STATISTICAL ANALYSIS OF END-POINTS IN CANCER CLINICAL TRIALS

IAN RONALD CAMPBELL
MA**,** MB, B.Chir, FRCS, FRCR.

CLATTERBRIDGE CENTRE FOR ONCOLOGY, MERSEYSIDE
AND
ST JOHN'S COLLEGE, CAMBRIDGE

DISSERTATION SUBMITTED FOR THE DEGREE OF
DOCTOR OF MEDICINE
UNIVERSITY OF CAMBRIDGE
1992

To my parents

CONTENTS

SUMMARY

The major end-points arising from cancer clinical trials are reviewed. These are: tumour response, treatment morbidity, survival with related data, and quality of life.

A survey of tumour response data from 81 published clinical trials found the most common statistical test in use to be a Chi squared test of the total response rate, but a total of 21 different statistical methods were used. The various statistical tests available are reviewed, including the Mann-Whitney test and the Chi squared test for trend which make use of all the categories of response and their intrinsic order. The assumptions underlying the tests are described. Theoretical considerations support the Mann-Whitney test as the optimum choice for the analysis of tumour response data.

Methods for comparing alternative statistical tests are summarised, and a new method is described which uses a number of typical sets of data to estimate the relative efficiency of two statistical tests by the median value of the square of the ratio of the *z*-values. Using this technique, and data from the 81 trials, the Mann-Whitney test is found to be around 40% more efficient than the Chi squared test of the total response rate (this increased efficiency is equivalent to increasing the recruitment to the trial by 40%).

This practical result is confirmed by mathematical modelling of tumour response using the power relation of the Mann-Whitney test for ordered categorical data, which is derived. Clinical data is found to fit best a shift model which assumes homogeneity of treatment effect across the different grades of response. On the basis of this model, the Mann-Whitney test is found to be 30% to 110% more efficient than a Chi squared test of the total response rate.

The similarities of acute morbidity data to tumour response data lead to similar general conclusions on the optimum method of statistical analysis. In a survey of 36 published clinical trials, the most common method of statistical analysis was again a Chi squared test of a dichotomy (such as no morbidity versus morbidity of any grade). Analysis of data from these trials shows the Mann-Whitney test to be more efficient by around 30%.

A survey of 81 papers reporting tumour response in clinical trials found that few of them used methods of estimation of the difference between the treatments, or derived confidence intervals of the size of such a difference. Methods of estimation and calculation of confidence intervals were found even less often in a survey of methods of presentation of morbidity results. The possible reasons for this are discussed.

It is concluded that the current methods of analysis of tumour response data and many sets of acute treatment morbidity data are not optimum, and a change should be made from the Chi squared test to the Mann-Whitney test. Such a change could be equivalent to an increase in recruitment into many cancer clinical trials of around 40%.

KEY WORDS

Neoplasms [therapy]; Clinical Trials [methods]; Research Design; Statistics; Efficiency; Mann-Whitney test

CHAPTER 1. GENERAL INTRODUCTION

1.1 TYPES OF END-POINT IN CANCER CLINICAL TRIALS

To cure sometimes, to relieve often, to comfort always.
(quoted in Strauss, 1968)

This folk saying dates back to the fifteenth century or earlier, but the aims of the treatment of cancer today are almost exactly the same. They are firstly cure, secondly induction of tumour remission with relief of symptoms and prolongation of life if cure is not possible, and thirdly palliation of those symptoms which cannot be controlled by specific cancer treatments.

Although these aims may be clear in the minds of the doctors treating cancer, none of them can easily be measured in a scientific manner when treatments are being assessed within a clinical trial. It is necessary instead to measure what can be measured and then make inferences concerning the primary aims. In practice, the four main types of end-point used in cancer clinical trials are

- Survival and related measures
- Tumour response
- Treatment morbidity
- Quality of life

SURVIVAL

For curative treatments, survival is the primary end-point, but the morbidity of treatment must also be considered, since it may be of no benefit to patients if a small increase in the cure rate is obtained at the expense of a large increase in long term morbidity. There are some difficulties in the precise definition of cure in cancer - in particular, recurrences may not develop until many years after diagnosis (as in breast cancer and melanoma, for example), and second primary tumours may occur and be difficult to distinguish from recurrences of the first primary (as in breast and bowel cancers). For some cancers, a proportion of patients can be identified several years after treatment, whose death rate is similar to that of the rest of the population (Haybittle, 1983), but in general, it is difficult to distinguish between a cured and non-cured fraction, and separate analyses are not usually done.

There are two particular characteristics of survival data which mean that specific statistical techniques are required for analysis. Firstly the observation of a patient is often incomplete (usually because the patient is still alive at the time of analysis and his duration of survival is therefore not known), and secondly the duration of follow-up is different for different patients, since entry into the clinical trial occurs over an accrual period, often of several years. These characteristics are shared by a number of other endpoints of cancer clinical trials, as follows.

- Time to any recurrence
- Time to local recurrence
- Time to distant recurrence
- Time to death from cancer (as distinct from death from any cause, e.g. where a large proportion of the patients die of non-malignant causes such as in carcinoma of the prostate gland)
- Time to development of late morbidity

These measures may be less important than the survival data, but can be useful to indicate where treatment is failing. Because of the similarity with survival data, these end-points are usually analyzed by the same methods as the overall survival data. Late morbidity is often measured as an ordered categorical variable, which makes analysis of data even more complex. Methods for the analysis of censored ordered failure-time data are not well developed, and will not be considered in this dissertation.

TUMOUR RESPONSE

When a cancer treatment is given, the first indication of the efficacy of the treatment is whether the tumour shrinks or continues to grow. This response of the tumour to treatment may be an indicator of both amelioration of symptoms and of long term cure. However, a higher response rate does not necessarily lead to a higher cure rate, and a higher response does not necessarily benefit the patients if it is obtained at the expense of severe morbidity. Response data give a much earlier indication of treatment efficacy than survival data, since response data are available within a period of months in a clinical trial, whereas estimation of cure usually requires follow-up for many years. Consequently, response data are most used in the early phases of treatment development.

TREATMENT MORBIDITY

Many cancer treatments are given close to their maximum tolerated doses and so observation of morbidity is important in cancer clinical trials. In general, frequency and severity of morbidity increase with the dose of chemotherapy or radiotherapy, but the relationship is not always a simple one. In some cases (e.g. the lung fibrosis caused by bleomycin and the cardiotoxicity caused by doxorubicin), there is a threshold, with little toxicity below a threshold dose and then a rapid rise in both frequency and severity. In other cases, e.g. myelosuppression, lesser grades of morbidity may cause insignificant

clinical problems whereas the more severe grades may be life-threatening - minor grades of white count and platelet suppression usually cause no clinical manifestations whereas severe grades can cause septicaemia and haemorrhage respectively and can be fatal.

In any clinical trial, it is important to identify the set of questions to be answered, and this will govern the kind of data to be collected, and the statistical analyses to be performed.

QUALITY OF LIFE

More and more cancer clinical trials are including assessments of quality of life. This results from the realisation that it is not sufficient to show that a treatment causes shrinkage of a tumour, or prolongs life; if these effects are obtained at the expense of severe morbidity, then the patients may not have received a net benefit from the treatment.

'Quality of life' cannot be defined precisely, and the term is used to cover three types of approach. Firstly, the control of symptoms due to a cancer can be assessed, for example control of cough and chest pain from a lung cancer can be measured, as in the recent MRC lung cancer radiotherapy trials. These end-points are relatively easy to measure and evaluate. Secondly, instruments have been developed to assess the effect of the disease and its treatment on a range of physical and psychological functions of the patient. These usually take the form of questionnaires (some self-administered and some administered by clinicians or interviewers) which may combine specific disease or treatment related questions (e.g. pain and nausea) as well as general questions (Fitzpatrick et al, 1992). Most give scores for a number of different dimensions, e.g. physical, emotional and social functions, and some provide a total score derived from the individual dimensions - although whether this kind of summation should be

recommended is questionable (Fitzpatrick et al, 1992; Spiegelhalter et al, 1992). Summation may result in contradictory trends for the different dimensions of quality of life being overlooked. General quality of life scales have been available for some time (Clark and Fallowfield, 1986) and measurement scales specific to cancer patients have more recently been developed (Aaronson et al, 1988; Maguire & Selby, 1989). These questionnaires may be quite time consuming to complete, which makes them less easy to use repeatedly during the course of an illness. There are also problems of analysis when patients withdraw from a study or are lost to follow-up.

The third approach is to evaluate quality of life in terms of a single summary figure which reflects the balance of gains and losses from treatments given. A number of different methods have been proposed, and have been reviewed by Kind (1988). For example, overall quality of life can be assessed by a single visual analogue scale. A second method is the time trade-off technique, where patients are asked to choose between two alternatives of (a) remaining in the current state of health for time *t* followed by death, or (b) returning to a perfect state of health for a shorter time *x* followed by death. The duration *x* is varied until the patient rates the two alternatives as equal and is unable to choose between them. The patient's assessment of his or her health state in then given by the ratio *x/t*. In the standard gamble, patients are asked to choose between another two hypothetical alternatives. In one alternative, the patient remains in the current state of health; the second alternative is a risky choice in which there is a possibility of returning to good health with a probability *P*, with a complementary probability (1 - *P*) of immediate death. The probabilities are varied until the patient is unable to choose between the certainty of remaining in the initial health state and the gamble. The patient's assessment of his or her health state is then given by *P*. A single quality of life figure can also be obtained from data expressed in terms of various rating scales by use of the Rosser disability-distress matrix (Gudex & Kind, 1988). Repeated estimation or measurement of this overall quality of life score over time will give a measure in terms

of quality of life years - the QALY index (Gudex and Kind, 1988; Maguire and Selby, 1989), or in terms of quality adjusted time without symptoms and toxicity – Q-TWiST (Gelber et al, 1991).

QALYs can be used to choose between alternative treatments for the same patient, or to choose among different interventions aimed at different patient groups. Increasing constraints on health care expenditure have led to a search for rational and accountable methods of allocating resources, and QALYs have been seized upon as possibly fulfilling this role. However, there are major doubts over whether the cost and outcome information and methodology are available yet for QALYs to perform this role effectively (Spiegelhalter et al, 1992)

SCOPE OF THE DISSERTATION

This dissertation will concentrate on the evaluation of two types of end-point in cancer clinical trials: firstly the evaluation of the efficacy of treatment on the basis of tumour response, and secondly on some kinds of acute treatment morbidity where the level of morbidity is graded by allocation to a number of categories. It will be shown in chapters 2 and 3 that there are close similarities between these two types of data, and this means that similar statistical techniques are appropriate for analysis. Survival data and related types of data require a different set of techniques (Kaplan and Meier, 1958; Peto et al, 1976; Peto et al 1977; Armitage & Berry, 1987; Altman, 1991), and will not be discussed further.

1.2 THE NEED FOR STATISTICAL METHODS IN CANCER CLINICAL TRIALS

Occasionally a new treatment is developed where it becomes obvious, as soon as a few patients have been treated, that the results are a large improvement on previous treatments. Here, there may be no need for comparative clinical trials or detailed statistical analysis. An example is the introduction of the drug cisplatin into combination cytotoxic treatment of disseminated testicular teratoma (Einhorn et al, 1989)

Much more commonly, the new treatment has results roughly similar to those of existing treatments, and the problem is to determine whether there is a modest improvement, no change, or a modest detriment in outcome. Compelling arguments can be put forward for comparison of treatments in clinical trials to be by random allocation of patients between treatments rather than by a comparison against historical controls (Staquet & Dalesio, 1984), but these will not be discussed further.

Although it is easy to draw conclusions on the treatment of future patients when a trial shows a large clearcut difference in results, it is less easy to do so when the difference is small. Clearly the ideal is to draw the correct conclusions and adopt a new treatment if it is really better than existing treatments and reject it if it is no better or worse than existing treatments (the designation of a treatment as "better" needs to include a consideration of morbidity and costs as well as beneficial outcome). Thus two kinds of error are possible in generalising from the results of a trial. The first is to adopt a treatment which is apparently better but in reality no better than existing treatments, and the second is to reject a new treatment which apparently is no better but in reality is an improvement. These errors may be termed "false positive conclusions" and "false negative conclusions" respectively.

Statistical methods are important because they help to minimize these errors in the application of trial results. Without them, the interpretation of results would be essentially a matter of guesswork, and would be subject to arguments that could not be resolved. To quantify the risk of false positive conclusions, statistical techniques are available to calculate *P* values from the observed clinical data; these are the probability of obtaining the observed results (or more extreme results) supposing that there is no difference between the treatments.

POWER AND SAMPLE SIZE

To quantify the risk of a false negative conclusion being drawn from the clinical data, the concept of power has been developed. This is the probability that the clinical trial would have detected a particular difference in outcome. Commonly, the planners of multicentre clinical trials aim for power of 80% or more. The power of a clinical trial depends on the specified difference in outcome (larger differences are easier to detect), and the statistical test used. It also depends on the level of statistical significance that will be used (usually 5%) and the numbers of patients entered in the trial (the power increases with the number of patients entered).

The relation of the power to the number of patients entered is crucial to the design of clinical trials. In many cancer trials, the numbers of patients entered are such that only large real differences in outcome would have been reliably detected, and the probability of small differences (a more realistic expectation) being detected is small. The small size of many trials is usually due to the difficulties in recruitment of adequate numbers of patients (Pocock, 1978; Pocock et al, 1978). Many other trials considered are never even started because of the limitations of patient numbers available for study. Thus recruitment is often the major limiting factor to the pace of advance of clinical knowledge through clinical trials.

Because the power of a clinical trial also depends on the statistical method used to analyze the clinical data, it is important that the statistical method used is the one with the greatest power. If there is a choice between two statistical methods, and there is general use of the less powerful statistical method, then more patients will be required in clinical trials and the pace of advance of clinical knowledge will be slower than it could be. This is a strong argument for clinicians to pay attention to statistical methodology and this will be discussed further in section 2.4 which will discuss the relative efficiency of statistical tests.

ESTIMATION AND CONFIDENCE INTERVALS

The discussion so far has been in terms of testing whether a difference exists between the results of treatments in a clinical trial (i.e. hypothesis testing). However, this approach has in the past been misused and significance levels from clinical trials have been reported without adequate discussion of the size of the differences found (Pocock et al, 1987). There are increasing calls (Gardner and Altman, 1988; Anonymous, 1987) for greater emphasis to be put on the size of the difference found in a clinical trial, and to give some idea of the precision of the estimate of the difference by the use of a confidence interval.

A 95% confidence interval can loosely be described as a range of values that includes the true value with a probability of 95% (Altman, 1991). More exactly, if a clinical trial were repeated 100 times and a 95% confidence interval were calculated each time, then 95 of these 100 confidence intervals would be expected to contain the true value.

CHAPTER 2. TUMOUR RESPONSE DATA

2.1 INTRODUCTION

METHODS OF MEASUREMENT

The response of solid tumours to treatment is assessed by observing the change of tumour size with time. The tumour size can be measured directly by clinical examination using a ruler or callipers, if the tumour is superficial, or indirectly by imaging techniques (combined with a magnification factor) if it is deep. For example, chest masses can be evaluated from chest X Rays and deep abdominal masses can be evaluated by computerised tomography (CT). Laboratory studies e.g. blood biochemistry are sometimes included in the evaluation of response e.g. HCG levels in teratoma of the testis.

PRECISION OF MEASUREMENT

The precision of measurement of solid tumour masses depends on their position, shape and degree of demarcation. Deep masses are measured less accurately than superficial masses, especially deep masses within the pelvis. With the advent of CT scanning this is less of a problem, although CT scanning cannot usually be performed as frequently as clinical examination, so that there may be only a few occasions on which the tumour size is accurately known.

Irregularly shaped masses are more difficult to assess than spherical masses. For example, the diameters of a rounded lung metastasis may be readily measured, but different observers will find difficulty in agreeing on the dimensions of a pleural effusion.

Although well demarcated tumours can be measured with a precision of a few millimetres, less well defined tumour masses such as diffusely infiltrating tumours or lymphangitis carcinomatosis can be difficult or impossible to measure. Involvement of the liver by multiple tumour masses is usually assessed only by the enlargement of the whole organ rather than measurement of the individual masses (Miller et al, 1981).

Little work has been done on the reliability of measurement of tumour size. However, Moertel & Hanley (1976) in a simulation used a series of spheres beneath a layer of foam rubber to assess the accuracy of clinicians' assessments of tumour size. One finding was that clinicians frequently reported a difference of 25% in tumour size (measured as the product of two perpendicular diameters) when presented with the same sized spheres. The conclusions of the study were that high accuracy is difficult to achieve, and it is necessary to define tumour response by a large estimated reduction in tumour size (50% or more in the product of two perpendicular diameters) in order to give a reliable indication that the real tumour size has in fact decreased. Furthermore, the error rate should be less if the same clinician assesses the tumour on all occasions.

A tumour measurement must be recorded accurately if it is to be compared with a subsequent measurement in order to assess tumour response. In an external review of assessment of response of breast cancer to chemotherapy, Sears and Olson (1980) found that 16% of cases were nonevaluable because of inadequate documentation. In another 5% of cases, reviewers disagreed with the investigators' assessment of whether there had been a response to treatment.

SYSTEMS OF CLASSIFICATION OF TUMOUR RESPONSE

Because of the inaccuracies in measuring tumour size, response of solid tumours to treatment is rarely reported in terms of a numerical change in size. Instead, patients are allocated into one of the 4 broad categories of complete response (CR), partial response (PR), no change (NC), and progressive disease (PD). Standard definitions for the assessment of response in breast cancer were published by the International Union Against Cancer (UICC) (Hayward et al, 1977). Representatives of the EORTC, the National Cancer Institute of the USA, the UICC and several other organisations met and agreed on standard definitions for response assessment in general, which were published by the World Health Organisation (WHO) (WHO, 1979). These recommendations have also been reported in a more accessible source (Miller et al, 1981). Table 2.1 gives the WHO definitions of response for measurable disease. Recommendations are also given for non-measurable disease and bone metastases, as well as for other aspects of reporting the results of cancer treatment.

Table 2.1. WHO Definition of Objective Response for Measurable Disease (Miller et al, 1981)

---

1. Complete response (CR).

   The disappearance of all known disease, determined by two observations not less than four weeks apart.

2. Partial response (PR).

   50% or more decrease in total tumour load of the lesions that have been measured to determine the effect of therapy by two observations not less than four weeks apart. Bidimensional: single lesion, greater than or equal to 50% decrease in tumour area (multiplication of longest diameter by the greatest perpendicular diameter); multiple lesions, a 50% decrease in the sum of the products of the perpendicular diameters of the multiple lesions. Unidimensional: greater than or equal to 50% decrease in linear tumour measurement. In addition there can be no appearance of new lesions or progression of any lesion.

3. No change (NC).

   A 50% decrease in total tumour size cannot be established nor has a 25% increase in the size of one or more measurable lesions been demonstrated.

4. Progressive disease (PD).

   A 25% or more increase in the size of one or more measurable lesions or the appearance of new lesions.

The application of these criteria is often not simple and has been extensively discussed (Smith, 1983; Israel, 1983; Parbhoo & Wahba, 1983; Hoogstraten, 1984). For example,

the volume of a tumour is not closely related to the number of clonogenic cells (i.e. cells with a potential to divide indefinitely) (Hall, 1978), or different masses within the same patient may respond differently. There may also be problems that tumour masses do not completely disappear after treatment and yet apparently contain no viable malignant cells, as may occur in testicular teratoma treated by chemotherapy (Tait et al, 1984). However, the classification systems have now been in routine use for many years and there seems to be reasonable satisfaction with their ease of use and reproducibility. Sometimes a fifth category is used of minimal response (of between a 25% and 50% reduction in the product of perpendicular diameters). However, this has been criticised (Hoogstraten, 1984) on the basis that the work of Moertel and Hanley (discussed above) implies that the designation of a minimal response is very unreliable.

TUMOUR RESPONSE: TYPE OF DATA

Although tumour size is a numerical quantity (in statistical terminology, a continuous quantitative variable), the classification system that is almost universally used reduces tumour response to an allocation into one of 4 categories (i.e. to a categorical variable). Thus the data consist of the numbers of patients in the 4 categories. A point that is crucial to much of the discussion of tumour response in this dissertation is that the 4 categories clearly fall into a natural order from the best to the worst, i.e. CR, PR, NC, and PD. The data is therefore termed ordered categorical data. This is in contrast to data such as histological type where there is no intrinsic order in the categories, for example, transitional cell carcinoma, squamous cell carcinoma, and adenocarcinoma have no intrinsic order.

The question of how to analyze response in cancer clinical trials is thus essentially one of how to analyze ordered categorical data. Because the categories are ordered, a particular set of statistical techniques can be applied with advantage, and this

has not always been appreciated. These statistical techniques will be discussed in section 2.3.

THE VALUE OF MEASUREMENT OF TUMOUR RESPONSE

There is no intrinsic value in tumour response (except perhaps for the patient's morale); response is a surrogate end-point and not an end in itself. Its value is as a predictor of survival or improved quality of life. Response has advantages over quality of life in that the latter is very subjective and methods of evaluation in Oncology are still under development (Aaronson et al, 1988). Response also has advantages over survival in that response data are usually available within a period of months of treatment, whereas survival data require follow-up over a period of years, a point which is particularly important in the early phases of treatment development. In addition to these general considerations, there are four areas where response data are particularly valuable.

Firstly in a crossover trial of two treatments, where all patients failing to respond to (or relapsing after) each treatment are then given the other treatment, a difference in survival may not be expected, and a difference in treatment efficacy may well show only by a difference in response or quality of life.

Secondly, where a treatment is a local treatment (e.g. radiotherapy or surgery) and the aim of treatment is local control of disease, there will be no effect of treatment on the metastatic disease that may coexist and which may well be the main determinant of survival. Here again a survival improvement may not be expected even when there is a difference in treatment efficacy. An example is locally advanced breast cancer where a large proportion of the patients will die because of metastatic disease which coexists at presentation (either symptomatic or occult), and yet an important part of the

overall management is the attempt at local control of the disease. Here, response data will be a useful indicator of treatment efficacy. A second measure in this situation could be the duration of local control, but there are difficulties in the definition of this (Parsons et al, 1990).

A third area is in advanced disease where attempts at a cure have been abandoned and the aim of treatment is purely palliative. The main end-points are symptomatic relief and quality of life, but if the patient has several problems and a treatment affects only one of them, any effect of the treatment on the overall quality of life may be masked by the coexisting problems. The response of the individual problems may well provide a better indication of treatment efficacy.

Lastly, a clinical trial comparing two treatments may show a survival difference which is of marginal statistical significance. Examination of whether or not the response data show a clearcut difference will aid the interpretation of the survival data.

2.2 A SURVEY OF STATISTICAL TESTS IN USE FOR TUMOUR RESPONSE DATA

INTRODUCTION

As already discussed, the response of tumours to treatment is usually assessed by allocation to one of the 4 categories of complete response (CR), partial response (PR), no change (NC), and progressive disease (PD). Several statistical methods are available for the analysis of tumour response data when two or more treatments are being compared, and these will be reviewed in the following section. Because the categories of response fall into a natural order, a case can be made for the use of the Mann-Whitney test (Moses et al, 1984; Morton & Dobson, 1990) or other tests that make use of the order of the categories (Moses et al, 1984; Bartolucci, 1984). However, Moses et al found that these tests were not being used in the analysis of ordered categorical data published in 1982 in articles in a general medical journal; instead, the most common statistical test used was the less efficient simple Chi squared (χ^2) test for heterogeneity.

This survey of published trials was undertaken to answer the question of which statistical tests are currently being used in the analysis of response data in Oncology.

METHODS

A survey was made of seven specialist Oncology journals, "British Journal of Cancer", "Cancer", "Clinical Oncology" (formerly a section in "Clinical Radiology"), "European Journal of Cancer" (formerly "European Journal of Cancer and Clinical Oncology"), "International Journal of Radiation Oncology Biology Physics", "Journal of

Clinical Oncology", and "Radiotherapy and Oncology". Papers were identified from the contents pages and were studied provided they fulfilled the criteria of randomised controlled trials of treatments for solid tumours in which response was reported. Papers on lymphoma were included, but papers on myeloma or leukaemia were not, since they are not solid tumours, and different response measures are used. Brief communications in letters and abstracts from meetings were excluded.

A decision on the size of the survey was made after the first 20 papers had been studied as a pilot study. Two criteria were used. Firstly a minimum of 50 papers was required in order to study the distribution of papers across the different statistical tests with reasonable precision. Secondly, a minimum of 20 papers having results significant at a 5% level by both a Chi squared test and a Mann-Whitney test was stipulated because the same series of papers was also to be used for the study of re-analysis of data that is presented in section 2.5. Starting with issues dated December 1990, journals were surveyed in reverse date order in batches of 6 months until the above criteria had been satisfied. This resulted in a total of 81 papers published between July 1988 and December 1990 inclusive. Each paper was analyzed using the form in Appendix C.

CHARACTERISTICS OF THE PAPERS STUDIED

Table 2.2 gives the distribution of tumour types in the 81 papers. The type of treatment comparison reported is shown in Table 2.3. The number of treatments compared was 2 in 67 papers (83%), 3 in 11 papers (13%), and more than 3 in 3 papers (4%). The numbers of patients included in the trials ranged from 12 to 859 with a median of 111.

Whether response was a major end-point in the study was judged from whether or not the abstract of the paper included data or conclusions concerning response. Response

was included in the abstract in 78 papers (96%), it was not included in 2 papers (3%) and one paper had no abstract.

Table 2.2. Distribution of tumour types

Tumour type	Number of papers
Breast Carcinoma	15
Bronchial Carcinoma, Non-Small Cell type	15
Bronchial Carcinoma, Small Cell type	9
Colorectal Carcinoma	13
Head & Neck Carcinoma	6
Prostatic Carcinoma	4
Multiple types, metastatic	4
Cervix Uteri Carcinoma	2
Ovarian Carcinoma	2
Urothelial Carcinoma	2
Other*	9
Total	81

*one paper each of Carcinoid Tumour, Germ Cell Tumours, Hepatocellular Carcinoma, Hodgkin's Disease, Melanoma, Non-Hodgkin's Lymphoma, Pancreatic Carcinoma, Salivary Gland Carcinomas, and Soft Tissue Sarcoma.

Table 2.3. Types of treatment comparison

Type of comparison	Number of papers
Chemotherapy versus Chemotherapy	51
Hormone therapy versus Hormone therapy	9
Radiotherapy versus Radiotherapy + Chemotherapy	4
Chemotherapy versus Immunotherapy	3
Chemotherapy versus Chemotherapy + Radiotherapy	2
Immunotherapy versus Immunotherapy	2
Radiotherapy versus Radiotherapy + Radiation sensitizer	2
Other	8
Total	81

The system used for classification of tumour response was reported as a standard one such as the WHO system (Miller et al, 1981) in 43 papers (53%), it appeared to be especially constructed in 28 papers (35%), and in 10 papers (12%) it was unstated. The number of categories used for recording response was 2 in 6 papers (7%), 3 in 17 papers (21%), 4 in 51 papers (63%) and 5 in 7 papers (9%). Thus the great majority of papers graded response in a way that would have allowed statistical analysis by a test that makes use of the order of the categories.

RESULTS

The methods of statistical analysis reported in the 81 papers are summarised in Table 2.4. It is possible that other tests were also applied to the data and not included in the published paper, but this issue was not mentioned in any paper.

The most common test reported was the simple Chi squared test, but there was considerable variation in how it was used. A single Chi squared test was used in 26 papers; two Chi squared tests were applied to the data in 6 papers, and three Chi squared tests were used in one paper. In some cases the Chi squared test was applied to the original table of frequencies and in some cases it was applied after collapsing the table to a dichotomy such as CR + PR versus NC + PD, as detailed in Table 2.4. In 5 papers the Chi squared test of a dichotomy was combined with multivariate analysis including factors such as performance status and age.

Fisher's exact test was used in 10 papers, again with variation in the precise details as shown in Table 2.4. The Chi squared test for trend was used in a total of 3 papers, and the Mann-Whitney test was one of the tests used in one.

Six different dichotomies were used in those papers where the original table was collapsed to a dichotomy prior to analysis by a Chi squared test, Fisher's exact test or multivariate analysis. The most common dichotomies were CR + PR versus NC + PD (41 instances), CR versus PR + NC + PD (7 instances), and CR + PR + NC versus PD (4 instances). There was one instance each of use of the other three dichotomies.

Altogether, including variations and combinations, 21 different statistical methods were applied to the reported data.

Table 2.4. Method of statistical analysis reported in the published paper

Method			Number of papers
Chi squared test			40
Single Chi squared test		26	
On overall table	3		
After collapsing to a dichotomy	19		
Method unclear	4		
Two Chi squared tests		6	
On overall table and one dichotomy	1		
On two dichotomies	5		
Three Chi squared tests (on three dichotomies)		1	
Chi squared test and multivariate analysis		5	
On one dichotomy	3		
On two dichotomies	2		
Multivariate analysis on one dichotomy		2	
Fisher's exact test			10
One test on one dichotomy		7	
Two tests on two dichotomies		2	
One test on one dichotomy with multivariate analysis		1	
Chi squared test for trend			2
Chi squared test for trend, and Chi squared test on one dichotomy			1
Mann-Whitney test, and Chi squared test on two dichotomies			1
No statistical test reported			19
Method unclear			8
Total			81

The *P* values from the statistical tests were reported in 43 papers (68% of the papers that reported statistical analysis). In 15 papers (24%) only "significant" or "not significant" was reported. In 4 papers (6%) multiple tests were done without a consistent pattern to the reporting. The statistical analysis was reported as two-sided (or two-tailed)

in 15 papers (24%), and as one-sided (or one-tailed) in 4 papers, but no mention of this aspect was made in 44 papers (70%).

DISCUSSION

This survey was not a fully comprehensive survey of all papers that include a comparison of tumour response data, since not all specialist Oncology journals were included in the survey, and since some Oncology papers are published in more general journals. However, the range of tumour types and the range of treatment comparisons shown in Tables 2.2 and 2.3 suggest that the papers studied are reasonably representative of the Oncology literature in general. In almost all cases, reference was made to response in the abstract of the paper and so response appears to have been judged a major endpoint by the authors. Therefore some kind of statistical analysis of response would be expected in the papers, either a statistical test of a null hypothesis or estimation of the difference in outcome together with the confidence interval.

Only 53% of the papers reported that assessment of response was by standard criteria. In some of the remainder the differences from the WHO criteria (Miller et al, 1981) were minor. In some papers there were good reasons for non-standard criteria being used as the problem being studied was not well covered by the standard criteria.

The variety of statistical techniques used (a total of 21 techniques including all variations and combinations) shows a need for standardisation of the method of statistical analysis. The great majority of studies assessed response using more than 2 categories of classification and so could have used a statistical technique that takes advantage of the order of the categories e.g. the Mann-Whitney test or the Chi squared

test for trend. Only 4 papers used one or other of these techniques despite recommendations for their use in the analysis of ordered categorical data (Moses et al, 1984; Armitage & Berry, 1987), of which response data are examples. These techniques will be discussed in detail in the following section, which describes the available statistical tests that can be applied to tumour response data.

There was a variety of dichotomies used when the 4 categories of tumour response were collapsed to 2 for analysis by the Chi squared test or Fisher's exact test. This illustrates the disadvantages of these techniques in that the choice of dichotomy is to some extent arbitrary, and this will be discussed further in the following section.

No statistical test was reported in 19 papers. In one of these a 95% confidence interval for the difference in the total response rate was calculated by the authors, but in 18 papers, no method at all was reported. (The use of confidence intervals is analyzed in more detail in section 2.8). It can be argued that statistical analysis is unnecessary where the outcome from a trial is clearcut, and this may occur either because the results from the treatments compared are exceedingly similar or because they are exceedingly dissimilar. In order to assess whether this could explain the lack of a statistical method in the 18 papers, the results of the papers were analyzed by statistical tests, and the papers were then arbitrarily divided into 3 groups on the basis of the z-values (briefly, the z-value is the number of standard errors, a z-value of 1.96 corresponding to a P value of 0.05; fuller details are given in Appendix B.) The three groups were (a) those where the z-values were all less than 1.0 i.e. two-sided P value was greater than 0.32 (whether analysis was done by the Chi squared test or the Mann-Whitney test), where the results might be judged by eye to be similar, (b) those where the z-values lay between 1.0 and 3.0, and (c) those where all z-values were over 3.0 ($P < 0.0027$), where the results might be judged by eye to be dissimilar. Examples of sets of data falling in these three groups are given in Table 2.5. The division of the 18 papers into these 3 groups was (a) 10,

(b) 6, (c) 2. Thus in at least 6 papers, it appears that some form of statistical analysis should have been reported, especially in the 3 papers out of these 6 where conclusions were drawn on the relative efficacy of the treatments studied.

Table 2.5. Example data sets comparing two treatments (denoted A and B) where the *z*-values are (a) less than 1.0 ($P > 0.32$), (b) between 1.0 and 3.0, or (c) greater than 3.0 ($P < 0.0027$).

	(a)		(b)		(c)	
	A	B	A	B	A	B
CR	40	38	0	1	0	0
PR	28	34	62	69	15	2
NC	9	8	26	28	29	33
PD	9	9	40	21	10	20

The failure to make clear the method of analysis used in 8 papers (10%) is disappointing. So also is the reporting of the analysis only in terms of "significant" or "not significant" in 15 papers (24%), especially as the derived statistic (which would have allowed readers to calculate a *P* value) was given in only 2 of these 15. Furthermore, only a minority of papers reported whether statistical tests were one- or two-sided. These failures illustrate that guidelines on the reporting of results (Altman et al, 1983; Gardner et al, 1986) are not always being followed.

2.3 A REVIEW OF THE STATISTICAL TESTS AVAILABLE FOR TUMOUR RESPONSE DATA IN CANCER CLINICAL TRIALS

INTRODUCTION

The preceding section showed considerable variation in the statistical techniques being used to analyze tumour response data in cancer clinical trials. The current section reviews the principal statistical tests that can be applied to such data. These are

- The Chi squared test of 4 categories of response
- The Chi squared test of the total response rate
- Fisher's exact test
- The Chi squared test for trend
- The *t* test after allocation of scores
- The Mann-Whitney test
- The Kruskal-Wallis test
- Regression methods

The tests will be reviewed in turn, and their underlying assumptions, their advantages and their disadvantages will be examined. A sample set of data, shown in Table 2.6, will be used to demonstrate the calculations for each of the tests. This set of data has been taken from Iversen et al (1990), which is one of the 81 articles studied in the preceding section. It reports a study of two hormonal therapies in carcinoma of the prostate – a comparison of orchidectomy with a combination of goserelin and flutamide. All calculated *P* values are two-sided.

The clinical question to be answered is whether the treatments have equal efficacy or appear to differ in efficacy. Most of the discussion in this section refers to a comparison of just two treatments in a clinical trial, but in a comparison of three or more treatments, the considerations are almost the same.

Table 2.6. Example tumour response data (Iversen et al, 1990). Numbers of patients in each response category for each treatment.

Category of response	Orchidectomy	Goserelin and flutamide (G + F)	Total
Complete response (CR)	0	1	1
Partial response (PR)	62	69	131
No change (NC)	26	28	54
Progressive disease (PD)	40	21	61
Total	128	119	247

THE CHI SQUARED TEST OF 4 CATEGORIES OF RESPONSE

The Chi squared test for heterogeneity is one of the best known statistical tests. It is widely available in textbooks of statistics and it will be described here only in outline. From a table of the observed (O) values, column and row totals are calculated (as in Table 2.6) and hence the values expected (E) in each cell of the table are obtained. The Chi squared (χ^2) statistic is calculated as

$$\chi^2 = \Sigma \ (O - E)^2 / E,$$

summing over all cells in the table, and this is compared with standard tables of the Chi squared distribution e.g. in Neave (1981).

