Analyzing Quality of Life (QOL) Endpoints in Clinical Trials via the SAS System

Jeff A. Sloan, Paul J. Novotny, Charles L. Loprinzi, Mayo Clinic

Jeff A. Sloan, 440 Plummer Building, Mayo Clinic Rochester, MN 55905

Key Words: Quality of life, Oncology, Clinical trials, Measurement, Reliability, Quality-Adjusted Life Years (QALY), Time Without Symptoms and Toxicity (TWiST)

ABSTRACT

The cure for many chronic diseases, such as cancer and AIDS, remains elusive. Clinical research increasingly has focused more on maximizing patient QOL because improvements in survival and response to treatment have become more difficult to achieve. Patient quality of life is rapidly becoming a standard endpoint in clinical treatment trials for such diseases. Analytical techniques for QOL endpoints differ from those used for the more traditional endpoints of survival and response to treatment. We will summarize the "state of the art" for the analysis of these psychosocial outcome measures and provide the requisite SAS algorithms. An overview of the technical background will precede a presentation of SAS macros for QOL tool performance assessment (reliability and validity) comparison among competing instruments (Bland and Altman, 1996), Q-TWiST analysis (Gelber, Goldhirsch and Cole, 1996), power analysis (Cohen, 1988), longitudinal analysis (Diggle, Liang and Zegher, 1994) and handling the problems of missing data in longitudinal studies. Example studies are drawn from Mayo/North Central Cancer Treatment Group experiences.

QOL IN HEALTH CARE RESEARCH

The purpose of this review article is to give the reader a broad overview of various statistical techniques that are presently being applied to QOL endpoints in health care research. The discussion will center largely in oncology research. The methodology, however, applies readily to other conditions, notably AIDS, in which the focus of the clinical effort is not so much to provide a cure but instead to enhance the quality of the time a patient has available.

The impetus for the recent upsurge in QOL research activity can be traced back to publications regarding the "war on cancer" declared by President Nixon . Several publications subsequently reported that the war on cancer was not being won, in terms of incidence and mortality rates (Beardsley, 1994 for example). An oft-quoted passage in a seminal article in the New England Journal of Medicine (1987) tacitly summarized the state of cancer research in saying that "when the cure remains elusive it is time to focus on the patient not the disease" (Tannock, 1997).

The amount of literature detailing QOL research has expanded enormously over the past five years (Sloan et al, 1998a). It is impossible to summarize the literature sufficiently in the space available here. Instead, we refer the reader to a small subset of literature that collectively would form an overview for the status of QOL research in 1997. If the reader could only have one book on QOL research it would have to be the tome of Spilker (1996) which provides a wealth of every aspect of QOL research. The book by Dimsdale and Baum (1995) would also serve the reader well in a concise format. There is a compendium for QOL literature and copies of tools on the worldwide web put together by Marcell Tamburini et al. at *http://www.glam.com/ql/url.htm*.

There are a number of key journal resources that the QOL researcher may find useful. The August 1997 issue of Controlled Clinical Trials was almost entirely dedicated to QOL research. In 1996, the American Society for Clinical Oncology (ASCO) published guidelines for incorporation of QOL endpoints into clinical trials by stating that QOL is secondary only to survival in importance as an endpoint and was more important than tumor response. In 1996, the National Cancer Institute published a monograph summarizing the use of QOL in Cancer Clinical Trials (Klausner et al, 1996). The goals of this document were "a) to define elements of QOL that are relevant to clinical decision-making and serve as endpoints in clinical trials, b) to evaluate currently available instruments...c) to identify site-specific questions of high priority and d) to examine...integration of findings from ....QOL measurements.

There is no generally agreed upon approach to QOL measurement or analysis (Aaronson, 1991; Cella, 1996; Sloan et al, 1998a 1998b). Leplege and Hunt (1997) argue that much of the QOL research to date has focused on what they term "subjective health status" rather than true QOL. They further argue that one cannot accurately measure the quality of one's life. While this may be merely a minor discrepancy in definition, the argument demonstrates how muddled the field of QOL research is at the present. Contradictory literature leaves the clinical researchers at a loss in regard to guidelines for the experimental design process. A first step is to put together a synopsis of the available literature in the hopes of finding some common signals among the noise. We have compiled a 30 page bibliography of major QOL references published in recent years (available from the first author).

