REVIEW Jpn. J. Cancer Res. 84, 929–937, September 1993

Guidelines for Analysis and Reporting of Clinical Trials in Oncology

Steven Piantadosi,[1] Nagahiro Saijo[2] and Tomohide Tamura[2]

[1]*Oncology Biostatistics, Johns Hopkins Oncology Center, 550 North Broadway, Suite 1103, Baltimore, Maryland 21205, USA and* [2]*Pharmacology Division and Medical Oncology, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104, Japan*

1. Introduction

This article is the second of two papers for practicing oncologists outlining fundamental considerations in the design, analysis, and reporting of clinical trials. This paper will follow the terminology and concepts in our earlier publication.[1)] As before, we discuss only studies in which the treatment or exposure, patient follow-up and data collection, and analysis are all controlled by the investigator, i.e., clinical trials. Among clinical trial methodologists, there is a wide variety of methods for analysis and reporting, and there is considerable variation in the content and level of detail in published reports of trial results. We believe that, because of the need to develop new cancer treatments in a logical fashion, there are a few basic guidelines for analysis and reporting that clinicians will find helpful.

As was the case with our first paper, this article is not intended to cover all that the oncologist should know about either biostatistics or clinical trials analysis. Our focus here is on broad concepts rather than methodological details. Most clinical trials will require analysis and reporting with the help of an expert trialist. However, the clinician can improve the inferences from such reports by understanding the basic statistical concepts and approaches to the analysis. For the more mathematically oriented reader, there are a number of technical sources that can provide more details. These include the analysis of prognostic factors,[2, 3)] survival models,[4)] general surveys,[5)] review articles,[6)] and binary data.[7)] The reader interested in a more comprehensive or technical analytic background is referred to these sources. For details on reporting recommendations and related topics, there are a few recent reviews.[8–11)] Discussions regarding the limitations of *P*-values are useful reading for clinicians.[12–22)]

To begin, we contrast "estimation" and "hypothesis testing." Then we review some basic concepts concerning the two commonest types of data that arise from clinical trials: proportions (e.g., response rates), and failure rates from time-to-event (survival) studies. We then discuss analysis and reporting concerns generally applicable to all clinical trials. Finally, we discuss some limitations of analysis and interpretation. There are many concerns about clinical trials, especially management, computer software, data management, interim reporting, and monitoring committees which have not been covered in these two articles. The reader is referred to general discussions in the references for these topics.

2. Estimation versus Hypothesis Testing

"Estimation" is the measuring of clinical effects (and confidence intervals) whereas "hypothesis testing" is the comparison of estimates with hypothetical values using probability statements (e.g., *P*-values). The approach we recommend to analysis and reporting follows estimation rather than hypothesis testing.[12–15)] Some editorial boards of medical journals have adopted general guidelines for reporting.[8, 11)] Similar guidelines have been proposed for statistical reporting[9, 10)] of clinical trials. The recommendations which follow are similar to those.

Measuring and reporting clinical effects and associated confidence intervals is more informative and useful than focusing attention on formal tests of statistical hypotheses and *P*-values. A simple example should make the difference clear. Suppose a clinical trial is performed comparing two treatments, A and B, and the major outcome is survival. When hypothesis testing is emphasized, investigators might report the results of a statistical test comparing treatments A and B and report "survival on treatment A is significantly longer than on B ($P<0.05$, logrank test)." Alternatively when emphasizing estimation, investigators might report "the estimated hazard ratio (A vs. B) for death was 2.0 (95% confidence limits 1.5–2.3)." In the first case, the reader is left only with a *P*-value to summarize the data whereas in the second case, the treatment difference is summarized more completely.

Although we have suggested some specific statistical methods and summaries for certain kinds of data, there are additional or alternative analytic procedures that need to be adopted in special cases and we do not seek to limit analyses or reports. However, the basic concepts and approaches outlined here should prove to be helpful both clinically and statistically for correctness, lack of bias, completeness, and consistency.