The Chi squared test can be applied to the whole 2 (treatments) × 4 (response categories) table. Where some of the values in the table are less than 5 (as in the CR category in Table 2.6), some categories should be combined according to standard criteria (Armitage & Berry, 1987; Altman, 1991). For example, in Table 2.6, the categories of CR and PR should be combined for each treatment. Application of the Chi squared test to the data of Table 2.6 then gives a Chi squared value of 6.16 with 2 degrees of freedom and a *P* value of 4.6%, i.e. just significant at the 5% level.

The Chi squared test has advantages of being well known and fairly simple to perform. In the preceding section, it was found to be currently one of the most used statistical tests of tumour response data. However, it can be criticised for not making full use of the information available (Moses et al, 1984). If the Chi squared test is applied to the 2 (treatments) × 4 (response categories) table, it makes no use of the intrinsic order in the 4 categories. The Chi squared test is a test of *any* departure from equality of CR, PR, NC and PD rates between the two treatments. Because of this, it is relatively insensitive to those departures from an equal distribution that indicate a difference in treatment efficacy. For example, by exchanging the data in rows in Table 2.6, a number of different contingency tables can be obtained, some of which are shown in Table 2.7. The derived Chi squared values of all these tables are the same, but the tables differ greatly in their implications for treatment efficacy.

Table 2.7. Three possible contingency tables formed by exchanging the rows of Table 2.6

	Original		New 1		New 2		New 3	
CR	0	1	0	1	62	69	0	1
PR	62	69	40	21	0	1	26	28
NC	26	28	62	69	40	21	40	21
PD	40	21	26	28	26	28	62	69

A further disadvantage of the Chi squared test occurs in relatively small trials in that several cells of the table may have expected frequencies less than 5, as already demonstrated in the CR category of Table 2.6. Because it is recommended that some categories are then combined to avoid this (Armitage & Berry, 1987), there may be some further loss of the information contained in the data.

THE CHI SQUARED TEST OF THE TOTAL RESPONSE RATE

As well as applying the Chi squared test to the whole 2 (treatments) × 4 (response categories) table, the Chi squared test can also be applied in a second way by combining the categories of complete and partial response (giving a category of total response), and by combining the categories of no change and progressive disease (giving a category of total non-response), to give a 2 (treatments) × 2 (response categories) table. With this technique, the value of Chi squared for the data of Table 2.6 is 2.27 with 1 degree of freedom and the *P* value is 13.2%. This *P* value is considerably different to the value of

4.6% for the Chi squared test of the whole table, and is no longer significant at the 5% level. This approach to tumour response data has been described by some authors (Gelman & Zelen, 1987; Leventhal & Wittes, 1988; Simon, 1989). If it is to be used in the analysis of the results of a clinical trial, then estimation of the numbers of patients to be entered in the trial is straightforward, since several sets of graphs and tables exist for this purpose, where the end-point of the trial is a dichotomy such as success or failure of some kind (Boag et al, 1971; Altman, 1980; George, 1984; Machin & Campbell, 1987; Leventhal & Wittes, 1988).

The Chi squared test of the total response rate was found in the preceding section to be currently the most commonly used statistical test of tumour response data. However, like the Chi squared test of the overall 4 category table, the Chi squared test of the total response rate has the disadvantage of not making full use of the information available. The change in the *P* value shown above from 4.6% to 13.2% (when the number of response categories is reduced to 2) is an illustration of how loss of information can affect the *P* value obtained and hence the conclusion of the study.

If the Chi squared test is applied to the 2 × 2 (treatment versus response) table, there is loss of the information contained in the data by the distinction between complete and partial response, and by the distinction between no change and progressive disease. If a treatment has a higher response rate, it may be expected that more of the responses will be complete responses, and also that fewer of the non-responses will be progressive disease. These additional indications of greater treatment efficacy are lost if the 4 response categories are combined to two. For example, in Table 2.6, the total response rate for G + F is higher than for orchidectomy (59% against 48%); this is associated with the proportion of total responses which are complete responses being marginally higher (1/70 against 0/62) and with the proportion of the non-responses which are progressive disease being considerably lower (21/49 = 43% against 40/66 = 61%).

This point is further illustrated in Table 2.8 which is another set of tumour response data taken from the series of 81 papers of the preceding section. The higher total response rate of the radiotherapy + ACNU arm (89% versus 76%) is accompanied by a higher proportion of the responses being complete responses (18/31 = 58% versus 8/29 = 28%). This is not mere manipulation of numbers for their own sake; the distinction between a complete and a partial response is very significant clinically, as is the distinction between no change and progressive disease. Ideally statistical analysis should reflect this.

It would be wrong to generalise from only the two example sets of data shown, but it will be shown in section 2.7 that the points that they illustrate are part of a more general pattern. The above arguments over what relation is to be expected between CR rates and PR rates, and so on, will be developed further in section 2.6 which considers three kinds of model of treatment effect.

Table 2.8. Second example of tumour response data, taken from Okawa et al (1989), comparing radiotherapy alone with radiotherapy + ACNU chemotherapy in non-small cell lung cancer.

Category of response	Radiotherapy alone	Radiotherapy + ACNU
Complete response (CR)	8	18
Partial response (PR)	21	13
No change (NC)	9	4
Progressive disease (PD)	0	0

A further difficulty with a Chi squared test of a 2 × 2 table is over the choice of dichotomy. In many clinical trials, it may appear that the best dichotomy to use is of total response against non-response. However, when the total response rate is very high, it may appear that a better dichotomy to test is of complete response versus absence of complete response, and similarly, at low response rates, it may appear that a better dichotomy to test is of progressive disease against freedom from progressive disease. This was illustrated in the preceding section where six different dichotomies were found in use in the 81 papers studied. The choice of dichotomy is thus rather arbitrary. Alternatively, if several tests are done using several dichotomies, there is the problem of multiple significance testing on the same set of data; this should be avoided if possible as it can be misleading (Pocock et al, 1987).

FISHER'S EXACT TEST

Full details of Fisher's exact test will again not be given because of its wide availability in text books e.g. Armitage & Berry (1987) and Altman (1991). It is a test used for 2 × 2 tables when the expected value in at least one of the cells of the table is very small, which is usually taken as meaning less than 5. In this situation it is normally recommended that the simple Chi squared test for heterogeneity is not used (Armitage & Berry, 1987; Altman, 1991). At the other extreme, for very large expected values, the calculations for Fisher's test become more difficult. In the intermediate range, where both a Fisher's test and a Chi squared test can be used, the *P* values obtained are similar (Altman, 1991).

Fisher's exact test is thus complementary to a Chi squared test of a 2 × 2 table and can substitute for it when a Chi squared test cannot be used. Fisher's test suffers from just the same disadvantages as the Chi squared test of a 2 × 2 table in the analysis of tumour response data, i.e. in making less than full use of the information, and in the choice of a dichotomy.

THE CHI SQUARED TEST FOR TREND

The Chi squared test for trend is a statistical test specifically designed for ordered categorical data (Cochran, 1954; Armitage, 1955; Armitage & Berry, 1987). Its use for tumour response data has been described by Bartolucci (1984) and he gives a worked example (although this contains arithmetical errors). Other worked examples are given in Armitage and Berry (1987) and in Altman (1991). The procedure is first to allocate a scoring system to the ordered categories; e.g. for tumour response data, CR = +2, PR = +1, NC = 0, and PD = -1. A formula is then applied resulting in the test statistic, $\chi^2_{1,t}$. The $\chi^2_{1,t}$ value (Chi squared value for trend) is compared with standard tables of the Chi squared distribution for one degree of freedom. When applied to the data in Table 2.6, the test gives a $\chi^2_{1,t}$ value of 5.37, with $P = 2.04\%$.

In some situations, a linear scoring system such as given above may appear a reasonable assumption, but in others (including tumour response data) the choice of the scoring system is more arbitrary. This is a major drawback in the use of the test for tumour response data. The scoring system given above implies that a complete response is twice as worthwhile with respect to no change as a partial response. In some clinical situations this 2:1 weighting may seem appropriate; but in others it is clearly inappropriate if all the patients with a partial response will eventually relapse and succumb from their disease, while many patients with a complete response are cured (such as in cancer of the cervix uteri treated by radiotherapy). There may well be disagreement amongst clinicians over the choice of the scoring system, and yet the derived statistic and therefore the *P* value and the conclusion of the study will depend on the scoring system used. There is also the undesirable possibility of multiple analyses being done (with different scoring systems), with the reported analysis perhaps chosen on the basis of minimizing the *P* value.

It could be argued that the most natural scoring scale for the Chi squared test for trend is one that is selected by the data, using the relative numbers of observations in the categories to define the separation between the scale values. If this is done, and the score chosen for each category is equal to the middle rank for that category, then the Chi squared test for trend is almost equivalent to the Mann-Whitney test (which will be discussed next). The only difference is a factor applied to the standardized deviate of $\sqrt{\{(n-1)/n\}}$ (Armitage, 1955) which is near unity for large samples. This equivalence is a strong argument for use of the Mann-Whitney test in preference to the Chi squared test for trend since the problem of the choice of a scoring system is obviated. In practice, the two tests will often give fairly close results (Armitage, 1955).

A further argument against the Chi squared test for trend is that the formula used and the theory behind the test will not be readily understandable to most clinicians (but see the next section). A fourth objection is that the test gives a *P* value, i.e. is a hypothesis test, but unlike the Mann-Whitney test, does not provide an estimate of the size of the difference between treatments.

One general advantage of the test is that, in addition to the test for trend, it allows a test for departure from linearity (Bartolucci, 1984; Armitage & Berry, 1987), but this does not have any great relevance to the study of tumour response.

THE *t* TEST AFTER ALLOCATION OF SCORES

This method is described by Moses et al (1984). As with the Chi squared test for trend, a scoring system is first devised. The mean scores of the patients for each of the two treatments are calculated and so are the variances, and a *t* test is applied. The procedure will often give a result similar to a Mann-Whitney test (Moses et al, 1984). For the data of Table 2.6, the test gives a value of *t* of 2.33 with 245 degrees of freedom (P = 1.96%).

The *t* test with allocated scores is closely related to the Chi squared test for trend (with the same allocated scores). The formula for the Chi squared test for trend can be interpreted as a *t* test with three minor modifications. Firstly, the standard error of the difference in the mean scores is calculated from the overall variance of the combined samples rather than from the pooled sample variances; secondly a divisor for the variance of $n_1 + n_2$ replaces $n_1 + n_2 - 2$; and finally the whole formula is squared to give a Chi squared rather than a *t* statistic.

Since the procedure is almost equivalent to the Chi squared test for trend, it suffers from all the objections outlined above except that the basis and workings of the test can be more readily appreciated.

THE MANN-WHITNEY *U* TEST

PRINCIPLE

The Mann-Whitney *U* test is a method of comparing two groups of observations. In theory, each observation of group A is compared with each observation of group B and a count is made of the number of comparisons where the former is less than the latter (Mann & Whitney, 1947; Armitage & Berry, 1987; Altman, 1991). Where the two observations in one of these comparisons are equal, a value of ½ is added to the count. The total value is termed U_{AB}. In practice, to facilitate the determination of U_{AB}, the observations of each group may be arranged in order and for each group A observation, a count is made of the number of higher group B observations; the total of these counts gives the value of U_{AB}. Alternatively all the observations are put in order and ranked, and the sum of the ranks of group A observations is used to calculate U_{AB}. This second method is known as the Wilcoxon rank sum test but is not really a different test, just a different way of performing the same test. Which method is quicker to perform depends on the kind of data being studied. Kendall's *S* test is also an equivalent test (Armitage & Berry, 1987).

For small samples, once the value of U_{AB} has been calculated, it is compared against standard tables e.g. Neave (1981). For large samples, where the size of the larger sample is greater than 15, and the size of the smaller is greater than 3, the value of U_{AB} is approximately normally distributed. Denoting the number of observations of group A by n_1, and the number of observations of group B by n_2, and using n to denote the sum of n_1 and n_2, the expected value of U_{AB} is $\frac{1}{2}n_1n_2$ and its variance is (Armitage & Berry, 1987)

$$n_1n_2(n+1)/12$$

Where there are many equal observations, i.e. tied observations, the formula for the variance must be adjusted by a factor of

$$1 - \{\Sigma(t^3 - t)\}/(n^3 - n)$$

where the summation is taken over all groups of tied observations, t being the number of observations in a particular group of ties (Armitage & Berry, 1987). For small samples, since this correction is not applied, the significance levels obtained may be conservative.

IMPORTANCE OF THE VALUE OF U

The calculated value of U_{AB} is the number of pairs of observations, one of group A and one of group B, where the group A observation is less than the group B observation. This can be used to give a prediction of the probability of a new group A observation being less than a new group B observation (Armitage & Berry, 1987; Altman, 1991). This will be discussed further in section 2.9.

USE OF THE MANN-WHITNEY TEST IN ORDERED CATEGORICAL DATA

The Mann-Whitney test was originally described for continuous numerical data with no tied observations (Mann & Whitney, 1947). However, it is not necessary for the observations to be of a numerical quantity such as age. It is only necessary that the observations can be put into a natural order, and so the Mann-Whitney test can be applied to ordered qualitative data. Furthermore, the test remains valid when most or all of the observations are tied with other observations (although clearly not when all are tied on the same value). Thus the Mann-Whitney test can be applied to data in categories, provided that the categories fall into a natural order (i.e. it can be applied to ordered categorical data). In fact, in the extreme situation, when all the observations fall into one or other of just two categories, the formula for the standardized deviate from the

Mann-Whitney test reduces to that for the Chi squared test for a 2 × 2 table except for a ratio of $\sqrt{\{n-1)/n\}}$ which is near unity for large samples (Armitage & Berry, 1987). It is therefore reasonable to apply the test to short measurement scales where there are only a small number of categories. The calculations required for a Mann-Whitney test of ordered categorical data may well be simpler than for non-categorical data since the effort involved in ordering and ranking is eliminated. For ordered categorical data, all the categories will contain tied data and $\Sigma t = n$, where t is the total number of observations per category. The formula for the variance of U_{AB} then reduces to

$$n_1 n_2 (n^3 - \Sigma t^3) / \{12 n (n-1)\}.$$

Although the Mann-Whitney test can readily be performed in the analysis of ordered categorical data, it has been little used in medical research. Moses et al (1984) found in a survey of 168 papers published in 1982 in the New England Journal of Medicine, 47 instances of ordered categorical data and not one instance of the use of the Mann-Whitney test or other test making use of the order of the categories.

USE OF THE MANN-WHITNEY TEST IN TUMOUR RESPONSE DATA

Since tumour response data is an example of ordered categorical data, the Mann-Whitney test is a valid test. There are none of the disadvantages of the simple Chi squared test discussed above; firstly, the Mann-Whitney test is applied to the whole contingency table without combination of categories (however small the expected frequencies might be) and secondly, the test makes use of the order of the categories. However, in the survey described in section 2.2, only one paper was found where the Mann-Whitney test was applied to tumour response data. Since the Mann-Whitney test appears to be underused in the analysis of data from ordered categories, a worked example is worthwhile.

WORKED EXAMPLE (USING TUMOUR RESPONSE DATA) OF THE MANN-WHITNEY TEST FOR ORDERED CATEGORICAL DATA

The data of Table 2.9 will be used for this worked example. This is taken from a trial of two chemotherapy regimes in advanced breast cancer. The calculations are shown in Table 2.10 and below. In outline, the procedure is to count for each patient given treatment A (here VAC chemotherapy) how many patients given Treatment B (VNC chemotherapy) had a better outcome in terms of response.

Consider first the 9 patients given treatment A who had a CR (Table 2.10). There were no patients given treatment B who had a better outcome but the 4 patients given treatment B with a CR had an equal outcome (scored ½). The contribution to U_{AB} from each of these 9 patients is ½ × 4, and the contribution from all 9 is

$$9 \times ½ \times 4 = 18.$$

Table 2.9. Example tumour response data (Green et al, in preparation). Numbers of patients in each response category for each treatment.

Response	Treatment A VAC chemotherapy	Treatment B VNC chemotherapy	Total
Complete response (CR)	9	4	13
Partial response (PR)	20	19	39
No change (NC)	14	17	31
Progressive disease (PD)	4	14	18
Total	47	54	101

Consider now the 20 patients given Treatment A who had a PR. The 4 patients given Treatment B who had a CR had a better outcome and the 19 with a PR had an equal outcome. The contribution of each of the 20 patients to U_{AB} is 4 + ½19, and the total contribution is

$$20 \times (4 + ½19) = 270.$$

Consider next the 14 patients given Treatment A who had NC. The 4 patients given Treatment B who had a CR had a better outcome as did the 19 with a PR, and the 17 patients with no change had an equal outcome. The contribution of each of the 14 patients to U_{AB} is 4 + 19 + ½17, and the total contribution is

$$14 \times (4 + 19 + ½17) = 441.$$

The total contribution from the 4 patients given Treatment A with PD is similarly

$$4 \times \{4(\text{CR}) + 19(\text{PR}) + 17(\text{NC}) + ½14(\text{PD})\} = 188.$$

Addition of these 4 contributions to U_{AB} gives the total value, 917. The number of tied observations in each group t is simply the sum of the numbers of patients given treatments A and B. The quantity Σt^3 needed in calculating the variance is thus calculated in the same Table 2.10.

Table 2.10. Worked example of the Mann-Whitney test in the analysis of ordered categorical data.

	Treatments				
	A (VAC)	B (VNC)	Contribution to U_{AB}	t	t^3
CR	9	4	$9 \times ½ \times 4 = 18$	$9 + 4 = 13$	2197
PR	20	19	$20 \times (4 + ½19) = 270$	$20 + 19 = 39$	59319
NC	14	17	$14 \times (4 + 19 + ½17) = 441$	$14 + 17 = 31$	29791
PD	4	14	$4 \times (4 + 19 + 17 + ½14) = 188$	$4 + 14 = 18$	5832

Totals:

n_1=47 n_2=54 U_{AB} = 917 Σt^3 = 97139

n = 47 + 54 = 101

The expected value of $U_{AB} = ½n_1n_2 = ½\ 47 \times 54 = 1269$

Therefore the difference from the expected value = 1269 - 917 = 352

The variance = $n_1n_2(n^3 - \Sigma t^3) / \{12n(n\text{-}1)\}$

$= 47 \times 54(101^3 - 97139) / (12 \times 101 \times 100) = 19541$

The standard error (SE) = √variance = √19541 = 139.79

Therefore the standardised deviate (the difference from the expected value divided by the SE) = 352 / 139.79 = 2.518

and this is compared with tables of the standard normal distribution e.g. Neave (1981) to give a 2-sided P value of 2 × (1 - 0.99408) = 1.2%.

There is thus strong evidence that Treatment A is better than Treatment B.

For the data of Table 2.6, the value of *U* is 8725, giving a *P* value of 3.0%, which is a similar result to the Chi squared test for trend.

Thus the Mann-Whitney test for ordered categories can be readily done with a hand calculator and is not appreciably more difficult to perform than a Chi squared test. It can also be done with any of the major statistical computing packages; data is entered by allocating values e.g. l, 2, 3 and 4 respectively to the category of response. The choice of values is unimportant (and does not affect the outcome of the test) as long as they are in numerical order. However, it is essential that the statistical program includes an allowance for ties, and not all do so (Altman, 1991).

EXTENSION OF THE MANN-WHITNEY TEST TO THREE OR MORE TREATMENTS: THE KRUSKAL-WALLIS TEST

The Mann-Whitney test can only be used for the comparison of two treatments. If three or more treatments are being compared in a clinical trial, a Kruskal-Wallis test can be applied to the data (Sprent, 1989; Altman, 1991). It is a ranking test exactly equivalent to the Mann-Whitney test, but in a more generalised form. When it is applied to ordered categorical data, a correction for ties is necessary (Sprent, 1989). When applied to two samples, it reduces to a Mann-Whitney test.

REGRESSION MODELS FOR ANALYSIS OF RESPONSE

In a clinical trial with random allocation of treatments, it is quite possible for imbalances to arise in prognostic factors between the treatment groups; for example, one treatment group may by chance contain more patients with advanced stages of disease

than the other(s), and a spurious impression of difference in treatment efficacy may arise. Information on known prognostic factors can be included in regression models, which can allow for imbalances. The standard multivariate regression technique for categorical data is logistic regression; details are given in Altman (1991). If applied to tumour response data, logistic regression requires the categories of response to be collapsed to a dichotomy (e.g. to total response versus non-response), and this results in some loss of information, as already discussed. However, work has also been done on developing models for the analysis of ordered categorical data (McCullagh, 1980; Anderson, 1984; Holtbrugge & Schumacher, 1991). This is a promising area of research for the future. However, in investigating analysis of tumour response, it seems sensible first to try to optimise the univariate method of analysis, and this dissertation will not consider multivariate methods in any more detail.

DISCUSSION

The principal tests that have been considered in this section are summarized in Table 2.11. This emphasizes that different *P* values can be obtained for the same set of data when analyzed by different statistical tests.

Table 2.11. Summary of the statistical tests reviewed with the *P* values of each for analysis of the data of Table 2.6.

Statistical test	*P* value	Comments
Chi squared test of 4 categories of response	4.6%	Loss of information Well known, simple test
Chi squared test of total response rate	13.2%	Greater loss of information Well known, simple test
Chi squared test for trend using linear scale	2.0%	Arbitrary scoring system Uses the order of the categories
t test using linear scale	2.0%	Arbitrary scoring system Uses the order of the categories
Mann-Whitney test	3.0%	Uses the order of the categories

The next section will discuss the general considerations in the choice of a statistical test, but first, it is worthwhile considering what clinical question is being asked by investigators when carrying out clinical trials and assessing tumour response. Although the preceding section showed that the commonest method of comparing

treatments is in terms of their total response rates, the clinical question is not usually one of which treatment has the higher total response rate. The definition of total response is an arbitrary one, and the real question is usually one of which treatment has a higher overall efficacy. The assessment of total response rates is thus just one (but perhaps the simplest) method of comparing treatment efficacies.

There are clear theoretical advantages to the tests that make use of the order of the response categories (the last three of the list of Table 2.11). Recommendations have been made for their use (Moses et al, 1984; Armitage & Berry, 1987; Altman, 1991), but the preceding section found them to be rarely used in analysis of tumour response data. It is unclear why the recommendations are not being followed, but there are a number of possible reasons. Firstly, there may be a need for wider dissemination of the recommendations. Secondly, there may be a failure of communication between the clinicians and statisticians involved in the studies about the nature of tumour response data, i.e. that the categories are ordered; in some of the 81 papers studied, morbidity of treatment was analyzed by a Chi squared test for trend, and yet in the same paper, tumour response was analyzed by the simple Chi squared test. Thirdly, it may be that simply stating an advantage for the tests is not sufficient to justify the slight increase in complexity, and quantification of the advantage is required.

A strong theoretical case can be made for use of the Mann-Whitney test in the analysis of tumour response data since it makes good use of the information available in the data and does not suffer from the arbitrariness of the scoring systems required by the Chi squared test for trend and the *t* test. Comparison of two statistical tests can be done in terms of their relative efficiency and this will be discussed in the following section. The need for quantification of the advantages of the Mann-Whitney test in the analysis of tumour response data is addressed in sections 2.5 and 2.6 in practice and in theory, respectively.

2.4 CONSIDERATIONS IN THE CHOICE OF A STATISTICAL TEST

It has been shown in the preceding section that a number of statistical tests can be applied to tumour response data. It is also true in general that a variety of statistical tests may be applied to any particular kind of data. It is thus necessary to make a choice between the available tests, and there are several considerations in this choice, which will be described. It is not satisfactory simply to apply several or all of the available tests and select the test reported on the basis of minimizing (or maximizing) the *P* value; this may be misleading and lead to errors in the application of the trial results to clinical practice.

SIMPLICITY

Statistical errors are common in medical journals (Altman and Bland, 1991). There is thus an advantage to a statistical method if it is a simple one, since errors are less likely to be made, and results can more easily be checked by a reader. There is also an advantage if a clinician is able or has been able at some time to perform the statistical test by hand. He is in a better position to appreciate the conclusions of the statistical analysis and may well be more likely to apply the conclusions of the study to his own clinical practice. The Chi squared test for trend is an example of a test where the formula used is not readily understandable; for many, it will be little more than a recipe, and this is a point against its use.

A more complex statistical method may obscure the poor quality of some of the clinical data collected. For example, the Kaplan-Meier and logrank survival analyses can readily be applied to data where many patients have been lost to follow-up, but the level of follow-up may not be made clear and this may allow bias to enter the results undetected.

AVAILABILITY OF COMPUTER SOFTWARE

Increasingly, statistical analysis is being done by computer and this has advantages of speed and generally increased reliability. Easy availability of computer software for a particular test is thus an advantage. However, not all software is free from errors (Altman, 1991) and there may be problems if the software is used by people not familiar with its operation or interpretation.

RELATIVE EFFICIENCY

Two different statistical tests applied to the same set of data will often give different results - if only slightly. This has been shown in the previous section where the *P* values from the different tests ranged from 2.0% to 13.2%. If one test is consistently better than another in the sense of being more likely to detect a difference when it exists, then it is said to be more efficient. It can do this by making more use of the information contained in the data.

The concept of efficiency is central to this dissertation. Two statistical tests (denoted test A and test B) can be compared by their relative efficiency which is denoted by $e_{A,B}$. If the two tests are equally efficient, then the efficiency of test A relative to test B $e_{A,B}$ will be 1.0. If test A is the more efficient, then $e_{A,B}$ will be greater than 1.0, and if test B is the more efficient then $e_{A,B}$ will be less than 1.0. Relative efficiency can be measured in terms of the ratio of the numbers of patients required by the two tests to answer the same clinical question. A formal definition can be given as follows (Pratt and Gibbons, 1981).

> If test A requires n_A observations to detect a difference δ at a significance level α with power $1-\beta$ and test B requires n_B observations to do the same, then the relative efficiency of test A relative to test B $e_{A,B}$ is given by the ratio n_B/n_A.

In general the relative efficiency of two tests will depend principally on the *type* of difference to be detected i.e. on the alternative hypothesis; it is much less dependent on the significance level a, the power 1- β and the size of the difference δ.

The consequence of using a less efficient test when a more efficient test is available can be explained in terms of the numbers of patients to be entered in clinical trials. As an example, if the relative efficiency $e_{A,B}$ is 1.5, then statistical test A requires only 1/1.5 as many patients, i.e. 2/3 as many patients to be entered in the clinical trial as test B. Since many clinical trials are limited by problems of patient recruitment, making maximum use of the available clinical information is important and can be achieved by selecting the statistical test which has the greatest efficiency.

ESTIMATION OF RELATIVE EFFICIENCY

A number of methods can be used for estimating the efficiency of one test relative to another:

- By comparison of the power functions for mathematical models of the data
- By computer simulation for models of data
- By performing the statistical tests on a number of typical sets of data and comparing the results in three ways (only the first of these has been previously described - the other two are new methods proposed here)
    - by counting the numbers of significant results
    - by a scatter plot of the *z*-values
    - from the ratios of the *z*-values.

In the first method, standard techniques are used to calculate the relation between the power of each statistical test and the number of patients entered (i.e. the power function). From the two power functions, the relative efficiency is calculated. The

limitation of this method is that assumptions are needed concerning what effects of treatment are to be expected - in statistical terminology, what alternative hypothesis is assumed. (For example, for tumour response data, it is necessary to make assumptions of the relation between CR, PR, NC and PD rates and formulate a model or models of treatment effect. This is done in this dissertation in the modelling in section 2.6.) It will then be necessary to test whether the assumed model is valid before giving weight to estimates of relative efficiency which have been derived from it. This method has been extensively used to calculate various values of relative efficiency of a number of statistical tests using different alternative hypotheses.

The second technique has been used where there are difficulties in evaluating the power functions of the two tests; a computer is used to repeatedly simulate the problem, typically 1000 times. This has been used to evaluate the power function of the logrank test (Breslow, 1984). Like the first technique, this is limited by the need to assume some model of treatment effect.

In the third set of methods, a number of sets of data are collected and analyzed using each of the two statistical tests. It is necessary to select sets of data that are alike in some way. No two sets of data are the same, but it will often be possible to define sets of data with features in common, e.g. tumour response data. The conclusions can then apply to similar sets of data, just as the conclusions from a study of patients with a condition can be applied to a future patient with a similar condition. This third set of methods has been little used, and it appears that only the first of them (counting the number of significant results) has ever been used previously. Why this approach has been little used in the past is not clear, but it may reflect the very large number of types of data to which statistical tests can be applied.

Counts of results significant at 5%

By simply counting the number of results significant at the arbitrary cutoff of 5% (or other significance level), a qualitative estimate of relative efficiency may be obtained. The statistical test with the higher count of significant results is likely to be the more efficient. Mann and Whitney (1947) used this technique to demonstrate the superiority of their test over an alternative. However, it gives no quantitative estimate of relative efficiency and it overemphasizes the arbitrary cutoff (e.g. 5%) or cutoffs used.

Scatter plot of *z*-values

The simple method just described can be extended by plotting the sets of data as a scatter diagram, plotting the *z*-value with corresponding *P* value for each set of data for one test against the *z*-value with corresponding *P* value for the same set of data for the other test. (The *z*-value from a statistical test is the equivalent value from a standard normal distribution; for example, a *z*-value of 1.96 is equivalent to a *P* value of 5%. A further description of *z*-values is given in Appendix B.)

As an example, Fig. 2.1 shows this technique for a single set of data, namely that of Table 2.6 (already extensively used as an example). In the previous section it was shown that the *P* value for this set of data analyzed by the Chi squared test of the total response rate was 13.2% (equivalent *z*-value 1.51) and the *P* value for a Mann-Whitney analysis was 3.0% (equivalent *z*-value 2.18). These values are represented by the cross on Fig. 2.1.

The sloping line of Fig. 2.1 is the line of equality between the tests. If the two tests are exactly equivalent, all the points will lie on the line of equality. If the tests are roughly equivalent, and there are only minor differences between the tests, the points will be clustered around this line. If a difference in efficiency exists, the points will have a

tendency to lie either above or below the line of equality, depending on which test is the more efficient.

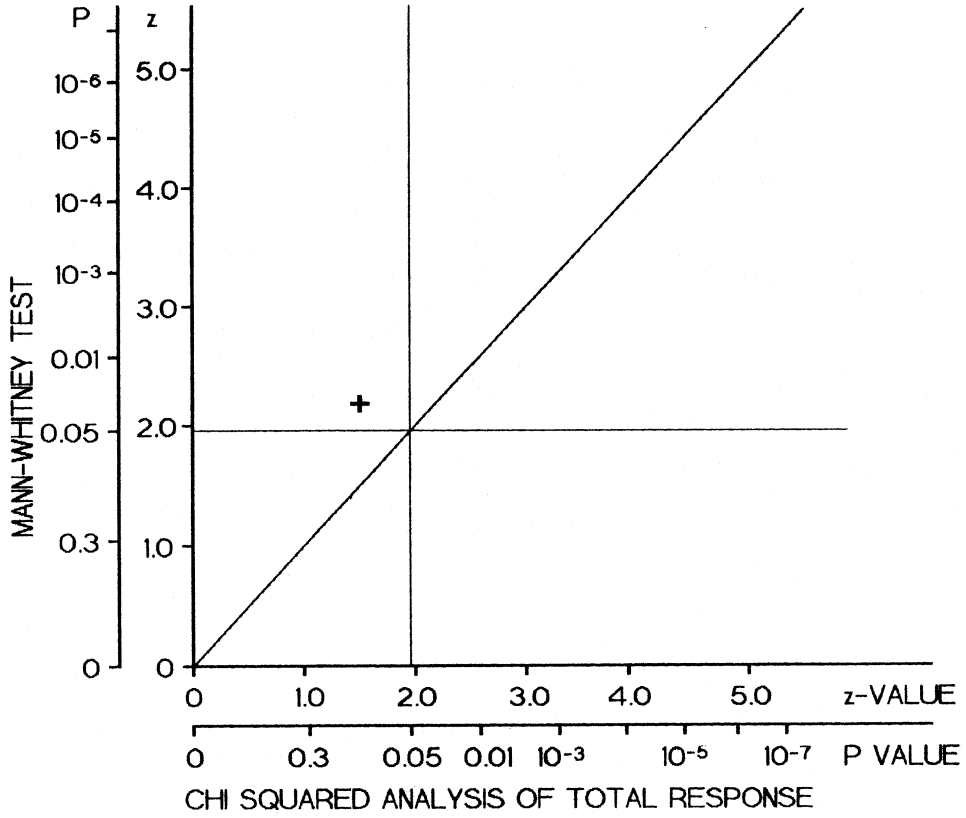


Fig. 2.1. Representation of the *z*-values (and corresponding *P* values) of the data of Table 2.6 analyzed by both the Mann-Whitney test and the Chi squared test of total response. The sloping line is a line of equality between the tests. The vertical and horizontal lines demarcate results significant at 5% from those which are not.

The ratio of the *z*-values

It is shown in Appendix D that an estimate of the relative efficiency for two statistical tests can be given by the square of the ratio of the *z*-values from analysis of the same set of data by the two tests. In the above example, the estimate for the efficiency of the Mann-Whitney test relative to the Chi squared test of the total response rate is $(2.18/1.51)^2$, i.e. 2.08. Furthermore, when a large number of data sets are analyzed by

each of two tests, giving a z-value denoted z_A for test A for each set of data, and a z-value denoted z_B for test B, an overall estimate of the relative efficiency $e_{A,B}$ can be obtained from the median value of $(z_A/z_B)^2$. This is a new result and is applied in sections 2.5 and 3.4 of this dissertation.

2.5 STATISTICAL ANALYSIS OF TUMOUR RESPONSE DATA: COMPARISON OF THE CHI SQUARED TEST WITH THE MANN-WHITNEY TEST

INTRODUCTION

As discussed in section 2.1, response of solid tumours to cancer treatment is usually assessed by allocation to one of 4 categories, namely complete response (CR), partial response (PR), no change (NC) and progressive disease (PD).

There are several statistical methods that can be applied to the analysis of the tumour response data from a clinical trial comparing two or more treatments, and these were discussed in section 2.3. Section 2.2 showed that a variety of different statistical techniques are being used in practice. The most commonly used method is to collapse the 4 categories to the 2 categories of responders (complete plus partial responses) and non-responders (no change plus progressive disease); and then assess the combined table by either the Chi squared test or Fisher's exact test (the latter being used when the expected frequencies in some of the categories are small).

A second commonly used method is to retain the original 4 categories of classification of response and to perform a Chi squared test on the 4×2 (response versus treatment) contingency table, or on the $4 \times a$ contingency table where a is the number of treatments being compared, i.e. the number of arms of the trial. As discussed in section 2.3, both these applications of the Chi squared test can be criticised because full use is not made of the information available in the clinical data. A third method, the Mann-Whitney test has theoretical advantages (which were discussed in section 2.3). However, section 2.2 showed that it is rarely used in the analysis of tumour response data in Oncology.

Whether these theoretical arguments are of any practical importance is addressed in the current section. A total of 74 sets of tumour response data were all analyzed by each of these 3 methods and the derived *P* values were compared. In section 2.3, a comparison was made of these 3 (and other) methods on the set of tumour response data shown in Table 2.6 and the current section extends this to a larger number.

The results of the comparison of statistical tests are shown graphically by plots of the *z*- and associated *P* values, as discussed in section 2.4. The results are also used to estimate the relative efficiencies of the 3 tests; relative efficiency was discussed in section 2.4, and can be briefly described as follows. If the relative efficiency $e_{A,B}$ for test A relative to test B is greater than 1.0, then fewer patients are needed in a clinical trial if analyzed by statistical test A instead of test B; and the ratio of the numbers of patients needed is $1/e_{A,B}$.

The Chi squared test for trend has also been advocated for use in the analysis of ordered categorical data in general (Armitage & Berry, 1987) and in the analysis of tumour response data in particular (Bartolucci, 1984). However, because of the difficulties in devising a scoring system, and because of the equivalence to the Mann-Whitney test on certain scoring systems, as discussed in section 2.3, the Chi squared test for trend is not included in this comparison of techniques. A further worthwhile comparison would be with regression methods of tumour response data, taking into account the prognostic factors. These may improve on the reliability of analyses done without prognostic factor information, but to include these methods in the work of this section would require the full data sets including prognostic factor information.