WHAT EXACTLY IS QOL?

The lack of a generally accepted concise definition of QOL has been one of the major stumbling blocks in its development. The WHO definition is simply that it is "the state of complete well-being" (Spilker, 1996). This vague definition has been compartmentalized by a number of authors into five dimensions: physical, emotional, social, economic and spiritual. Operationally, QOL has been defined in clinical trials as a buzzword representing anything other than response or survival. Spilker (1996) presents a complex theoretical framework model of how the various aspects of QOL interact with clinical outcomes to produce a measure of the patient's overall sense of well-being (QOL).

Numerous authors have discussed methods for deciding upon appropriate QOL tools to use (Cella, 1996, Aaronson, 1993, Sloan et al, 1998b are just a few examples). There exists a wide variety of tools available, but there is no uniformly accepted optimal tool or approach (Cella, 1996). Even the most seasoned veteran among these tools is little more than a decade away from its original inception. There is no agreed upon measurement approach (Likert, visual analogue, yes/no, etc.). There are no accepted guidelines for timing, either in terms of when the patient should be measured or how often. There is agreement that it is useful to use a global measure of QOL supplemented by domain-specific tools for a more complete description of QOL.

QOL TOOL SELECTION

The basic recipe for deciding which QOL tools to include in a clinical trial is to 1) define the intangible constructs which are likely to be affected in the trial (e.g., pain); 2) perform a structural decomposition of the intangible constructs so as to describe the specific characteristics of each construct to be measured (e.g. pain location, intensity, duration, description); 3) produce operational definitions for each of the components (e.g. pain location will be operationalized by major body part, pain duration by a measure of how long it takes for the patient to be able to get out of bed again and return to normal functioning) and; 4) construct measurement items corresponding to each definition (e.g. list of adjectives to be circled to describe the pain, picture for pain location and Likert scale for pain intensity). Answering these questions guide the process of choosing and/or developing trial-specific QOL components.

RELIABILITY AND VALIDITY

In general, the researcher wants to have a QOL tool that provides a measurement consistently across people and time (referred to as reliability) and that presents a score for the construct that it is purported to measure.

There are too many forms of measurement instrument validity to allow for a detailed discussion of the methods here. Many of the forms of validity are handled by standard SAS procedures. For example, criterion validity (does the QOL scale correlate with measures that theoretically it should such as hematologic data) is readily handled by PROC CORR's simple statistics. Predictive validity (does the scale have any power to predict events it should effect like survival) are ably handled by PROC REG for complete data or PROC LIFETEST for censored data. T-tests and ANOVAS, routinely produced by PROC NPAR1WAY or PROC GLM, can be used to confirm that different patient subpopulations score differently as they should if the theoretical framework is correct.

Testing reliability is akin to asking the question "If I were to use this tool under the same conditions on the same person would I get the same score?". Test-retest reliability is easily accomplished by PROC CORR by examining the intrapatient correlation between timepoints with associated scatterplots. The methods of Bland and Altman (1996), which are applied to choosing among competing QOL instruments later in this paper, also apply to the examination of test-retest reliability.

The other primary form of reliability, namely Internal consistency, however, is typically measured by Cronbach's alpha coefficient which is produced as an option within PROC CORR. Unfortunately, Cronbach's alpha has some serious mathematical flaws including the possibility of inadmissible negative values for the coefficient (Sloan and Yeung, 1997). Confidence intervals and hypothesis testing and the probability of a negative value for Cronbach's alpha were implemented in SAS and the ALFAINF code is available from the authors. The confidence intervals produced typically indicate a very narrow range of likely values for a given experiment, even for a study of modest sample size. The macro ALFAPLOT produces the distribution of Cronbach's alpha for a given experimental design and typically indicates that the distribution is extremely leptokurtotic and resembles a mirror image of an F distribution. The mathematical details of this work are contained in Sloan and Yeung,1997.

GENERAL ANALYSIS CONSIDERATIONS

In essence, from a statistical standpoint, QOL may be considered as an endpoint itself or as a covariate that may be confounded with or cause a modification of the true effect of a therapeutic treatment on survival or tumor response. This broad definition places QOL in the appropriate context for statistical analysis as usually having to be considered in concert with more traditional and objective endpoints.