3. Data Yielding Proportions

For many research questions in phase II and III trials, the outcome for each patient is dichotomous, i.e., can only have two possible values. For example, a common

Table I. Number of Responding Patients on a Clinical Trial Comparing Treatment A versus B

		Treatment	
		A	B
Response	Yes	a	b
	No	c	d

dichotomous outcome variable often equals 1 for a "yes" (or response) and 0 for a "no" (or non-response). We often summarize such data by calculating the proportion of responders, P, or the odds of response $P/(1-P)$. Often, the association between a dichotomous response and a predictor variable is most conveniently summarized as an odds ratio, OR, which is used extensively in epidemiologic studies. For example, in Table I we see the results of a study comparing the response on two treatments. The odds ratio for response on treatment A vs. B is calculated as $OR=(a \cdot d)/(b \cdot c)$. We do not intend to fully explain the OR, its estimation, limitations, or use. Details of these are available in standard textbooks.[5, 7]

The OR is the measure of effect that clinicians will encounter most frequently in summarizing the effect of a prognostic factor on a dichotomous response. An OR close to 1.0 implies no effect, while values different from 1.0 indicate some effect of the prognostic factor on the probability of response. Confidence limits are usually placed on the estimated OR and a formal test comparing it to the null value, 1.0, is often performed yielding a P-value.

Although the outcome variable may be dichotomous, the predictor or prognostic variables may not be. For example, we might be interested in the association between age (measured on a continuum) and response (yes or no). In this circumstance, using estimation methods more sophisticated than those described above, the OR can be expressed as the change in odds of response per unit change in the prognostic variable. Thus, we might observe an OR of 1.10 per year of age, implying that the odds of response increases 10% with each additional year of age. The risk ratio is assumed to be the same for low ages as it is for high ages. To determine the effect of 5 years age difference, we would calculate effect $= 1.10^5 = 1.61$.

Many clinicians and some statisticians do not emphasize the estimated odds ratio(s), but proceed directly to statistical tests (P-values) in contingency tables. Omitting the OR estimate and confidence interval is an unfortunate and unnecessary error because P-values do not adequately summarize the relevant clinical effects. For example, suppose we are told that "increased age is associated with significantly increased risk, $P<0.01$." The effect of age on risk could be large or small. Examination of odds ratios and confidence intervals is much more informative.

Estimated odds ratios generalize easily, through the use of statistical models, to situations where two or more prognostic variables must be considered simultaneously. This method allows estimation of "adjusted" odds ratios, i.e., those for which the effects of other prognostic factors are controlled. Multiple logistic regression is the commonest method for accomplishing this and yields estimates of adjusted odds ratios, confidence limits, and P-values with the same interpretations as above.

4. Failure Time Data

Failure time or "survival" data are the most common type of longitudinal outcomes that arise in comparative cancer clinical trials. The analysis of failure time data requires methods and reasoning analogous to those for dichotomous response data. Because failures are not observed in all subjects under study (censoring), survival outcomes must be summarized by at least two observations on each subject. The first variable is the at-risk or event time and the second is a dichotomous variable which indicates whether the event time is a failure or a censored observation. Also because of censoring, survival data are often summarized as a cumulative distribution, i.e., the probability of surviving at least as long as a specified time. The lifetable and survival curve are products of this type of analysis.

Survival distributions can be described by one of three summaries: 1) the survival distribution (described above), which is most often used for graphical displays of data, 2) the frequency distribution of deaths (death density), which is not usually used by clinicians, or 3) the hazard[4] which is the most compact, natural, and clinically useful summary of risk. The hazard is defined as the probability of failure in the next instant of time, given that no failure has yet occurred. The hazard is also related to the slope of the survival curve and to the median survival. In many cases, the risk of failure is nearly constant over time (simple exponential survival). Even if the hazard changes with time, the "average" hazard is a useful summary.

Many clinicians are accustomed to thinking about differences between survival distributions in terms of differences in median survival. However, the hazard ratio is often a more informative summary of survival differences. Even when hazards change with time, the hazard ratio is often nearly constant. Another advantage is that the hazard ratio and confidence limits can always be estimated from survival data whereas estimating the median survival requires special methods and may not even be possible for some data. Other difficulties arise when trying to place confidence limits on the median.

Finally, the hazard ratio is "natural" because the prognostic effects from commonly used survival models are usually expressed in terms of relative hazards, analogous to odds ratios described above for dichotomous outcomes. In other words, models such as the proportional hazards model provide the clinician with estimates of hazard ratios between subsets of patients defined by levels of the prognostic variables. Generalization to more than one prognostic factor is straightforward, as with odds ratios.

