supported by awards AI27664, AI27757, AI30731, DK49477 (R.W. Coombs), AI35161 (L. Kalish), AI95030 (I. Marschner), AI35043 (A. Muñoz) and AI29196 (J. Taylor) from the National Institutes of Health. Thanks go to Andrew Kaplan, Janis Giorgi and Pamela Hartigan for providing the data analysed on the setpoint hypothesis presented by Jeremy Taylor. Figure 5 was based on data provided by John Mellors and the Multicenter AIDS Cohort Study with principal investigators John Phair, Roger Detels, Janis Giorgi, Joseph Margolick, Alvaro Muñoz, and Charles Rinaldo. The authors also thank the editor and referees for their helpful comments.

REFERENCES

1. Taylor, J. M. G., Fahey, J. L., Detels, R. and Giorgi, J. V. 'CD4 percentage, CD4 number and CD4 : CD8 ratio in HIV infection: Which to choose and how to use', *Journal of Acquired Immune Deficiency Syndromes*, **2**, 114–124 (1989).
2. Malone, J. L., Simms, T. E., Gray, C. G., Wagner, R. F., Burge, J. R. and Burke, D. S. 'Sources of variability in repeated T-helper lymphocyte counts from human immunodeficiency virus type 1-infected patients: Total lymphocyte count fluctuations and diurnal cycle are important', *Journal of Acquired Immune Deficiency Syndromes*, **3**, 144–151 (1990).
3. Prentice, R. L. 'Surrogate endpoints in clinical trials: definitions and operational criteria', *Statistics in Medicine*, **8**, 431–440 (1989).
4. Vandamme, A. M., Schmit, J. C., Van Dooren, S., Van Laethem, K., Grobbers, E., Kok, W., Goubau, P., Witrouw, M., Peetermans, W., DeClerq, E. and Desmyter, J. 'Quantification of HIV-1 RNA in plasma: comparable results with the NASBA HIV-1 RNA QT and the AMPLICOR HIV monitor kit', *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **13**, 127–139 (1996).
5. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
6. Ioannidis, J. P. A., Cappelleri, J. C., Sacks, H. S. and Lau, J. 'The relationship between study design, results and reporting in randomized trials of HIV infection', *Controlled Clinical Trials*, **18**, 431–444 (1997).
7. Hogan, C. H., Hodges, J. S., Mugglin, A., Peterson, P. M., Abrams, D. and Saravolatz, L. 'The perils of visit-driven endpoints in antiretroviral trials'. In: Abstracts of the XI International Conference on AIDS, Vancouver, Canada, July 7–12, 1996 (abstract TUB.522).
8. Choi, S., Lagakos, S. W., Schooley, R. T., and Volberding, P. A. 'CD4 + lymphocytes are an incomplete surrogate marker for clinical disease progression in persons with asymptomatic HIV infection', *Annals of Internal Medicine*, **118**, 674–680 (1993).
9. Babiker, A. for the Delta Co-ordinating Committee and Virological Group. 'Can HIV–1 RNA viral load be used as a surrogate for clinical outcome in HIV disease?', 6th European conference on Clinical Aspects and Treatment of HIV Infection, Hamburg, Germany, 11–15 October 1997 (Abstract 103).
10. O'Quigley, J. and Flandre, P. 'Predictive capability of proportional hazards regression', *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 2310–2314 (1994).
11. Mellors, J. W., Rinaldo, C. R., Gupta, P., White, R. M., Todd, J. A. and Kingsley, L. A. 'Prognosis of HIV-1 infection predicted by the quantity of virus in plasma', *Science*, **272**, 1167–1170 (1996).
12. Henrard, D. R., Phillips, J. F., Muenz, L. R., Blattner, W. A., Wiesner, D., Eyster, M. E. and Goedert, J. J. 'Natural history of HIV-1 cell-free viremia', *Journal of the American Medical Association*, **274**, 554–558 (1995).
13. Phillips, A. N., Eron, J. J., Bartlett, J. A., Rubin, M., Johnson, J., Price, S., Self, P. and Hill, A. M. 'HIV-1 RNA levels and the development of clinical disease', *AIDS*, **10**, 859–865 (1996).
14. Kuhn, A. M. and DeMasi, R. A. 'Empirical power for tests on longitudinal data with non-informative missingness with applications to AIDS trials', Presentation at the International Biometric Society Eastern North American Region Meetings, Memphis, TN, March, 1997.