We have found it useful to transform all QOL scores into percentages of the theoretical range of the instrument. This provides a unitless score of 0-100% which can be used for comparison across tools or studies. For example, if a QOL tool is comprised of a summative scale involving k separate questions, each of which is scored with a minimum value of MIN and a maximum value of MAX, the percentage score Y is produced from the original scale score X by the simple formula:

Y = [X - (k*MIN)]/[k*(MAX-MIN)]*100%.

Performing this transformation on each QOL total score and subscale allows for a ready interpretation of where the individual is relative to the potential range of the instrument used. Hence a transformed score of 57 means that an individual is roughly halfway between the minimum and maximum possible score. Similarly comparing across means between two instruments can be made directly. A difference of 10 units across treatment groups does accurately reflect that the patients are 10% apart along the potential range of the instruments. Analytically, the statistical methods required are no different than what would be used for the raw scale scores.

MISSING DATA

Missing data is one of the more difficult problems involved in QOL measurement. Patients will omit questions and complete questionnaires even in the most closely monitored study. There is no satisfactory method for addressing the issue although there are several alternatives. One may, for example, prorate the scale using the information that was provided by the patient. This assumes that what was missed is the same as what was provided which may or may not be tenable. One may throw out the cases with missing data and perform a complete cases analysis, but this should only be used as a validation of alternative approaches. The loss of data experienced by using only complete cases typically has serious ramifications in terms of the power of the subsequent statistical procedures.

Imputation is also an option, and there are many approaches suggested in the literature. One may insert the average value of data obtained for each patient (referred to as the AVCF method). Alternatively, especially in studies where much of the missing data is due to attrition or censoring, the last value may be carried forward to the end of the study or the minimum value carried forward (the LVCF and MVCF methods respectively). The Generalized Estimating Equations (GEE) of Liang and Zegher (1996) have also been used to impute missing data in longitudinal studies.

The methods for imputing missing data are case-specific and need to be used with care if the amount of missing data is more than 5%. All approaches involve some very strong assumptions about the mechanism by which the data came to be missing. One solution is to use a number of approaches to perform a sensitivity analysis. The SAS code for implementing these approaches is quite straight forward and is typically accomplished in the DATA step. Example code is available from the authors.

POWER ANALYSIS

Power and effect size calculations for QOL endpoints is difficult because of the "soft" nature of the construct under study. The optimal approach would be to incorporate effect sizes from the literature, but typically such information is either lacking or not directly relevant to the study under consideration. A key component that is missing from the literature for most tools is the minimum clinically significant difference.

We have developed a framework for estimating effect sizes for the purposes of study design power calculations. The method is adapted from an approach due to Cohen (1988) which involves classifying effect sizes as either small, moderate or large as a function of the number of standard deviations that can be detected by a given test. We also make use of the empirical rule that indicates the standard deviation for a variable is roughly one sixth of the range. This approach produces relatively conservative estimates of the power available for a given effect size. There is inherent measurement error in the measurement of QOL, however, so it is prudent to not accept as clinically relevant any difference in QOL value unless it represents a considerable change. In terms of the transformed theoretical range variables these can be used to make general statements across studies. For example, if we consider the range of the percentage scores (0-100) as roughly equivalent to 6 standard deviations then an estimate of the tool's standard deviation is 16%. Cohen indicates that for a two sample t-test, small moderate and large effect sizes correspond to 0.2, 0.5 and 0.8 standard deviations. This would mean that small moderate and large effect sizes for a two-sample t-test on the transformed QOL scores would be 3.2%, 8% and 12.8% respectively. These provide a context for the examination of sample size calculations and can be used to test the veracity of the small, moderate and large effect sizes for the particular trial. Our experience has been that if a tool does not move by 10% of its theoretical range, the difference is not clinically important. This corresponds to a moderately large effect size by Cohen's vernacular. This is a comfortable level for many researchers who consider QOL measures to be soft endpoints fraught with potentially large inherent measurement error.