5. *Basic Steps in Analysis*

Having described the two most common outcome measures in clinical trials, we now consider some details of analysis and reporting. The exact procedure for analyzing a clinical trial depends on the design and purposes of the study. For example, phase I trials might require pharmacologic modeling and estimation of physiological parameters in each patient to meet their objectives, whereas phase III trials usually require summaries of relative treatment effects and confidence intervals. Analyses for phase I and III studies seem to have very little in common. However, when we consider that both types of trials should inform us about the population being studied, the need for unbiased statistical estimation, and information that is of clinical utility to the medical community, much common ground is evident.

In this spirit, we offer the basic steps in analysis of clinical trials in Table II. We emphasize that these steps are conceptual and do not necessarily occur in chronological order. Also, some steps are only relevant to randomized or comparative trials.

5.1 *"Intention to Treat"*

A clinical trial is a test of treatment policy, not a test of treatment received. This is the "intention to treat" principle. It is unfortunate that investigators conducting clinical trials cannot guarantee that the patients who participate will finish or even receive the treatment assigned. Many factors contribute to patients failing to complete the intended therapy including unexpectedly severe toxicity, disease progression, unstated preference for another treatment, and a change of mind. However, once patients are selected for participation in a clinical trial, they represent the population defined by the eligibility criteria. If some participants are excluded from analysis on any basis other than eligibility, the trial results will no longer represent the population intended. In fact, the results will represent an unknown population, perhaps with characteristics and prognosis substantially different from that intended.

From the clinical perspective, when selecting a treatment for a new patient, the physician has no knowledge of whether or not the patient will complete the treatment intended. The physician is primarily interested in the probability that the treatment will benefit the patient. It is not helpful to know that the treatment might work depending on some events in the patient's future. Consequently, the physician should be most interested in clinical trial results that include all patients who were assigned to the therapy.

On the practical side, to be certain that the trial results closely reflect the effect of the treatment, the eligibility

Table II. Basic Steps of Analysis

1. Approach the analysis as a test of treatment policy, not a test of treatment received. This is the "intent to treat" principle.
2. Plan to include all patients registered/randomized regardless of post entry events.
3. Examine the data.
4. Describe the cohort.
5. Verify the effectiveness of randomization.
6. Estimate the effect of each prognostic factor on the major outcome (univariate analyses). One of these will be the treatment effect of the trial.
7. Using standard multivariate statistical methods or models (e.g., linear, logistic, or proportional hazards regression), re-estimate the treatment effect while adjusting for:
   a) statistically significantly imbalanced prognostic factors,
   b) strong or influential prognostic factors, whether imbalanced or not,
   c) any prognostic factor for which it is important to demonstrate convincing control.
8. Repeat steps 1–6 after excluding ineligible patients, i.e., those patients who are ineligible based on pre-entry criteria.
9. Consult the biostatistician concerning special methods to address specific clinical questions. Any analyses not protected by the randomization should correspond to clinical hypotheses stated as study objectives and use multivariate control of prognostic factors.
10. Cautiously conduct exploratory or hypothesis generating analyses:
    a) any analysis suggested by the data and not by hypothesis (*P*-values will be wrong),
    b) any analysis which excludes patients based on post-entry criteria (results will be biased),
    c) subset analyses (prognostic factors will not be controlled).

criteria should exclude patients with characteristics that might prevent them from completing the therapy. For example, if the therapy is lengthy, perhaps only good performance status patients should be eligible. If the treatment is highly toxic, only patients with normal function in major organ systems will be likely to complete the therapy.

Based on these considerations, the first and most reliable analysis will include all patients registered or randomized on the trial regardless of post entry events. This analysis is the intention to treat analysis. It is possible to exclude patients who were retrospectively found not to meet the eligibility criteria, i.e., those who were mistakenly placed on study, without creating bias. Ideally, such patients would not have been entered into the study because they were ineligible. Only eligibility or pre-entry criteria should be used to make such exclusions. If patients are excluded based on "evaluability" or other post-entry criteria, the possibility of bias becomes larger. This occurs because "evaluability" criteria are outcomes, no matter how reasonably defined clinically, and if we exclude subjects based on outcomes, the potential for bias is great.