METHODS

By a review of seven specialist Oncology journals, 81 articles were identified, published between July 1988 and December 1990 inclusive, where tumour response was compared between two or more treatments in a randomised trial. These articles are the same series as were analyzed in the survey described in section 2.2 and further details are given there.

In 11 papers, response data were reported by allocation into only 2 categories (although in several cases it had been collected according to a 3 or 4 category system of classification). These 11 papers were therefore not suitable for re-analysis since for 2 categories, a Mann-Whitney test is almost exactly identical to a Chi squared test (Armitage & Berry, 1987). Four of the remaining 70 papers contained two separate sets of data, giving a total of 74 sets of data for re-analysis.

Several papers excluded some patients as being non-evaluable. Some authorities urge that patients unable to complete therapy should be judged as failures of treatment (Hoogstraten, 1984), but for the purposes of this comparison, the rates were used as reported.

Although most papers reported tumour response rates according to the standard 4 categories (Miller et al 1981), 7 papers used an additional category such as "minimal response" (a subdivision of no change which, for example, was defined as a decrease in the sum of the products of diameters of masses of between 25% and 50%). Use of such categories can be criticised on the basis that distinction from no change can be very unreliable (Moertel & Hanley, 1976). Since the aim of this section is to compare use of 2 categories of response with use of the standard 4 categories, any data reported in a

category of minimal response was combined with that in the category of no change for this re-analysis.

In 18 papers, the results were reported by allocation to 3 rather than 4 categories. In 4 of these papers, the clinical situation was such that very few if any patients would be expected in the 4th category (the tumour was very sensitive or very insensitive to the treatments given). In the remaining 14 papers, it would be expected that all the standard 4 categories of response would have contained appreciable numbers of patients. All of these 18 papers were included in the main analyses.

STATISTICAL ANALYSES

Three statistical approaches were used for re-analysis of each of the 74 sets of data. Firstly, the 4 categories of classification were reduced to 2 by combination of the categories of complete response with partial response and by combination of the categories of no change with progressive disease to give the total responders and total non-responders respectively. Then the 2 categories were compared between the 2 or more treatments by a Chi squared test or (in 11 cases) by Fisher's exact test, when one or more of the expected values was less than 5, according to the recommendations of Armitage & Berry (1987). The Chi squared test in all 2 × 2 tables included a correction for continuity as recommended by Armitage & Berry (1987) and Altman (1991).

In the second approach, the response versus treatment contingency tables were analyzed by a Chi squared test for heterogeneity using the standard 4 categories of classification or by combining categories to ensure that most or all expected frequencies were at least 5 according to the recommendations of Armitage & Berry (1987). In 20 sets of data, in order to achieve expected frequencies of at least 5, it was necessary to combine categories in pairs so that the 4 categories reduced to 2 categories

- the intended Chi squared test of 4 categories of classification in these cases thus became identical to the Chi squared test of the total response rate. These 20 sets of data were included in all the analyses. Thus the comparison between the tests was on an "intention to analyze" basis, analogous to "intention to treat" analyses in clinical trials.

The third approach was to analyze the response versus treatment contingency tables by a Mann-Whitney test using an allowance for ties as recommended by Armitage & Berry (1987). If there were more than two treatments compared in the reported trial, a Kruskal-Wallis test was used instead (this is the direct extension of the Mann-Whitney test when the number of treatments is more than two).

Except for the exact test which was done by hand, all calculations were done using a program which was written for the purpose in Mallard BASIC extended by Lightening Basic software (CP Software, UK), and run on an Amstrad PCW 8512 personal computer. The program was verified against 16 published worked examples (10 examples of a Chi squared analysis, 5 examples of a Mann-Whitney analysis for ordered categorical data, and one example of a Kruskal-Wallis analysis) in 6 publications (Neave, 1981; Moses et al, 1984; Armitage & Berry, 1987; Sprent, 1989; Morton et al, 1990; Altman, 1991).

For each of these statistical approaches, the test statistic was converted to a 2-sided z-value (the standardised normal deviate). For a Chi squared test on a 2 × 2 table, the z-value was obtained as the square root of $\Sigma(O\text{-}E)^2/E$ since the χ_1^2 distribution is the square of the standard normal distribution. For a Chi squared test on a larger table, where the number of degrees of freedom v is greater than one, the z-value was obtained by linear interpolation using the empirical approximately linear relationship between χ_v^2 and z^2 together with tabulated values of χ_v^2 at particular levels of significance (see Appendix E). The inaccuracy introduced to the value of z by this linear interpolation was

estimated as a maximum of 0.04 (equivalent to a maximum inaccuracy of 0.5% for any *z*-value greater than 1.96). For a Mann-Whitney test the *z*-value was obtained by the standard formulae for *U* and its variance with allowance for ties (Armitage & Berry, 1987) as discussed in section 2.3. In the cases of 3 sets of data, the number of patients per treatment in the study was less than 16 and the normal approximation for the distribution of *U* may not have been accurate (Armitage & Berry, 1987). However, no great changes to the conclusions of this section would result if these 3 sets of data were excluded.

The three statistical approaches were compared in three pairs by scatter plots of the *z*-values obtained (an example is shown in the preceding section). They were also compared in terms of their relative efficiency by the method derived in Appendix D, which may be summarized as follows. Firstly, a number of data sets are collected from a particular field of research and each data set is analyzed by each of two statistical tests A and B. The data sets that are significant at 5% are selected and the relative efficiency of test A relative to test B, $e_{A,B}$, can then be estimated by the median value of $(z_A/z_B)^2$ from the significant sets of data, where z_A is the *z*-value obtained from test A, and z_B is the *z*-value obtained from test B. When only two tests are being compared, it seems reasonable to include data sets significant at 5% by *either* of the two tests (as opposed to only those data sets significant by *both* tests). This raises a problem in the current section, which is a comparison of *three* statistical approaches, since this policy would result in the selection of three different series of data sets for the three pairwise comparisons. It was felt preferable to obtain a single series of data sets, and consequently the data sets were selected for estimation of the relative efficiency if the result was significant by the test which performed best on the preliminary analyses, which was the Mann-Whitney test.

RESULTS

NUMBERS OF SIGNIFICANT RESULTS AT 5%

Of the 74 sets of data analyzed, 19 sets were significant at 5% by a Chi squared test of the total response rate, 23 were significant by a Chi squared test of 4 categories of response and 27 were significant by a Mann-Whitney test. Only 3 or 4 significant results would have been expected by chance (i.e. 1 in 20).

All of the 19 sets significant by a Chi squared test of total response were significant by both a Chi squared test of 4 categories of response and by a Mann-Whitney test; and thus the numbers of additional significant sets by these two tests were 4 and 8 respectively. Of the 23 sets significant by a Chi squared test of 4 categories of response, all but one were also significant by a Mann-Whitney test; a further 5 were significant only by the latter test. These points will be illustrated by the figures of the following subsection.

RELATIONSHIPS BETWEEN *z*-VALUES AND BETWEEN *P* VALUES

Figure 2.2 plots the relationship between the *z*-values by the Mann-Whitney test and by the Chi squared test of total response. Each of the points represents one of the 74 sets of data, plotting the *z*-value for the Mann-Whitney test against the *z*-value for the Chi squared test of total response. The corresponding *P* values are also given on each axis.

The sloping line is a line of equality between the tests. If there were only minor differences between the tests, the points would be clustered around this line. However, there is a clear tendency for the points to lie above this line i.e. for the *z*-values by the Mann-Whitney test to be greater and the *P* values to be more extreme.

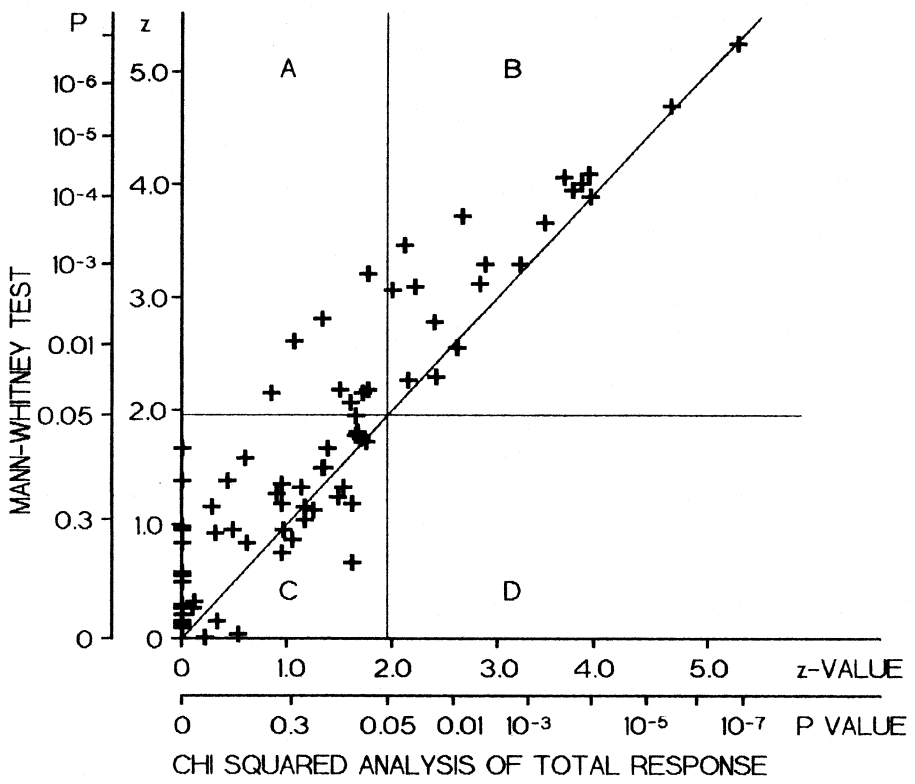


Figure 2.2. Comparison of the *z*-values (and the corresponding *P* values) for 74 sets of tumour response data analyzed by both the Mann-Whitney test and the Chi squared test of total response (Fisher's exact test was substituted for the Chi squared test in 11 cases because of small numbers). The sloping line is a line of equality between the tests. The vertical and horizontal lines demarcate results significant at 5% for each of the two tests from those which are not.

The points can be considered in 4 quadrants demarcated by *z*-values of 1.960 (equivalent to a two-sided *P* value of 5%), and denoted by zones A, B, C and D in Fig. 2.2. In zone B are the 19 points representing the sets of data significant by both the Chi squared test of total response and the Mann-Whitney test. The tendency for the points to lie above the line of equality is especially marked here; 15 of the 19 do so. This scatter around the line is significantly different from an equal distribution at $P = 2\%$ by the sign test. This is good evidence for the superiority of the Mann-Whitney test over the Chi squared test of total response.

In zone A are the 8 points representing the 8 sets of data significant at 5% by the Mann-Whitney test but not by the Chi squared test of total response. There are no points in zone D illustrating the lack of converse sets of data. In zone C are points representing sets of data not significant at 5% by either test. These are roughly evenly distributed about the line of equality except for a number of *z*-values of zero for the Chi squared test, reflecting the effect of the continuity correction of the Chi squared test which will give a Chi squared value of zero whenever *O*-*E* is less than 0.5.

Figure 2.3 in the same way plots the *z*-values for the Chi squared test of 4 categories of response against the *z*-values for the Chi squared test of the total response rate. There is some tendency for the points to lie above the line of equality and in the right upper quadrant, 9 lie above the line, 3 lie below it and 7 lie on the line (for these 7 points, the Chi squared test on 4 categories reduced to a Chi squared test of total response because small numbers forced amalgamation of categories). There is thus some evidence that the distribution differs from an even distribution ($P = 15\%$ by the sign test).

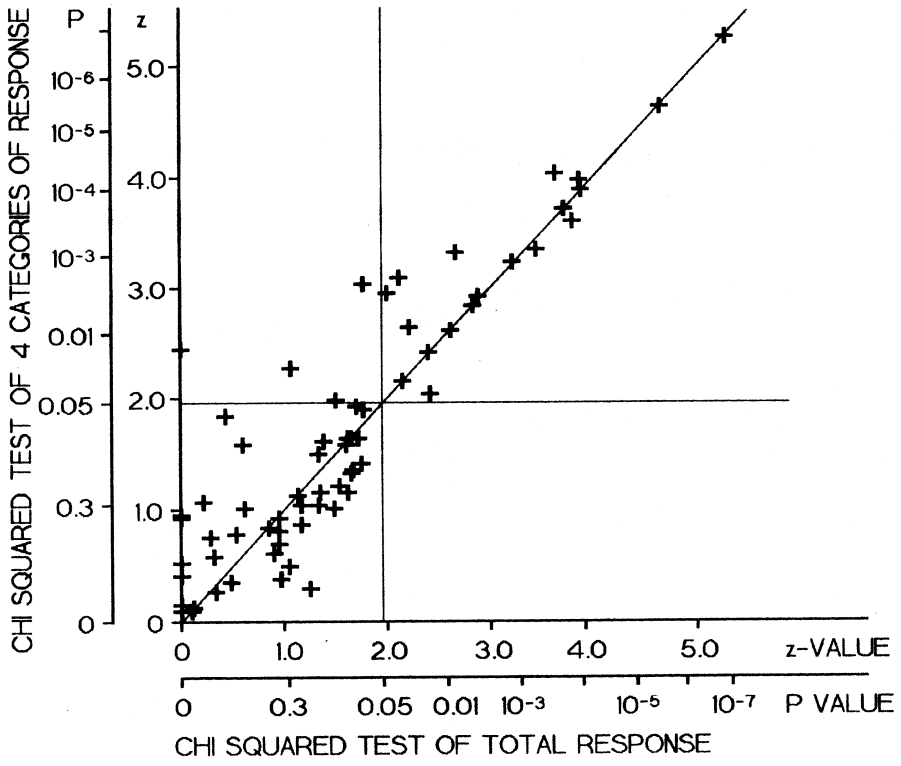


Figure 2.3. Comparison of the *z*-values (and the corresponding *P* values) for 74 sets of tumour response data analyzed by both the Chi squared test of 4 categories of response and the Chi squared test of total response (or Fisher's exact test in 11 cases). The sloping line is a line of equality between the tests. The vertical and horizontal lines demarcate results significant at 5% for each of the two tests from those which are not.

Figure 2.4 is the third comparison of tests, plotting the *z*-values for the Mann-Whitney test against the *z*-values for the Chi squared test for 4 categories of response. Again, there is a tendency for points to lie above the line of equality and in the right upper quadrant, 19 of the 22 points lie above the line ($P = 0.1\%$ by the sign test). The figure shows the one set of data which was significant by a Chi squared test of 4 categories of response but not by the Mann-Whitney test. On inspection of the trial from which this data set was taken, there were no special features to indicate why this was so, and it is concluded that the cause is merely random variability between data sets.

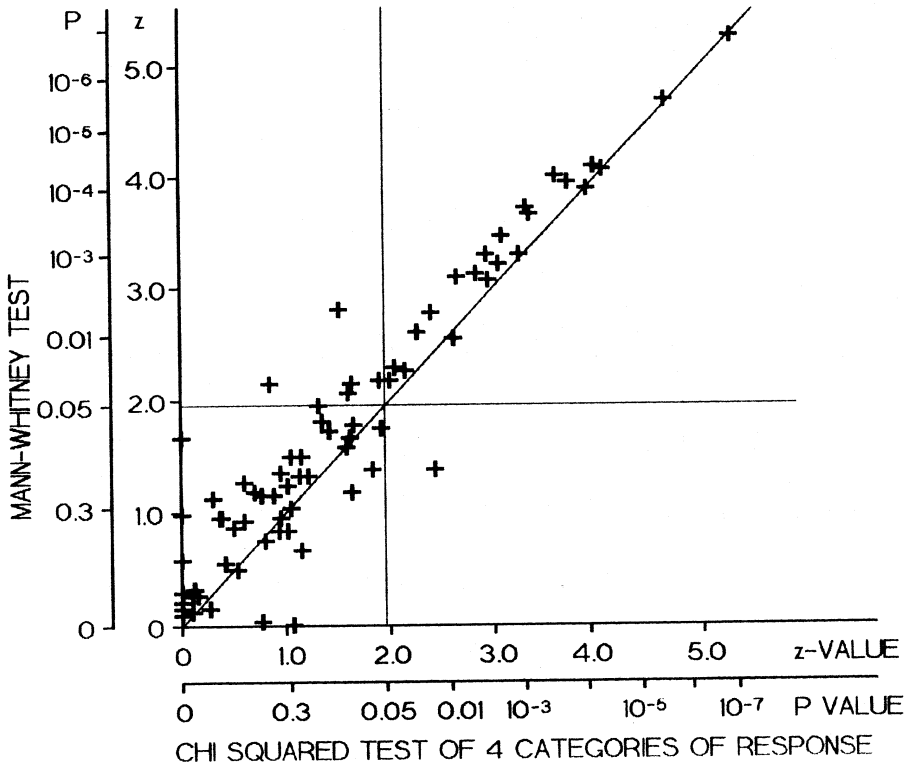


Figure 2.4. Comparison of the *z*-values (and the corresponding *P* values) for 74 sets of tumour response data analyzed by both the Mann-Whitney test and the Chi squared test of 4 categories of response. The sloping line is a line of equality between the tests. The vertical and horizontal lines demarcate results significant at 5% for each of the two tests from those which are not.

ESTIMATION OF RELATIVE EFFICIENCY

Since there is good evidence from the results already presented that the Mann-Whitney test is more efficient than either of the applications of the Chi squared test, the relative efficiencies of the three tests were estimated from the 27 sets of data significant at 5% by the Mann-Whitney test.

The first comparison is of the Mann-Whitney test with the Chi squared test of the total response rate. The relative efficiency is estimated from the median value of the

square of the ratio of the *z*-values (Appendix D) giving an estimate of the efficiency of the Mann-Whitney test relative to the Chi squared of the total response rate of 1.32 (95% confidence interval 1.11 to 1.94).

In the same way, an estimate for the efficiency of the Chi squared test of 4 categories of response relative to the Chi squared test of total response was 1.01 (95% confidence interval 1.00 to 1.31); and the efficiency of the Mann-Whitney test relative to a Chi squared test of 4 categories of response was 1.21 (95% confidence interval 1.09 to 1.32).

EFFECTS ON THE CONCLUSIONS OF THE STUDIES

As noted above, there were 8 sets of data (reported in 8 papers) containing results which were significant at 5% by a Mann-Whitney test but not significant by a Chi squared test of the total response rate. As detailed in Table 2.12, one paper reported no statistical analysis, and in a second, it was unclear. Two papers reported non-significant analyses, and one paper incorrectly calculated a significant result. Only 3 of the 8 papers (data sets numbered 36, 47, and 49) correctly reported significant statistical analyses; one reported a Chi squared test of 4 categories of response, the second reported a Chi squared test of the total response rate after adjustment for prognostic factors by multivariate analysis, and the third reported both a Mann-Whitney test and a Chi squared test of the dichotomy of CR versus absence of CR. In all these 3 papers, other analyses of tumour response not significant at 5% were also done and the overall conclusions were equivocal. Thus in all of these 8 papers, the overall conclusion concerning response would have been changed if the Mann-Whitney test had been used as the principle method of analysis.

Table 2.12 Details of the 8 data sets significant at 5% by the Mann-Whitney test but not by the Chi squared test of the total response rate. Three of the papers reported several analyses. Data set number 36 is the one shown in Table 2.6.

Data set	*P* value by		Reported in publication		
	MW test	Chi squared test*	*P* value	Analysis done	Conclusion given
26	3.2%	40.0%*	0.39%**	Chi squared 4 cats.	sig. higher PR and SD
28	0.5%	18.5%*	none	none	objective response better
36	3.0%	13.2%	4.7%	Chi squared 4 cats.	(11% difference in response rate, 95% CI ±13%) Overall equivocal
47	2.9%	7.6%	6%	Chi squared response	approached stat. sig.
			5%	multivariate analysis	reached stat. sig.
49	0.9%	28.9%	NS	Chi squared response	Overall: CR rates sig. different
			< 5%	Chi squared of CR rate	
			< 0.1%	MW test	
64	3.1%	8.5%	18%	Chi squared response	no sig. difference
65	4.0%	11.0%*	none	unclear	equivocal
66	0.1%	7.6%	7%	Chi squared response	... appears less active ...

Abbreviations: MW = Mann-Whitney, cats. = categories, sig. = significant(ly), CI = confidence interval, stat. = statistical, NS = not significant.
* *P* values marked with an asterisk were derived by Fisher's exact test in place of the Chi squared test as described in the text.
** This reported *P* value of 0.39% appears to be an error through inappropriate application of a Chi squared test to the entire contingency table where the counts of PD, NC, PR, and CR were 3, 5, 0, 0 and 0, 5, 2, 0 respectively in the two treatment groups.
The discrepancy in data set numbered 47 between the calculated *P* value by the Chi squared test of response (7.6%) and the published value (6%) is explained if the published value was obtained without a continuity correction. The discrepancies between the calculated and published *P* values in data set 49 by the Mann-Whitney test and in data set 64 by the Chi squared test cannot be explained other than by errors or misprints in the publication.

DISCUSSION

The sets of data used in this comparative re-analysis include a range of tumour types and treatment comparisons, as discussed in section 2.2, and may be taken to some extent as representative of tumour response data in the literature.

The simplest method of comparison of statistical tests is by a count of results significant at 5% in analysis of typical sets of data. This has historical justification (Mann & Whitney, 1947), and assumes that in at least some of the sets of data, the null hypothesis was not true, i.e. there was a real difference in treatment effect. Using this method, the most efficient statistical test for the tumour response data was the Mann-Whitney test with 27 significant results, the Chi squared test of 4 categories of response was less efficient with 23 significant results and the Chi squared test of the total response rate was least efficient with 19 significant results. These findings are in line with the theoretical arguments based on how much of the available information is used by the tests, discussed in section 2.3.

However, a count of results significant at 5% overemphasizes the 5% cutoff, and a better comparison of the tests is by a plot of all the *z*-values as in Figs 2.2 to 2.4. For example, Fig. 2.2 shows that there is a general tendency for *z*-values (and therefore *P* values) to be more extreme for a Mann-Whitney test than a Chi squared test of the total response rate, and this is not limited to around the 5% cutoff. All these three figures show a tendency for the points to lie above the line of equality indicating an advantage for the test plotted as the y axis in each case. This confirms the findings from the simple counts of significant results of the most efficient method of analysis being a Mann-Whitney test, followed by a Chi squared test of 4 categories of response, followed by a Chi squared test of the total response rate.

The advantage of one statistical test relative to another can be quantified by the relative efficiency, which was here estimated from the ratio of the *z*-values of the 27 significant sets of data. The principal finding was that the Mann-Whitney test had an efficiency of 1.32 relative to a Chi squared test of the total response rate (95% confidence interval 1.11 to 1.94). This again implies that the Mann-Whitney test is more efficient than the Chi squared test of the total response rate. The other calculations indicated the efficiency of the Chi squared test of 4 categories of response to be intermediate.

These estimates of relative efficiency are probably underestimates since in 9 of the 27 sets of data significant by the Mann-Whitney test, the results were reported (and available for this comparative re-analysis) in terms of 3 rather than 4 categories. If these 9 sets of data are excluded, a higher estimate is obtained of 1.45 for the relative efficiency of the Mann-Whitney test relative to the Chi squared test of the total response rate. It is unwise to put too much weight on subgroup analyses and so the best estimate for the value of this relative efficiency is probably around 1.4. Exclusion of the 9 sets of data reported by 3 categories gives similar small increases in the other estimates of relative efficiency.

The practical relevance of a relative efficiency of 1.4 will be discussed more fully in section 2.10, but a clear illustration is the 8 papers where the data set was significant at the 5% level by the Mann-Whitney test but not by the Chi squared test of the total response rate. These 8 papers represented 11% of the total number of data sets, and none of them reported unequivocally that the results were significant at the 5% level. It thus seems that treatment differences are currently being under-reported due to suboptimal statistical analysis.

2.6 THE EFFICIENCY OF STATISTICAL ANALYSIS OF TUMOUR RESPONSE DATA: RELATION TO CATEGORIES OF CLASSIFICATION AND MODELS OF TREATMENT EFFECT

INTRODUCTION

A number of statistical techniques can be applied to tumour response data as detailed in section 2.3. The conclusion of that section was that the Mann-Whitney test has theoretical advantages that could make it the preferred choice. However, section 2.2 found that the Mann-Whitney test was rarely used in practice; the test most often used was the Chi squared test after collapsing the 4 standard categories of response to a dichotomy of total response (CR + PR) versus total non-response (NC + PD). There seems to be a need to quantify the advantages of the Mann-Whitney test over a simple Chi squared test. Section 2.5 estimated the advantage of the Mann-Whitney test by re-analysis of a number of published sets of tumour response data, and the current section does so by consideration of a number of theoretical models of tumour response.

The primary aim of the modelling in this section is thus to compare the Mann-Whitney test using 4 categories of response with the Chi squared test using 2 categories of response. In fact a Chi squared test and a Mann-Whitney test can be shown to be identical when there are only 2 categories of classification, provided that the number of observations is large (Armitage & Berry, 1987). The problem therefore reduces to a comparison of classification by 4 categories with classification by 2 categories when all analyses are done by a Mann-Whitney test. The standard way to compare two statistical tests mathematically is in terms of their relative efficiency. The current section compares statistical methods of analysis of tumour response by their asymptotic relative efficiency i.e. the value that the relative efficiency approaches when the number of patients is very large, and the difference in outcome to be detected is very small. The reason for

calculating the relative efficiency under asymptotic conditions is that this simplifies the calculations. The value of the relative efficiency under more practical conditions is often taken to be close to the asymptotic value.

In order to calculate the value of the asymptotic relative efficiency, it is necessary to make some assumptions of the relationships between rates of CR, PR, NC, and PD when the efficacy of treatment changes. Each set of assumptions will form one model of treatment effect.

Some subsidiary aims of the modelling in this section are to compare the standard 4 categories for assessment of response with other systems of assessment (with all analyses done by a Mann-Whitney test); the number of categories of assessment could be increased beyond 4 and the effect of this on the efficiency of analysis is explored, and some work is done on how the categories are best defined (given a fixed number of categories). This section is the most technical in the dissertation (except for the Appendices), but this is unavoidable.

METHODS

NOTATION OF MODELS

Three types of model were considered and are described here mainly by a graphical representation of the interrelation of PD, NC, PR and CR rates (or whichever system of classification is considered).

In biological systems in general, the graph of the proportion of individuals showing an effect when plotted against the intensity of the stimulus is often found to be a sigmoid shape similar to that shown in Fig. 2.5 (Armitage & Berry, 1987).

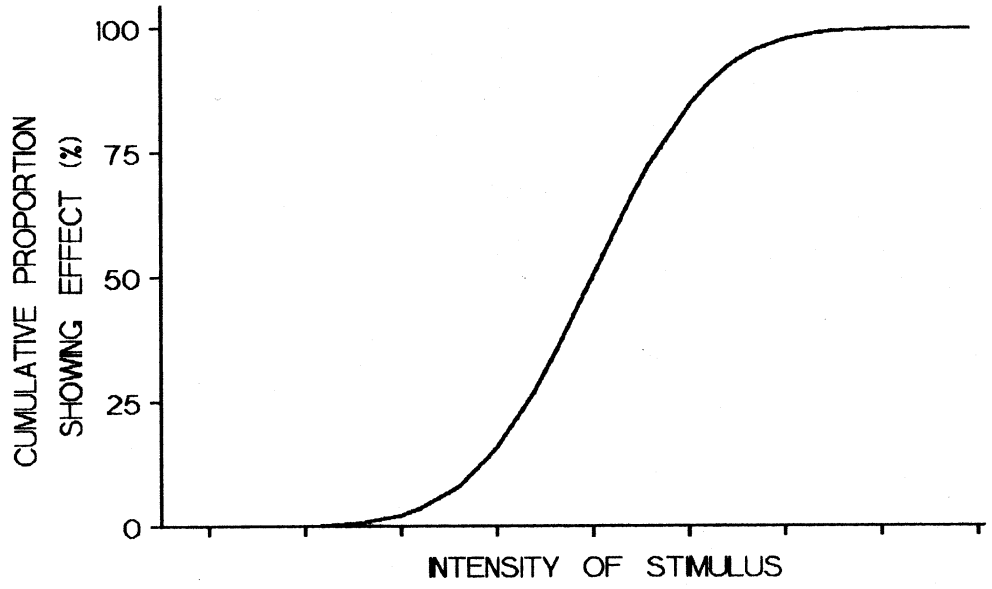


Figure 2.5. A common biological relation between the proportion of individuals showing an effect and the intensity of the stimulus.

In Oncology, the probability of achieving a particular effect on a tumour is often assumed to have a similar sigmoid relation to dose of radiation (Sutton & Hendry, 1985) or dose of drug (Wilkinson & Fox, 1985). There is some clinical evidence supporting this, for example in the control rates of Hodgkin's disease with radiation (Kaplan, 1966). The

graph drawn in Fig. 2.5 is in fact the cumulative normal distribution function (see Appendix B) with the intervals on the x axis each one standard deviation. In this section the proportion of patients having a response (complete or partial) by standard criteria, is represented graphically against treatment efficacy on the x axis; and treatment efficacy E is defined such that the graph of the total response rate plotted against E is the cumulative normal distribution curve. The total response rate is represented by $f_2(E)$, and thus

$$f_2(E) = \Phi(E)$$

($\Phi(E)$ being the cumulative normal distribution function of efficacy E).

Where a model contains 4 categories of response, it is defined by two other functions (represented graphically by the two other curves in Fig. 2.6). The function $f_1(E)$ is the sum of the proportions of no change, partial response and complete response. It will clearly be greater than or equal to $f_2(E)$ for all values of treatment efficacy E and will progressively increase as E increases. The function $f_3(E)$is the proportion of complete responses. It will be less than or equal to $f_2(E)$ but will similarly progressively increase as E increases.

Where more than 4 categories of response are considered in a model, the notation is extended in an obvious way. If the number of categories is k, there will be k-1 boundaries between the categories which will be represented by $f_1(E)$ to $f_{k-1}(E)$.

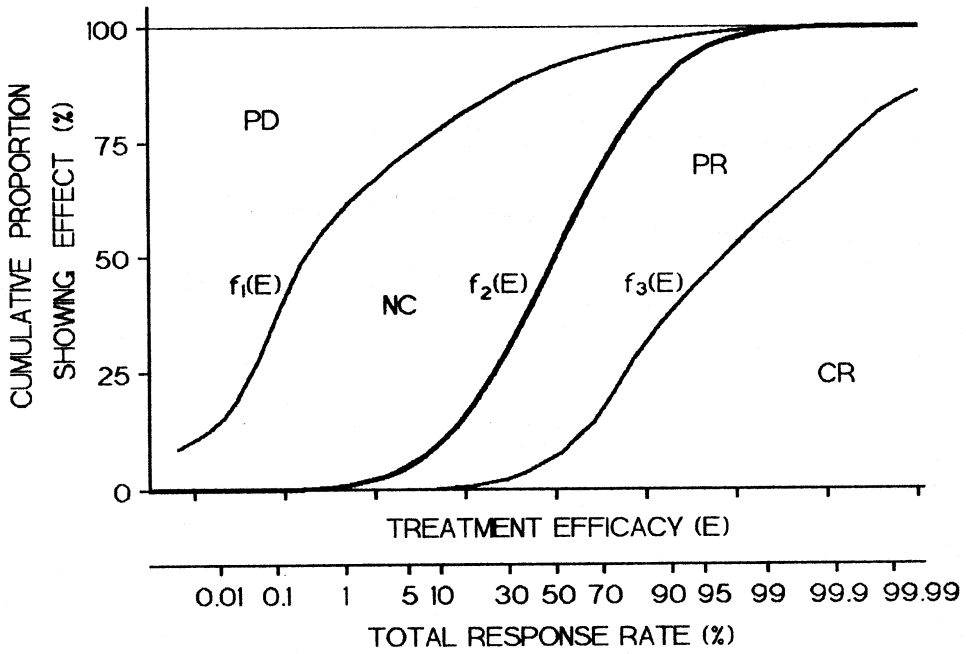


Figure 2.6. Format for specifying models. For 4 categories, $f_2(E)$ is the sum of PR and CR rates, and is fixed for all models as the cumulative normal distribution curve. The scale of the x axis (treatment efficacy, E) is defined such that this is so. A model is specified by $f_1(E)$ (the sum of the NC, PR, and CR rates) and $f_3(E)$ (the CR rate). Thus

the PD rate = $1-f_1(E)$,

the NC rate = $f_1(E) - f_2(E)$,

the PR rate = $f_2(E) - f_3(E)$, and

the CR rate = $f_3(E)$.

For example, at a response rate of 30% in the diagram, the PD rate is 1 - 88% = 12%, the NC rate is 88% - 32% = 56%, the PR rate is 32% - 3% = 29%, and the CR rate is 3%.

The whole of these functions would not normally be demonstrated clinically; very low response rates will not be seen in very active drugs, and very high response rates will not be seen with resistant tumours, where the dose of treatment will be limited by toxicity. However, the whole of these functions potentially exist, and in a clinical trial comparing two treatments, a particular section of the diagram will be relevant (as in Fig. 2.7), corresponding to the range of efficacies of the treatments being compared.

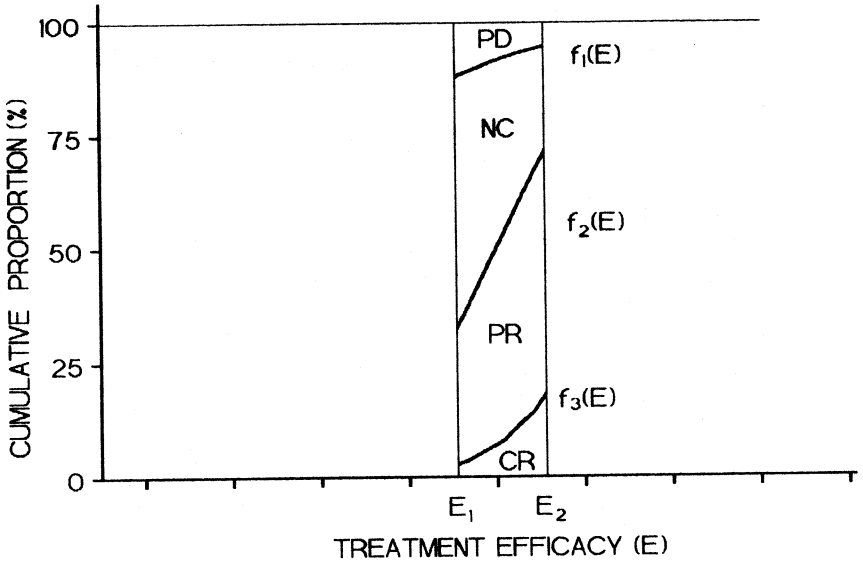


Figure 2.7. The portion of Fig. 2.6 relevant in a clinical trial comparing two treatments with two different efficacies E_1 and E_2.

MODELS EXAMINED AND ANALYSES PERFORMED

Model 1: Lateral shift model

The model is defined in Fig. 2.8 and its legend. The model implies that for any given effect of treatment (such as complete response), there will be a low level of treatment efficacy at which no individuals show the effect, and also a high level of treatment efficacy at which all the individuals show the effect. These seem to be reasonable assumptions; they apply to *in vitro* systems for many effects e.g. the relation between the dose of a drug and cell death. One consequence of this model is that as the total response rate increases, the proportion of responses which are complete response increases, and this seems reasonable *a priori*. The corresponding implication for the non-responses is that as the non-response rate decreases, fewer and fewer of the non-responses are progressive disease.