Figure 1 contains required sample size and power curves for the two-sample t-test varying the effect size present, from small to large, in terms of average QOL score. The graphs are easily produced by SAS/GRAPH and the QOLPOWER2SAM macro code is available from the authors. It is important to note that this approach is tool-independent. In other words, the SAS code produces sample size requirements to detect a small, medium or large effect size for any QOL tool that has been transformed onto a 0-100 theoretical range. If the standard deviation for a given tool is well documented in the literature, then our approach may be overly conservative due to an overestimation of the tool's standard deviation. As previously mentioned, however, estimates for the standard deviations of QOL tools is relatively sparse at present in the literature and typically are reported for specific patient populations with limited generalizability.

Q-TWiST ANALYSIS

The Q-TWiST, "Quality-adjusted Time Without Symptoms of disease and Toxicity", method is a relatively recent development intended to incorporate quality of life information into survival analysis so that survival may be expressed as "Quality-adjusted Life Years or QALY's (Gelber et al, 1993). To date, Q-TWiST has not been extensively used in medical research perhaps in part due to the fact that convenient software is not available.

The basic idea is that the overall survival (OS) is divided into three parts: time with toxicity (TOX), Time Without Symptoms and Toxicity (TWiST) and time after the progression or relapse (REL). Relative weightings are then assigned to each of the above health states by considering how valuable a day with toxicity or after relapse is to each patient. TOX and REL can be weighted by utility coefficients referred to notationally as Ut and Ur , which take values between 0 and 1 depending upon the individual's circumstance. Using these weightings, Q-TWiST is defined as

Q-TWiST = Ut x TOX + TWiST + Ur x REL

We use an example dataset from an NCCTG clinical trial regarding treatment of melanoma with either interferon or standard methods.

The original S-Plus code for the Q-TWiST analysis was generously provided by Richard Gelber. We translated the code for the SAS system, partially because it had not yet been done (Sheri Gelber, personal communication ASCO, 1997), and partially because we wanted to explore the development/use of some other graphics tools. Figure 2 shows Gelber's (1993) partitioned survival plots for interferon-2a produced by SAS/GRAPH. Survival curves are shown for overall survival, progression and toxicity (from top to bottom), with the shaded areas indicating the time to toxicity (TOX), TWiST and REL, the sum of which is the overall survival time. The plot shows that a substantial amount of toxicity was present for patients receiving interferon.

Our SAS code also produces a partitioned scatter plot of utility constants Ut vs. Ur to show where each treatment was favored in terms of mean survival similar to those produced by Gelber et al (not shown). This graph is useful but difficult to interpret in terms of precise survival benefit. Further output is produced via SAS PROC G3D and ANNOTATE to show in another fashion the difference in mean and median Q-TWiST between treatment groups at various levels of Ut and Ur.

A NEW GRAPHIC: THE QALY PLOT

Figure 3 illustrates a new graphic for QALY analysis, the QALY plot. The goal of this graphic is to portray a complete QALY analysis on one plot. As such, the graphic is not one that can be deciphered instantly and requires some explanation. However, the plot is rich in information, and we have been surprised at the number of parsimonious comparative data that can be drawn from the graphic.

The graphic presents a comparison of Kaplan-Meier estimates for median QALY between two treatment groups for all combinations of utility attributions Ut and Ur. The individual points on the plot present QALY for a given Ut, Ur utility penalty combination. At the bottom left, the "0" under the column labelled as Ur=0.0 represents the median QALY comparison if we totally penalize (give no credit for) days that have toxicity or days after progression. This comparison then is only based on TWiST days (without toxicity and before progression). In this plot, under these utilities interferon-2a has a median survival of 250 days, compared to 725 days survival on standard treatment. The fact that the columns are almost perfectly vertical indicates that changing Ut has almost no effect on median QALY in the control group.

Altering the penalty (Ur) for days after progression does not necessarily have a proportional impact on QALY. Notice that the gaps are wider between the second and third (Ur=0.1, 0.2) and fourth and fifth (Ur=0.3,0.4) columns of numbers than any others. This indicates that choice of penalty amount (Ur) impacts QALY to a varying degree. At the other end of the penalty spectrum, the top right "1" point under the column Ur=1.0 gives medium QALY for the case where we assign no penalty to either toxicity or progression (Ur,Ut=1.0). Hence this point actually gives median OS with no discounting for toxicity or progression. Here we see that median QALY for interferon-2a is around 2100 days, compared to 1630 days for the standard treatment. Each point plotted between these two extremes gives the median QALY comparison for all combinations of penalization (Ut and Ur) in increments of 0.1.