15. Raboud, J. M., Montaner, J. S., Conway, B., Maley, L., Sherlock, C., O'Shaughnessy, M. V. and Schechter, M. T. 'Variation in plasma RNA levels, CD4 cell counts, and p24 antigen levels in clinically stable men with human immunodeficiency virus infection', *Journal of Infectious Diseases*, **174**, 191–194 (1996).
16. Ho, D. D. 'Viral counts count in HIV infection', *Science*, **272**, 1124–1125 (1996).
17. Mellors, J. W., Kingsley, L. A., Rinaldo, C. R., Todd, J. A., Hoo, B. S., Kokka, R. P. and Gupta, P. 'Quantitation of HIV-1 RNA in plasma predicts outcome after seroconversion', *Annals of Internal Medicine*, **122**, 573–579 (1995).
18. Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M. and Markowitz, M. 'Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infections', *Nature*, **373**, 123–126 (1995).
19. Hamilton, J. D., Hartigan, P. M., Simberkoff, M. S., Day, P., Diamond, G., Dickinson, G., Dursano, G., Egorin, M., George, W., Gordin, F., Hawkes, C., Jensen, P., Klimas, N., Labriola, A., Lahart, C., O'Brien, W., Oster, C., Weinhold, K., Wray, N., Zolla-Pazner, S. and the Veterans Affairs Cooperative

Study Group on AIDS treatment, 'A controlled trial of early versus late treatment with Zidovudine in symptomatic HIV infection', *New England Journal of Medicine*, **326**, 437–443 (1992).

20. O'Brien, W. A., Hartigan, P. M., Martin, D., Esinhart, J., Hill, A., Benoit, S., Rubin, M., Simberkoff, M. F., and Hamilton, J. D. 'Changes in Plasma HIV-1 RNA and CD4 + lymphocyte counts and the risk of progression to AIDS', *New England Journal of Medicine*, **334**, 426–431 (1996).
21. Paxton, W. B., Coombs, R. W., McElrath, M. J., Keefer, M. C., Hughes J., Sinangil, F., Chernoff, D., Demeter, L., Williams, B. and Corey, L. 'Longitudinal analysis of quantitative virologic measures in HIV infected subjects with <400 CD4 lymphocytes: implications for applying measurements to individual patients', *Journal of Infectious Diseases*, **175**, 247–254 (1997).
22. Shi, M., Taylor, J. M. G. and Muñoz, A. 'Models for residual time to AIDS', *Lifetime Data Analysis*, **2**, 31–49 (1996).
23. Giorgi, J. V., Liu, Z., Hultin, L.E., Cumberland, W. G., Hennessey, K. and Detels, R. 'Elevated levels of CD38 + CD8+ T cells in HIV infection add to the prognostic value of low CD4+ T cell levels: Results of 6 years of follow-up', *Journal of Acquired Immune Deficiency Syndrome*, **6**, 904–912 (1993).
24. Shi, M., Currier, R. J., Taylor, J. M. G., Tang, H., Hoover, D. R., Chmiel, J. S. and Bryant, J. 'Replacing time since HIV infection by marker values in predicting residual time to AIDS diagnosis', *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **12**, 309–316 (1996).
25. Yong, F. H. L., Taylor, J. M. G., Bryant, J. L., Chmiel, J. S., Gange, S. J. and Hoover, D. 'Dependence of the hazard of AIDS on markers', *AIDS*, **11**, 217–228 (1997).
26. Shi, M., Taylor, J. M. G., Fahey, J. L., Hoover, D. R., Munoz, A. and Kingsley, L. A. 'Early levels in CD4, neopterin and b2 microglobulin indicate future HIV disease progression', *Journal of Clinical Immunology*, **17**, 43–52 (1997).
27. Ioannidis, J. P. A., Cappelleri, J. C., Lau, J., Sacks, H. S. and Skolnik, P. R. 'Predictive value of viral load measurements in asymptomatic, untreated HIV-infection: a mathematical model', *AIDS*, **10**, 255–262 (1996).
28. Ioannidis, J. P. A., Cappelleri, J. C., Schmid, C. H. and Lau, J. 'Impact of epidemic and individual heterogeneity on the population distibution of disease progression rates. An example from patient populations in trials of human immunodeficiency virus infection', *American Journal of Epidemiology*, **144**, 1074–1085 (1996).