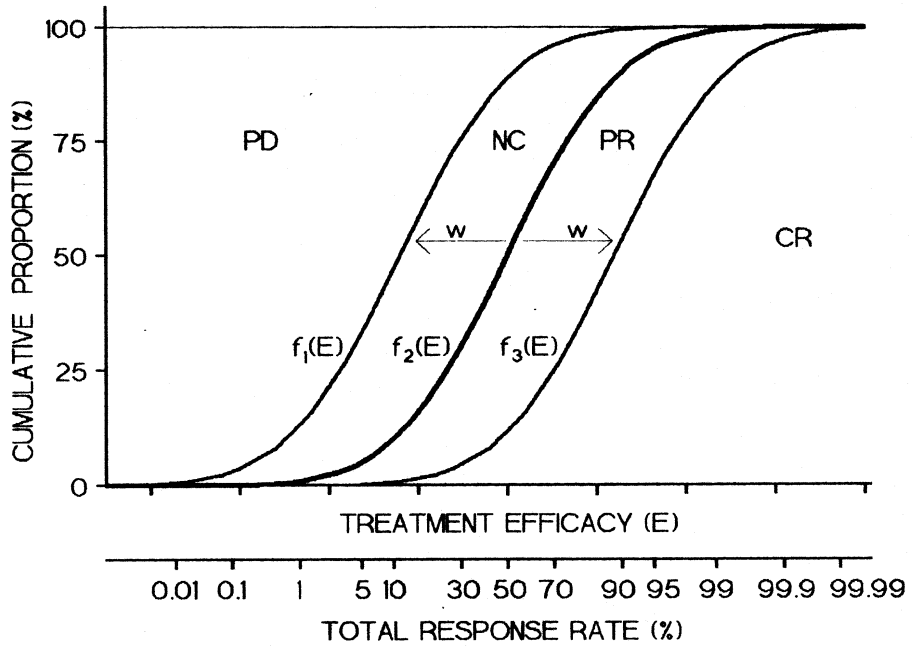


Figure 2.8. Lateral shift model. The total response curve $f_2(E)$ is shifted laterally by w units to the left to give $f_1(E)$ and by w units to the right to give $f_3(E)$. The category "width" is thus measured by w. The units are those of a standardised normal deviate. Thus

$$f_1(E) = \Phi(E+w) \text{ and}$$

$$f_3(E) = \Phi(E-w)$$

In statistical terms, the sigmoid shape of the relation implies that the threshold level of treatment efficacy at which the various individuals show an effect is normally distributed. The similarity of the sigmoid shapes of all of the curves implies a similar effect of treatment at all levels of effect (in statistical terms, the variances of the thresholds of all the effects are equal). This assumption is not an obvious one and requires to be confirmed or refuted by clinical data experimentally. Some evidence in favour of this assumption arises in the following section.

To reduce the number of alternatives considered, the left shift (to give $f_1(E)$) was always kept equal to the right shift (to give $f_3(E)$). Furthermore, when a system of more then 4

categories was considered, the number of categories was kept even, with half of the categories being divisions of the total responses, and half of the categories being divisions of non-response. Clearly very many more alternatives could be considered by relaxing these constraints, but it seems unlikely that the conclusions of this section will be greatly altered by doing so.

Three analyses were done within this model. Firstly, keeping the number of categories of assessment fixed at 4, the width of the central 2 categories (of PR and of NC) *w* was varied from 0.1 standardised deviate units (graphs la and 3a in Fig. 2.9) through intermediate values of w in 0.1 unit intervals (e.g. graphs 1b and 3b) to a maximum of 4.0 units (graphs 1c and 3c).

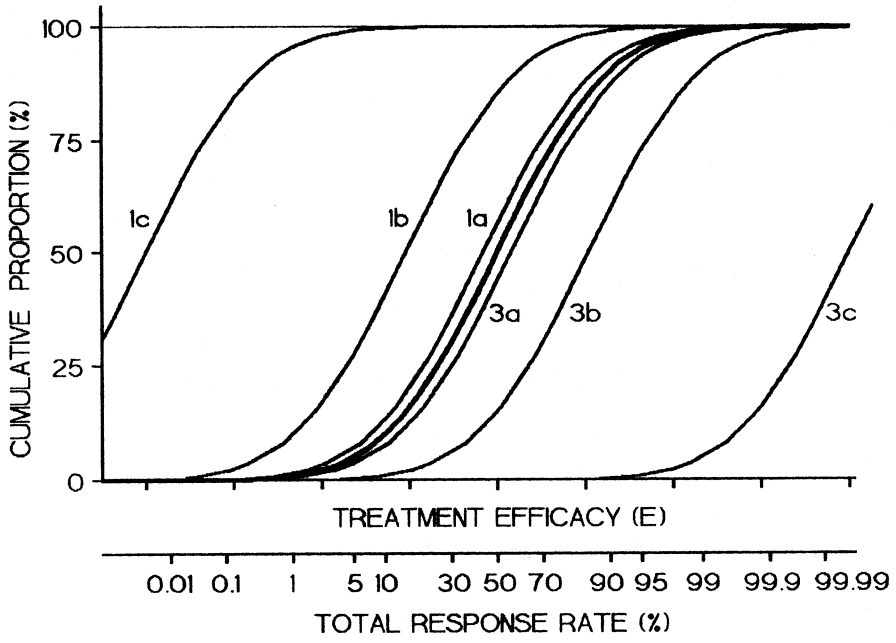


Figure 2.9. First analysis using the lateral shift model. The category width *w* was varied from 0.1 units (graphs la and 3a) to 4.0 units (graphs 1c and 3c) through intermediate values (e.g. 1.0 units in graphs 1b and 3b) in 0.1 unit increments.

For each value of *w*, the asymptotic relative efficiency was calculated for a system of 4 categories of classification relative to a system of 2 categories (CR + PR versus NC + PD) by the method detailed below.

In a second analysis, a similar procedure was carried out with the number of categories of assessment *k* equal to 6, 8, 10 and so on up to a maximum of 32 categories. The width of each of the central *k*-2 categories (*w*) was again varied from a minimum of 0.1 units in steps of 0.1 units to a maximum value governed by a limitation to the value of $w \times (k\text{-}2)$ of 8.4. The data obtained were used to assess the maximum asymptotic efficiency (over the range of values of w considered) for each value of *k*.

In the third analysis, the data derived in the first two analyses was also used to assess the proportions of observations in the different categories at peak asymptotic efficiency.

Model 2: Equal subdivision model

The model is defined in Fig. 2.10 and its legend. The CR and PR rates are each half of the total response rate for all response rates, and the NC and PD rates are each half of the rate of non-response. The asymptotic relative efficiency for 4 categories relative to *2* categories was calculated.

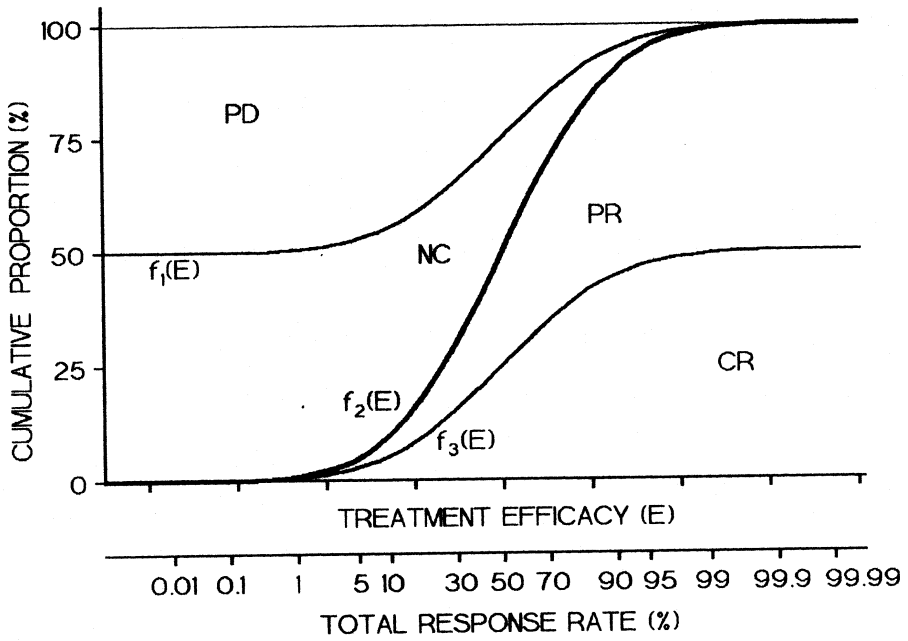


Figure 2.10. Equal subdivision model. The proportion of non-responders 1 - $\Phi(E)$ is divided equally into the PD and NC categories, and the proportion of responders $\Phi(E)$ is divided equally into PR and CR. Thus

$$f_1(E) = \Phi(E)/2 + 0.5 \text{ and}$$

$$f_3(E) = \Phi(E)/2$$

Model 3: Vertical shift model

The model is defined in Fig. 2.11 and its legend. There is a definite threshold in treatment efficacy before any complete responses are seen, and once this is reached, there is a rapid rise in the CR rate with the PR rate remaining constant. The relation of the NC and PD rates is the converse. The asymptotic relative efficiency for a system of 4 categories relative to a system of 2 categories was calculated for the value of p of 0.5.

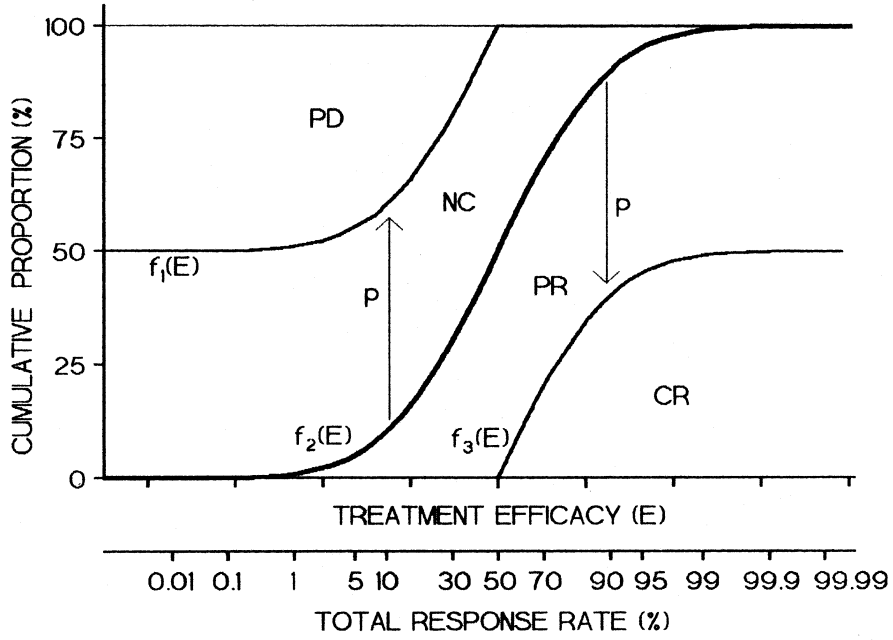


Figure 2.11. Vertical shift model. The total response curve $f_2(E)$ is shifted vertically upwards by proportion p to give $f_1(E)$ and is shifted vertically downwards by proportion p to give $f_3(E)$. Thus

$f_1(E) = \Phi(E) + p$ (subject to a maximum value of 1.0) and

$f_3(E) = \Phi(E) - p$ (subject to a minimum of 0).

CALCULATION OF ASYMPTOTIC RELATIVE EFFICIENCY

Calculations of the relative efficiency for all the models were made using expression 2.1 which is the power relation for the Mann-Whitney test when applied to ordered categorical data, and which is derived in Appendix F.

$$n_k = \frac{4 \sum_{r=1}^{k-1} f_{r-1}(E)\, f_r(E)\, (f_{r-1}(E) - f_r(E))\, (z_{\alpha/2} + z_\beta)^2}{\{ \sum_{r=1}^{k-1} (f_{r-1}(E) - f_{r+1}(E))\, f'_r(E) \}^2 \quad \delta^2} \qquad (2.1)$$

where

n_k is the total number of patients required in a clinical trial of two treatments

δ is the difference in efficacy to be detected

β is 1- the power

α is the two-sided significance level

E and $E + \delta$ are the true efficacies of the two treatments

$f_r(E)$ is the function representing the boundary between the r^{th} and the $(r + 1)^{th}$ categories of classification, and

$f'_r(E)$ is the derivative of $f_r(E)$, i.e. the gradient of the curve when $f_r(E)$ is plotted against E.

For 2 categories ($k = 2$), there is only one boundary between the categories, which is the cumulative normal distribution function $\Phi(E)$, and expression 2.1 reduces to

$$n_2 = \frac{4\, \Phi(E)\, (1 - \Phi(E))\, (z_{\alpha/2} + z_\beta)^2}{\varphi^2(E) \quad \delta^2} \qquad (2.2)$$

$\varphi(E)$ being the derivative of $\Phi(E)$.

The relative efficiency (denoted $e_{k,2}$) of the use of k categories of classification relative to the use of 2 categories is given by n_2/n_k (as discussed in section 2.4). Thus

$$e_{k,2} = \frac{\varphi^2(E) \sum_{r=1}^{k-1} f_{r-1}(E)\, f_r(E)\, (f_{r-1}(E) - f_r(E))}{\Phi(E)\,(1 - \Phi(E))\,\{\sum_{r=1}^{k-1} (f_{r-1}(E) - f_{r+1}(E))\, f'_r(E)\}^2} \qquad (2.3)$$

and this is thus independent of α, β and δ and depends only on the functions representing the category boundaries.

For model 1, substitution of the values of $f_r(E)$ in terms of $\Phi(E)$ and $\varphi(E)$ was used to calculate the relative efficiency. Expression 2.3 is algebraically complicated but can be readily calculated by computer using tabulated values of $\Phi(E)$ and $\varphi(E)$ e.g. as given in Neave (1981). A computer program was written in BASIC to perform this task, and details of this program together with a sample output from it are given in Appendix G.

For models 2 and 3, the calculation of $e_{k,2}$ was simplified by first carrying out a transformation to a linear scale of response, since this converts all the functions $f_r(E)$ into simple linear functions and evaluation of $e_{k,2}$ reduces to simple algebra. The values of $e_{k,2}$ were then calculated using a computer program written in BASIC with calculations done at 0.1 intervals of treatment efficacy.

RESULTS

MODEL 1: LATERAL SHIFT MODEL

The relative efficiency for 4 versus 2 categories of classification

The first analysis was described in Fig. 2.9, and the results are shown in Fig. 2.12. Consider firstly the curve for a response rate θ of 50%. At very small values of *w* (corresponding to Fig. 2.9 curves 1a and 3a), the relative efficiency $e_{4,2}$ is close to 1.00, i.e. there is little gain in efficiency. This would be expected since the central two categories contain only a small proportion of the total observations, and the difference from a 2 category classification is small. At high values of *w* (corresponding to Fig. 2.9 curves lc and 3c), the gain in efficiency is also small. Again this is expected, as the central 2 categories at a response rate of 50% contain almost all of the observations, and the difference from a 2 category classification is again slight. Between these extremes, there is a gain in efficiency of analysis, with a maximum relative efficiency of 1.37 for a response rate of 50%. This occurs at a value of *w* of 0.9. The same pattern is repeated at other values of the total response rate θ, except that firstly the relative efficiency is higher, and secondly the peak relative efficiency occurs at larger values of the category width *w*.

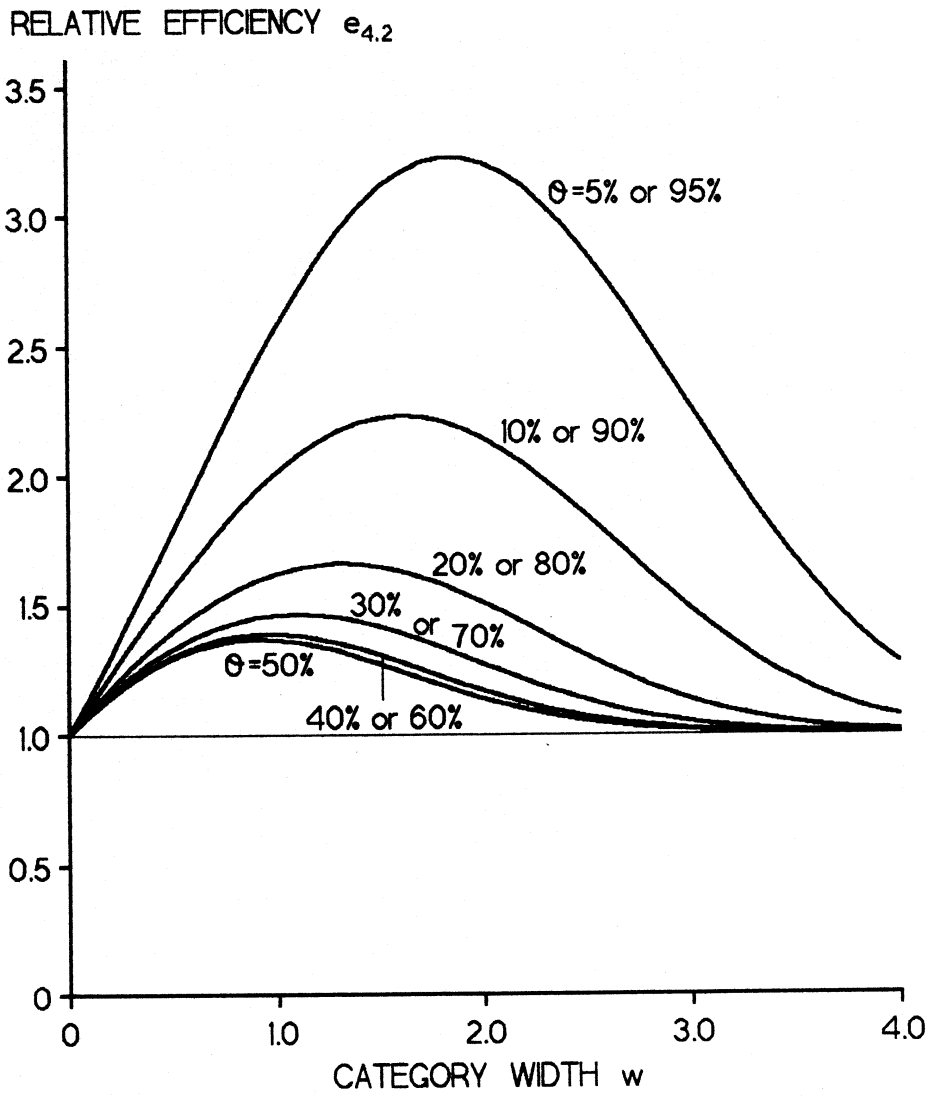


Figure 2.12. The lateral shift model. The relative efficiency for the Mann-Whitney test of tumour response data is shown for a 4 category system of classification relative to a 2 category system ($e_{4,2}$). The value of $e_{4,2}$ is plotted against the width of each of the two central categories w. Curves are shown for values of the total response rate θ of 5%, 10%, 20%, 30%, 40% and 50%. The curves for values of θ of over 50% are identical to those for θ less than 50% with symmetry around θ of 50%, e.g. curves for θ of 60% and of 40% are identical.

The relative efficiency for multiple versus 2 categories of classification

Fig. 2.12 showed the variation of relative efficiency with category width w for 4 categories of classification (i.e. for k equal to 4). Fig. 2.13 replots the greater parts of the curves of Fig. 2.12 and in addition plots the corresponding curves for larger numbers of categories i.e. with k equal to 6, 8, 16 and 32.

It can be seen that for $\theta = 50\%$, the relative efficiency at low w is slightly higher at $k = 6$ than at $k = 4$, and the peak value of $e_{6,2}$ of 1.44 (at $w = 0.6$) is slightly higher than that of $e_{4,2}$ (1.37) At high values of w, the curves for $k = 4$ and $k = 6$ come together (which is expected because the additional 2 categories contain very few observations). The values of $e_{8,2}$ are again slightly higher than those of $e_{6,2}$ at low w, but with each further increase in the number of categories, there is less and less gain in efficiency. The same pattern is repeated at all values of θ (except that the proportionate increase in efficiency is greater at the more extreme values of θ e.g. 5% and 95%). The peak value of $e_{k,2}$ appears to reach a limit as $k \to \infty$ and $w \to 0$ and this limit can be shown to be equal to $3\theta(1-\theta)/\{\pi\varphi^2(\Phi^{-1}(\theta))\}$ (see Appendix H). The values plotted adjacent to the y axis in Fig. 2.13 are calculated by this formula.

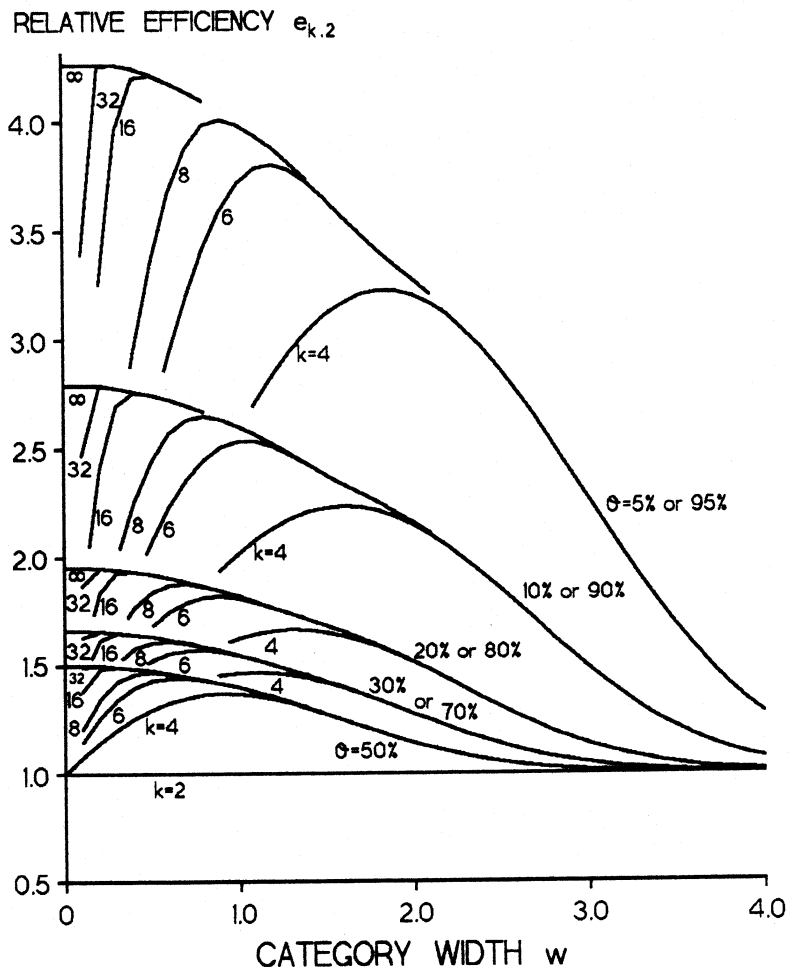


Fig. 2.13. The lateral shift model. The relative efficiency for the Mann-Whitney test of tumour response data for k categories of classification is shown relative to 2 categories of classification ($e_{k,2}$). Values of $e_{k,2}$ are plotted against the width w of each of the central categories, for values of k of 4, 6, 8, 16 and 32. The curves for $k = 4$ are replotted from Fig. 2.12. Curves are shown for values of the total response rate θ of 5%, 10%, 20%, 30% and 50% (the curves for θ = 40% are omitted for clarity). The curves for values of θ of over 50% are identical to those for θ less than 50%, with symmetry around θ of 50%. The line for $k = 2$ is at unity, by definition.

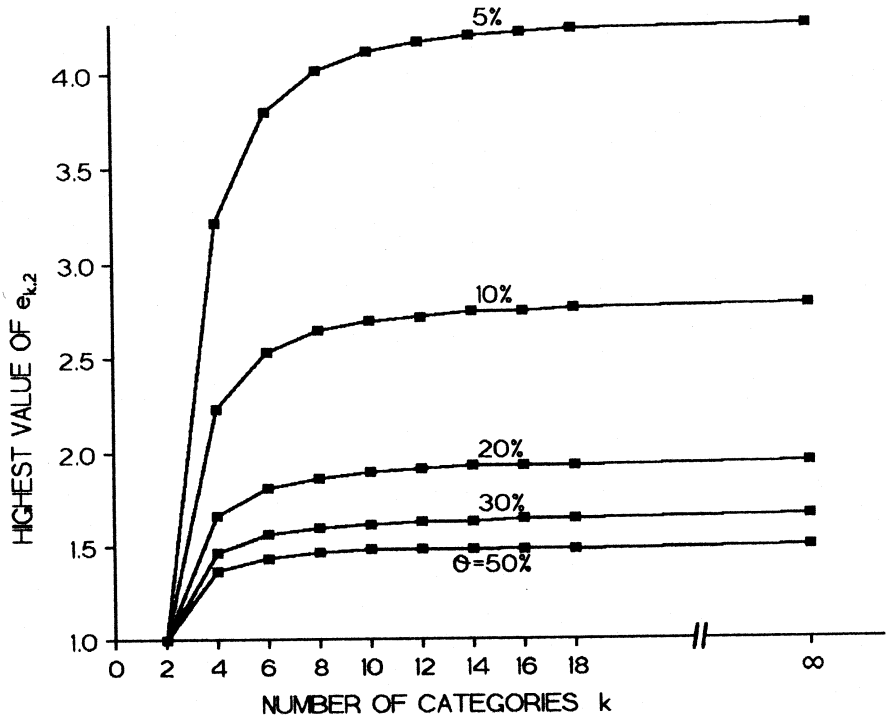


Figure 2.14. The peak relative efficiency for the Mann-Whitney test of a k category system of classification relative to a 2 category system ($e_{k,2}$) under the lateral shift model. For different values of the response rate θ, the highest value of $e_{k,2}$ (for any width of the categories w) is plotted against the number of categories k. The value for an infinite number of categories is calculated from the relation established in Appendix H. The curves for values of θ of over 50% are identical to those for θ less than 50%, with symmetry around θ of 50%.

The relationship between peak efficiency and the number of categories is shown more clearly in Fig. 2.14. From this it can be concluded that there is a limit to the gain in efficiency when the number of categories is increased from 2 to a large number, and that there is little (less than 5%) gain in efficiency when the number of categories is increased beyond 6, for response rates of 50%, or beyond 10, for response rates of 5%. Furthermore, out of 10 categories at a response rate of 5%, 5 are subdivisions of the 95% of non-

responders, and 5 are subdivisions of the 5% of responders and so contain very few observations. Hence, only 5 or 6 categories contain an appreciable proportion of the observations. Thus it may be concluded, on this model, that classification and analysis by 6 categories is more efficient than 4 categories but there is little gain beyond 6 categories (if chosen appropriately).

Optimum definitions of 4 categories of response

When a 4 category system of classification is being devised, there may be some freedom in the definitions of category boundaries. It has been assumed here that the central boundary is first fixed, i.e. the boundary between response and non-response. Then there may well be freedom to decide on the exact definitions of the boundary between PR and CR, and of the boundary between NC and PD. For example, the minimum duration of CR is usually taken as 4 weeks, but this could be varied.

Consider firstly a total response rate of 50%. The 50% of responders will be equally divided between CR and PR by a lateral shift w of 0.67, and the 50% of non-responders will be equally divided between NC and PD by the converse shift w of -0.67 (these values are obtained as $z_{50\%} - z_{25\%}$, and so on). This value of 0.67 matches roughly the value of w of 0.9 at which the relative efficiency peaks (see Fig. 2.12). Table 2.13 shows that this relationship is a general one for other values of the response rate θ of 50% or more, i.e. the value of w that results in division of the total response rate equally into CR and PR is close to the value of w at which the relative efficiency peaks. There is symmetry around θ of 50%, so that for θ less than 50%, the value of w that results in equal division of the non-response rate into PD and NC is close to that at which the relative efficiency peaks.

Table 2.13. Comparison of the lateral shift model that results in equal division of the total response rate, and the model at which the relative efficiency peaks.

Total response rate	*w* that divides response equally into CR and PR	*w* of peak relative efficiency
50%	0.67	0.9
60%	0.78	1.0
70%	0.91	1.1
80%	1.09	1.3
90%	1.41	1.6
95%	1.71	1.8

MODEL 2: EQUAL SUBDIVISION MODEL

In this model (defined in Fig. 2.10), the proportion of non-responders is (at all response rates) divided equally into the PD and NC categories, and the proportion of responders is (at all response rates) divided equally into the PR and CR categories. By some simple algebra, expression 2.3 reduces to $4\theta(1-\theta)/(1+\theta-\theta^2)$, and this is plotted in Fig. 2.15 as curve 1.

It can be seen that the value of $e_{4,2}$ for the equal subdivision model is less than 1.0 for all values of response, with a maximum of 0.8 and minimum of zero. There is thus a disadvantage to the use of 4 categories relative to the use of 2 categories of classification at all response rates.

MODEL 3: VERTICAL SHIFT MODEL

This model was defined in Fig. 2.11. For $p = 0.5$, expression 2.3 reduces to $9\theta/(1+2\theta)$ for $\theta \leq 50\%$, and to $9(1-\theta)/(3-2\theta)$ for $\theta \geq 50\%$. This is plotted in Fig. 2.15 curve 2. For response rates θ of 15% to 85%, the relative efficiency $e_{4,2}$ is greater than 1 with a peak of 2.25, and classification by 4 categories is more efficient than classification by 2. Outside these limits, $e_{4,2}$ is less than 1, which is consistent with this model resembling the equal subdivision model in these regions, as can be appreciated from Figs. 2.11 and 2.10.

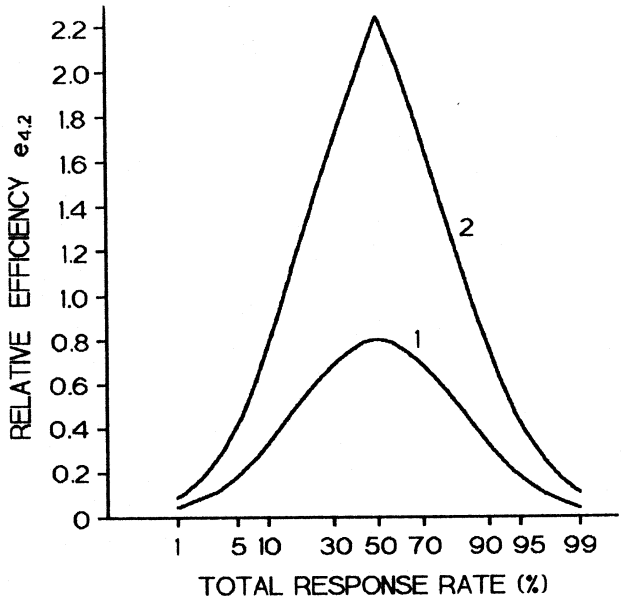


Figure 2.15. The equal subdivision and vertical shift models. The relative efficiency for the Mann-Whitney test of tumour response data for a 4 category system of classification relative to a 2 category system ($e_{4,2}$) is plotted against response rate θ for the equal subdivision model (curve 1) and the vertical shift model (curve 2).

DISCUSSION

Before the power of a statistical test of tumour response can be calculated, some assumptions must be made about the effect on rates of CR, PR, NC, and PD of a change in treatment efficacy, i.e. it is necessary to formulate a model of treatment effect. Any conclusions are then applicable only to the model assumed. However, the conclusions from a model may be generally applicable if the model approximates to the real situation.

In this section, three types of model have been considered. The lateral shift model seems to be the most plausible since it implies that an increase in the response rate due to an improvement in treatment efficacy is accompanied by an increase in the proportion of responses which are complete responses, and also a decrease in the proportion of nonresponses which are progressive disease. The equal subdivision model seems to be less plausible since it has no similar implication. The vertical shift model seems to be the least plausible, but has been included here in order to widen the range of models considered. The following section explores which model best fits real tumour response data, and in fact finds the lateral shift model to be the best fit.

The lateral shift model shows an advantage for 4 categories of classification for all values of the response rate 8 and for all values of the category width w (since the relative efficiency $e_{4,2}$ is in all cases greater than 1.0). Response rates θ outside of the range of 10% to 90% are rare in clinical practice. The following section will consider which values of w are the most likely to occur in practice; but on theoretical grounds, values of w less than 0.5 or greater than 1.5 are unlikely to occur in an internationally adopted 4 category assessment system since such values would lead to only small proportions of observations in one or more categories and a recognition that the category is redundant. For values of θ of 10% to 90% and values of w of 0.5 to 1.5, $e_{4,2}$ lies in the range 1.3 to 2.0 (as can be seen in Fig. 2.12). Thus on the lateral shift model, there is an appreciable

gain in efficiency of between 30% and 100% in a 4 category relative to a 2 category system of classification.

The practical importance of such a gain in efficiency will be discussed in section 2.10, but briefly, it implies a considerable saving in the numbers of patients to be entered in a clinical trial.

The data presented in Figs. 2.13 and 2.14 on classification using 6 or more categories of classification show a further theoretical gain in efficiency when the number of categories of classification is increased from 4 to 6, and little further gain thereafter.

These conclusions are in broad agreement with those of Husler and Riedwyl (1989). They approached the general problem of grouping of data for analysis from the opposite direction in calculating the efficiency of the Mann-Whitney test for grouped data relative to ungrouped data. They considered a range of probability distributions, including the normal distribution considered here, using a lateral shift model. They found for the various distributions that a grouping of the data into 8 to 14 groups is for large samples as efficient as the ungrouped rank test, and furthermore that such grouping is more efficient than the ungrouped rank test if the underlying density distribution has large tails. On the whole the work of Husler and Riedwyl is not directly comparable with this work as a different grouping of data is assumed, so that all of the groups contain equal numbers of observations. For two categories, this translates into consideration only of response rates of 50% in terms of the present work. Where the two sets of work are most comparable, which is in the lowest curve of Fig. 2.14, there is close numerical agreement.

These theoretical considerations must be combined with the practicalities of response assessment. Moertel and Handley (1976) showed that there is considerable

variability between observers using the standard 4 categories of response (section 2.1), and it is clear that as the number of classification categories increases, the observer variability will increase. The theoretical gain of efficiency of 5% to 18% (depending on the response rate) in increasing the number of categories from 4 to 6 may well therefore not be seen in practice. The calculations using the equal subdivision model indicate that subdivision of a category by the addition of a new category boundary will actually decrease the efficiency of analysis if it provides no information on treatment efficacy. An alteration of the standard 4 category classification system to include additional categories such as "minimal response" may thus possibly lead to a loss rather than a gain in efficiency and cannot be generally recommended from this model. An exception to this might be at very high or at very low response rates when there are only two or three categories that contain appreciable proportions of the observations. Subdivision of the largest category may then be worthwhile.

The findings concerning the optimal definition of 4 categories of response are not surprising. It might be expected *a priori* that the statistical efficiency is highest when the categories are of roughly equal size and this was found in the modelling. This conclusion is in broad agreement with the work of Connor (1972) who investigated the optimum grouping of data for a number of continuous distributions, when performing a test for trend.

CONCLUSIONS

The work of this section can be summarized as follows. From consideration of models of tumour response:

- Response data should be analyzed using the standard 4 categories of response rather than only the dichotomy of total response versus non-response.
- There is little advantage in the use of more than 4 categories of response.
- If possible, the categories should be defined such that there are roughly equal numbers in each.
- Subdivision of the standard 4 categories of response e.g. use of a category of "minimal response" is unlikely to increase and may actually reduce the efficiency of statistical analysis.

2.7 A SURVEY OF THE DISTRIBUTION OF TUMOUR RESPONSE DATA IN PUBLISHED CLINICAL TRIALS AND COMPARISON WITH MODELS OF TREATMENT EFFECT

INTRODUCTION

In the preceding section, a number of models of tumour response data were considered in order to calculate the relative efficiency of 4 or more categories of classification relative to 2 categories of classification (with analysis done by a Mann Whitney test). Each model related total response rate, complete response rate, and progressive disease rate. In general, this relative efficiency was greater than 1.0, i.e. there was an advantage to the use of 4 or more categories, but the conclusions depended on the model assumed. The current section addresses the question of which model provides the best fit to real clinical data.

METHODS

The clinical data studied were the 71 sets of tumour response data discussed in section 2.5, and further details are given there. The interrelation of rates of CR, PR, NC, and PD are shown in two plots which are really two variations of the same method. The example set of data already used extensively will be used to demonstrate the two techniques. It is repeated in Table 2.14 together with the rates expressed as percentages.