The precise characters plotted warrant further explanation. Each vertical "column" of digits (01234567891) represents the median QALY comparison for a given penalty (0%-100%) of each day after disease progression (Ur). Keep in mind that the greater the penalty, the lower the utility value is for each day in that situation. For example at the "middle" combination of discounting days with toxicity or after disease progression by 50% (Ut, Ur both 0.5) both treatments give around 1300 days of QALY survival. The solid line on the plot represents equality of median QALY between the two treatment groups. Either treatment arm may provide a greater median QALY depending upon how much you penalize OS by assigning utilities to days with toxicity and after progression. Points above the solid line indicate that the median QALY is greater for interferon-2a than for the standard treatment while points below the solid line indicate that the standard regimen has larger median QALY.

If we were to produce graphic for a study that had one treatment with uniformly better QALY irrespective of the amount of toxicity or progression penalty, then all characters plotted would be on one side of the equality line. Thus this aspect of the graph provides an ad hoc sign test for median QALY equality. Sweeping across the graphic from left to right, between the two extreme points of TWiST and OS respectively, we see that increasing the penalty for toxicity and, to a much lesser extent, for disease progression produce QALY that favors the standard treatment. The magnitude of adjustment is the same at the two extremes of Ur (0.0 and 1.0) as 3 of the points are above and below the line of equality respectively in each case. This also indicates that interferon-2a has more toxicity, since the plotted points go from the largest benefit for interferon-2a for the case of no penalty for toxicity to the largest benefit for the standard treatment when the total penalty for toxicity is applied.

Finally the dotted line indicates the value of median QALY for the interferon-2a group that would be required to a statistically significant difference at the 0.05 level for each value of median QALY in the control group, based on the logrank test for the given sample size of the study and the number of deaths in the control group. This part of the SAS code was achieved by using PROC LIFETEST to obtain the critical values. In essence the upper and lower bounds are equivalent to a 90% confidence interval for the median difference in QALY. Only with severe penalizing of both toxicity and progression does interferon-2a provide statistically inferior QALY. Hence, we can say that in terms of QALY, the two treatments have the same survival benefit. This allows for a comprehensive examination of QALY for all possible utility values. If a large number of points had fallen outside the dotted boundary, then we could discuss how toxicity and progression impact the survival comparison.

A final visual guide to the QALY analysis produced by our SAS code takes a subset of the QALY utility combinations and plots the relevant survival curves in comparison to the OS curves. PROC GPLOT and ANNOTATE once again are used. For our example, the comparison of Ut, Ur both set to zero was of interest. These are plotted in figure 4 showing that while there was no difference in OS between the two treatments, the curves for TWiST show a substantial differential in favor of the standard treatment.

COMPARING COMPETING QOL TOOLS

In this section, we present results of SAS code developed to carry out the methods of Bland and Altman (1996) as well as Poon et al. (1989) for comparing the relative accuracy and precision of competing QOL tools when both are administered to each patient within a study. The code produces a substantial amount of tabular and graphical output using a wide array of SAS procedures.

We use as an example an NCCTG clinical trial developed to compare and contrast the utility of four different instruments (Sloan et al, 1998a). Each tool provides a single score representing overall QOL ranging from a very simple, one-item instrument to more detailed instruments. Each patient and physician separately completed the single item Spitzer Uniscale (UNISCALE) rating the patient's QOL at baseline and monthly (Spitzer, 1981). One hundred and twenty-eight patients were randomly assigned to complete, in addition, either the 22-item Functional Living Index-Cancer (FLIC), the 5-item Spitzer QOL index (QLI), a single item graduated picture face scale (PICT), or nothing else (Spitzer, 1981, Schipper et al, 1986) .

A difference of 10% or more of the theoretical range between tools was operationally defined as clinically significant. QOL scores obtained by the various instruments are compared in a number of ways by SAS procedures. Regression analysis (PROC REG) and subsequent equality of regression line slopes testing for similar profiles over time between tool scores within each treatment group, are carried out via standard independent t-test and Wilcoxon procedures (PROC NPAR1WAY). Area under the curves (AUC) of QOL scores over time are calculated for each patient and the average intrapatient difference for each tool is compared via paired t-test or Wilcoxon test, dependent upon the result of Shapiro-Wilk normality testing from PROC UNIVARIATE. Selected results from each procedure were output to SAS datasets and then collected into tabular format for presentation (Sloan et al, 1998a). All of this work was combined into a single macro called QOLANAL.