5.2 Examine the Data

The first practical step in any analysis is to examine the data. This includes looking at formatted printouts and other simple tabulations that might highlight incorrect data values. Many problems in analyzing clinical trials can be prevented by correcting errors that become apparent when simply looking at the data. This is also a step that physicians, who are often very knowledgeable about the data, can perform. With the widespread use of computers to manage clinical information and automated analysis procedures, it is possible to produce results from clinical studies without carefully examining the data. Even a cursory examination of raw data by a technically knowledgeable person can detect many errors of importance to the analysis. For example, suppose we record serum calcium which is expected to be normal. Some of the kinds of errors that are amenable to detection by inspection include 1) incorrectly missing data (patient had level measured but not recorded in the database), 2) incorrect decimal points (80 recorded instead of 8.0), 3) failure to convert numerical codes for special values, (calcium becomes 99 instead of "missing"), 4) out of range or impermissible values (0.0 recorded instead of 8.0), 5) mis-labeled variables (age is mistaken for calcium and vice-versa), and 6) coding and recoding errors (0 should mean normal and 1 should mean abnormal, but values are reversed).

Inspection of the data is particularly important for small or single investigator studies in which the data management techniques are not subject to regular quality control procedures as might be the case in multi-institutional cooperative group studies. Errors in small studies can be particularly influential. Many times, small studies are recorded entirely on paper with transcription to a computer at a later time, creating an opportunity for errors. Other times, data are stored on computers using convenient but unsophisticated software such as spreadsheets rather than database management programs which permit validation and checking. Fortunately, the quantity of data from such small studies is often very amenable to checking by inspection.

Although the statistician can produce helpful charts, tables, and graphs of aggregate data, the additional insight of the clinical investigator is essential to correct some errors. Furthermore, the clinician will be reassured that all data are correct by personally inspecting all items. It is embarrassing, frustrating, and poor for morale to have to ask that a series of analyses be repeated because data errors were discovered late. After the statistician puts the data in analysis-ready form, it is often useful to pause for several days or weeks while the clinical investigators assure themselves that all the information is correct. As a consequence of this care, any subsequent unusual findings in the data can not be attributed to data error.

5.3 Describe the Cohort (Study Population)

All clinical trials are studies of particularly well defined and relatively small cohorts. Although the eligibility criteria define a population of particular interest, the patients actually accrued on a trial may differ as a result of chance or institutional characteristics. Investigators will want to describe the observed cohort, particularly with regard to important prognostic factors. Simple population measures and summary statistics usually suffice for this purpose. This process is also valuable in error checking.

5.4 Verify the Effectiveness of Randomization

In reports of many randomized studies, the first table presented is often intended to show the comparability of treatment groups. The lack of statistically significant differences between treatment groups dose not guarantee the absence of influential imbalances. It only demonstrates the effectiveness of randomization. Even so, this is important because investigators will have more confidence in the findings if imbalances are either absent or controlled in the analyses. Statistically non-significant imbalances in strong prognostic factors can influence treatment comparisons. This is discussed more completely in deciding when to adjust (below). Conversely, statistically significant imbalances are not necessarily influential — the imbalance may occur in an inconsequential factor. From a clinical perspective, the magnitude of the

difference between groups and the strength of the imbalanced factor are more important that the *P*-value.

5.5 Estimate Prognostic Effects

The portion of the analysis plan draws directly from the earlier discussion of odds and hazard ratios. The most useful clinical information is how the outcome (e.g., odds of response, or risk of death) changes with each unit change in the value of a prognostic factor. Statistical models such as logistic regression for dichotomous outcomes and survival regression for event times are likely to be of greatest use. Provided the assumptions of these models are met, they provide estimates of the appropriate relative risk parameter(s), confidence limits, and *P*-values.

Each prognostic factor can be studied independently using these models (univariate analyses). For a comparative clinical trial, the prognostic factor of most interest is likely to be treatment assignment. However, other factors may also be prognostic and perhaps also unequally distributed in the two treatment groups. For this reason, adjusted or multivariate analyses are often helpful.