29. Ioannidis, J. P. A., Sacks, H. S., Cappelleri, J. C. and Lau, J. 'Clinical efficacy of antiretroviral changes in treatment-experienced HIV-infected patients. A meta-analysis', *Online Journal of Current Clinical Trials*, 15 May 1997; Doc No 204 (1997).
30. Havlir, D. V. and Richman, D. D. 'Viral dynamics of human immunodeficiency virus: implications for drug development and therapeutic strategies', *Annals of Internal Medicine*, **124**, 984–994 (1996).
31. Marschner, I. C., Collier, A. C., Coombs, R. W., D'Aquila R. T., DeGruttola, V., Fischl, M. A., Hammer, S. M., Hughes, M. D., Johnson, V. A., Katzenstein, D. A., Richman, D. D., Smeaton, L. M., Spector, S. A. and Saag, M. S. 'The use of changes in plasma levels of human immunodeficiency virus type 1 RNA levels to assess the clinical benefit of antiretroviral therapy', *Journal of Infectious Diseases*, **177**, 40–47 (1998).
32. Hughes, M. D., Johnson, V. A., Hirsch, M. S., Bremer, J. W., Elbeik, T., Erice, A., Kuritzkes, D. R., Scott, W. A., Spector, S. A., Basqoz, N., Fischl, M. A. and D'Aquila R. T. 'Monitoring plasma human immunodeficiency virus (HIV-1) RNA levels in additions to CD4+ lymphocyte count improves assessment of antiretroviral therapeutic response', *Annals of Internal Medicine*, **126**, 929–938 (1997).
33. Emini, E. A. 'Resistance to anti-human immunodeficiency virus therapeutic agents', *Advances in Experimental Medicine and Biology*, **390**, 187–195 (1995).
34. Rosenbaum, P. R. 'The consequences of adjustment for a concomitant variable that has been affected by the treatment', *Journal of the Royal Statistical Society*, *Series A*, **147**, 656–666 (1984).
35. Fleming, T. R. 'Surrogate markers in AIDS and cancer trials', *Statistics in Medicine*, **13**, 1423–1435 (1994).
36. Daniels, M. J. and Hughes, M. D. 'Meta-analysis for the evaluation of potential surrogate markers', *Statistics in Medicine*, **16**, 1965–1982 (1997).
37. De Gruttola, V., Fleming, T. R., Lin, D. Y. and Coombs, R. 'Validating surrogate markers – are we being naive?', *Journal of Infectious Diseases*, **175**, 237–246 (1997).
38. Lin, D. Y., Fleming, T. R. and DeGruttola, V. 'Estimating the proportion of treatment effect explained by a surrogate marker', *Statistics in Medicine*, **16**, 901–910 (1997).

39. Volberding, P. A., Lagakos, S. W., Koch, M. A., Pettinelli, C., Myers, M. W., Booth, D. K., Balfour. H. H. Jr, Reichman, R. C., Bartlett, J. A., Hirsch, M. S., *et al.* 'Zidovudine in asymptomatic human immunodeficiency virus infection: a controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter', *New England Journal of Medicine*, **322**, 941–949 (1990).
40. Lin, D. Y., Fischl, M. A. and Schoenfeld, D. A. 'Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials', *Statistics in Medicine*, **12**, 835–842 (1993).
41. Miller, R. G., Jr. *Simultaneous Statistical Inference*, 2nd edn, Springer-Verlag, New York, 1981.
42. Meinert, C. L. *Clinical Trials: Design, Conduct, and Analysis*, Oxford Univeristy Press, New York, 1986.
43. Welles, S. L., Jackson, J. B., Yen-Lieberman, B., Demeter, L., Japour, A. J., Smeaton, L. M., Johnson, V. A., Kuritzkes, D. R., D'Aquila, R. T., Reichelderfer, P. A., Richman, D. D., Reichman, R., Fischl, M., Dolin, R., Coombs, R. W., Kahn, J. O., McLaren, C., Todd, J., Kwok, S. and Crumpacker, C. S. for the AIDS Clinical Trials Group Protocol 116A/116B/117 Team. 'Prognostic value of plasma human immunodeficiency virus type 1 (HIV-1) RNA levels in patients with advanced HIV-1 disease and with little or no prior zidovudine therapy', *Journal of Infectious Diseases*, **174**, 696–703 (1996).
44. Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. *Graphical Methods for Data Analysis*, Belmont, CA, Wadsworth International Group (Duxbury Press), 1983.