A second set of SAS code called QOLPLOT was created to produce a series of graphical representations. Scores for baseline and first month of treatment along with the 10% clinically significant error bounds are produced comparing each tool with the patient UNISCALE scores (not shown).

Bland and Altman plots involving differences between intrapatient pairs of QOL tools scores are plotted via PROC GPLOT versus the average level of these same pairs of QOL scores (with the plus or minus 10% operationalized error bounds). Figure 5 presents an example plot comparing the UNISCALE score provided by the patient and physician. If between any pair of tools differences are due merely to chance then the plot should appear as a random scattering within the error bounds. Figure 5 displays another marked pattern with the patient-physician UNISCALE differences showing relatively good agreement between patients and physicians (most points within the error bounds) for very high or very low QOL. The marked discrepancy between patients and physicians is seen for moderate QOL (35-75%of maximum) scores with the majority of points indicating a physician underestimate.

LONGITUDINAL ANALYSIS

Another perplexing challenge to analyzing QOL endpoints is to investigate the effect of time on the QOL a patient experiences. Repeated measures analysis of variance modelling has been the primary method of choice, including profile analysis (Morrison, 1976, pp. 205-216). These procedures involve testing three separate hypotheses regarding the "profiles" of the patient QOL scores over time between treatment groups. The first hypothesis is referred to as one dealing with "parallelism" and tests whether the two treatment groups move in tandem over time. In essence this is a test for time by treatment group interaction via analysis of variance models. The second hypothesis is referred to as the "levels" test and is a direct comparison of the average scores for the treatment groups, given that the two groups have been demonstrated to move in parallel from the results of the first hypothesis test. The final test is one of "flatness" and is conditional on the results of the preceding hypothesis tests so that an overall test for an effect due to time can be carried out. The SAS code for this work was gathered into a single macro QOLPROFILE and includes the standard multivariate analysis of variance test for the equality of mean vectors.

More recent work involving the GEE model have application to longitudinal QOL studies (Diggle, Liang and Zegher, 1994). Code for producing the analysis is available under the macro GEEQOL. Primarily we have found the GEE models as useful sensitivity analysis tools for testing variance structure assumptions and to confirm results obtained from the simpler procedures described earlier.

CONCLUSIONS

The purpose of this paper was to review the "state of the art" of QOL research in the context of clinical trials and to share some ideas and experiences of using the SAS system to overcome some of the hurdles involved therein. There is much more work to be done in the development of QOL research and the versatility of the SAS system is key to the ease with which the above analytical procedures was implemented.

ACKNOWLEDGEMENTS

The material involved herein represents countless hours of programming by numerous personnel within the Mayo Cancer Center Statistics Unit. The authors would like to express sincere thanks to Pamela Atherton Skaff, Cristine Allmer and Lei Zhang for programming work above and beyond the call of duty.

The QOL bibliography, references and SAS code or macros are available from either of the first two authors at the address above or via email at jsloan@mayo.edu or novotny@mayo.edu.

REFERENCES

Aaronson Neil K. Methodological Issues in Assessing the Quality of Life of Cancer Patients. Cancer 1991(suppl); 67:844-850.

American Society for Clinical Oncology. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. Journal of Clinical Oncology 14:671-679, 1996.3.

Beardsley T. A war not won. Scientific American. January: 130-138,1994.4.

Cella DF. Quality of life outcomes: measurement and validation. Oncology 11(S): 233-246, 1996.

Cohen J. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1988.

Gelber Richard D Goldhirsch Aron, Cole Bernard F. Evaluation of effectiveness: Q-TWiST. Cancer Treatment Reviews 1993; 19 (Supplement A): 73-84.

Klausner RD Hubbard SM Chappell J. Quality of Life in Clinical Cancer Trials. Journal of the NCI Monograph 20, 1996.2.

Leplege A Hunt S. The problem of quality of life in medicine. JAMA 278: 47-50, 1997.