5.6 Adjustments

Not all clinical trial statisticians agree on the need for adjusted analyses in properly designed clinical trials. However, we believe, as many investigators do, that it is often useful to learn about differences in estimated treatment effects before and after adjustment. Furthermore, observational studies in which randomization is not employed invariably are analyzed with adjustment procedures. The kinds of systematic errors that can arise in observational studies can arise by chance in clinical trials. This seems to provide a firm rationale for examining the results of adjusted analyses.

One of the principal advantages of the model based approach outlined above is its simple generalization to analyses using adjustment. Using the method in the analysis of a comparative clinical trial, investigators re-estimate the treatment effect while accounting for prognostic factors that meet one of three criteria:

a) statistically significantly imbalanced prognostic factors,

b) strong or influential prognostic factors, whether imbalanced or not,

c) to prove that a particular prognostic factor does not artificially create the treatment effect.

The underlying philosophy in adjusting in these circumstances is to be certain that the observed treatment effect is not due to imbalances in influential prognostic factors. Changes in relative risk parameters (more than changes in *P*-values) with adjustment are the effects of clinical interest.

5.7 Repeat Analyses

Repeat steps 1–6 after excluding ineligible patients, i.e., those patients who are ineligible based on pre-entry criteria.

5.8 Special Methods

A skilled trial methodologist should be consulted when special analyses are needed to address specific clinical questions. For example, in phase I studies, pharmacologic models and specialized parameter estimation may be needed to study pharmacokinetic end points. In prognostic factor studies, special relative risk regression models may be needed to investigate the effects of prognostic factors that change over time (time-dependent covariates). For comparative studies, analyses not protected by the randomization (e.g., subset analyses) should correspond to clinical hypotheses stated as study objectives and should use adjustment methods to control prognostic factors.

5.9 Data Exploration

Clinicians generally need very little encouragement to conduct exploratory analyses of their data. By exploratory analyses, we mean those which do not follow directly from the design of the experiment. Such analyses are neither automatically inappropriate not wrong. However, the conclusions derived from these analyses are often unreliable. Therefore, they should serve only to generate hypotheses to be tested more rigorously in the future. Reasons why exploratory analyses may be unreliable include the following.

a) A comparison suggested by the data and not by prior hypothesis is likely to have a type I error larger than the nominal *P*-value.

b) An analysis which excludes patients based on post-entry criteria (responses) will likely produce biased results.

c) Subset analyses are likely to be influenced by uncontrolled prognostic factors. Investigating large numbers of subsets can lead to "significant" differences purely by chance (i.e., inflated type I error).

By relying on estimation of clinical effects rather than unplanned tests of statistical hypotheses, errors resulting from these exploratory analyses might be decreased. Investigators might be less likely to misinterpret the results or to exaggerate their clinical utility. In any case, these types of exploratory analyses should never be the primary analysis of a clinical trial.

6. Basic Steps in Reporting

As with analysis, the most informative summaries and amount of detail to report from a clinical trial will depend largely on the nature of the clinical hypotheses

Table III. Basic Steps of Reporting

1. Report all clinically relevant descriptions of the cohort including patients who met the eligibility criteria but chose not to participate.
2. Report those patients who were retrospectively found to have failed the eligibility criteria and those patients who failed to complete the assigned treatment.
3. Report all statistical methods and assumptions made.
4. For univariate analyses, report estimated treatment effects (log-odds ratios or hazard ratios), confidence intervals, and significance levels of tests of no treatment effect (*P*-values).
5. For adjusted analyses, report adjusted estimates of treatment effects, confidence intervals, and *P*-values.
6. When no treatment effect is found, do not report the power of the study. Instead use point estimates and confidence limits.
7. Report any differences between "intention to treat" analyses and "eligible patient" analyses.
8. Results with strong biological or clinical justification and *P*-values near 0.05 should be called "statistically significant."
9. Results with no biological or clinical justification or those which seem paradoxical should be reported and interpreted with caution.
10. Only informally report exploratory or hypothesis generating analyses.

being studied. This section will outline basic reporting guidelines that follow the estimation analytic approach discussed above and that should be helpful for reporting many types of clinical trials. A summary is given in Table III. These guidelines should also be useful for reviewing and interpreting published reports of trials and prognostic factor analyses.