In both techniques, rates of complete response (CR) and rates of freedom from progressive disease (CR + PR + NC) are plotted against the total response rate in the same format as Fig. 2.6 of the preceding section, i.e. using a non-linear scale for the total response rate on the *x* axis, such that the plot of the total response rate is the sigmoid shape of the cumulative normal distribution curve. In statistical terminology,

this is done by a probit transformation of the total response rate. The first technique is demonstrated in Fig. 2.16.

Table 2.14. Example tumour response data (from Iversen et al, 1990).

Category	Counts		Individual rates		Cumulative rates	
	O	G+F	O	G+F	O	G+F
CR	0	1	0%	1%	0%	1%
PR	62	69	48%	58%	48%	59%
NC	26	28	21%	23%	69%	82%
PD	40	21	31%	18%	100%	100%
CR + PR	62	70	48%	59%		
Total	128	119				

O = orchidectomy; G + F = goserulin and flutamide.

When the results from several trials are plotted on the same graph by the first technique, the information of which groups of points arise from the same clinical trial is lost and this is why the second technique is also used. This retains this information but shows less clearly the positions of the points. It is illustrated in Fig. 2.17 for the same set of data. The data for a two arm trial are represented by a line joining the CR rates for the two treatments, and by a second line joining the proportions free of progressive disease. These lines are the experimental counterparts of the population curves $f_3(E)$ and $f_1(E)$ of

Fig. 2.7 of section 2.6. When the number of arms of the trial *a* was greater than 2, the data are similarly represented, but by *a* - 1 lines joining points ordered according to the total response rates of each of the treatment arms.

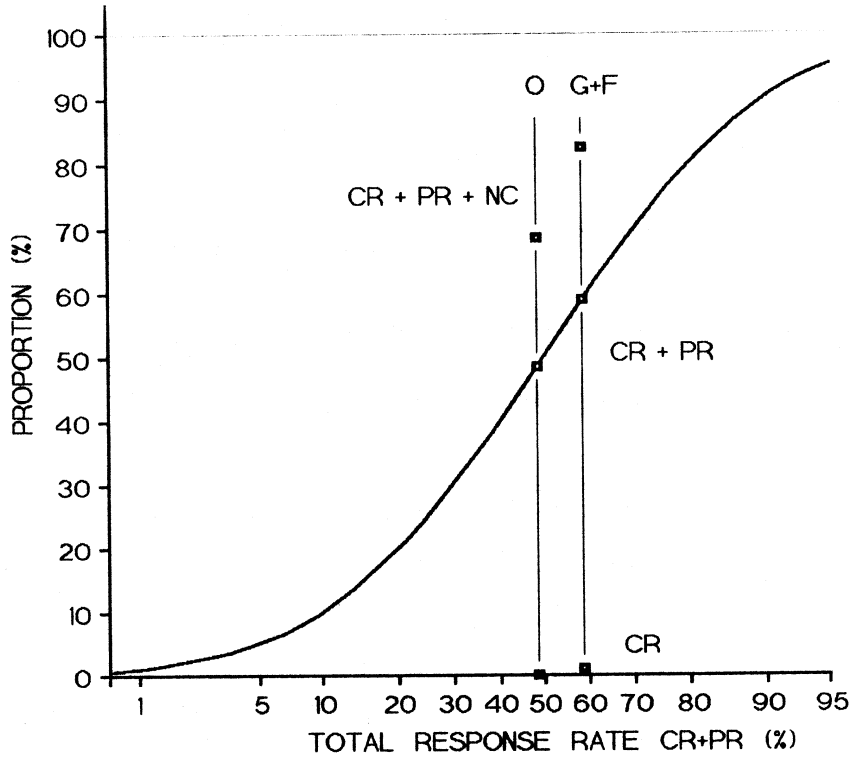


Figure 2.16. First technique of displaying the interrelation of rates of CR, PR, NC and PD from published clinical trials. The rates of complete response (CR), and freedom from progressive disease (CR + PR + NC) for the trial of O versus G + F (Table 2.14) are plotted by the square boxes against the total response rate on a probit scale on the *x* axis.

Some or all of the data from 5 of the 71 sets of data were excluded because of observed total response rates of either 0% or 100% which cannot properly be represented on a probit scale such as in Fig. 2.16. These observed total response rates of 0% and 100% do not necessarily imply that the format of Fig. 2.16 is too restrictive in disallowing total response rates of 0% and 100%, since in all 5 cases, the group sizes were small (ranging

from 8 to 45), the 95% confidence intervals of the total response rate were therefore wide, and the true total response rates could well have been other than 0% or 100%.

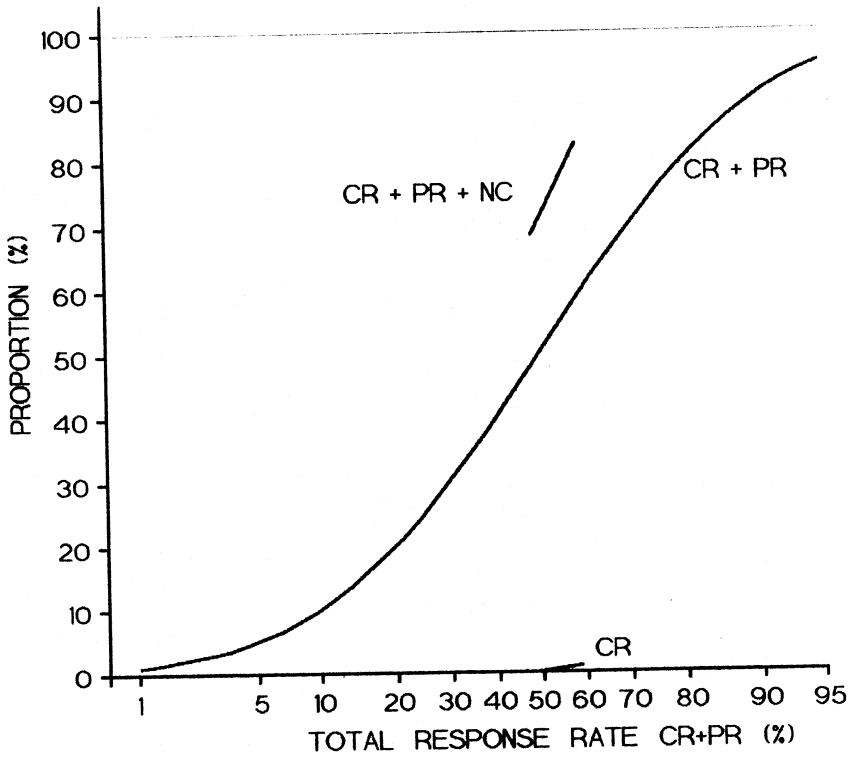


Figure 2.17. The second technique of displaying the interrelation of rates of CR, PR, NC and PD from published clinical trials using the same sample data as Fig. 2.16. The rates of complete response (CR) and freedom from progressive disease (CR + PR + NC) are plotted against the total response rate on the *x* axis on a probit scale.

The group sizes ranged from 6 to 296 with a median of 61. This median value implies 95% confidence intervals in all the rates of around ± 13%. There is thus some uncertainty in both the *x* and the *y* coordinates of the points plotted by the first technique and similarly of the ends of the lines drawn by the second.

RESULTS AND DISCUSSION

Fig. 2.18 plots the relation between complete response, freedom from progressive disease and the total response rates for the 71 sets of tumour response data by the first technique. There is considerable vertical scatter of the points representing complete response, and also of the points representing freedom from progressive disease. There are several potential reasons for the scatter. Firstly, there is necessarily some sampling variability due to the relatively small sample sizes. This will give variability in both the x and the y coordinates (and this variability will be linked). The 95% confidence intervals will be around ± 10% to 20%, which is sufficient to explain some but not all of the scatter on the diagram. The other sources of scatter are the differences between the trials in tumour type, classification system, stage of disease, treatment type, amount of pretreatment, and so on.

Fig. 2.19 displays the same data as Fig. 2.18 by the second technique (which was illustrated in Fig. 2.17). Despite the scatter of the points, something can be learned from these plots. They can be compared with the three models of tumour response represented in Figs 2.8, 2.10 and 2.11 of section 2.6. By eye, the data fit the lateral shift model of Fig. 2.8 better than either of the other two models, and this is particularly so of the complete response rates. It is interesting that the lateral shift model is the model for which the Mann-Whitney test is optimally efficient. A fit by eye of the lateral shift model using shifts of +1.1 and -l.l is shown in Fig. 2.20. Taking these shifts together with the relation shown in Fig. 2.12 gives an estimate of the value in practice of the relative efficiency $e_{4,2}$ for 4 categories relative to 2 categories of classification. This estimated value ranges from 1.3 at a total response rate of 50% to 2.1 at a total response rate of 10% or 90%.

This range of values includes the value of 1.4 estimated for the same relative efficiency by the re-analysis of the data sets in section 2.5. The practical implication of these values of the relative efficiency are discussed in section 2.10.

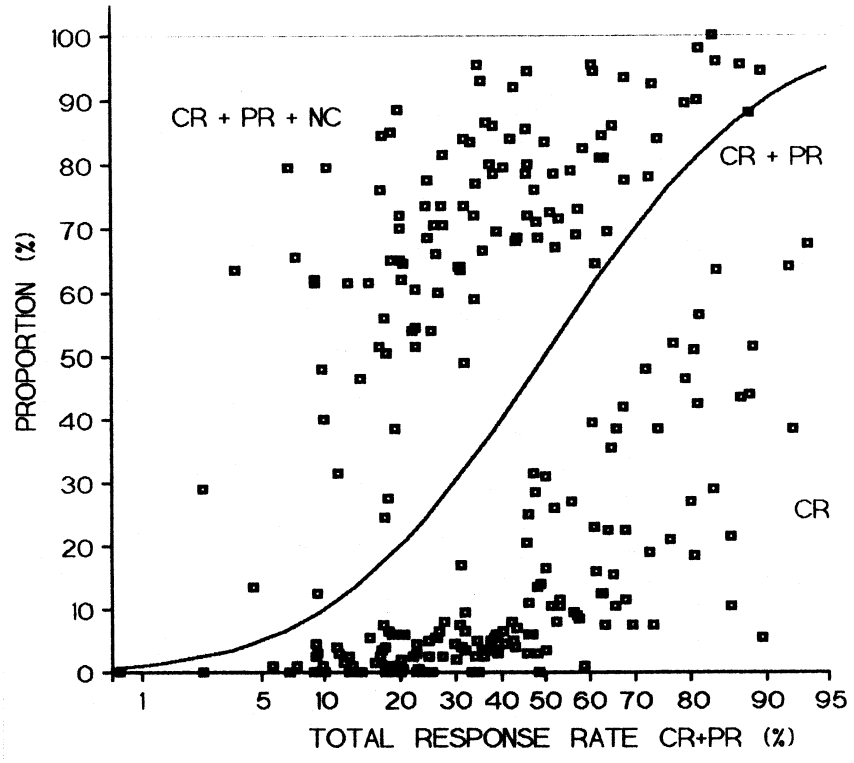


Figure 2.18. Data on the relation between the complete response, freedom from progressive disease and total response rates in 71 clinical trials. Complete response rates are plotted against total response rates in the right lower part of the figure. Freedom from progressive disease (equalling the sum of the CR + PR + NC rates) are similarly plotted in the left upper part of the diagram.

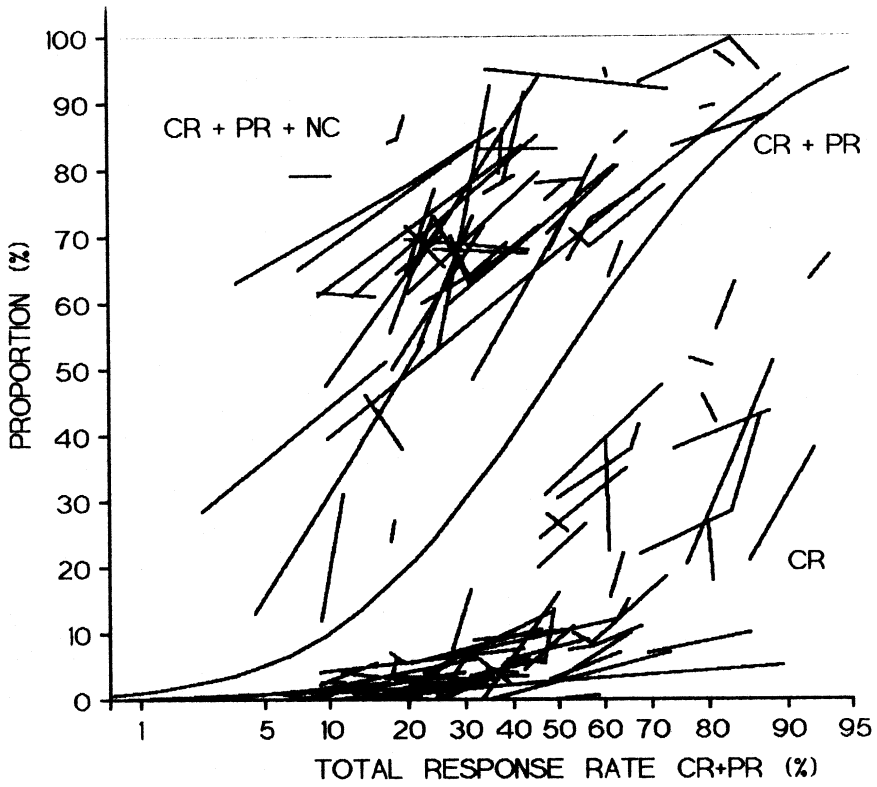


Figure 2.19. Data on the relation between the complete response, freedom from progressive disease and total response rates in 71 clinical trials. The data points from within each trial are plotted by a series of lines. Complete response rates are plotted against total response rates in the right lower part of the figure. Freedom from progressive disease (equalling the sum of the CR + PR + NC rates) are similarly plotted in the left upper part of the diagram.

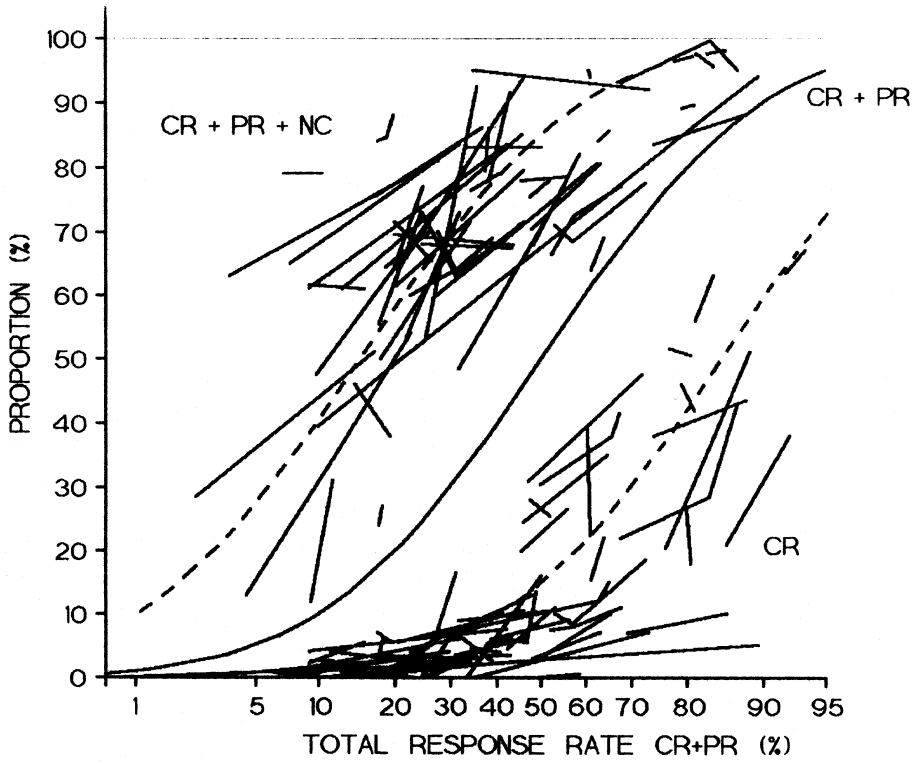


Figure 2.20. Fit by eye of the lateral shift model to the data of Fig. 2.19. Shifts are of + 1.1 for the PD/NC boundary and -1.1 for the PR/CR boundary.

2.8 A SURVEY OF METHODS OF ESTIMATION OF TUMOUR RESPONSE IN CANCER CLINICAL TRIALS

INTRODUCTION AND METHODS

There are increasing calls for clinical research to estimate the sizes of differences between treatments and not merely test whether or not a difference exists (as discussed in section 1.2). This survey aimed to assess the extent to which methods of estimation of differences in outcome were used in recent cancer clinical trials comparing tumour response. The series of 81 clinical trials discussed in section 2.2 was surveyed and the use of methods of estimation and calculation of confidence intervals were recorded. Further details of the trials are given in section 2.2.

RESULTS AND DISCUSSION

Table 2.15 shows the methods used to summarize the results and estimate the difference in outcome. Most of the papers calculated response rates as a percentage, although there were variations in whether this was done for each individual treatment, for the overall combined results, or for both. Few papers (7) calculated the arithmetic difference between the response rates although this is such a simple calculation that in many papers it was done implicitly rather than explicitly. Only two papers used a more elaborate technique of estimation. The following section will review a number of methods of estimation that can be applied to tumour response data, and will discuss possible reasons why they are not being used in publications.

Table 2.15. Summary estimates calculated in the 81 publications

Method of estimation		Number of papers
Total response rates or complete response rates		77
Overall, combining treatments	5	
For each treatment	61	
Both overall and for each treatment	11	
Difference in total response rate		7
Difference in complete response rates		0
Other		2
Difference in response rate adjusted for prognostic factors by logistic regression	1	
Odds ratio for response	1	

Table 2.16 details the use of confidence intervals in the 81 papers. Only 27 papers reported confidence intervals at all and in only 7 was there calculation of a confidence interval of a difference between treatments. This lack of use of confidence intervals may reflect either a lack of awareness of the techniques, or limited usefulness, and this will be discussed in chapter 4.

Table 2.16. Calculation of confidence intervals in the 81 publications

Method	Number of papers
No confidence interval calculated	54
Confidence interval calculated	27
of overall response rate	3
of individual response rate	20
of difference in response rate	6
of difference in response rate adjusted for prognostic factors by logistic regression	1
Confidence interval used	
95%	26
90%	1

2.9 A REVIEW OF METHODS OF ESTIMATION OF TUMOUR RESPONSE IN CANCER TRIALS

There are increasing calls for statistical tests of significance of clinical trial data to be supplemented by estimates of the magnitude of the differences found. This is because the degree of statistical significance is a poor indicator of the degree of clinical significance of the trial result, since the former depends also on the trial size - a clinically insignificant difference of a few percent in success rates may be highly significant statistically if the trial is very large, and similarly, a large difference in success rates may be non-significant statistically if the trial is very small. The principle methods of estimation that can be applied to tumour response data are reviewed in this section. These are:

- The arithmetic difference in the response rate
- The relative risk of response
- The odds ratio
- The Mann-Whitney estimator
- The rank sum estimator

The preceding section showed these techniques to be little used in recent cancer clinical trials. In addition to calculation of the best estimate of the difference in two treatments, methods are available for the calculation of confidence intervals to indicate the precision of the calculated estimates. Again, these techniques appear from the preceding section to be little used in recent cancer clinical trials. The data of Table 2.6 will again be used to demonstrate the methods and it is repeated (with the columns reversed for convenience) as Table 2.17.

Table 2.17. Example tumour response data (Iversen et al, 1990). Numbers of patients in each response category for each treatment.

Category of response	Goserelin and flutamide (G + F)	Orchidectomy (O)	Total
Complete response (CR)	1	0	1
Partial response (PR)	69	62	131
No change (NC)	28	26	54
Progressive disease (PD)	21	40	61
Total response (CR+PR)	70 (58.8%)	62 (48.4%)	
Non-response (NC+PD)	49 (41.2%)	66 (51.6%)	
Total	119	128	247

THE DIFFERENCE IN PROPORTIONS

This is the simplest technique and was found to be the one most commonly used (even though infrequently) in the survey of the preceding section. The success rate (in this context, the total response rate or the complete response rate) is calculated for each treatment and the arithmetic difference is obtained by subtraction. For the data of Table 2.17, the difference in the total response rate is 58.8% - 48.4%, i.e. 10%.

The 95% confidence interval for the difference in the response rates can be obtained from the difference ± 1.96 × the standard error (Altman, 1991), using the formula for the standard error *(SE) of* the difference in two proportions p_1 and p_2 (Altman, 1991):

$$SE(p_1 - p_2) = \sqrt{\{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2\}}$$

where

n_1 is the total number of observations in group 1, and

n_2 is the total number of observations in group 2.

In the current example,

$$SE(p_1 - p_2) = \sqrt{\{0.588(1 - 0.588)/119 + 0.484(1- 0.484)/128\}}$$

$$= 6.32\%$$

and the 95% confidence interval for the difference in proportions is calculated as

$$p_1 - p_2 - 1.96 \times SE(p_1 - p_2) \text{ to } p_1 - p_2 + 1.96 \times SE(p_1 - p_2)$$

which in this example equals -2.0% to +22.8%

By the simple interpretation of confidence intervals, the trial can be said to have found a range of values in which the true difference between G + F and orchidectomy probably lies, i.e. from G + F being 2% worse, to orchidectomy being 23% better.

Calculation of the confidence interval for the difference in the total response rate was the principle method of confidence interval estimation found in the survey reported in the preceding section.

THE RELATIVE RISK

This method was developed in epidemiology but can be applied to other areas. The rate of occurrence of a particular outcome is calculated for each group, and the ratio is taken. For tumour response data, the outcomes of greatest interest are usually the total response rate and the complete response rate. For the data of Table 2.17, the relative risk of a response (complete or partial) is

$$\frac{58.8\%}{48.4\%} = 1.21$$

A method for the calculation of a confidence interval is available (Altman, 1991). This method of estimation was not found to be in current use in cancer clinical trials in the survey of section 2.8, and it is hard to see any merit in it.

THE ODDS RATIO

This method is similar to the calculation of the relative risk, but the odds of occurrence of a particular outcome are calculated for each group and the ratio is taken; if the outcome occurs in a proportion p, then the odds are

$$p / (1 - p).$$

Again, odds ratios could be applied either to the total response rate or the complete response rate. For the data of Table 2.17, the odds ratio for a response is

$$\frac{58.8\% / 41.2\%}{48.4\% / 51.6\%} = 1.52$$

Again, a method is available for the calculation of a confidence interval (Altman, 1991).

The calculation of odds ratios occurred in only one instance in the 81 articles surveyed in the preceding section. However, there is arguably more to recommend it than

the calculation of relative risks. As an example, the odds ratios for the following improvements in the response rate in a clinical trial are all approximately equal to 2.0.

5% to 10%,

10% to 18%,

41% to 59%,

82% to 90%, and

90% to 95%

Although the odds ratios for these improvements in response are all equal, the arithmetical differences in response rates are not, ranging from 5% to 18%. Many oncologists would recognise an approximate equivalence between the improvements listed above, or at least would be familiar with the concept that it is easier by developments in treatment to increase a response rate of 50% to one of 60% than it is to increase a response rate of 90% to one of 100%. Thus it may be that the odds ratio is a better index of differences in treatment efficacy than the arithmetical difference in response rates.

An extension of this is in multiple logistic regression, which can correct for the influence of prognostic factors such as stage, histological grade and performance status. The model underlying logistic regression assumes an equivalence of odds ratios e.g. it assumes equivalence of the differences in success rates such as given in the list above, and this does not seem unreasonable. (In fact, a logistic transformation is very closely similar (Armitage & Berry, 1987) to the probit transformation that was found in section 2.7 to be a reasonable fit in conjunction with a lateral shift model to real tumour response data. There are thus strong parallels between the lateral shift model explored in this dissertation and the model that underlies logistic regression.) Thus an advantage of the use of odds ratios is that correction can be made for imbalances in prognostic factors between the treatment groups.

THE MANN-WHITNEY ESTIMATOR

As well as being used for a statistical test, the Mann-Whitney test procedure can give an estimate of the difference between two groups. The calculated value of U_{AB} is the number of pairs of observations, one of group A and one of group B, where the group A observation is less than the group B observation. Since the total number of these pairs of observations is $n_1 n_2$, the value of $U_{AB} / (n_1 n_2)$ gives the proportion of pairs of observations in which the group A observation was less than the group B observation. This can be used as a prediction of the probability of a new group A observation being less than a new group B observation (Armitage & Berry, 1987; Altman, 1991). Procedures are available for calculating a confidence interval for this probability, although computers are needed for their evaluation (Morton & Dobson, 1990).

As an example, the value of U from the data of Table 2.17 is 8725, giving a value of $U / n_1 n_2$ of

$$8725 / (128 \times 119) = 0.57.$$

Thus if two new patients are treated, one by orchidectomy and the other by G + F, there is an estimated 57% probability that the patient treated by G + F will have a better response, and an estimated 43% probability that the patient treated by G + F will have a worse response (these probabilities are distinct from the probability that an individual patient will have a better response from G + F than from orchidectomy).

These probabilities with their confidence intervals are a way of viewing the relative efficacy of two treatments. They may be a better reflection of the relative efficacy of two treatments than a comparison of the response rates since they make use of all the response data rather than the arbitrary cutoff of whether or not there is at least a partial response. This is particularly so when nearly all the patients respond (and the difference between the treatments is shown mainly in the proportion of complete responses) and

also when almost none of the patients respond (and the chief difference between the treatments is in the relative frequencies of no change and progressive disease).

It is possible to extend the concept of relative efficiency from hypothesis testing to methods of estimation, in that a more efficient method of estimation will give a relatively narrower confidence interval than a less efficient method of estimation (Pratt and Gibbons, 1981). In fact, the relative efficiency of two methods of estimation is equal to the relative efficiency of the two equivalent statistical tests (Pratt and Gibbons, 1981). This means that the conclusions in the previous sections concerning the superiority of the Mann-Whitney test over the Chi squared test of 2 categories of response also imply an advantage for the Mann-Whitney estimator over the comparison of two proportions either by the arithmetic difference or by odds ratios.

THE RANK SUM ESTIMATOR

The rank sum estimator (δ) has been described by Morton and Dobson (1990). It is closely related to the Mann-Whitney estimator, and is not really a different estimator - it just presents the same information in a different form. In the context of the above example, the rank sum estimator gives an estimate of the difference between the probability that a patient given G + F will have a better response, and the probability that a patient having an orchidectomy will have a better response, i.e. of 5'7% - 43% = 14%. This estimator seems to have less clinical relevance than the Mann-Whitney estimator for tumour response data (or similarly for morbidity data recorded by categories).

DISCUSSION

This section has reviewed the principle methods for estimating the size of difference in tumour response between the treatments in a clinical trial. Some advantages have been discussed for the more complex methods of estimation using odds ratios and the Mann-Whitney estimator. However, it must be remembered that the aim of methods of estimation is to inform oncologists as to the size of difference in outcome that may be expected in their patients. Most practising oncologists are not familiar with the more complex concepts, and are unlikely ever to become familiar with them, given the many demands on their time. Thus it seems that the outcome of a clinical trial is best presented in terms of the differences in total and complete response rates (or possibly in terms of the proportions free of progression depending on the clinical situation). The more complex methods could be used more to correct the response rates either for prognostic factors (e.g. by logistic regression methods), or by incorporation of information from CR and PD rates using a lateral shift model as discussed in this dissertation.

2.10 GENERAL DISCUSSION

The survey described in section 2.2 showed that the most commonly used statistical test in the analysis of tumour response data is the Chi squared test of the total response rate. However, the theoretical arguments detailed in section 2.3 support the Mann Whitney test as probably the best test for the analysis of tumour response data in clinical trials. Furthermore, section 2.5 found an advantage in practice for the Mann-Whitney test with an increase in the efficiency of analysis by a factor of around 1.4; and the modelling of section 2.6 combined with the study of distribution of response data of section 2.7 supported this with an estimate of between 1.3 and 2.1 (depending on the rate of response) for the efficiency of the Mann-Whitney test relative to the Chi squared test of the total response rate.

What is the practical relevance of this relative efficiency of 1.4? For two equivalent tests, the relative efficiency is 1.00. A relative efficiency greater than 1.00 can be looked at in two ways. Firstly (by definition), it implies that more patients are required by the less efficient statistical test to give a conclusion with the same level of certainty (the same statistical power). In the case under discussion, the implication is that a Chi squared test of the total response rate requires around 40% more patients in a clinical trial than a Mann-Whitney analysis for the same power. Putting it another way, a switch from a Chi squared analysis of total response to a Mann-Whitney test is equivalent to increasing recruitment to the trial by around 40%. Use of a less efficient test is equivalent to discarding the results from a proportion of the patients prior to the analysis.

A second way of viewing the same situation is that the use of the more efficient test will give a more extreme *z*-value (and more extreme *P* value) if there is a true underlying difference in treatment effect, and it will thus show stronger evidence against the hypothesis of no treatment difference. For a relative efficiency of around 1.4, if the *P*

value from analysis of a set of data by the less efficient test was 10%, it will be expected to fall to 5% on analysis by the more efficient test. A *P* value of 5% will be expected to fall to 2%, and a *P* value of 1% will be expected to fall to 0.3% on changing from the less to the more efficient test. This follows from the relation between *z*-values and relative efficiency that is discussed in Appendix D.

Since section 2.2 found that currently the most common test used in the analysis of tumour response data is the Chi squared test of the total response rate, it follows that results significant at the 5% level are currently probably being missed and going unreported. This was confirmed in section 2.5 where 5 of the 27 results significant by a Mann-Whitney test were not reported as significant in the original publication, and a further 3 were reported in an equivocal way. It is clearly wrong to overemphasize the 5% cutoff, but these figures are an illustration that extend to other significance levels; all levels of conclusion will tend to be strengthened by the use of the more efficient test.

The results in this chapter provide very strong support for the routine use of the Mann-Whitney test in the analysis of tumour response data. It would be wrong to perform a number of statistical tests, for example including the Chi squared and the Mann-Whitney tests, and select the test to be reported on the basis of minimizing the *P* value since such multiple testing can be misleading. Where a general advantage has been shown for a statistical test in a particular area of research (such as here for the Mann-Whitney test in tumour response data), its use should be routine.

The surveys in this chapter have also shown that the current analysis of tumour response data emphasizes statistical testing of hypotheses as opposed to methods of estimation and calculation of confidence intervals. The possible reasons for this are discussed in Chapter 5.

CHAPTER 3. TREATMENT MORBIDITY

3.1 INTRODUCTION

Ideally, cancer treatments would cause no appreciable morbidity, but in practice the treatments often exploit only small differences between malignant cells and normal cells, and toxic effects are common. Many cancer treatments, especially radical treatments given with the aim of cure, are given at their maximum tolerated doses; and there are only a few exceptions: for example, seminoma and lymphoma can usually be controlled by relatively low doses of radiotherapy, and choriocarcinoma can usually be cured by relatively low doses of chemotherapy. In the development of new drugs, the phase II clinical trials are aimed at identifying the maximum tolerated dose of a new drug or regime.

The recent UK trials of neutron irradiation may be cited as an example of the importance of the study of morbidity. In the Hammersmith trial of low energy neutron against photon radiotherapy, the initial enthusiasm over higher local control rates was later tempered by an appreciation of higher morbidity in the neutron treated patients (Duncan, 1983). However, higher morbidity was not found at the doses studied in the Clatterbridge trial of high energy neutrons in advanced pelvic malignancies (Errington et al, 1991). The study of morbidity is thus essential in the context of cancer clinical trials. Despite this, many radiotherapy trials have not included a proper assessment of morbidity (Dische et al, 1989).

SCALES OF MEASUREMENT OF MORBIDITY

There are many parallels between morbidity data and tumour response data. Just like tumour response, the morbidity of treatment is in general difficult to measure exactly. Most types of morbidity are qualitative although a few are quantitative e.g. depression of white cell and platelet counts, or of renal function, or elevation of liver enzyme levels. However, even for the few quantitative types of morbidity, the accuracy of measurement is not great - white cell and platelet counts may be measured only two or three times weekly and the true nadir may occur between measurement days; and estimation of renal function is well known to be unreliable. Because of these inaccuracies, and for ease of presentation of results, all types of morbidity including quantitative variables are usually recorded by allocation to one of five broad categories. Standard scales for morbidity have been devised, for example those of the WHO (Miller et al, 1981), the Eastern Cooperative Group of the USA (Monfardini et al, 1987), the EORTC/RTOG (Radiation Therapy Oncology Group), and a working group sponsored by the Medical Research Council (Dische et al, 1989).

Morbidity is frequently divided into *early* and *late* according to how soon it develops after treatment. Separate scales for early and late toxicity have been developed in the EORTC/RTOG system.

TREATMENT MORBIDITY: TYPE OF VARIABLE

For acute morbidity, it is usual to record the most severe grade of morbidity experienced by each patient and to compare treatments in terms of these most severe grades. The treatment morbidity data is thus in the form of counts of patients in each of the categories. Like tumour response categories, treatment morbidity categories form a

natural order from the best to the worst, and thus the data again form an ordered categorical variable.

This conventional analysis takes no account of the duration of morbidity. When the treatments are similar in type, this may not matter, but where they are dissimilar, a more complex method of recording may be needed. For example, in the current Continuous Hyperfractionated Accelerated Radiotherapy (CHART) trials, where radiation treatments with overall times of 12 days and 6 weeks are being compared, a comparison of the peak morbidity experienced will not be adequate, as the durations are likely to be very different.

When the onset of morbidity is delayed, e.g. in the fibrosis that may follow radiotherapy, there is a further complication in that the length of follow-up will be different for different patients, and the calculation of incidence rates becomes more complex.

THE VALUE OF MEASUREMENT OF TREATMENT MORBIDITY

Before going on to explore the statistical techniques available for the assessment of treatment morbidity in the following sections, it is worthwhile considering exactly why assessment of morbidity is important, and this really comes down to what clinical questions are to be answered by the collection of the morbidity data.

In many cancer clinical trials, the clinical question is whether the morbidity of the treatments compared are equal. For many chemotherapy and radiotherapy regimes, the response rates, the cure rates and the morbidity rates all tend to increase as the dose of treatment increases. An improvement in response rate, local control rate or even cure rate may not in fact benefit the patient if it is at the expense of a large increase in

morbidity. Before one treatment can be considered superior to another, it is necessary to demonstrate a higher efficacy for the same morbidity, or equivalently less morbidity at the same efficacy. For example, in the second BIR laryngo-pharynx trial, where no difference in efficacy was found between short and long radiotherapy fractionation regimes (Wiernik et al, 1991), an important question was whether the morbidity (acute and late) was equal in the two treatment groups.

Other clinical trials may investigate the management of morbidity, for example, the optimum control of vomiting in patients receiving chemotherapy. Here, the principle clinical question is again whether there is a difference in morbidity in the groups studied. The size of any difference found is a secondary question, but this will be particularly important if there is a large difference in the cost or morbidity of the treatments studied.

In some clinical trials, it is already well known that two treatments have different patterns of morbidity, and when the treatments are compared in a trial, the main questions are about response and survival; for example, it is usually not useful to compare the effects on the bone marrow in a trial of radiotherapy versus chemotherapy as it is so well known that the effects are different. A second situation where morbidity data do not require a test of statistical significance is in a dose ranging study; the intention is to find the maximum tolerated dose, and it is known in advance that the morbidity levels at different doses will be different. A third situation is where morbidity levels are very small and occurring only at placebo rates. In statistical terms, the problems here are of estimation rather than hypothesis testing.

3.2 THE STATISTICAL TESTS AVAILABLE FOR ACUTE TREATMENT MORBIDITY DATA

Treatment morbidity data can be divided into acute morbidity data (usually categorical but sometimes continuous) and late morbidity data. The simple statistical methods available for continuous data such as weight loss are the *t* test and the Mann Whitney test; this does not seem to be a contentious issue, and will not be discussed further. Data on delayed morbidity closely resemble survival data, from the point of view of statistical analysis; similar methods of analysis are appropriate, and these will not be considered further. The rest of this chapter will concentrate on acute morbidity recorded by categories.