45. Muñoz, A. and Sunyer, J. 'Comparison of semiparametric and parametric survival models for the analysis of bronchial responsiveness', *American Journal of Respiratory and Critical Care Medicine*, **154**, S234–S239 (1996).
46. Mellors, J. W., Muñoz, A., Giorgi, J. V., Margolick, J. B., Tassoni, C. J., Gupta, P., Kingsley, L. A., Todd, J. A., Saah, A. J., Detels, R. R., Phair, J. P. and Rinaldo, C. R., Jr. 'Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection', *Annals of Internal Medicine*, **126**, 946–954 (1997).
47. Yusuf, S., Held, P., Teo, K. K. and Toretsky, E. R. 'Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria', *Statistics in Medicine*, **9**, 73–86 (1990).
48. Saravolatz, L., Neaton, J. D., Sacks, L., Deyton, L., Rhame, F. and Sherer, R. 'CD4+ T lymphocyte counts and patterns of mortality among patients infected with human immunodeficiency virus who were enrolled in Community Programs for Clinical Research on AIDS', *Clinical Infectious Diseases*, **22**, 513–520 (1996).
49. Kozal, M. J., Shafer, R. W., Winters, M. A., Katzenstein, D. A., Aquiniga, E., Halpem, J. and Merigan, T. C. 'HIV-1 syncytium inducing phenotype, codon 215 reverse transcriptase mutation and CD4 cell decline in zidovudine-treated patients', *Journal of Acquired Immune Deficiency Syndromes*, **7**, 832–838 (1994).
50. Holodnyi, M., Katzenstein, D., Mole, L., Winters, M. and Merigan, T. 'HIV reverse transcriptase codon 215 mutations diminish virologic response to didanosine-zidovudine treatment in subjects with non-syncytium inducing phenotype', *Journal of Infectious Diseases*, **174**, 854–857 (1996).
51. Dean, M., Carrington, M., Winkler, C., Hutley, G. A., Smith, M. W., Allikmets, R., Goedert, J. J., Buchbinder, S. P., Vittinghof, E., Gomperts, E., Donfield, S., Vlahov, D., Kalsow, R., Saah, A., Rinaldo, C., Detels, R. and O'Brien, S. J. 'Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene', *Science*, **273**, 1856–1862 (1996).
52. Mocroft, A., Johnson, M. A. and Phillips, A. N. 'Factors affecting survival in patients with the acquired immune deficiency syndrome' [editorial], *AIDS*, **10**, 1057–1065 (1996).
53. Wong, J. K., Gunthard, H.F., Havlir, D. V., Haase, A. T., Zhang, Z. Q., Kwok, S., Ignacio, C.C., Keating, N. A., Chodakewitz, J., Emini, E., Meibohm, A., Jonas, L. and Richman, D. D. 'Reduction of HIV blood and lymph nodes after potent antiretroviral therapy', Presentation at 4th Conference on Retroviruses and Opportunistic Infections, 22–26 January 1997.
54. Freedman, L. S., Graubard, B. L. and Schatzkin, A. 'Statistical validation of intermediate endpoints for chronic diseases', *Statistics in Medicine*, **11**, 167–178 (1993).
55. Fleming, T. R. and DeMets, D. L. 'Surrogate endpoints in clinical trials: are we being misled?', *Annals of Internal Medicine*, **125**, 605–613 (1996).
56. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. 'Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction', *New England Journal of Medicine*, **321**, 406–412 (1989).
57. Sperling, R. S., Shapiro, D. E., Coombs, R. W., Todd, J. A., Herman, S. A., McSherry, G. D., O'Sullivan, M. J., VanDyke, R. B., Jimenez, E., Rouzioux, C., Flynn, P. M. and Sullivan, J. L. 'Maternal viral load,

zidovudine treatment, and the risk of transmission of human immunodeficiency virus type 1 from mother to infant', *New England Journal of Medicine*, **335**, 1621–1629 (1996).

58. DeGrutolla, V., Wulfsohn, M., Fischl, M. A. and Tsiatis, A. 'Modeling the relationship between survival and CD4 lymphocytes inpatients with AIDS and AIDS related complex', *Journal of Acquired Immune Deficiency Syndromes*, **6**, 359–365 (1993).
59. Tsiatis, A. A., DeGruttola, V. and Wulfsohn, M. 'Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS', *Journal of the American Statistical Association*, **90**, 27–37 (1995).