6.1 Describe the Study Population

Clinically relevant descriptions of the cohort and population targeted by the eligibility criteria should be reported. In phase I studies, especially those that are drug oriented, the target population and trial cohort may not have as much clinical relevance as in a phase II or III study. It is also sometimes important to discuss patients who met the eligibility criteria but chose not to participate, when this information is available. The need for this might arise when patients from a large group are asked to participate but many refuse. Adequate descriptions of the cohort will aid generalizations of the results.

6.2 Treatment and Eligibility Failures

As mentioned above, it is acceptable to perform statistical analyses on only the subset of eligible patients, even when eligibility is corrected in retrospect. This does not create bias in the estimate of comparative effects. Investigators should report those patients who were retrospectively found to have failed the eligibility criteria as well as those patients who failed to complete the assigned treatment. This latter group is not excluded from the analysis.

6.3 Statistical Methods and Assumptions

It may seem obvious that readers of trial reports should be aware of any assumptions made in both the design and analysis of a clinical trial. For a discussion of some practical issues, see DerSimonian *et al.*[19] Many common statistical procedures, their assumptions and limitations, are well understood by clinicians. However, the readers of clinical trial reports should be convinced that the data analyst has verified all assumptions and reported the methods for less familiar statistical procedures. Examples of assumptions that are often made in analysis, often violated by the data, and also likely to be consequential are distributional assumptions underlying the t test and proportionality of hazards in lifetable regressions.

6.4 Univariate Analyses

Univariate analyses will likely be conducted to test the effect of all potentially important prognostic variables on the major outcomes. For univariate analyses, investigators should report estimated treatment effects (odds ratios or hazard ratios), confidence intervals, and significance levels of tests of no treatment effect (*P*-values). This does not preclude presenting other displays of univariate analyses (e.g., survival curves or 2×2 tables) if these analyses are especially relevant. However, the investigators should keep in mind that univariate analyses, particularly in prognostic factor studies, are subject to confounding. Consequently, these analyses should not be emphasized or presented in excessive detail.

6.5 Adjusted Analyses

The best style of reporting adjusted analyses is the same as or similar to that for univariate effects. However, the results of adjusted analyses are usually interpreted only after several analyses have been performed. That is, the best adjusted analyses to report should be selected from a larger set of preliminary results. As an example,

consider a lifetable regression model attempting to predict time to cancer recurrence. The "best" (most predictive but parsimonious) model might be built using a step-down procedure from a large set of potential prognostic factors. Each step in the analysis need not be reported, but the final model is a major objective of the analysis.

For adjusted analyses, investigators should report adjusted estimates of treatment effects, confidence intervals, and *P*-values. Not all prognostic factors retained in multiple regression models must be "statistically significant." As stated above, it is often useful to keep non-significant effects in an adjusted model to demonstrate convincingly that the treatment effect persists in the presence of particular covariates.

6.6 Negative Findings

In comparative trials, when no statistically significant treatment difference is found, investigators will want to convince readers that this also means the absence of important clinical effects. Because clinical effects are measured by risk ratios rather than *P*-values, guidelines given above emphasizing estimated treatment differences rather than hypothesis tests are important. Helpful advice regarding negative clinical trials is provided by Detsky and Sackett.[20] Power calculations performed after the study is completed are rarely, if ever, helpful. This is true because 1) the power of a non-significant study against the observed difference will be low, and 2) the alternative hypothesis used when designing the trial is probably no longer supported by the data.

6.7 Differences between Analyses

Although we have emphasized the value of the intent to treat principle and related analyses, in practice, many exploratory analyses will be done. Investigators should report any differences between "intention to treat" analyses and "eligible patient" analyses. If subset analyses are performed, discrepancies between these and the major analyses of the clinical trial should be reported.

6.8 What is Significant?

The *P*-value should not be the only criterion for significance. Results with strong biological or clinical rationale and *P*-values near 0.05 are "statistically significant." Although conventional methods use a cutoff of 0.05 as the significance level, there is no reason to expect this to suffice for all circumstances. Specifically, when biological justification is strong, effect estimates are large, and confidence intervals or *P*-values indicate significance near conventional levels, it seems appropriate to label these results as significant.