There are close similarities between acute morbidity data and tumour response data in that both form ordered categories. The statistical tests that can be applied to acute treatment morbidity are thus the same as the tests for tumour response data that were discussed in section 2.3. These are:

- The Chi squared test of all categories of morbidity,
- The Chi squared test of a dichotomy,
- Fisher's exact test,
- The Chi squared test for trend,
- The *t* test after allocation of scores,
- The Mann-Whitney and Kruskal-Wallis tests,
- Regression methods.

The discussion of section 2.3 on the relative merits of these tests will not be repeated, but the conclusion are similar, i.e. that in many cases, the Mann-Whitney test makes more use of the information contained in the data and should be a more efficient test, without the need for a scoring system of the Chi squared test for trend and the *t* test.

The difficulties with Fisher's test or the Chi squared test of a dichotomy may be greater than for tumour response data, since the choice of dichotomy may be more difficult. It is possible to divide the grades of morbidity into a dichotomy in a number of ways: no morbidity versus any, the most severe grade of morbidity versus all lesser grades, or by a grouping of several lesser grades of morbidity versus a grouping of the remaining more severe grades (e.g. grades 0 - 2 versus grades 3 - 5). Theoretical work (Connor, 1972) supports use of a dichotomy that divides the total number of patients into two roughly equal groups. The exceptions to these general arguments are where the relation between the dose of treatment and the clinical significance of the morbidity is not a simple one. As detailed in Chapter 1, there may be a threshold effect, e.g. in myelosuppression, minor levels of morbidity may have very little clinical significance whereas severe levels may cause septicaemia or haemorrhage, which may be fatal. Here, it will be important to assess the important dichotomy of whether or not the clinically significant event (e.g. neutropenic fever) has occurred and the Chi squared test (or Fisher's test) is the appropriate technique. The selection of a statistical method depends on identifying the clinical question to be answered in the clinical trial - whether interest centres on minimizing all grades of the morbidity (as in nausea and vomiting), or on minimizing the frequency of the most severe grade(s). Even in this second case, it may be that a Mann-Whitney test of all categories of morbidity is helpful; for example, it may be that the frequency of neutropenic fever is low, and the power of the trial to detect a difference in myelosuppression is consequently low; but by including all grades of myelosuppression in the comparison, a better indication of the relative degrees of myelosuppression may be obtained, although this depends on the relationship between the frequencies of different grades of myelosuppression, and this remains to be investigated.

3.3 A SURVEY OF METHODS OF ANALYSIS OF ACUTE MORBIDITY DATA IN PUBLISHED CANCER CLINICAL TRIALS

INTRODUCTION AND METHODS

Section 2.2 found that tumour response data is currently most commonly analyzed by the Chi squared test, although sections 2.6 and 2.5 found the Mann-Whitney test to be superior both in theory and in practice. Since acute morbidity data are usually recorded in ordered categories, like response data, the statistical problem is generally the same. Thus a Mann-Whitney test making more use of of the information in the data may be more efficient than a Chi squared test. This section reports a survey of the statistical tests used for acute treatment morbidity.

In contrast to response data, morbidity data often do not require a statistical test of a null hypothesis, as discussed in section 3.1. A study such as this one of how morbidity data are being analyzed (and the study in the next section of how they should be analyzed) must therefore be selective and concentrate on the morbidity data in the literature where a statistical test is most appropriate. This judgement of the relevance of a statistical test is clearly a subjective one, and consequently two clearcut criteria were employed in this survey to define a set of papers where statistical analysis of morbidity was most appropriate; firstly where the authors themselves carried out a statistical analysis, and secondly where they reported the data in detail by categorising it by grade and treatment.

Papers were studied from the same seven journals as provided the tumour response data for chapter 2. Papers were studied using the same criteria of randomised studies of solid tumours, and were included if firstly the morbidity data was subjected to a statistical test by the authors themselves, or secondly if the morbidity data was presented

as a contingency table (or the equivalent in text) categorised by grade of morbidity and by treatment. Dose ranging studies aimed at finding the acute normal tissue tolerance were excluded. Data were collected by an adaptation of the form used for the response data survey of section 2.2.

A total of 36 papers were identified, published between January and December 1990 inclusive. Most of the papers also qualified for inclusion in the series of papers of the response survey of section 2.2, and so the characteristics of the papers in the two series are broadly the same.

In 8 of the 36 papers, the clinical trial concerned the control of morbidity of treatment e.g. control of vomiting. In 4 of these papers, in addition to the main morbidity data, (e.g. vomiting) there were data on the side-effects of the treatment aimed at minimizing that morbidity (e.g. side-effects of the anti-emetic drugs). In these cases the two types of morbidity were analyzed separately as if they had been reported in separate papers. This gave a total of 40 sets of morbidity data for analysis. Table 3.1 shows the distribution of the 36 papers by tumour type.

Table 3.1. Distribution of tumour types

Tumour type	Number of papers
Breast Carcinoma	7
Head & Neck Carcinoma	6
Bronchial Carcinoma, Non-Small Cell type	4
Bronchial Carcinoma, Small Cell type	2
Colorectal Carcinoma	3
Multiple types	8
Other	6
Total	36

RESULTS AND DISCUSSION

Table 3.2 shows the statistical tests reported. The data were mostly reported by categories rather than as continuous data. Several of the papers reported more than one test and so the total number of tests shown in Table 3.2 exceeds the number of sets of morbidity data. The findings are broadly similar to those of the survey of statistical tests reported for tumour response data of section 2.2. There was again a wide variety of tests in use, and the commonest test was a Chi squared test of a dichotomy. The Mann-Whitney test was used in two sets of categorical data.

Table 3.2. The types of statistical test reported for the 40 sets of morbidity data. Several papers reported more than one test.

Statistical test	Number of sets of data	
Categorical data		22
Chi squared test	14	
of a dichotomy	10	
of the full table	0	
unclear / varied	4	
Fisher's exact test of a dichotomy	3	
Multivariate analysis of a dichotomy	3	
Mann-Whitney test	2	
Continuous data		6
t test	3	
Mann-Whitney test	2	
Logrank test of time to development	1	
Test unclear		8
No test reported		10

In a total of 14 papers, the analysis was of a dichotomy. In 4 of these 14 papers, the method of presentation of the morbidity data was as either present or absent. In the remaining 10, the morbidity data were detailed by treatment and by grade, but for

analysis, the grades were combined to a dichotomy. The distribution of these 10 dichotomies was: none versus any in 3, lesser grades versus more severe grades in 4, unclear in 1, and varied in 2.

3.4 STATISTICAL ANALYSIS OF ACUTE TREATMENT MORBIDITY DATA: COMPARISON OF THE CHI-SQUARED TEST WITH THE MANN-WHITNEY TEST

INTRODUCTION AND METHODS

The survey reported in the preceding section showed the Chi squared test of a dichotomy to be the statistical test most commonly used in the analysis of acute treatment morbidity data, and there was little use of the Mann-Whitney test which makes more use of the information in the data. This section explores how much difference the choice of statistical test makes in practice, by analysis of a number of sets of real morbidity data by several statistical techniques. It uses the same methods to compare statistical tests as were used in section 2.5 describing re-analysis of 74 sets of real tumour response data.

The statistical techniques compared were the same as for the tumour response data of section 2.5, namely the Chi squared test of a dichotomy, the Chi squared test of all categories of morbidity, and the Mann-Whitney test of all categories. In the preceding section, no consistent policy was found over the choice of a dichotomy, with some analyses being carried out of the least severe grade of morbidity observed versus more severe grades, and some analyses combining lesser grades and testing the combination against more severe grades (e.g. grades 0 - 2 versus grades 3 - 5). Usually these latter combinations seemed to divide all the observations into roughly equal groups, but this was never explicitly stated as the rationale. Formation of a dichotomy so as to divide the total observations equally into two groups has theoretical justification (Connor, 1972). A third possible technique (not encountered in the preceding section) is to test the most severe grade of morbidity experienced against any lesser grade. In this section, all of these three variants of the Chi squared test were included.

Thus, all of the sets of data were each analyzed by the following 5 methods (all two-sided):

A. a Chi squared test of the dichotomy of the least severe grade of morbidity versus more severe grades,

B. a Chi squared test of the dichotomy which divided the total observations most nearly equally into two halves,

C. a Chi squared test of the dichotomy of the most severe grade of morbidity versus lesser grades,

D. a Chi squared test of all grades of morbidity subject to the rules avoiding small category sizes described in section 2.3, and

E. a Mann-Whitney test of all grades of morbidity.

In the cases of techniques A and C, Fisher's exact test was substituted where necessary because of small numbers, and categories were combined where necessary to give a total of at least 6 patients in the smaller category (since with less than 6 patients, Fisher's test cannot normally give a *P* value less than 5%). For technique E, the Kruskal-Wallis test was substituted where more than two treatments were compared. The methods were compared on an intention-to-analyze basis. For 51 cases of technique A, 30 cases of technique C, and 30 cases of technique D, the Chi squared test reduced to the same test as technique B, by forced combination of categories to avoid small numbers per category.

Where the outcomes of the treatments in a clinical trial are very different, it can be argued that no statistical test is necessary. It therefore seemed inappropriate to include in this re-analysis sets of morbidity data where a difference in outcome was obvious. To define such sets of data, a cutoff of a *z*-value of 4.0 ($P = 0.006\%$) was arbitrarily chosen as the upper limit for inclusion. As in section 2.5, the study size was planned to be such as to give at least 20 sets of data significant at 5%, and with *z*-values less than 4.0, i.e. with *z*-values between 1.96 and 4.0. The minimum of 20 significant sets of data was

chosen (as in section 2.5) in order to give a reasonably narrow 95% confidence interval of the square of the median *z*-value ratio (see Appendix D). Papers were studied in 3-month batches of the same 7 journals as used previously and morbidity data were recorded until this target was achieved. Data were not recorded for re-analysis if the total number of complications (of any grade) was less than 6 for two treatment groups, or less than 8 for 3 treatment groups. This gave the same set of 36 papers as were studied in the preceding section.

In 15 papers, the data were presented only as dichotomies which were thus unsuitable for re-analysis. The remaining 21 papers contained data on a total of 92 types of morbidity (up to 9 types per paper), presented as a table by grade and treatment (or the equivalent in text). In 4 of these 92 data sets, there were three or more treatments, and there were so few patients in the categories outside of one large category that a Chi squared test was not appropriate, and these 4 data sets were excluded. A further 8 sets of data were excluded because the numbers of patients outside one large category were too small to allow a Chi squared test by standard criteria (Armitage & Berry, 1987); a Mann-Whitney test may also be inappropriate in this situation. Of the remaining 80 sets of data, there were 17 types of morbidity where the *z*-values were greater than 4.0 by at least two of the techniques detailed above, and the remaining 63 sets of data were subjected to detailed analysis. The types of morbidity in these 63 sets of data are shown in Table 3.3. The numbers of patients included in the 63 data sets ranged from 35 to 2740 with a median of 251.

Table 3.3. Types of morbidity in the 63 sets of data used for re-analysis.

Type of morbidity	Number of sets of data
Nausea / Vomiting	13
Leucocyte nadir	6
Platelet nadir	3
Other haematological	4
Diarrhoea	4
Neurotoxicity	4
Renal	3
Other	26
Total	63

RESULTS AND DISCUSSION

NUMBERS OF SIGNIFICANT RESULTS

The numbers of results significant at 5% and at 1% by the five statistical methods are shown in Table 3.4. By chance, 3 or 4 results significant at 5% would be expected (i.e. 5% of 63). These results are similar to those for the tumour response data of section 2.5, i.e. the Mann-Whitney test again gave the highest numbers of significant results. This supports the Mann-Whitney test as being the most efficient test, in line with the

expectation from theory. Relatively little difference was found between the three variants of the Chi squared test of a dichotomy (techniques A, B and C). This is not surprising because in 51 of the 63 data sets, technique A became identical to technique B by the amalgamation of categories forced by small numbers, and in 30 of the data sets, technique C similarly became identical to technique B. For subsequent comparisons in this section between the Chi squared test of a dichotomy and techniques D and E, it was necessary to select one of the techniques A, B and C. Since little difference was found using the data studied between techniques A, B and C, the version chosen was technique B, as this has theoretical justification (Connor, 1972).

Table 3.4. Numbers of the 63 data sets with significant results by the 5 statistical techniques.

Technique	Number at 5%	Number at 1%
A. Chi squared test - no versus any morbidity	22	14
B. Chi squared test - even dichotomy	26	13
C. Chi squared test - severe versus less morbidity	27	15
D. Chi squared test - all grades of morbidity	28	15
E. Mann-Whitney test of all grades	28	18

RELATIONSHIP BETWEEN *z*-VALUES

Figure 3.1 displays the *z*-values for the 63 sets of data analyzed by the Mann-Whitney test and the Chi squared test of the dichotomy most equally dividing the patients. There is a general tendency for the points to lie above the line of equality. Of the 29 sets of data significant at 5% by either test, 7 lie below and 22 lie above the line of equality ($P = 0.5\%$ by the sign test). There is thus good evidence that the Mann-Whitney test is the more efficient.

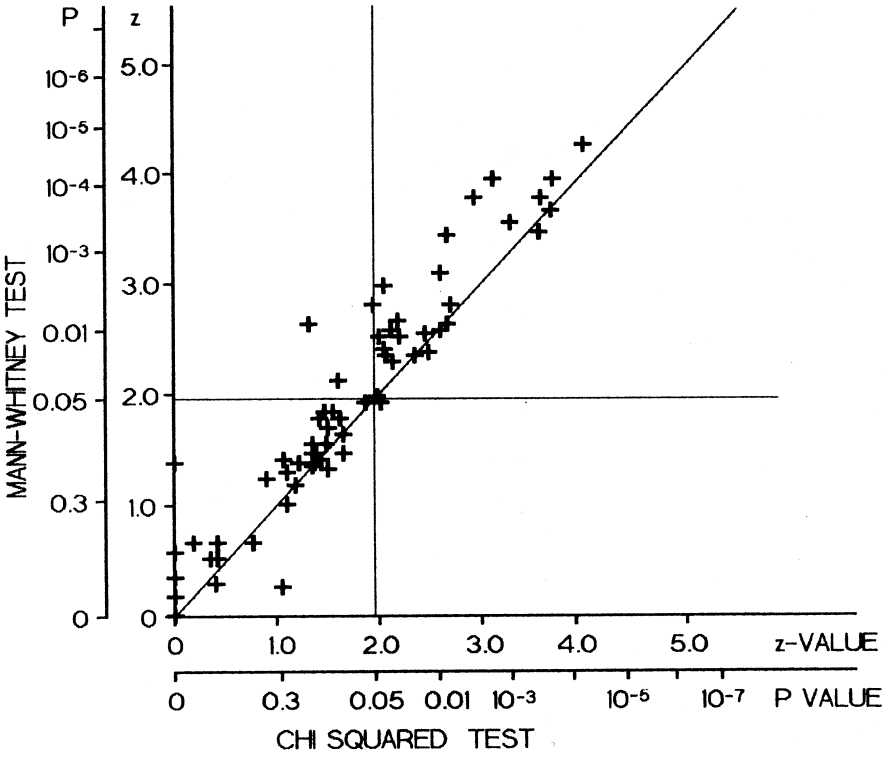


Figure 3.1. Comparison of the *z*-values (and the corresponding *P* values) for 63 sets of acute treatment morbidity data analyzed by both the Mann-Whitney test and the Chi squared test of the dichotomy most equally dividing the patients.

ESTIMATION OF RELATIVE EFFICIENCY

By the method of Appendix D, using the 29 sets of data significant by either test, the overall relative efficiency of the Mann-Whitney test relative to the Chi squared test of the dichotomy most equally dividing the patients is estimated as 1.18 (95% confidence interval 1.09 to 1.51). This overall estimate of the relative efficiency may be an underestimate, as in many sets of data, the results presented and available for this reanalysis had categories partially combined, e.g. results were presented for the 3 combined categories of 0 - 1, 2 - 3 and 4. By excluding such sets of data, which were presented using less than the number of categories in the standard morbidity scoring systems, a revised estimate of the relative efficiency of 1.31 is obtained.

For those morbidity data sets which were the primary end-point of the trial, e.g. where control of morbidity was the subject of study of the trial, the relative efficiency for the Mann-Whitney test relative to the Chi squared test of a dichotomy was estimated as 1.67 (from 7 significant data sets). These sets of data were more likely to have a wide distribution of patients across the morbidity categories, whereas when morbidity was a secondary endpoint, the data were more often largely limited to one or two grades of morbidity.

The best overall estimate for the advantage of the Mann-Whitney test in the analysis of acute morbidity of treatment over the Chi squared test of a dichotomy is thus around 1.3, or possibly more where morbidity is the primary end-point of the trial. This means that the majority of investigators, currently using the Chi squared test, can effectively increase the recruitment to their clinical trials by around 30% or more by simply switching their method of analysis to a Mann-Whitney test.

The distribution of the *z*-values indicated the efficiency of the Chi squared test of all grades of morbidity to be intermediate between the Chi squared test of a dichotomy and the Mann-Whitney test. This finding is consistent with expectation from theory and with the similar findings for the tumour response data of section 2.5.

3.5 STATISTICAL TESTING OF ACUTE MORBIDITY DATA: CONSIDERATIONS FROM MATHEMATICAL MODELLING

Sections 2.5 and 2.6 explored what can be learned from the construction of models concerning statistical analysis of tumour response data from clinical trials. Since there are strong parallels between tumour response data and acute morbidity data, the conclusions from sections 2.5 and 2.6 can be tentatively extended to acute morbidity data.

A priori, a lateral shift model again seems the most reasonable model for the effect of an increase in treatment efficacy on the distribution of patients in the various categories, i.e. grades of morbidity. This model would imply a similar effect of treatment at all grades of morbidity, which would follow if the pathogenesis at each of the grades of morbidity was similar in type, but differing in degree - which is usually the case. On this basis, the implications are qualitatively the same as for tumour response data, i.e. that analysis using several categories of classification is more efficient than the analysis of a dichotomy (but there is little gain in efficiency when the number of categories increases beyond 6). These considerations would again support the analysis of acute morbidity data by a Mann-Whitney test using all the categories of the standard assessment systems.

3.6 METHODS OF ESTIMATION OF ACUTE MORBIDITY DATA: A SURVEY AND DISCUSSION

INTRODUCTION

The statistical techniques available for hypothesis testing and for estimation of differences in tumour response were discussed in Chapter 2. The survey of section 2.8 confirmed that it is common practice for tumour response data to be summarised in publications by the total response rates or the complete response rates (or both), and that differences between treatments are summarised most commonly by the arithmetical differences in the total response rates. Confidence intervals were calculated only rarely. The preceding sections of this chapter have discussed methods of hypothesis testing in treatment morbidity. For completeness, this section considers methods of estimation of the difference in morbidity in a clinical trial of two treatments, although the clinical relevance of such methods is unclear.

A SURVEY

A survey was made of all trials reporting treatment morbidity in 1990 in the 7 journals used previously. The 55 papers found were largely the same as those of the previous two sections. For each paper, the most complex method used to summarize the morbidity data was recorded (there were often several types of morbidity presented and for obvious reasons, some types were presented in more detail than others). The use of confidence intervals was also recorded. Of the 55 papers, 51 reported categorical data only, 2 reported continuous data only, and 2 reported both. The 4 papers reporting continuous data all used means or medians as summary measures - this will not be discussed further. Of the 53 papers reporting categorical data, 23 summarized the data by calculating the percentage of patients experiencing each grade of morbidity for each

treatment. None of the papers explicitly calculated the arithmetic difference in the proportion of patients experiencing morbidity, but this was sometimes done implicitly. There was no use of more complex methods of estimation such as calculation of odds ratios.

A total of three confidence intervals were calculated (in two of the 55 papers). One was of a difference in a continuous measure (difference in a visual analogue score between treatments), and two were of data reduced to dichotomies. One of these was calculated incorrectly, and in the other, the percentage of the confidence interval was not stated.

DISCUSSION

In section 2.9, there was a discussion of the methods of estimation of the difference between treatments that can be applied to tumour response data. These same methods (difference in proportions, odds ratios, etc) can be applied to acute morbidity data, but again, it is not clear that they will assist oncologists in the interpretation of the data. On theoretical grounds, the best method of estimation may be by the Mann-Whitney estimator, with the advantages of making use of all the grades of morbidity, and of avoiding the need to select a dichotomy. For treatment morbidity data, there is no summary measure that parallels the total response rate for tumour response data, and how such data are best presented is not clear. It may be that the best simple summary measure is in terms of the dichotomy that most equally divides the total observations into two groups.

CHAPTER 4. GENERAL DISCUSSION

Four surveys have been presented in this dissertation - on statistical tests and methods of estimation for both tumour response and acute treatment morbidity data - and all have shown a wide gap between recommendations in the literature and the procedures in practice.

Considering firstly the recommendations on statistical testing, considerable evidence has been presented in favour of the recommendations that appear in several sources for use of more complex methods than the simple Chi squared test in analyzing ordered categorical data, such as tumour response and treatment morbidity. For the slight increase in complexity of the Mann-Whitney test, the gains in practice are not insignificant - a switch in the method of analysis from the Chi squared test to a Mann-Whitney test is equivalent to increasing recruitment to the clinical trial by around 40% for tumour response data and by around 30% or more for treatment morbidity data.

The improved efficiency of analysis of the Mann-Whitney test has a number of practical consequences. If the number of patients entered in a clinical trial is unchanged, the improved efficiency means that small differences in outcome between the treatment groups are more readily detected (the trial is more likely to find a difference significant at 5%), and also large differences in outcome will be detected more conclusively (the *P* value from the trial is more likely to be small enough for the result to be regarded as conclusive). Alternatively, the improved efficiency of the Mann-Whitney test could be exploited by reducing the numbers of patients entered, so that trials could be completed

in a shorter time and at less cost than previously. A further consequence is that it may be possible to carry out some studies which otherwise would not be commenced through lack of numbers of patients.

Why the Mann-Whitney test is not being used in the analysis of tumour response and treatment morbidity data is not clear. It may be that there is a need for better education and wider dissemination of the recommendations for the use of the Mann Whitney test, such as those of Moses et al (1984). Secondly, there may be a need for better communication between clinicians and statisticians as to the nature of tumour response and treatment morbidity data. The third and perhaps the most likely reason is that there may be a general reluctance to use a more complex statistical test unless it has been shown to be a definite improvement on the simpler alternative. Such a reluctance is reasonable, especially as it has been shown in this dissertation in the modelling of section 2.6 that the Mann-Whitney test is not more efficient than a Chi squared test for all conceivable models of treatment effect - the Mann-Whitney test was actually less efficient than a Chi squared test on the assumption of the equal subdivision model (this model was in fact subsequently shown in section 2.7 to be a poor fit to clinical data). Thus it appears to be necessary to test general recommendations (such as for the use of the Mann-Whitney test) in each field of scientific research. The techniques used in this dissertation for the comparison of statistical tests deserve wider application.

It is less clear that a change is needed in current practice in methods of estimation and calculation of confidence intervals. The calculation of simple response rates and morbidity rates as percentages provides the information that oncologists require in their clinical practice - a change to more elaborate methods such as use of odds ratios appears hard to justify, in view of the difficulties in understanding what they mean.

USE OF CONFIDENCE INTERVALS IN CANCER CLINICAL TRIALS

The use of confidence intervals in cancer clinical trials will be discussed in more detail because of the current clamour for them (Gardner and Altman, 1988; Anonymous, 1987). When a study is aimed at estimating a single value e.g. the response rate to a particular drug, it is important to have a measure of the reliability of the estimate, and the calculation of 95% confidence intervals achieves this purpose. The use of standard errors and of error bars on charts has been long established in scientific work for a similar end.

The extension of use of confidence intervals to comparative clinical trials is less clear. Here, it is not a single estimate that is important, but the difference in outcome between 2 or more treatments. To concentrate on the *P* value from a clinical trial to the exclusion of the size of the difference found is wrong. There are also certain kinds of cancer clinical trial where a test of a null hypothesis is not appropriate, such as dose ranging studies (which may be randomised) where the intention is to find the maximum tolerated dose; a test of whether different doses result in differences in response or morbidity is inappropriate - it is known that there are differences and the clinical question concerns how large the differences are. Here there is a need for the calculation of confidence intervals.

There are, however, problems with extending the calculation of confidence intervals to cancer clinical trials in general. In many trials, the outcome in any patient depends on many prognostic factors, e.g. the response to chemotherapy in metastatic breast cancer depends on performance status, number of disease sites, and so on (Henderson et al, 1989). These prognostic factors may in fact be more important than the choice of treatment given in determining the response rate. Here, the difference in outcome e.g. the arithmetic difference between response rates between the treatments is likely to depend heavily on the prognostic factors. There is thus not one single difference

to be estimated. In contrast, whether or not a difference exists (and in which direction) may well be constant across all prognostic factors, and there is no corresponding problem with hypothesis testing. This problem for estimation where prognostic factors are important may be overcome by the use of more complex models, but what they estimate may not be meaningful to clinicians.

A further difficulty is that many trials compare more than two treatments. Here, it is possible to perform a statistical test of the hypothesis of no difference between any of the treatments, whereas confidence intervals of the difference between treatments can only be done between pairs of treatments. Furthermore, when a new treatment is being developed, it may not matter if the difference from existing treatments is only small, since it may well be possible to modify the new treatment to optimise it and enlarge on that difference; the important question may be whether any difference in efficacy exists. The development of radiation sensitizers is an example where the difference in outcome over conventional treatment is small, but sufficient to encourage further research.

There are also problems over the interpretation of confidence intervals. The more correct interpretation is of less use in clinical practice, and the simpler (incorrect) interpretation, that the 95% confidence interval is a range of values that includes the true value with a probability of 95%, is the one that most clinicians will adopt. However, the actual frequency depends on what kinds of true difference are likely to occur, and this has been little studied.

THE IMPORTANCE OF STATISTICS IN MEDICINE

It is well established that clinicians' understanding of statistics could be improved (Altman & Bland, 1991). For some clinicians, statistical methods are a routine to be performed at the end of a clinical study. The work of this dissertation may help to establish that statistical issues should be considered throughout the planning and performance of clinical trials, and not just in the final analysis - in particular, data should be recorded in sufficient detail to enable an efficient analysis to be performed.

The conclusions of this dissertation on the greater efficiency of the Mann-Whitney test compared to the Chi squared test in ordered categories may well also apply to other measurement scales in medicine. In particular, in Oncology, quality of life is often recorded as ordered categorical data, and it may well be that the Mann-Whitney test should be used in this area as well.

STATISTICAL DEVELOPMENTS

There are some disadvantages to the Mann-Whitney test. It is less widely known and less available on computer statistical software. Routines for use of the Mann-Whitney test on ordered categorical data should be added to computer software packages. However, the Mann-Whitney test can be done on ordered categorical data using a hand calculator in only a few minutes more than a Chi squared test; and these disadvantages of the Mann-Whitney test are minor compared with the effective increase in recruitment to clinical trials of 30% to 40%.

Although the Mann-Whitney test has been established as an appropriate statistical test for ordered categorical data for large sample sizes, the limitations of the test where there are both small numbers of categories and small samples has not been explored. As

an example, consider the data of Table 4.1, which is from one of the 81 papers surveyed in section 2.2. Standard criteria (Altman, 1991) prevent the application of the Chi squared test to such data either to the whole table or after collapsing to a dichotomy, because the numbers outside of the CR category are too small. It seems likely that the Mann-Whitney test is also not appropriate. This difficulty with the Mann-Whitney test does not weaken the conclusions of this dissertation since of the 27 significant sets of response data used in section 2.5, only one (which is shown in Table 4.1) was of this form, and exclusion of this data set results in only a slight change in the estimated relative efficiency; this problem was minimized in section 3.4 re-analyzing morbidity data by excluding all data sets where the Chi squared test is not valid by standard criteria. Further work on the small sample limitations of the Mann-Whitney test is required.

Table 4.1. Tumour response data from Cimino et al (1989) comparing a chemotherapy regime with a radiotherapy technique in the treatment of Hodgkin's disease.

Category of response	Chemotherapy	Radiotherapy
Complete response	40	45
Partial response	0	0
No change	1	0
Progressive disease	3	0

CHAPTER 5. CONCLUSIONS

This dissertation has studied a number of aspects of the design and analysis of cancer clinical trials. It is well established (Altman & Bland, 1991) that statistical errors are common in medical research in general. This dissertation has quantified the extent of misapplication of statistical methods in Oncology, in particular in the analysis of tumour response and acute treatment morbidity. It may be expected on theoretical grounds that the Mann-Whitney test is more efficient than the Chi squared test in the analysis of tumour response and treatment morbidity. This dissertation has both confirmed and quantified this increased efficiency over a considerable number of published data sets - derived from 81 trials that included analysis of tumour response, and from 36 trials that included analysis of acute treatment morbidity.

The dissertation also includes a new method for comparing the performance of statistical tests in practice - by estimating the relative efficiency from the median value of the square of the *z*-values. In contrast to theoretical derivation of relative efficiency, this new method can be used when the model of treatment effect (i.e. the alternative hypothesis) is unclear, which is the case in very many situations. This new method was used in the dissertation to quantify the advantage of the Mann-Whitney test over the Chi squared test in the types of clinical data considered, and this quantification of the advantage of the Mann-Whitney test may help to encourage its wider use.

It may be possible to refine this new method for estimating the relative efficiency of two statistical tests. The method uses the median value of the square of the ratio of the

z-values of the significant data sets. It does not take account of the number of observations made (e.g. the number of patients in the trial) or the size of the departure from the null hypothesis (e.g. the size of the difference in response or morbidity). It may be that a refinement leading to a better estimator of relative efficiency would be to include a weighting for the number of observations, or the size of difference, or the size of the *z*-values themselves (which are a reflection of the first two factors).

The dissertation also investigated the relationship between the four categories of tumour response - which has not previously been studied - and this relationship was found to fit best to a lateral shift model.

The weaknesses of the dissertation are that some parts of it, working in new areas, are rather preliminary. The comparison of statistical tests of tumour response included all types of solid tumour, and further work may be worthwhile, looking at individual tumour types or particular clinical situations to see whether the general conclusions of this dissertation regarding tumour response apply in all situations. Similarly, the comparison of statistical tests of acute treatment morbidity in this dissertation covers a range of types of morbidity, a range of treatment comparisons and also several different reasons for measuring morbidity. Again, further work may clarify whether there are any exceptions to the general conclusions of the dissertation. However, these weaknesses should not delay adoption of the Mann-Whitney test as the standard method of analysis for tumour response and acute treatment morbidity since the practical evidence collected in this dissertation is backed by theoretical arguments. If the clinical situation does have any bearing, it is likely to be merely on the size of the advantage of the Mann-Whitney test. In support of this view, it can be seen from the figures of sections 2.5 and 3.4 that the advantage of the Mann-Whitney test was consistent across almost all of the data sets examined (in that almost all the points lie above the line of equality). The exceptional data sets may well have arisen through random effects.

A number of recommendations can be made on the basis of the work of this dissertation.

DESIGN OF CANCER CLINICAL TRIALS

- Tumour response data should be recorded using the 4 standard categories of assessment
- The category of "minimal response" should not be added to the standard 4 category classification system without good justification, as it may actually decrease the efficiency of analysis.
- Acute treatment morbidity data should be recorded using the standard 5 or 6 categories of classification.
- When measurement scales are being developed for comparison of treatments, the aim should be for 4 to 6 categories of measurement to each contain appreciable numbers of observations..

ANALYSIS OF CANCER CLINICAL TRIAL DATA

- Tumour response data should be analyzed by the Mann-Whitney test.
- Certain acute treatment morbidity data should be analyzed by the Mann-Whitney test.
- Further research is required on more elaborate methods of analysis using regression methods and ordered categorical regression models, to assess whether these can further improve the efficiency of analysis.
- Questions over the optimal choice of statistical methods could and should be answered more often by empirical comparison of the methods in practice using real sets of data.
- The facility to perform the Mann-Whitney test on ordered categorical data should be added to those statistical software systems where it is currently lacking.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support and encouragement of Prof. GRH Sealy. I am also very grateful to Dr MKB Parmar, MRC Cancer Trials Office, Cambridge, who has acted as supervisor, to the Clatterbridge Cancer Research Trust for a grant which enabled the bulk of the work to be done, and to my wife for proof-reading.

REFERENCES

AARONSON, N. K., BULLINGER, M., & AHMEDZAI, S., 1988.
A modular approach to quality-of-life assessment in cancer clinical trials. *Recent Results in Cancer Research, 111,* 231-248.

ALTMAN, D. G., 1980.
Statistics and ethics in medical research. III How large a sample? *British Medical Journal, 281*, 1336-1338.

ALTMAN, D. G., 1991.
Practical Statistics for Medical Research. (Chapman and Hall, London).

ALTMAN, D. G., & BLAND, J. M., 1991.
Improving doctors' understanding of statistics. *Journal of the Royal Statistical Society A, 154*, 223-267.

ALTMAN, D. G., GORE, S. M., GARDNER, M. J. & POCOCK, S. J., 1983.
Statistical guidelines for contributors to medical journals. *British Medical Journal, 286*,1489-1493.

ANDERSON, J. A., 1984.
Regression and ordered categorical variables (with discussion). *Journal of the Royal Statistical Society B, 46*, 1-30.

ANONYMOUS, 1987.
Report with confidence [Editorial]. *Lancet, i,* 488.

ARMITAGE, P., 1955.
Tests for linear trends in proportions and frequencies. *Biometrics, 11,* 375-386.

ARMITAGE, P. & BERRY, G., 1987.
Statistical Methods in Medical Research. (Blackwell Scientific Publications, Oxford).

BARTOLUCCI, A. A., 1984.
Estimation and comparison of proportions. In *Cancer Clinical Trials.* Ed. by M. E. Buyse, M. J. Staquet and R. J. Sylvester (Oxford Medical Publications, Oxford), pp. 337-360.

BOAG, J. W., HAYBITTLE, J. L., FOWLER, J. F., & EMERY, E. W., 1971.
The number of patients required in a clinical trial. *British Journal of Radiology, 44,* 122-125.

BRESLOW, N., 1984.
Comparison of survival curves. In *Cancer Clinical Trials.* Ed. by M. E. Buyse, M. J. Staquet and R. J. Sylvester (Oxford Medical Publications, Oxford), pp. 381-406.

CIMINO, G., BITI, G. P., ANSELMO, A. P., MAURIZI-ENRICI, R., BELLESI, G. P., BOSI, A., CIONINI, L., MUNGAI, V., PAPA, G., & PONTICELLI, P., 1989.
MOPP chemotherapy versus extended-field radiotherapy in the management of pathological stages I - IIA Hodgkin's disease. *Journal of Clinical Oncology, 7,* 732-737.

CLARK, A., & FALLOWFIELD, L. J., 1986.
Quality of life measurements in patients with malignant disease: a review. *Journal of the Royal Society of Medicine, 79,* 165-169.

COCHRAN, W. G., 1954.
Some methods for strengthening the common χ^2 tests. *Biometrics, 10,* 417-451.

CONNOR, R. J., 1972.
Grouping for testing trends in categorical data. *Journal of the American Statistical Association, 67,* 601-604.

DISCHE, S., WARBURTON, M. F., JONES, D., & LARTIGAU, E., 1989.
The recording of morbidity related to radiotherapy. *Radiotherapy and Oncology, 16,* 103-108.

DUNCAN, W., 1983.
A clinical evaluation of fast neutron therapy. In *The Biological Basis of Radiotherapy.* Ed. by G. G. Steel, G. E. Adams and M. J. Peckham (Elsevier, Oxford), pp. 277-286.

EINHORN, L. H., CRAWFORD, E. D., SHIPLEY, W. U., LOEHRER, P. J., & WILLIAMS, S. D., 1989.
Cancer of the testes. In *Cancer: Principles and Practice of Oncology.* Ed. by V. T. DeVita, S. Hellman and S. A. Rosenberg (Lippincott, Philadelphia), pp. 1071-1098.

ERRINGTON, R. D., ASHBY, D., GORE, S. M., ABRAMS, K. R., MYINT, S., BONNETT, D. E., BLAKE, S. W., & SAXTON, T. E., 1991.
High energy neutron treatment for pelvic cancers: study stopped because of increased mortality. *British Medical Journal, 302,* 1045-1051.