60. Gulick, R., Mellors, J., Havlir, D., Eron, J., Gonzalez, C., McMahon, D., Richman, D., Valentine, F., Jonas, L., Meibohm, A., Chiou, R., Deutsch, P., Emini, E. and Chodakewitz, J. 'Potent and sustained antiretroviral activity of indinavir in combination with zidovudine and lamivudine', In: Abstracts of the 3rd Conference on Retroviruses and Opportunistic Infections, Washington, D.C., 28 January to 1 February 1996 (abstract LB7).
61. Markowitz, M., Cao, Y., Vesanen, M., Tala, A., Nixon, D., Hurley, A., O'Donovan, R., Racz, P., Tenner-Racz, K. and Ho, D.D. 'Recent HIV infection treated with AZT,3TC, and a potent protease inhibitor', In: Abstracts of the 4th conference on Retroviruses and Opportunistic Infections, Washington, D.C., 22–26 January 1997 (abstract LB8).
62. Pantaleo, G., Menzo, S., Vaccarezza, M., Graziosi, C., Cohen, O. J., Demarest, J. F., Montefiori D., Orenstein, J. M., Fox, C., Schrager, L. K. *et al.* 'Studies in subjects with long-term non-progressive human immunodeficiency virus infection', *New England Journal of Medicine*, **332**, 209–216 (1995).
63. Peters, B. S., Coleman, D. G., McGuiness, O., Pinching, A. J., Wadsworth, M. J., and Harris, J. R. 'Changing disease patterns in patients with AIDS in a referral center in the United Kingdom: the changing face of AIDS', *British Medical Journal*, **302**, 203–207 (1991).
64. Lederman, M., Connick, E., Landay, A., Kessler, H., Kuritzkes, D., Rousseau, F. and Spritzler, J. 'Partial immune reconstitution after 12 weeks of HAART. Preliminary results of ACTG315', In: Abstracts of the 4th Conference on Retroviruses and Opportunistic Infections, Washington, D.C., 22–26 January 1997 (abstract LB13).
65. Phillips, P., Zala, C., Rouleau, D. and Montaner, J. S. 'Mycobacterial lymphadenitis: can highly active antiretroviral therapy unmask subclinical infection?' In: abstracts of the 4th National Conference on Retroviruses and Opportunistic Infections, Washington, DC, 22–26 January 1997 (abstract 351).
66. Race, E., Adelson-Mitty, J., Barlam, T. and Japour, A. 'Focal inflammatory lymphadenitis (FIL) and fever following initiation of protease inhibitor in patients with advanced HIV-1 disease' In: Abstracts of the 4th National Conference on Retroviruses and Opportunistic Infections, Washington, DC, 22–26 January 1997.
67. Katzenstein, D. A., Hammer, S. H., Hughes, M. D., Gundacker, H., Jackson, J. B., Fiscus, S., Rasheed, S., Elbeik, T., Reichman, R., Japour, A., Merigan, T. C. and Hirsch, M. S. 'The relationship of virologic and immunologic markers to clinical outcomes after nucleoside therapy in HIV-infected adults with 200 to 500 CD4 cells per cubic millimeter', *New England Journal of Medicine*, **335**, 1091–1098 (1996).
68. Kleinbaum, D. G., Kupper, L. L. and Morganstern, H. *Epidemiologic Research: Principles and Quantitative Methods*, Van Nostrand Reinhold, New York, 1982.
69. Conway, B., Montaner, J. S. B., Cooper, D., Vella, S., Reiss, P., Lange, J., Harris, M., Wainberg, M., Kwok, S., Sninsky, J., Hall, D., Myers, M. and the INCAS study group. 'Randomized double-blind one year study of the immunologic and virologic effects of nevirapine, didanosine and zidovudine combinations among antiretroviral naïve AIDS-free patients with CD4 200–600', *AIDS*, **10**, S15 (1996).
70. Mulder, J., Resnick, R., Saget, B., Scheibel, S., Herman, S., Payne, H., Harrigan, R., and Kwok, S. 'A rapid and simple method for extracting human immunodeficiency virus type 1 from plasma: enhanced sensitivity', *Journal of Clinical Microbiology*, **35**, 1278–1280 (1997).
71. Conway, B., Shillington, A., Fransen, S., Sninsky, J., Kwok, S. and Montaner, J. S. G. 'Application of Ultra-Sensitive PCR to the analysis of plasma viral load testing in clinical trials', In: Abstracts of the 4th Conference on Retroviruses and Opportunistic Infections, Washington, D.C., 22–26 January 1997 (abstract 629).