Conversely, results with no biological or clinical justification or those which seem paradoxical should be reported and interpreted with caution, even when *P*-values are smaller than 0.05. There is no way to separate type I errors from truly significant results, except to rely on additional evidence and biological rationale. It is wise to report cautiously results which seem not to be consistent with other more reliable evidence.

6.9 Exploratory Analyses

Exploratory or hypothesis generating analyses should be only informally reported.

7. Some Limitations of Analysis

In some circumstances, a special statistical analysis appears to address an important clinical question, even though the procedure violates recommendations made above. Sometimes, such analyses will be appropriate and clinically useful. In general, analyses are likely to be useful if they are relevant to the treatment of new patients. Analyses which require information not yet available when seeing a new patient are not likely to be helpful. A good example of this is "evaluability" or completion of therapy. When deciding whether or not to administer a treatment to a new patient, "evaluability" is unknown. Consequently, results of trials that depend on "evaluability" are not very helpful.

Clinicians often would like to estimate the risk of a specific outcome while excluding other outcomes. For example, we might wish to know the risk of death from cancer excluding non-cancer causes of death. Although this seems like a reasonable question, it can be difficult or impossible to answer. In a population of patients with cancer, causes of death are probably not independent of one another. However, this is precisely the assumption made by censoring other causes of failure. For example, if patients who died from cardiovascular disease are censored, the resulting estimate of the cancer specific death rate is valid only if cancer death and cardiovascular death are independent of one another. Because this is not likely to be the case, the most reliable analyses will probably be those looking at composite events (e.g., time to death from any cause).

If patients with some characteristic are lost to follow-up (or censored) just before they fail, we will not observe all of the failures that are taking place. Consequently, we might observe a falsely low failure rate. This is "informative censoring" and can create bias. There is very little that can be done about this problem after it occurs. Therefore, clinical trials and prognostic factor studies must rely upon active follow-up to ascertain events and be sure that informative censoring has not occurred.

Statistical analysis cannot conclusively prove cause-effect relationships. At best, well designed experiments can demonstrate that associations in the data are unlikely to have arisen by chance. When all sources of bias have

been controlled and chance is not a likely explanation, investigators may reasonably attribute differences in the outcomes to treatment effects.

Even so, concerns about interpretation can arise, especially when several factors are investigated simultaneously as in multiple regression models. For example, risk ratios may change substantially or even become nonsignificant when adjusted for prognostic factors. Furthermore, when data on numerous prognostic factors are available, several sets of predictors may explain the outcome equally well. Sometimes, clinicians are confused or troubled by such occurrences because they seem to indicate that there is not one "best" answer.

Most of this type of behavior is due to correlations between predictor variables. Predictors are frequently "positively" correlated with one another so that inclusion of several effects in a multivariate model will reduce risk ratios (towards 1.0) and increase *P*-values. However, this is not required to happen, and it is possible that adjusting for one prognostic factor will increase the significance of another.

Sometimes, two prognostic factors are so highly correlated with one another that they cannot both be included in the same model. In this circumstance, investigators should not feel uncomfortable choosing the factor with the most meaningful clinical interpretation. The same can be said for entire multiple regression models, assuming that either model fits the data equally well. The point is that clinical interpretation has a vital role to play in the statistical modeling effort.

8. Summary and Conclusions

When analyzing and reporting the results of clinical trials, investigators should follow a simple approach. The purpose of a trial is to estimate an effect or treatment difference, which if present would have clinical utility when treating new patients. Procedures or methods that do not facilitate precisely and impartially estimating and reporting the treatment effect are likely to mislead investigators. Most often in clinical trials, investigators are interested in estimates of risk ratios (specifically odds or hazard ratios) between the treatment groups or levels of a prognostic factor.

These simple ideas suggest that the most useful results from clinical trials will be estimated risk ratios and their confidence limits. Especially in cancer, where disease progression, recurrence, and death are common events following treatment, estimates of risk difference are very relevant. Hypothesis tests and associated *P*-values, although often (or exclusively) reported, are of lesser utility because they do not fully summarize the data. These recommendations may be seen by some investigators to be contrary to accepted practice. It is true that they are somewhat contrary to common practice but their general acceptance is evident in many journals and presentations by clinical trial methodologists.