FITZPATRICK, R., FLETCHER, A., GORE, S., JONES, D., SPIEGELHALTER, D., & COX, D., 1992.
Quality of life measures in health care. I: Applications and issues in assessment. *British Medical Journal, 305,* 1074-1077.

GARDNER, M. J., & ALTMAN, D. G., 1988.
Estimating with confidence. *British Medical Journal 296,* 1210-1211.

GARDNER, M. J., MACHIN, D., & CAMPBELL, M. J., 1986.
Use of check lists in assessing the statistical content of medical studies. *British Medical Journal, 292,* 810-812.

GELBER, R. D., GOLDHIRSCH, A., & CAVALLI, F., 1991.
Quality-of-life-adjusted evaluation of adjuvant therapies for operable breast cancer. *Annals of Internal Medicine, 114,* 621-628.

GELMAN, R., & ZELEN, M., 1987.
Interpreting clinical data. *In Breast Diseases.* Ed. by J. R. Harris, S. Hellman, I.C. Henderson and D. W. Kinne (J. B. Lippincott Company, London), pp. 697-731.

GUDEX, C., & KIND, P., 1988.
The Qaly Toolkit. (Centre for Health Economics, York).

HALL, E. J., 1978.
Radiobiology for the Radiologist. (Harper and Row, Philadelphia).

HAYBITTLE, J. L., 1983.
What is cure in cancer? In *Cancer Treatment: End-point Evaluation.* Ed. by B. A. Stoll (John Wiley & Sons, Chichester), pp. 3-21.

HAYWARD, J. L., CARBONE, P. P., HEUSON, J.-C., KUMAOKA, S., SEGALOFF, A., & RUBENS, R. D., 1977.
Assessment of response to therapy in advanced breast cancer. *European Journal of Cancer, 13,* 89-94.

HENDERSON, I. C., HARRIS, J. R., KINNE, D. W., & HELLMAN, S., 1989.
Cancer of the breast. In *Cancer: Principles and Practice of Oncology.* Ed. by V. T. DeVita, S. Hellman and S. A. Rosenberg (Lippincott, Philadelphia), pp. 1197-1268.

HOLTBRUGGE, W., & SCHUMACHER, M., 1991.

A comparison of regression models for the analysis of ordered categorical data. *Journal of the Royal Statistical Society (Series C), 40*, 249-259.

HOOGSTRATEN, B., 1984.

Reporting treatment results in solid tumours. In *Cancer Clinical Trials.* Ed. by M. E. Buyse, M. J. Staquet and R. J. Sylvester (Oxford Medical Publications, Oxford), pp. 139-156.

HUSLER, J., & RIEDWYL, H., 1989.

Comparison of the efficiency of nonparametric location tests based on grouped and individual data. *Statistics and Probability Letters, 7,* 287-291.

ISRAEL, L., 1983.

Evaluating response in soft tissue tumour. In *Cancer Treatment: End-point Evaluation.* Ed. by B. A. Stoll (John Wiley & Sons, Chichester), pp. 67-73.

IVERSEN, P., CHRISTENSEN, M. G., FRIIS, E., HORNBOL, P., HVIDT, V., IVERSEN, H. G., KLARSKOV, P., KRARUP, T., LUND, F., MOGENSEN, P., PEDERSEN, T., RASMUSSEN, F., ROSE, C., SKAARUP, P., & WOLF, H., 1990.

A phase III trial of Zoladex and flutamide versus orchidectomy in the treatment of patients with advanced carcinoma of the prostate. *Cancer, 66,* 1058-1066.

KAPLAN, E. L. & MEIER, P., 1958.

Nonparametric estimation from incomplete observations. *American Statistical Association Journal, 53,* 457-481.

KAPLAN, H. S., 1966.

Evidence for a tumoricidal dose level in the radiotherapy of Hodgkin's disease. *Cancer Research, 26,* 1221-1224.

KIND, P., 1988.

The Design and Construction of Quality of Life Measures. (Centre for Health Economics Discussion Paper 43, York).

LACHIN, J. M., 1981.

Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials, 2,* 93-113.

LEVENTHAL, B. G., & WITTES, R. E., 1988.

Research Methods in Clinical Oncology. (Raven Press, New York).

McCULLAGH, P., 1980.

Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society B, 42,* 109-142.

MACHIN, D. & CAMPBELL, M. J., 1987.
Statistical Tables for the Design of Clinical Trials. (Blackwell Scientific Publications, Oxford).

MAGUIRE, P. & SELBY, P., 1989.
Assessing quality of life in cancer patients. *British Journal of Cancer. 60,* 437-440.

MANN, H. B. & WHITNEY, D. R., 1947.
On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18,* 50-60.

MILLER, A. B., HOOGSTRATEN, B., STAQUET, M., & WINKLER, A., 1081.
Reporting results of cancer treatment. *Cancer 47,* 207-214.

MOERTEL, C. G. & HANLEY, J. A., 1976.
The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer, 38,* 388-394.

MONFARDINI, S., BRUNNER, K., CROWTHER, D., ECKHARDT, S., OLIVE, D., TANNEBERGER, S., VERONESI, A., WHITEHOUSE, J. M. A., & WITTES, R., 1987.
Manual of Adult and Paediatric Medical Oncology. (Springer-Verlag, London).

MORTON, A. P. & DOBSON, A. J., 1990.
Analyzing ordered categorical data from two independent samples. *British Medical Journal, 301,* 971-973.

MOSES, L.E., EMERSON, J.D. & HOSSEINI, H., 1984.
Analyzing data from ordered categories. *New England Journal of Medicine, 311,* 442-448.

NEAVE, H. R., 1981.
Elementary Statistics Tables. (Unwin Hyman, London).

NOETHER, G. E., 1987.
Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association, 82,* 645-647.

OKAWA, T., AND OTHERS OF THE JAPAN RADIATION-ACNU STUDY GROUP, 1989.
A randomized prospective study of radiation versus radiation plus ACNU in inoperable non-small cell carcinoma of the lung. *Cancer, 63,* 249-254.

PARBHOO, S., & WAHBA, A., 1983.
Evaluating response in lesions of lung, liver, and central nervous system. In *Cancer Treatment: End-point Evaluation. Ed. by B.* A. Stoll (John Wiley & Sons, Chichester), pp. 75-112.

PARSONS, J. T., McCARTY, P. J., RAO, P. V., MENDENHALL, W. M., & MILLION, R. R., 1990.
On the definition of local control. *International Journal of Radiation Oncology Biology Physics, 18*, 705-706.

PEARSON, E. S. & HARTLEY, H. O., 1976
Biometrika Tables for Statisticians Volume 1. (Biometrika Trust, University College, London).

PETO, R., PIKE, M. C., ARMITAGE, P., BRESLOW, N. E., COX, D. R., HOWARD, S. V., MANTEL, N., McPHERSON, K., PETO, J., & SMITH, P. G., 1976.
Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer, 34*, 585-612.

PETO, R., PIKE, M. C., ARMITAGE, P., BRESLOW, N. E., COX, D. R., HOWARD, S. V., MANTEL, N., McPHERSON, K., PETO, J., & SMITH, P. G., 1977.
Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer, 35*, 1-39.

POCOCK, S. J., 1978.
Size of cancer clinical trials and stopping rules. *British Journal of Cancer, 38,* 757-766.

POCOCK, S. J., 1983.
Clinical Trials: A Practical Approach. (John Wiley & Sons, New York).

POCOCK, S. J., ARMITAGE, P. & GALTON, D. A. G., 1978.
The size of cancer clinical trials: An international survey. In *Methods and Impact of Controlled Therapeutic Trials in Cancer. Part I.* (UICC Technical Report Series - Volume 36, UICC, Geneva), pp. 5-32.

POCOCK, S. J., HUGHES, M. D. & LEE, R. J., 1987.
Statistical problems in the reporting of clinical trials. *New England Journal of Medicine, 317,* 426-432.

PRATT, J. W., & GIBBONS, J. D., 1981.
Concepts of Nonparametric Theory. (Springer-Verlag, New York).

RANDLES, R. H., & WOLFE, D. A., 1979.
Introduction to the Theory of Nonparametric Statistics. (John Wiley & Sons, New York).

SEARS, M. E., & OLSON, K. B., 1980.
Extramural review of clinical response of breast cancer to cytotoxic chemotherapy. *Cancer, 46,* 2928-2929.

SIMON, R. M., 1989.

Design and conduct of clinical trials. In *Cancer: Principles and Practice of Oncology*. Ed. by V. T. DeVita, S. Hellman and S. A. Rosenberg (Lippincott, Philadelphia), pp. 396-420.

SMITH, I. E., 1983.

Measuring response in incurable cancer. In *Cancer Treatment: End-point Evaluation. Ed. by B. A.* Stoll (John Wiley & Sons, Chichester), pp. 23-42.

SPIEGELHALTER, D. J., GORE, S. M., FITZPATRICK, R., FLETCHER, A. E., JONES, D. R., & COX, D. R., 1992.

Quality of life measures in health care. III: resource allocation. *British Medical Journal, 305*, 1205-1209.

SPRENT, P., 1989.

Applied Nonparametric Statistical Methods. (Chapman and Hall, London).

STAQUET, M., & DALESIO, O., 1984.

Designs for phase III trials. In *Cancer Clinical Trials.* Ed. by M. E. Buyse, M. J. Staquet and R. J. Sylvester (Oxford Medical Publications, Oxford), pp. 261-275.

STRAUSS, M. B., 1968.

Familiar Medical Quotations. (J. and A. Churchill, London).

SUTTON, M. L., & HENDRY, J. H., 1985.

Applied radiobiology. In *The Radiotherapy of Malignant Disease.* Ed. by E. C. Easson and R. C. S. Pointon (Springer-Verlag, Berlin), pp. 33-55.

TAIT, D., PECKHAM, M. J., HENDRY, W. F., & GOLDSTRAW, P., 1984.

Post-chemotherapy surgery in advanced non-seminomatous germ-cell testicular tumours: The significance of histology with particular reference to differentiated (mature) teratoma. *British Journal of Cancer, 50*, 601-609.

WHO, 1979.

WHO Handbook for Reporting Results of Cancer Treatment. (WHO Offset Publication No. 48, WHO, Geneva).

WICHMANN, B. A. & HILL, I. D., 1982.

An efficient and portable pseudo-random number generator. *Journal of the Royal Statistical Society (Series C), 31*, 188-190.

WIERNIK, G., ALCOCK, C. J., BATES, T. D., BRINDLE, J. M., FOWLER, J. F., GAJEK, W. R., GOODMAN, S., HAYBITTLE, J. L., HENK, J. M., HOPEWELL, J. W., HUNTER, R. D., LINDUP, R., PHILLIPS, D. L., & REZVANI, M., 1991.
Final report on the second British Institute of Radiology fractionation study: short versus long overall treatment times for radiotherapy of carcinoma of the laryngo-pharynx. *British Journal of Radiology, 64,* 232-241.

WILKINSON, P. M., & FOX, B. W., 1985.
Principles of chemotherapy. In *The Radiotherapy of Malignant Disease.* Ed. by E. C. Easson and R. C. S. Pointon (Springer-Verlag, Berlin), pp. 57-83.

WOODWARD, M., & FRANCIS, L. M. A., 1988.
Statistics for Health Management and Research. (Edward Arnold, London).

YATES, F., 1948.
The analysis of contingency tables with groupings based on quantitative characters. *Biometrika, 35,* 176-181.

APPENDICES

APPENDIX A: DETAILS OF WORK DONE

This work was carried out during the post of Senior Registrar in Clinical Oncology at Mersey Regional Centre for Radiotherapy and Oncology, Clatterbridge Hospital, Merseyside, and the post of Research Fellow attached to the J. K. Douglas Laboratories, Clatterbridge Hospital.

It is entirely my own work. Dr M.K.B. Parmar of the MRC Clinical Trials Office, Cambridge, acted as supervisor, and assisted through discussion of the general direction of the work and discussion of the results obtained.

Publications relating to the dissertation are:

CAMPBELL, I. R., 1992. Confidence intervals. *The Royal Statistical Society News and Notes, 18.7,* 3.

CAMPBELL, I. R., & PARMAR, M. K. B. Estimation of the relative efficiency of two statistical tests from the ratio of the *z*-values from typical data sets. Submitted 1992.

APPENDIX B: LIST OF SYMBOLS USED.

The following mathematical symbols are used in the text

P The probability of the observed results occurring (or more extreme results) assuming the null hypothesis of no difference between treatments. In some journals, the symbol p is preferred. The P value may be expressed as a percentage (e.g. 4.2%) or equivalently as a decimal (e.g. 0.042). The former is used in this dissertation for ease of comparison of different P values.

χ^2 Chi squared

$e_{A,B}$ Relative efficiency. If statistical test A requires n_A observations to detect a difference δ at a significance level α with power $1-\beta$, and test B requires n_B observations to do the same, then the relative efficiency of test A relative to test B $e_{A,B}$ is given by the ratio n_B / n_A .

E The efficacy of a treatment.

$\varphi(x)$ The normal distribution probability density function. The formula for the inverted U-shaped graph of the standard normal distribution shown in Fig. B.1 is

$$\frac{e^{-x^2/2}}{\sqrt{(2\pi)}}$$

The symbol usually used to represent this is $\varphi(x)$ and this is termed the probability density function of the normal distribution.

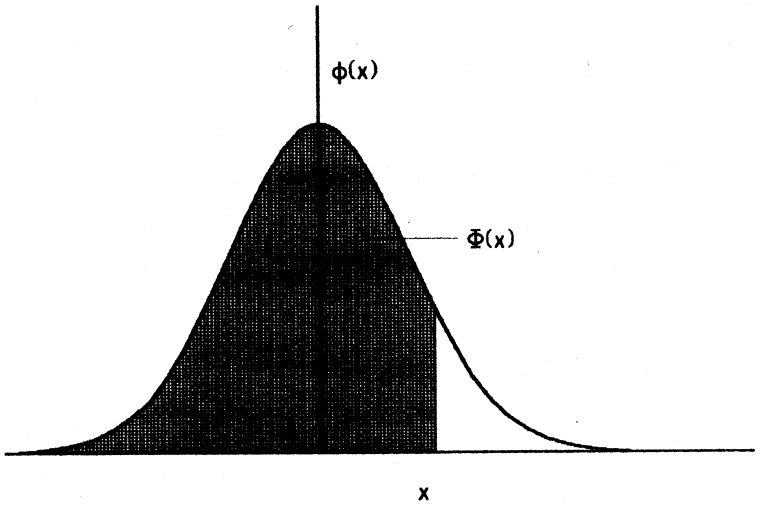


Fig. B.1. The standard normal distribution

$\Phi(x)$ The cumulative normal distribution function, i.e. the area under the standard normal distribution curve up to the value x (as demonstrated in Fig. B.1). The formula is

$$\int_{-\infty}^{x} \frac{e^{-x^2/2}}{\sqrt{(2\pi)}}\, dx$$

z-value The standard normal deviate from a statistical test (the deviation of the test statistic from its null hypothesis value divided by the standard error). For example, a z-value of 1.960 is just significant at 5%.

z_x The 100(1-x) percentile of the standard normal distribution, e.g. $z_{0.025} = 1.96$. Thus

$$\Phi(z_x) = 1 - x$$

k The number of categories of a classification system.

θ The total response rate

σ	Standard deviation
$\sum_{r=1}^{k}$	An instruction to add all the values of the expression to the right of the symbol by systematically changing the index r from a starting point of 1 to finish at k.
n_1	The number of patients or observations in group 1.
n_2	The number of patients or observations in group 2.
n	The total number of patients or observations = $n_1 + n_2$.

APPENDIX C: THE FORM USED FOR THE RECORDING OF THE STATISTICAL ANALYSIS OF TUMOUR RESPONSE DATA IN PUBLISHED ARTICLES

FORM FOR SURVEY OF METHODS OF ANALYSIS OF RESPONSE DATA v1.7

Number in series _____ Paper photocopied and labelled "RA *n*" Y N

First author _____ Year ___ Journal _____ Vol ___ P ___

Tumour _____

Response system used Standard Specially constructed Unstated

Type of comparison C/T vs C/T C/T vs H H vs H S abl vs H/C/T Other

Number of arms __ Treatments

Number of categories __

Divisions used PD/NC Y N

MR/PR Y N

PR/CR Y N

Numbers per category:		arm1	arm2	arm3	arm4	arm5	Meaning of category
worst	cat1	.	.	.	.	.	
first	cat2	.	.	.	.	.	
	cat3	.	.	.	.	.	
	cat4	.	.	.	.	.	

Analysis in paper:

Method None χ^2 MW Multivariate analysis of response Unclear

Was the contingency table collapsed Y N Unclear

If Y, to how many categories ___

Derived statistic reported Y N value ___

P value reported Y N value ____

One-sided Two-sided Unstated

One-tailed Two-tailed Unstated

Was response a major end-point of the study (included in abstract) Y N

Conclusion "

Re-analysis:

χ^2 test of response

Appropriate Y N Result ___ df__ Equiv z ___ *P* ___

χ^2 test of full table or table after partially collapsing because of numbers

Appropriate Y N No of cats. ___ Result ___ df__ Equiv *z* ___ *P* ___

Nonparametric test

Mann-Whitney Result *U* _____ *z* ____ *P* ___

Kruskal-Wallis Result _____ df ___ Equiv *z* ___ *P* ___

Comments

APPENDIX D: ESTIMATION OF THE RELATIVE EFFICIENCY OF TWO STATISTICAL TESTS FROM THE RATIO OF THE *z*-VALUES FROM TYPICAL DATA SETS

SUMMARY

The standard method for deriving the relative efficiency of two statistical tests is through the power functions of the two tests. These functions, and therefore the derived relative efficiency, depend critically on the assumed alternative hypothesis. This appendix presents a new method of estimating relative efficiency which avoids the need to assume a particular alternative hypothesis. It uses a number of data sets typical of a field of scientific research. Each data set is analyzed by each of the two statistical tests, and *z*-values are derived from the test statistics. The relative efficiency of the two tests is then estimated by the median value (over the data sets that are statistically significant) of the square of the ratio of the *z*-values. This estimate of the relative efficiency may be used to indicate which of the statistical tests is the more efficient (and by how much) in the analysis of other data sets in the same field of scientific research.

INTRODUCTION

The estimation of relative efficiency is important in the selection of a statistical test for data analysis. Selection of the most efficient test will minimize the sample size required in a study, or (equivalently) will maximize the power of a study for a particular sample size.

The efficiency of a statistical test A relative to a second test B can be defined as follows. If test A requires n_A observations to detect a difference δ at a significance level

α with power 1-β and test B requires n_B observations to do the same, then the relative efficiency of test A relative to test B $e_{A,B}$ is given by the ratio n_B/n_A. In general the relative efficiency of two tests will depend principally on the type of difference to be detected i.e. the alternative hypothesis; its dependence on the values of α, β and δ is relatively small. A detailed discussion is available in Pratt and Gibbons (1981).

The principle method of estimation of relative efficiency utilises the power functions of the two tests to derive the asymptotic relative efficiency. These power functions depend on the assumptions made about the distribution of the data under the alternative hypothesis. There is thus an uncertainty in the calculated relative efficiency if the distribution of the data under the alternative hypothesis in a particular area of research is unclear.

This appendix describes a technique which avoids the need to assume a distribution for the data, or equivalently the types of alternative which arise, since it uses real sets of data typical of a particular field of scientific research. The theory of the technique is first described for a single alternative hypothesis, making use of a simulation example. This is followed by extension to a series of similar data sets involving a range of alternative hypotheses.

THE METHOD FOR A SINGLE ALTERNATIVE HYPOTHESIS

The general formula (Lachin, 1981) relating the sample size and power for a two-sided statistical test is

$$n = \frac{(\sigma_0 z_{\alpha/2} + \sigma z_\beta)^2}{\delta^2}$$

where n is the number of observations, $\sigma/\sqrt{n}$ is the standard deviation of the test statistic, $\sigma_0/\sqrt{n}$ is the standard deviation of the test statistic on the assumption of the null hypothesis, z_x is the $100(1-x)$ percentile of the standard normal distribution, α is the level of statistical significance, β is the type II error (i.e. 1 - power), and δ is the true population difference from the null hypothesis. Both σ and σ_0 are related to (or identical to) the standard deviation of the individual measurements, and so are independent of n. It is often reasonable to assume that σ and σ_0. Then

$$n = \frac{\sigma^2 (z_{\alpha/2} + z_\beta)^2}{\delta^2} \qquad \text{(D.1)}$$

This relation can be grasped from consideration of Fig. D.1. Curve 1 represents the distribution of the z-values from the test statistic on the basis of the null hypothesis. This is by definition a Normal distribution with mean 0, and variance 1. The shaded area labelled $\alpha/2$ represents the (upper-tailed) proportion significant at the α-level if the null hypothesis is true. Curve 2 represents the distribution of the z-values (denoted by the variable Z) from the test statistic assuming that there is some deviation from the null hypothesis. This is usually taken to be a Normal distribution with the same variance 1, and mean denoted by $\bar{z}$. The shaded area β represents the proportion of these z-values that do not reach the critical value $z_{\alpha/2}$, i.e. the type II error. It can be seen from Fig. D.1 that

$$\bar{z} = z_{\alpha/2} + z_\beta$$

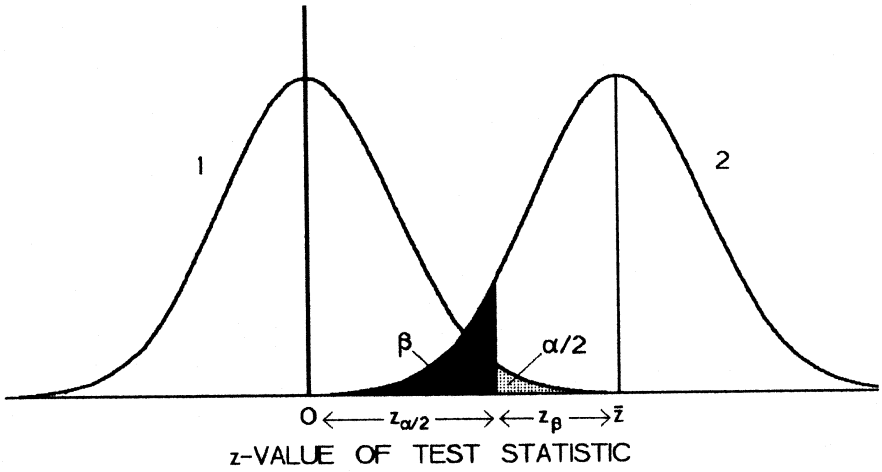


Fig. D.1. Demonstration of the general power formula.

Thus

$$n = \frac{\sigma^2 \quad \bar{z}^2}{\delta^2}$$

For simplicity, consider first one-sample statistical tests, and consider two statistical tests A and B being applied to samples from a population where the null hypothesis is known to be false and the alternative hypothesis is fixed. For statistical test A, applied to a sample of size n_A, using subscript A to denote the values of σ^2 and δ,

$$n_A = \frac{\sigma_A^2 \quad \bar{z}_A^2}{\delta_A^2} \tag{D.2}$$

For test B, applied to a sample of size n_B, the value of σ is different from σ_A and is denoted by σ_B. The difference between the alternative hypothesis and the null hypothesis δ_B is equivalent to the difference δ_A but may not be identical to δ_A because each of the statistical tests may express the alternative hypothesis in a different way. Thus

$$n_B = \frac{\sigma_B^2 \quad \bar{z}_B^2}{\delta_B^2} \tag{D.3}$$

Re-arranging equations (D.2) and (D.3) we obtain

$$\frac{\bar{z}_A^2}{\bar{z}_B^2} = \frac{n_A \ \delta_A^2 \ \sigma_B^2}{n_B \ \delta_B^2 \ \sigma_A^2}$$

For a given alternative hypothesis, $\delta_A^2 \ \sigma_B^2 / (\delta_B^2 \ \sigma_A^2)$ is a fixed constant, K say, which is independent of sample sizes. Thus

$$\frac{\bar{z}_A^2}{\bar{z}_B^2} = \frac{n_A}{n_B} \ K \qquad \text{(D. 4)}$$

We can allow the ratio of n_A to n_B to vary until $\bar{z}_A = \bar{z}_B$. Then the powers of the two statistical tests at the same significance level are equal (since $\bar{z} = z_{\alpha/2} + z_\beta$). In these circumstances the ratio n_B / n_A is defined as the relative efficiency of test A relative to test B ($e_{A,B}$) and thus from (D.4),

$$K = e_{A,B}$$

Thus, for a given alternative hypothesis,

$$\frac{\bar{z}_A^2}{\bar{z}_B^2} = \frac{n_A}{n_B} \ e_{A,B}$$

If the two statistical tests are applied to samples of the same size, so that $n_A = n_B$,

$$\frac{\bar{z}_A^2}{\bar{z}_B^2} = e_{A,B} \qquad \text{(D.5)}$$

It is this result that forms the basis of the technique described in this appendix.

For statistical tests which are two-sample rather than one-sample tests, the theory and the result are the same except that firstly δ represents the difference between the two populations, and secondly an additional variable to be considered is the relative size of the samples from the two populations, and it will be convenient to keep these always in the same proportion.

SIMULATION EXAMPLE FOR A SINGLE ALTERNATIVE HYPOTHESIS: THE EFFICIENCY OF THE WILCOXON ONE-SAMPLE TEST RELATIVE TO THE SIGN TEST FOR NORMALLY DISTRIBUTED DATA

Using the Wichmann and Hill (1982) pseudorandom number algorithm, one thousand samples of size 100 were randomly taken from a Normal distribution with mean 0.385, and variance 1. Interest lies in testing whether the mean of the sample is different from zero. The value of 0.385 was chosen so that 65% of the population distribution is positive, and 35% is negative, in order that the expected *z*-value under the sign test , $\bar{z}_s$ is (65 - 50) / √(100 × ½ × ½) = 3.00. Each sample was analyzed both by the Wilcoxon one-sample test and by the sign test (subscripts W and S respectively are used to identify the test statistics) giving two *z*-values for each sample. The sign test included the standard correction for continuity (Altman, 1991).

The distribution of the *z*-values by the sign test, $\bar{z}_S$ in the simulation is shown by the bar chart of Fig. D.2 (the distribution is discrete because, in the sign test, only integer counts of positive values can occur). The mean of the observed *z*-values $\bar{z}_{OS}$ = 2.956 agrees well with the expected value of 3. The standard deviation of 0.930 is less than unity, and a consideration of the function of the sign test gives two reasons. Firstly the standard deviation of a proportion of 0.65 differs from that for a proportion of 0.5 (by a factor of 0.95), and secondly the binomial distribution differs slightly from the Normal.

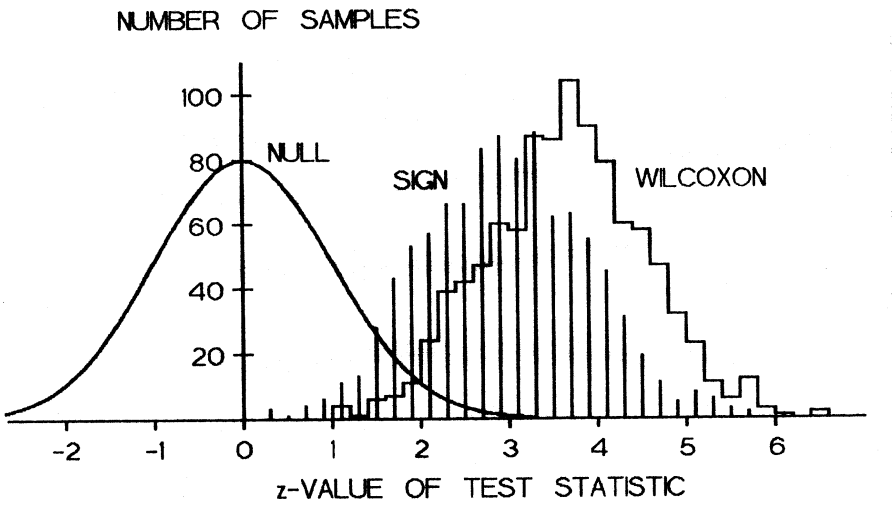


Fig. D.2. The bar chart shows the distribution of the *z*-values by the sign test in 1000 simulated samples of size 100 taken from a Normal distribution (0.385, 1). The histogram shows the distribution of the *z*-values by the Wilcoxon one-sample test for the same 1000 samples. The expected distribution on the null hypothesis is shown by the continuous curve on an equivalent scale.

The distribution of the *z*-values by the Wilcoxon test, Z_w in the simulation is shown as the histogram of Fig. D.2. The mean of the observed *z*-values, $\bar{z}_{OW} = 3.630$, and the standard deviation of 0.89 is again slightly less than unity. This gives a value for $(\bar{z}_{OW} / \bar{z}_{OS})^2$ of $(3.630 / 2.956)^2 = 1.508$, for an estimate of the relative efficiency for the Wilcoxon one-sample test relative to the sign test, e_{ws}. The asymptotic relative efficiency of the Wilcoxon one-sample test relative to the sign test for Normally distributed data is known to be 1.500 (Noether, 1987), and so there is good agreement between the theory in this paper and the known asymptotic value.

The joint distribution of the *z*-values by the Wilcoxon and sign tests, Z_w and Z_S is shown by the scatter diagram of Fig. D.3, with each sample represented by one point.

Some correlation between Z_W and Z_s would be expected, and Fig. D.3 in fact does show considerable correlation between the observed values of z_{ow} and z_{os}.

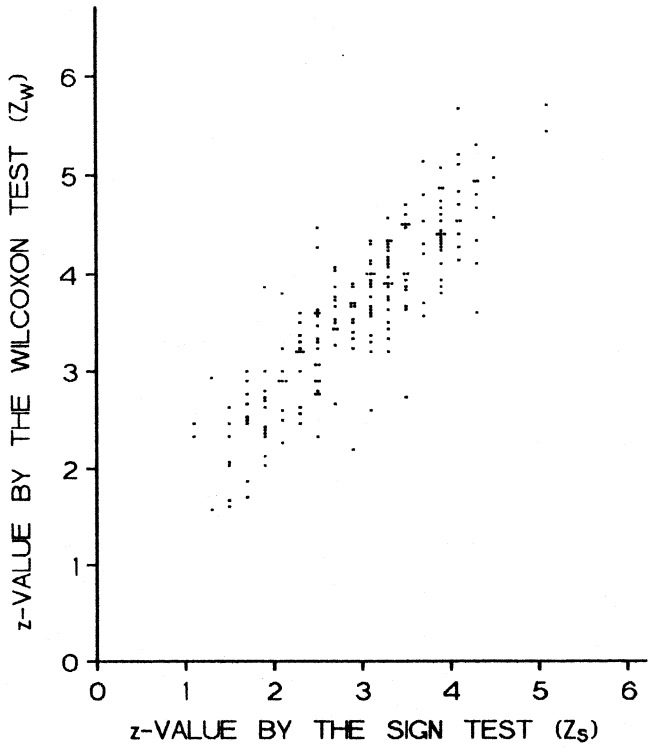


Fig. D.3. The joint distribution of the *z*-values by the sign and Wilcoxon tests in the simulated samples (for clarity, only the first 200 points in the simulation are shown).

The observed distribution of $(z_{ow} / z_{os})^2$, denoted Z_R^2, is shown in Fig. D.4 on a linear scale. The median value is 1.497, which is close to the known asymptotic relative efficiency e_{WS} of 1.500. The interquartile range of Z_R^2, is 1.23 to 1.92. It is interesting to note that a logarithmic transformation of Z_R^2, (or equivalently $1/Z_R^2$)gives an approximate Normal distribution.

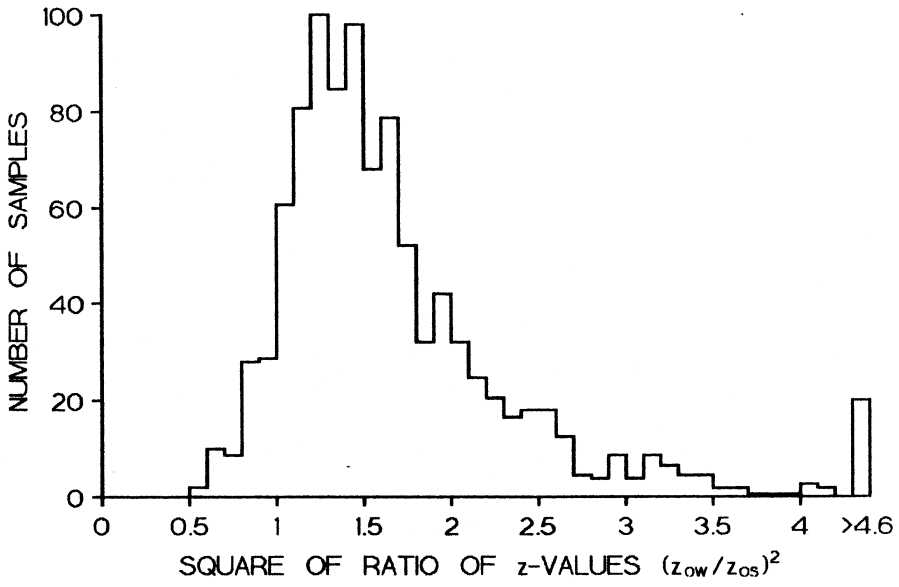


Fig. D.4. The distribution of the square of the ratio of the *z*-values from the Wilcoxon and sign tests $(z_{ow} / z_{os})^2$ for the 1000 samples in the simulation.

DERIVATION OF RELATIVE EFFICIENCY FROM REAL DATA SETS

It has been shown that the statistic $(\bar{z}_A / \bar{z}_B)^2$ is a good estimate of the relative efficiency of test A to test B (equation (D.5)). The simulation of the previous section gave an illustration that when there are a very large number of sample data sets from a population, the observed value of $(\bar{z}_{OA} / \bar{z}_{OB})^2$ gives a good estimate of the relative efficiency of test A to test B. However, in scientific research, it is rare for many data sets from the same population to be available, and usually there is only one. The value of $(z_{OA} / z_{OB})^2$ from a single data set may give only a poor estimate of the relative efficiency (the interquartile range in the preceding simulation was quite wide - 1.23 to 1.92). However, if we have a series of similar data sets which can be regarded as a representative sample of data sets that occur in practice in a particular field of research, and if the null hypothesis can be assumed not to be true in the data sets, then the

empirical distribution of the observed relative efficiency $(z_{OA} / z_{OB})^2$ will give an estimate of the underlying distribution of the relative efficiency of the two competing tests. By using a number of similar data sets, each from a different, but related population, an overall estimate of the relative efficiency may be made by taking the median value of the squares of the observed ratios $(z_{OA} / z_{OB})^2$, and a 95% confidence interval can be calculated by the usual technique for a median (Altman, 1991). It seems preferable to take the median value rather than the mean as the distribution of $(z_{OA} / z_{OB})^2$ may well be skewed in general, as it was in the preceding simulation.

Different data sets will differ in the sample size n and in the size of the deviation from the null hypothesis δ, and consequently will differ in the values of $\bar{z}$, but the relative efficiency of two statistical tests is comparatively independent of these quantities. Just as no two patients are exactly the same, and yet patients are grouped together for the purposes of comparison of medical treatments and prediction of the best treatment for future similar patients, so are no two sets of data the same, but sets of data may be grouped together for the purposes of comparison of statistical tests and prediction of the best test for future similar sets of data.

It is necessary for the sets of data used in this procedure to reflect nonzero population treatment differences (otherwise both Z_A and Z_B will be distributed around zero, and the median value of $(z_A/z_B)^2$ will be a meaningless quantity). It is usually not possible to know in which sets of data there are underlying nonzero differences in treatment effect (especially in clinical trials), but one reasonable approach is to take trial results significant at 5% by either test as likely to reflect nonzero population differences. A proviso is that the proportion of significant results should be large enough to swamp the results that are likely to be significant by chance.