Diggle PJ Liang KY Zegher SL. Analysis of longitudinal data. Clarendon Press, 1994

Dimsdale and Baum. Quality of Life in Behavioral Medicine Research. Hillsdale, NJ: Lawrence Erlbaum Assoc.; 1995.

Morrison DF. Multivariate Statistical Methods. McGraw-Hill, New York, 1976.

National Cancer Institute. Incorporating Quality of Life into Oncology Clinical Trials. NCI Monograph #20, Baltimore Maryland, 1996.

Poon MA O'Connell MJ Moertel CG et al. Biochemical modulation of fluorouracil: evidence of significant improvement of survival and quality of life in patients with advanced colorectal cancer. Journal of Clinical Oncology 7:1407-1418,1989.

Schipper H Clinch J McMurray A Levitt M. Measuring the quality of life of cancer patients: the functional living index-cancer: development and validation. Journal of Clinical Oncology 2:472-483,1986.

Sloan JA Yeung AY. Reformulations of Cronbach's alpha coefficient. Proceedings of the Biometrics Section of the American Statistical Association 367-371,1997.

Sloan JA Loprinzi CL Kuross SA Miser AW O'Fallon JR Mahoney MR Heid IM Bretscher ME Vaught NL. (1998a, in press). A randomized comparison of four tools measuring overall quality of life in patients with advanced cancer. Journal of Clinical Oncology.

Sloan JA O'Fallon JR Suman VJ Sargent DJ. Incorporating quality of life measurement into oncology clinical trials. Proceedings of the American Statistical Association (1998b, to appear).

Spilker B. Quality of Life pharmacoeconomics in Clinical Trials. Lippincott-Raven, 1996.

Spitzer W Dobson AJ Hall J Chesterman E Levi J Shepherd R, Battista N Catchlove B. Measuring the quality of life of cancer patients. A concise QL-index for use by physicians. Journal of Chronic Disease 34:585-597, 1981.

Tannock IF. Treating the patient, not just the cancer. N Engl J Med 1987; 317: 1534-1535.

Figure 1 Power Example for Varying n

n = observations per sample

one-sided two sample t-test

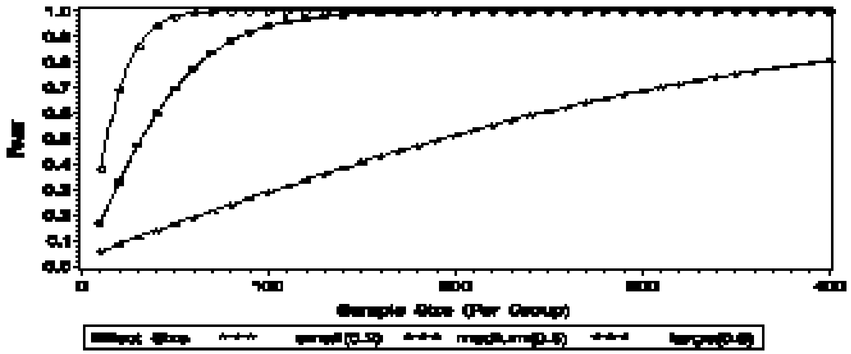
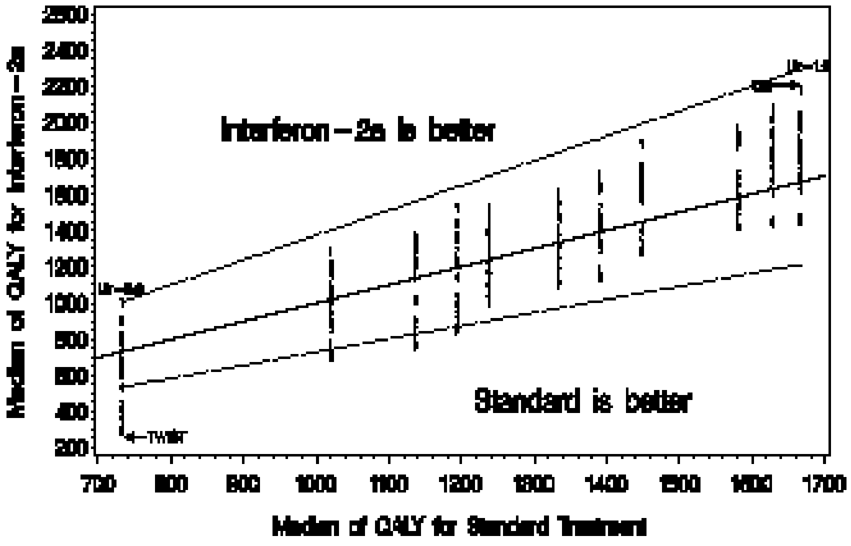