Despite some disagreement among statisticians regarding the need for adjustment of analyses for imbalanced prognostic factors, it is helpful to see if treatment effects change after accounting for imbalances. When this occurs, it may be of clinical interest. Although we discourage analyses that exclude any patients who meet the eligibility criteria, some circumstances will require that this be done (e.g., when a patient refuses to participate after randomization). Investigators should report, and emphasize as primary, those analyses that include all eligible patients. It is our hope and belief that analysis and reporting of trial results along the guidelines suggested here will result in impartial and useful information for journal readers.

ACKNOWLEDGMENTS

Supported in part by a Foundation for the Promotion of Cancer Research Fellowship. Thanks to Helen Cromwell for secretarial assistance.

(Received March, 24, 1993/Accepted June 1, 1993)

REFERENCES

1) Piantadosi, S., Saijo, N. and Tamura, T. Basic design considerations for clinical trials in oncology. *Jpn. J. Cancer Res.*, **83**, 547–558 (1992).
2) Byar, D. P. Identification of prognostic factors. *In* "Cancer Clinical Trials: Methods and Practice," ed. M. E. Buyse, M. J. Staquet and R. J. Sylvester, pp. 210–222 (1984). Oxford Univ. Press, Oxford.
3) George, S. L. Identification and assessment of prognostic factors. *Sem. Oncol.*, **15**, 462–471 (1988).
4) Kalbfleish, J. D. and Prentice, R. L. "The Statistical Analysis of Failure Time Data" (1980). John Wiley & Sons, New York.
5) Armitage, P. and Berry, G. The design of experiments. *In* "Statistical Methods in Medical Research," pp. 172–175 (1987). Blackwell Scientific Publications, Oxford.
6) Simon, R. M. "Design and Conduct of Clinical Trials," pp. 329–350 (1985). J. B. Lippincott Co., Philadelphia.
7) Cox, D. R. and Snell, E. J. "Analysis of Binary Data," 2nd Ed. (1989). Chapman and Hall, London.
8) International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *Ann. Intern. Med.*, **108**, 258–265 (1988).
9) Bailar, J. and Mosteller, F. Guidelines for statistical reporting for medical journals: amplifications and explana-

tions. *Ann. Intern. Med.*, **108**, 266–273 (1988).
10) Altman, D., Gore, S., Gardner, M. and Pocock, S. Statistical guidelines for contributors to medical journals. *Br. Med. J.*, **286**, 1489–1493 (1983).
11) Mosteller, F., Gilbert, J. and McPeek, B. Reporting standards and research strategies for controlled clinical trials; agenda for the editor. *Cont. Clin. Trials*, **1**, 37–58 (1980).
12) Berry, G. Statistical significance and confidence intervals. *Med. J. Aust.*, **144**, 618–619 (1986).
13) Simon, R. Confidence intervals for reporting results of clinical trials. *Ann. Intern. Med.*, **105**, 429–435 (1986).
14) Gardner, M. J. and Altman, D. G. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br. Med. J.*, **292**, 746–750 (1986).
15) Braitman, L. Confidence intervals extract clinically useful information from data. *Ann. Intern. Med.*, **108**, 296–298 (1988).
16) Berger, J. Are P-values reasonable measures of accuracy? *In* "Pacific Statistic Congress," ed. B. I. F. Manly and F. C. Lam, pp. 21–27 (1986). Elsevier, North Holland.
17) Berger, J. and Sellke, T. Testing a point null hypothesis: the irreconcilability of p-values and evidence. *J. Am. Stat. Assoc.*, **82**, 112–139 (1987).
18) Berry, G. Statistical guide-lines and statistical guidance. *Med. J. Aust.*, **146**, 408–409 (1987).
19) DerSimonian, R. Charette, L. J., McPeek, B. and Mosteller, F. Reporting on methods in clinical trials. *N. Engl. J. Med.*, **306**, 1332–1337 (1982).
20) Detsky, A. S. and Sackett, D. L. When was a 'negative' clinical trial big enough? *Arch. Intern. Med.*, **145**, 709–712 (1985).
21) Goodman, S. and Royall, R. Evidence and scientific research. *Am. J. Public Health*, **78**, 1568–1574 (1988).
22) Rothman, K. Significance questing. *Ann. Intern. Med.*, **105**, 445–447 (1986).