Figure 2 Quality-Adjusted Median Survival (KM estimates)

Interferon-2a vs. standard treatment

for varying utility of days with toxicity (Ut) and days after progression (Up)



Character plotted represents the utility value of a day with toxicity (Ut) from 0.0 to 1.0. Each column of characters (0123456789T) represents QALY for a given value of a day after progression (Up) from 0.0 to 1.0

dotted lines represent the critical value for one-sided logrank test (5% type I error)

Figure 3 Partitioned Survival Curves With Ut=Ur=0.5
for Interferon-2a

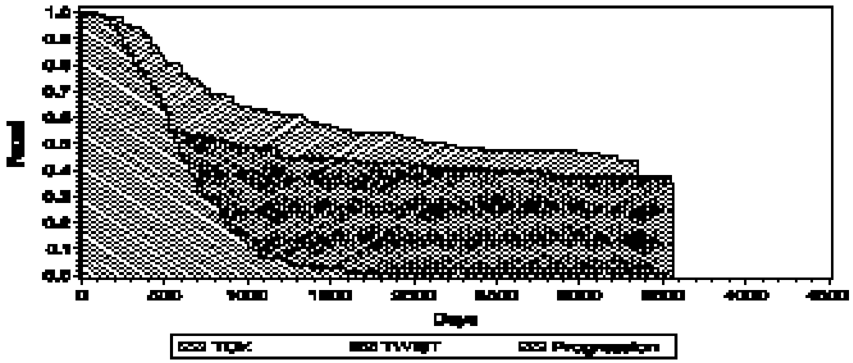


Figure 4 Overall Survival and TWIST
for Interferon-2a and Standard Treatment

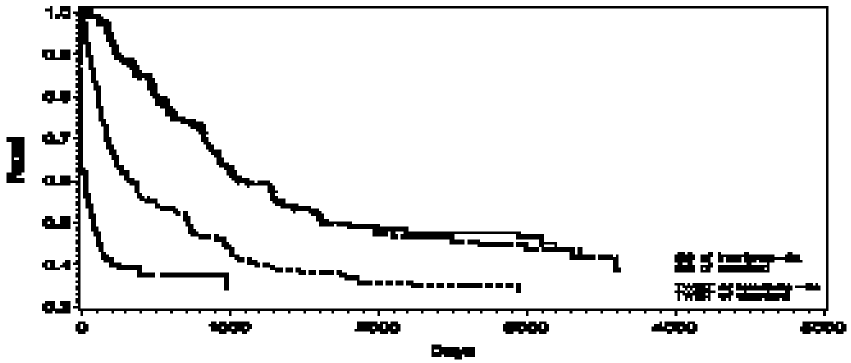
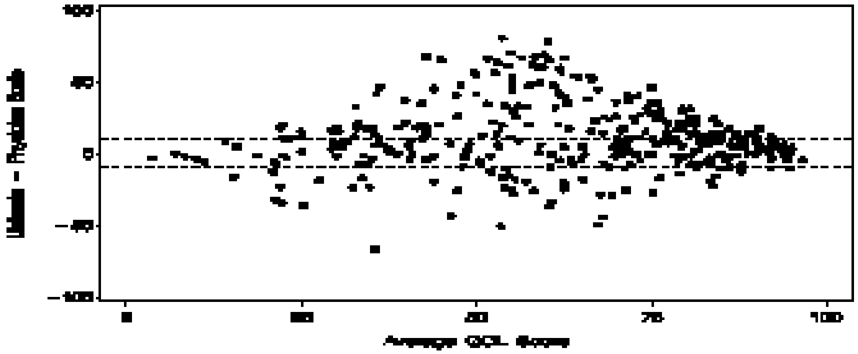


Figure 5 QOL Score Differences vs. Average for Each Patient
All Months


Average QOL Score