DISCUSSION

The method described in this paper is potentially of wide application. Conventionally, statistical tests are compared by a consideration of their power functions, but there are often difficulties in the exact calculation of these, and furthermore, models must be assumed for the alternative hypothesis. The method described here avoids these difficulties by directly comparing the tests in practice. It is not common for statistical tests to be compared by analysis of a series of real data sets, but one example was presented by Mann and Whitney (1947). In this paper the authors compared their test with that of Wald and Wolfowitz in analysis of 62 data sets, and rather crudely considered the numbers of the data sets significant at 5% and at 1%.

In the past, it has been rare for a number of data sets in the same area of research to be collected together, but with the increasing numbers of meta-analyses being performed, this is becoming more common. Those involved in meta-analyses will thus be in a position to answer questions of optimum methods of analysis. For example, in cancer research, a number of methods of survival analysis are available (Breslow, 1984), but a decision on the most efficient method depends on which model of survival distribution is assumed for the alternative hypothesis. The selection of the most efficient method could be done more directly by analysis of a number of clinical trial data sets by each of the various methods available, and comparison of the *z*-values as described in this appendix.

There are potential limitations to the method due to the various assumptions made. Firstly, the assumption of equal variance of the test statistic under both the null and the alternative hypothesis can be questioned, and in the computer simulation example, this was found to be invalidated to a small extent. However, the theory can accommodate a difference between σ and σ_0 with no change in the final conclusion provided that the ratio of σ/σ_0 is the same for each test.

The distribution of Z_A (or Z_B) will become skewed when $\bar{z}_A$ (or $\bar{z}_B$) approaches the maximum z-value given the sample size; for example, in the computer simulation, the maximum z-value possible is (100 - 50)/√(100 × ½ × ½) i.e. 10.0. There will thus be limitations to the method when sample sizes are small, or the size of the true difference is large.

It may be felt that the uncertainties in the method of this paper make it unreliable. However, it must be remembered that the alternative method of calculation of relative efficiency using calculated power functions also has uncertainties (which are probably greater) in the need to assume a model of treatment effect, and the method of this paper can thus be put forward as a more reliable alternative.

The precision required in the estimation of relative efficiency is in any case not great. The important question will usually be whether the relative efficiency is greater than or less than unity, i.e. which of the statistical tests is the more efficient. The actual value of the relative efficiency may be important only if there are other important factors in the choice of the statistical method such as one of the tests being much simpler to perform or understand, or a difference in the availability of computer software.

APPENDIX E: THE EMPIRICAL RELATION BETWEEN THE CHI SQUARED FUNCTION AND THE SQUARE OF THE EQUIVALENT *z*-VALUE

The empirical relation between χ_v^2 and the square of the equivalent *z*-value z^2 is shown in Fig. E.1, for degrees of freedom *v* from 1 to 6 inclusive. For degree of freedom $v = 1$, $\chi_1^2 = z^2$ by definition, so giving a line of equality. Values are taken from Pearson and Hartley (1976), Armitage and Berry (1987), Woodward and Francis (1988), and Altman (1991). The relation is approximately linear for values of z^2 greater than 1.0, and this was utilised in sections 2.5 and 3.4.

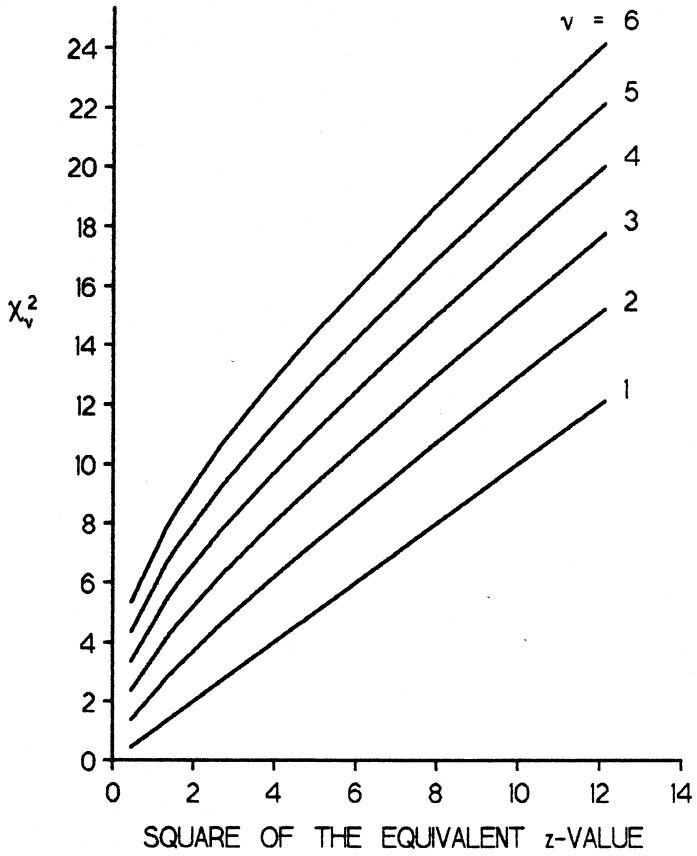


Fig. E.1. The empirical relation between χ_v^2 and z^2.

APPENDIX F: THE POWER RELATION OF THE MANN-WHITNEY *U* TEST IN THE ANALYSIS OF ORDERED CATEGORICAL DATA

INTRODUCTION

Consider a set of data that is classified into a number of categories where these categories fall into a natural order. Examples of such ordered categorical data from cancer clinical trials are the degree of response to a cancer treatment, or the level of morbidity from a cancer treatment. Examples from other fields of medicine are the components of APGAR scores and the standard grading of muscle strength. It has been recommended that such data are analyzed by the Mann-Whitney *U* test rather than the Chi squared test (Moses et al, 1984; Morton & Dobson, 1990) because of the greater power of the Mann-Whitney test. A general formula for the power of the Mann-Whitney test for ordered categorical data is required for the modelling of section 2.6, but does not appear in the literature. It is therefore derived in this Appendix.

NOTATION

In order to derive the formula, some assumptions are required about the distribution of the population between the categories. Suppose that there are k categories. In a number of situations it is reasonable to assume that the distribution of the whole population between the k categories depends on another (hidden) variable, which will be termed X, and which is continuous. Thus, returning to the examples from Oncology, the numbers of patients falling into the different response categories following a cancer treatment will depend on a variable which may be termed "treatment efficacy"; and the numbers of patients in the morbidity categories will depend on a variable which may be termed "effect on normal tissue". In a clinical trial comparing two or more treatments, the type of difference that concerns us (in statistical

terminology, the alternative hypothesis of interest) is of some difference in treatment efficacy, or some difference in effect on normal tissue respectively.

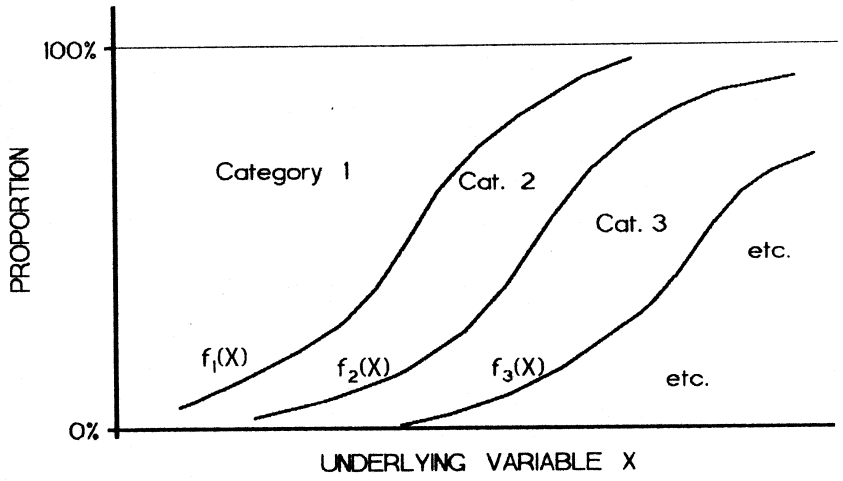


Fig. F.1. The distribution of the population between the *k* categories (e.g. categories of morbidity) and the relation to the underlying variable *X* (e.g. the effect of cancer treatment on normal tissue).

This may be represented graphically as in Fig. F.l, where efficacy or effect is represented by the variable *X*; and $f_1(X)$ represents the proportion of the population falling into a category of greater number than category 1, $f_2(X)$ represents the proportion falling into a category above category 2, and in general $f_r(X)$ represents the proportion falling into a category above category *r*, i.e. in categories $r+1$ to *k* inclusive. The proportion falling in category 2 is $f_1(X) - f_2(X)$ and in general the proportion falling into the r^{th} category will be $f_r(X) - f_{r-1}(X)$. Thus at low values of *X* most of the population may fall in the lowest category, but at higher values, the total proportion of observations falling in the second or higher categories may progressively increase. For example, the

total proportion of patients exceeding a particular level of morbidity would increase progressively as the effect on normal tissues increases.

DERIVATION OF FORMULA

In general, the power of a statistical test in analyzing a trial of two treatments is given by the formula (Noether, 1987):

$$\frac{\mu(T) - \mu_0(T)}{\sigma_0(T)} = z_{\alpha/2} + \rho z_\beta \qquad \text{(F.1)}$$

where

T is the numerical value resulting from the statistical test, i.e. the test statistic,

$\mu(T)$ is the expected value of the test statistic T assuming a difference between the treatments,

$\mu_0(T)$ is the expected value of the test statistic T assuming no treatment difference, i.e. assuming the null hypothesis,

$\sigma(T)$ is the standard deviation of the test statistic assuming a difference between the treatments,

$\sigma_0(T)$ is the standard deviation of the test statistic assuming the null hypothesis,

z_x is the 100(1-x) percentile of the standard normal distribution,

β is the type II error = 1 − power

α is the two-sided significance level, and

$\rho = \sigma(T)/\sigma_0(T)$ which will be close to 1 for treatment differences which are not large and which will tend to 1 as the treatment difference tends to zero and $n \to \infty$.

Equation F.1 will then simplify to

$$\frac{\mu(T) - \mu_0(T)}{\sigma_0(T)} = z_{\alpha/2} + z_\beta \qquad \text{(F.2)}$$

This is applicable to the Mann-Whitney test as the test statistic is approximately normally distributed, especially for large samples (Mann and Whitney, 1947).

To derive a formula for ordered categorical data using the Mann-Whitney test, the first step will be derivation of the quantity $\mu(U) - \mu_0(U)$ which is the deviation of the Mann-Whitney test statistic U from its value assuming the null hypothesis; and the second step will be the calculation of the quantity $\sigma_0(U)$ which is the standard deviation of U assuming the null hypothesis.

DEVIATION OF *U* FROM ITS VALUE ASSUMING THE NULL HYPOTHESIS

Suppose that two samples are taken from the population, a sample of n_1 at $X = x$ and a second sample of n_2 at $X = x+\delta x$. Suppose that there are k ordered categories and that $n = n_1 + n_2$. Then the expected number of observations in the r^{th} category in sample 1 is

$$n_1 \times \{f_{r-1}(x) - f_r(x)\}$$

and the expected number in sample 2 is

$$n_2 \times \{f_{r-1}(x+\delta x) - f_r(x+\delta x)\}.$$

It can be assumed that the functions f are smooth and so

$$f_r(x+\delta x) \rightarrow f_r(x) + \delta x \, {}^{d}/_{dx} f_r(x)$$

when δx is made small.

The terminology can be simplified by putting

$$f_r = f_r(x), \qquad f'_r = {}^{d}/_{dx} f_r(x), \qquad \text{and} \quad \delta = \delta x.$$

For completeness, define $f_0 = 1$ and $f_k = 0$, and thus $f'_0 = 0$ and $f'_k = 0$.

Consider the r^{th} category. The expected number of observations of sample 1 in this category (E_1) is $n_1\{f_{r-1} - f_r)$, and similarly there are an expected $n_2\{f_{r-1} + \delta f'_{r-1} - f_r - \delta f'_r)$

observations of sample 2 (E_2). The expected number of observations in sample 2 in categories higher than category r (E_3) is

$$n_2\ f_r(x+\delta x)\ =\ n_2\ (f_r\ +\ \delta f'_r).$$

The contribution to U from any one of the observations in sample 1 category r is E_3 + ½E_2 , calculated as the number of observations in sample 2 of higher value + half the number of ties in sample 2 (Armitage & Berry, 1987). Thus the expected contribution from all the E_1 observations in sample 1 is

$$E_1(E_3\ +\ ½\,E_2)$$
$$=\ n_1\ (f_{r-1}\ -\ f_r)\ n_2\ \{f_r\ +\ \delta f'_r\ +\ ½\,(f_{r-1}\ +\ \delta f'_{r-1}\ -f_r-\delta f'_r)\}$$
$$=\ ½\ \ n_1\ \ n_2\ \ (f_{r-1}\ -\ f_r)\ (f_{r-1}\ +\ f_r\ +\ \delta f'_{r-1}\ +\ \delta f'_r)\ .$$

Strictly this equality is only approximate since

$$\text{Expectation(P×Q)} = \text{Expectation(P)} \times \text{Expectation(Q)} + \text{Covariance(P,Q)}$$

but the covariance terms here are of order n, whereas the other terms are of order n^2, and asymptotically, as $\delta \to 0$ and $n \to \infty$ the covariance terms can be discounted. Summing over all categories, the total expected value of U is

$$\sum_{r=1}^{k} ½\,n_1\,n_2\ \ \{f^2{}_{r-1}\ -\ f^2{}_r\ +\ \delta\ (f_{r-1}\ -\ f_r)\ (f'_{r-1}\ +f'_r)\}$$

$$=\ ½\,n_1\,n_2\ \sum_{r=1}^{k}(f^2{}_{r-1}\ -\ f^2{}_r\ )+½\,n_1\,n_2\,\delta\{\sum_{r=1}^{k}(f_{r-1}\ -\ f_r)f'_{r-1}+\sum_{r=1}^{k}(f_{r-1}\ -\ f_r)\,f'_r\}$$

Now

$$\sum_{r=1}^{k} f^2{}_{r-1}\ -\ \sum_{r=1}^{k} f^2{}_r\ =\ \sum_{r=0}^{k-1} f^2{}_r\ -\ \sum_{r=1}^{k} f^2{}_r$$

$$=\ f^2{}_0\ -\ f^2{}_k\ =\ 1$$

and

$$\sum_{r=1}^{k}(f_{r-1}\ -\ f_r)\,f'_{r-1}\ =\ \sum_{r=0}^{k-1}(f_r\ -\ f_{r+1})\,f'_r\ \ ,$$

$$=\ \sum_{r=1}^{k-1}(f_r\ -\ f_{r+1})\,f'_r\ ,\ \text{as}\ f'_0=0$$

and

$$\sum_{r=1}^{k} (f_{r-1} - f_r) f'_r = \sum_{r=1}^{k-1} (f_{r-1} - f_r) f'_r \ , \ \text{as} f'_k = 0$$

Therefore the total expected value of U i.e. $\mu(U)$ is

$$\tfrac{1}{2} n_1 n_2 + \tfrac{1}{2} n_1 n_2 \delta \{ \sum_{r=1}^{k-1} (f_r - f_{r+1}) f'_r + \sum_{r=1}^{k-1} (f_{r-1} - f_r) f'_r \}$$

$$= \tfrac{1}{2} n_1 n_2 + \tfrac{1}{2} n_1 n_2 \delta F_1$$

where

$$F_1 = \sum_{r=1}^{k-1} (f_{r-1} - f_{r+1}) f'_r \ .$$

The difference from the mean value of U expected on the null hypothesis ($\tfrac{1}{2} n_1 n_2$) is thus

$\tfrac{1}{2} n_1 n_2 \delta F_1$, i.e.

$$\mu(U) - \mu_0(U) = \tfrac{1}{2} n_1 n_2 \delta F_1 \ .$$

THE VARIANCE OF U

The variance of the sampling distribution of U will now be calculated, on the assumption of the null hypothesis. From Armitage & Berry (1987) this is given by

$$\frac{n_1 n_2}{12n(n-1)} \{ n^3 - n - \Sigma_t (t^3 - t) \}$$

where the summation is taken over all groups of tied observations, t being the number of ties in a particular group. This reduces to

$$\frac{n_1 n_2}{12n(n-1)} \{ n^3 - \Sigma_t t^3 \}$$

for ordered categorical data where t *is* the total number of observations in each category since each category contains tied data, and thus $\Sigma t = n$.

The expected number of tied observations t in the r^{th} category on the null hypothesis is $E_1 + E_2$ when δx is 0, which equals $n(f_{r-1} - f_r)$. Since the number of ties is distributed according to the Binomial distribution, the Expectation of t^3 can be shown to be equal to (Expectation of t)^3 plus other terms of order n^2 and below - and these other terms become relatively insignificant as δ becomes asymptotically small and n becomes large.

Therefore

$$\sum_t t^3 \approx \sum_{r=1}^{k} n^3 (f_{r-1} - f_r)^3$$

Now

$$\sum_{r=1}^{k} (f_{r-1} - f_r)^3 = \sum_{r=1}^{k} f^3{}_{r-1} - 3\sum_{r=1}^{k} f^2{}_{r-1}\, f_r + 3\sum_{r=1}^{k} f_{r-1}\, f^2{}_r - \sum_{r=1}^{k} f^3{}_r$$

$$= \sum_{r=1}^{k} f^3{}_{r-1} - \sum_{r=1}^{k} f^3{}_r - 3\sum_{r=1}^{k} f_{r-1}\, f_r\, (f_{r-1} - f_r)$$

But

$$\sum_{r=1}^{k} f^3{}_{r-1} - \sum_{r=1}^{k} f^3{}_r = \sum_{r=0}^{k-1} f^3{}_r - \sum_{r=1}^{k} f^3{}_r$$

$$= f^3{}_0 - f^3{}_k = 1$$

and

$$3\sum_{r=1}^{k} f_{r-1}\, f_r\, (f_{r-1} - f_r) = 3\sum_{r=1}^{k-1} f_{r-1}\, f_r\, (f_{r-1} - f_r) \text{ since } f_k = 0$$

$$= 3\, F_2, \text{ say.}$$

Therefore the variance equals

$$\frac{n_1\, n_2}{12n\,(n-1)}\, \{n^3 - n^3\,(1 - 3\, F_2)\}$$

i.e.

$$\sigma_0^2(U) = n_1\, n_2\, n^2\, F_2\, /\{4\,(n-1)\}$$

Then substituting in equation F.2,

$$\frac{(\,½\, n_1\, n_2\, \delta\, F_1\,)^2}{n_1\, n_2\, n^2\, F_2\, /\{4\,(n-1)\}} \;=\; (\,z_{\alpha/2}\,+\,z_{\beta}\,)^2$$

and

$$\mathrm{n} \;\approx\; \frac{n^2\, F_2\;(\,z_{\alpha/2}\,+\,z_{\beta}\,)^2}{n_1\, n_2\, F_1^2\;\delta^2}$$

Thus when the sample sizes are equal and $n_1 = n_2 = n/2$, the relation between the number of observations n and the functions of the boundaries between the categories f in order to find a difference δ by the Mann-Whitney test with power $1-\beta$ at a two-sided significance level α is given by

$$n \;\approx\; \frac{4\sum_{r=1}^{k-1} f_{r-1}\, f_r\, (f_{r-1}\,-\,f_r\,)\;(\,z_{\alpha/2}+\,z_{\beta}\,)^2}{\{\;\sum_{r=1}^{k-1}\,(f_{r-1}\,-\,f_{r+1}\,)\;f'_r\;\}^2\qquad\delta^2} \qquad \text{(F.3)}$$

and this approximation tends to an equality as $\delta \rightarrow 0$ and $n \rightarrow \infty$ since all the approximations in its derivation tend to equalities.

DISCUSSION

As a consistency check, consider the situation where there are only 2 categories, and the variable X is the proportion of the population in category 2, and this is denoted by θ. Then

$$f_0 = 0, \quad f_1 = \theta, \quad f_2 = 0, \quad \text{and} \quad f'_1 = 1$$

and equation F.3 becomes

$$n = \frac{4 \; f_0 \, f_1 \, (f_0 - f_1) \; (z_{\alpha/2} + z_{\beta})^2}{\{ (f_0 - f_2) \; f'_1 \}^2 \quad \delta^2}$$

$$\text{i.e.} \quad n = 4 \; \theta (1 - \theta) \; (z_{\alpha/2} + z_{\beta})^2 / \; \delta^2 \qquad \text{(F.4)}$$

Equation F.4 is in fact the asymptotic power relation for a comparison of two independent proportions (Lachin, 1981; Machin & Campbell, 1987). This result is expected, i.e. it is expected that the power relation for the Mann-Whitney test should reduce to that for two proportions since for two categories, the Mann-Whitney test becomes equivalent to a test of two proportions, when the samples are large (Armitage & Berry, 1987).

APPENDIX G: COMPUTER PROGRAM TO CALCULATE THE EFFICIENCY OF THE MANN-WHITNEY TEST IN ANALYSIS OF TUMOUR RESPONSE DATA

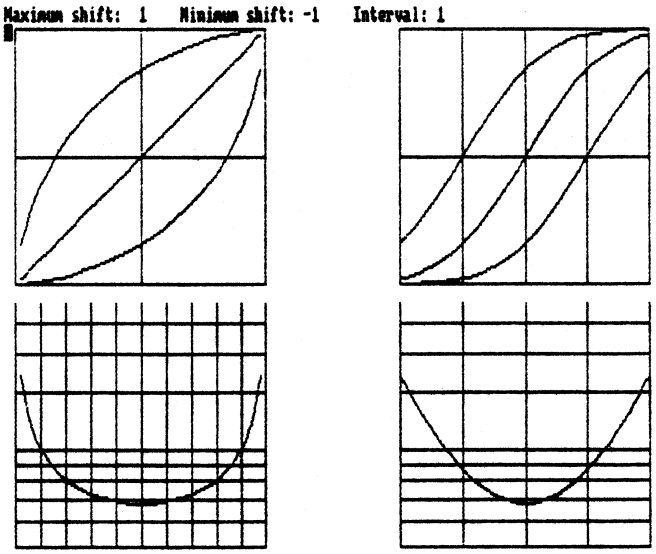
This BASIC program was written to calculate the relation between the numbers of patients required in a clinical trial and the number and separation between the ordered categories used to assess the tumour response data. It was run on an Amstrad PCW 8512 personal computer using Mallard BASIC extended by Lightening Basic software (CP Software, UK) to supply graphics commands. The results obtained are discussed in section 2.6. Some explanatory comments are given in "REM" (meaning remark) lines of the program.

PROGRAM INPUT

In the program, category boundaries can be specified by two alternative methods. Either individual boundaries can be specified as a series of shifts from the baseline total response sigmoid curve, or the extreme right hand and left hand curves can be specified in terms of their shift together with the number of categories between them (equal to two less than the total number of categories k). Values from standard tables of the normal distribution (Neave, 1981) are entered as DATA statements in lines 1150 to 1260.

PROGRAM OUTPUT

The program gives a graphical and a numerical output, the latter being in two parts. A sample occupies the next two pages. The four graphs in the first part of the output are a printer dump of graphics drawn on the monitor screen. The upper right graph corresponds to Fig. 2.8 (illustrating the lateral shift model) with scale lines on the horizontal scale at intervals of 1.0 standard deviations. The lower right graph has a corresponding horizontal scale. Values of the relative efficiency $e_{k,2}$ are calculated and



Maximum shift: 1 Minimum shift: -1 Interval: 1

z	response	n for 2 cats	n for expt	Pitman e.
-2	2,280003E-02	8,912075E-02	2,602299E-02	3,424693
-1,9	2,869999E-02	0,1115052	3,550692E-02	3,14038
-1,8	0,0359	0,1384448	4,792489E-02	2,888786
-1,7	4,460001E-02	0,1704434	6,355137E-02	2,681978
-1,6	5,479997E-02	0,2071877	8,335911E-02	2,485484
-1,5	0,0668	0,249351	0,1076253	2,316844
-1,4	0,0808	0,2970854	0,1369634	2,169087
-1,3	9,680003E-02	0,3497191	0,1717505	2,036205
-1,2	0,1151	0,407408	0,211946	1,922226
-1,1	0,1357	0,469142	0,2574105	1,822544
-1	0,1587	0,5340572	0,3077662	1,735269
-0,9	0,1841	0,6008287	0,361819	1,660578
-0,8	0,2119	0,6679935	0,418723	1,595311
-0,7	0,242	0,733744	0,4766831	1,53927
-0,6	0,2743	0,796238	0,533407	1,49274
-0,5	0,3085	0,853311	0,5871272	1,453367
-0,4	0,3446	0,9034034	0,6356367	1,421257
-0,3	0,3821	0,9443983	0,6758164	1,397419
-0,2	0,4207	0,9748461	0,7064136	1,379993
-0,1	0,4602	0,9936638	0,7256069	1,369425
0	0,5	1	0,7319504	1,366213
0,1	0,5398	0,9936638	0,7256069	1,369425
0,2	0,5793	0,9748461	0,7064136	1,379993
0,3	0,6179	0,9443983	0,6758164	1,397419

0,4	0,6554	0,9034084	0,8356367	1,421257
0,5	0,6915	0,853011	0,7871272	1,453267
0,6	0,7257	0,796238	0,733407	1,49274
0,7	0,758	0,743744	0,4766831	1,538927
0,8	0,7881	0,6679935	0,418723	1,595312
0,9	0,8159	0,6008287	0,3618191	1,660578
1	0,8413	0,5340572	0,3077662	1,735269
1,1	0,8643	0,469142	0,2574105	1,822544
1,2	0,8849	0,407408	0,211946	1,922226
1,3	0,9032	0,3497191	0,1717505	2,036205
1,4	0,9192	0,2970854	0,1369634	2,169087
1,5	0,9332	0,249351	0,1076253	2,316844
1,6	0,9452	0,2071877	8,335911E-02	2,485484
1,7	0,9554	0,1704434	6,355137E-02	2,681978
1,8	0,9641	0,1384448	4,792491E-02	2,888785
1,9	0,9713	0,1115052	3,550692E-02	3,14038
2	0,9772	8,912075E-02	2,602299E-02	3,424692

response	n for 2 cats	n for expt	Pitman e,
0,05	0,1898963	7,403785E-02	2,577952
0,1	0,3598067	0,1787791	2,016274
0,2	0,6392432	0,3943648	1,623249
0,3	0,8391262	0,5737757	1,463152
0,4	0,9585179	0,6900052	1,389338
0,5	1	0,7319504	1,366213
0,6	0,9585178	0,6900052	1,389338
0,6999999	0,8391262	0,5737757	1,463152
0,8	0,6392431	0,3943648	1,623249
0,9	0,3598068	0,1787792	2,016274
0,95	0,1898964	7,403786E-02	2,577951

plotted against this on a logarithmic scale (base 10) with scale lines at 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 3.0, 4.0, and 5.0. The left upper and lower graphs correspond to the right upper and lower graphs except for a transformation of the horizontal scale. This is a probit transformation in reverse, so that the horizontal scale is a linear scale of the total response rate $f_2(E)$ which equals $\Phi(E)$, rather than efficacy E.

The second part of the program output gives numerical results of the calculations at values of efficacy E from -2 to +2 standard deviations at intervals of 0.1. The column headed "response" is the value of $\Phi(E)$ i.e. the total response rate. The column headed "n for 2 cats" gives the value of expression 2.2 after omitting the two terms $(z_{\alpha/2} + z_{\beta})^2$ and δ^2 and after multiplying by $\varphi^2(E)$. The reason for the inclusion of this factor $\varphi^2(E)$ is to facilitate comparison with standard tables for the comparison of two proportions e.g. Machin & Campbell (1987) in which δ is defined in terms of the proportion of successes. The column headed "n for expt" gives the corresponding quantity for expression 2.1 i.e. omitting $(z_{\alpha/2} + z_{\beta})^2$ and δ^2 and multiplying by $\varphi^2(E)$. The final column headed "Pitman e." gives the ratio of these last two columns, which is the relative efficiency $e_{k,2}$, and which was plotted in the first part of the output.

The third part of the program output gives the values the last three columns of the second part for some values of total response that are round numbers (5%, 10% to 90% in 10% intervals, and 95%). These are obtained by linear interpolation. For example, the second part of the program does not give values for a total response rate of exactly 40%, but instead for the nearest values of 38.21% and 42.07%, corresponding to E of -0.3 and -0.2 respectively. The values for a total response rate of 40% are estimated by linear interpolation from the values for 38.21% and 42.07%.

VERIFICATION

In addition to arithmetical checks, several verification checks of the program were made. Firstly, when the number of boundaries was set at 1, with no shift, which is identical to a two category classification, the value of $e_{k,2}$ calculated by the program was 1.0 for all values of θ, as expected. Secondly, the values obtained in the second column of the third part of the program output were checked against the values in Table 3.1 of Machin and Campbell (1987) with close agreement. Finally, for each value of the total response rate, the peak calculated relative efficiency approached a limiting value as the number of categories k became large; and this limiting value (taken at the maximum value of k utilised of 86) was equal to the theoretical relative efficiency at an infinite number of categories (Appendix H), with an agreement of within 1%.

PROGRAM LISTING

```
10    REM 7.3.91
20    REM SAVE "rescal2" : rem on B12 disc
30    REM
40    REM response calculations - to calculate the numbers of patients
50    REM needed and Pitman efficiency for different assumptions
60    REM
70    INPUT "Advance printer? How many lines"; n%: IF n% = 0 THEN GOTO 90
80    FOR j %=1 TO n%: LPRINT : NEXT: GOTO 70
90    DIM j(120), jd(120), zshift(122) : REM j contains the cumulative frequency
      distribution of the standard normal distribution from z = -0 to +6 in 0.1
      intervals, jd contains the probability density
100   DIM f(122), fd(122) : REM f contains the value of the function defining the
      boundary between categories, and fd is its derivative
110   LEB xm
120   REM LEB xm turns drive status line off
130   REM contolling section ————————————
140   LEB c
150   REM LEB c clears screen
160   GOSUB 1140: REM reading data
170   zl=-2 : zh=2 : REM z1= z lower extreme to be displayed
180   x1z%=10 : x2z%=500 : y2z%=131 : y4z%=254 : y4max%=140 : x1s=240 :
      x2s=60 : y2s=120 : y4s=150 : REM xlz% is x zero for left graphs, x2z% is x
      zero for right graphs, x1s is x scale for left graphs etc
190   GOSUB 670: REM inputting shifts in z to cum freq curves
200   GOSUB 830: REM drawing axes and scales for graph 4
210   GOSUB 270: REM calculating and drawing Pitman efficiency
220   GOSUB 960: REM drawing axes and scales for graph 2
230   GOSUB 1300: REM drawing cum freq distributions
240   a$ = INKEY$ : IF a$ = "" THEN GOTO 240: REM loop to allow screen dump
250   GOSUB 410: REM printing numbers needed, Pitman efficiency, etc
260   STOP
270   REM calculating and drawing Pitman efficiency = graph 4 ——————————
280   z10%=z1*10 : GOSUB 1440: GOSUB 1400: yprev=LOG10(n2/n) : REM npts
290   FOR z10% = z1*10+1 TO zh*10 : GOSUB 1440: GOSUB 1400: ynew=
      LOG10(n2/n) : REM calculating npts
300   yg1%=y4z%-yprev*y4s : yg2%=y4z%-ynew*y4s : IF MIN(yg1%, yg2%) <
      y4max% THEN GOTO 340
310   z=z10%/10
320   LEB 1 x2z%+(z-0.1)*x2s, yg1%, x2z%+z*x2s, yg2%
330   REM LEB 1 draws line x1, y1 to x2, y2
340   yprev=ynew : NEXT
350   z10%=z1* 10 : GOSUB 1440: GOSUB 1400: yprev=LOG10(n2/n) : REM npts
360   FOR z10% = z1*10+1 TO zh*10 : GOSUB 1440: GOSUB 1400: ynew= LOG
      10(n2/n) : REM calculating npts
370   ygl%=y4z%-yprev*y4s : yg2%=y4z%-ynew*y4s : IF MIN(yg1%, yg2%) <
      y4max% THEN GOTO 390
380   LEB 1 x1z%+j(z10%+59)*x1s, y4z%-yprev*y4s, x1z%+j(z10%+60)*x1s,
      y4z%-ynew*y4s
390   yprev=ynew : NEXT
400   RETURN
410   REM printing numbers needed, Pitman efficiency, etc ——————————
420   INPUT "print to printer?", a$ : IF a$ = "y" THEN LEB le
```

```
430   PRINT
440   IF i$="r" THEN PRINT "Maximum shift: "; zshmax; " Minimum shift: ";
      zshmin; " Interval:"; zshint
450   IF i$="i" THEN PRINT "shifts: "; : FOR j%=1 TO n%: PRINT zshift(j%); :
      NEXT
460   PRINT
470   PRINT " z", "response", "n for 2 cats", "n for expt", "Pitman e."
480   FOR z10% = z1*10 TO zh*10 : z=z10%/10 : GOSUB 1440: GOSUB 1400:
      REM calculating npts
490   PRINT z, j(z10%+60), n2, n, n2/n
500   NEXT z10%
510   REM calculating for interpolated values of response
520   PRINT
530   PRINT "response", "n for 2 cats", "n for expt", "Pitman e."
540   r100%=5 : REM r100% is response rate x100
550   z10%=z1*10 : GOSUB 1440: GOSUB 1400: r1=j(z10%+60) : n21=n2 : n01=n
      : P1=n2/n : REM r1 is response rate immediately below interpolated value
560   FOR z10%=z1*10+1 TO zh*10 : z=z10%/10 : GOSUB 1440: GOSUB 1400:
      r2=j(z10%+60) : n22=n2 : n02=n : P2=n2/n : REM r2 is response rate
      immediately above
570   r=r100%/100
580   IF r2 < r THEN 640: REM continuing until r1 and r2 are on either side of
      interpolated value
590   fr1=(r-r1)/(r2-r1) : fr2=(r2-r)/(r2-r1)
600   PRINT r1*fr2+r2*fr1, n21*fr2 + n22*fr1, n01*fr2 + n02*fr1, P1*fr2 + P2*fr1
610   IF r100%=5 THEN r100%=10: GOTO 640
620   IF r100%=90 THEN r100%=95: GOTO 640
630   r100%=r100%+10
640   r1=r2 : n21=n22 : n01=02 : P1=P2 : NEXT
650   LEB n
660   RETURN
670   REM inputting shifts in z to cumulative freq curves
680   INPUT "regular shifts or irregular - enter r or i"; i$
690   IF i$ = "r" THEN GOTO 720
700   IF i$ = "i" THEN GOTO 790
710   GOTO 680
720   INPUT "maximum shift"; zshmax : REM largest positive shift (i.e. to the
      right) in cum freq curve
730   INPUT "minimum shift"; zshmin
740   INPUT "interval"; zshint
750   ncat%=(zshmax-zshmin)/zshint + 2
760   FOR j%=1 TO ncat%-1 : zshift(j%) = zshmax -zshint*(j%-1) : NEXT
770   LEB c
780   PRINT "Maximum shift: "; zshmax; "   Minimum shift: "; zshmin; "
      Interval:"; zshint : RETURN
790   INPUT "number of boundaries"; n% : ncat%=n%+1
800   FOR j%=1 TO n% : INPUT "key in next shift in decreasing order"; zshift(j%)
      : NEXT j%
810   LEB c
820   PRINT "shifts: "; : FOR j%=1 TO n%: PRINT zshift(j%); : NEXT: RETURN
830   REM drawing axes and scales for graph 4
840   x=x2z%+z1*x2s : GOSUB 1070: REM drawing vert axes for graph 4
850   FOR j%=z1*10+1 TO zh*10-1 : z =j% / 10 : x=x2z%+z*x2s : jrem% j% MOD
      10 : IF jrem% = 0 THEN GOSUB 1070: REM drawing vertical axis
860   NEXT j %
870   x=x2z%+zh*x2s : GOSUB 1070: REM drawing vert axes
```

```
880  FOR j%=0 TO 10: x=x1z%+j%*x1s/10 : GOSUB 1070: NEXT
890  FOR k = 1 TO 2 STEP 0.2: y=y4z%-y4s*LOG10(k) : IF y > y4max% THEN
     GOSUB 1100
900  NEXT
910  FOR k% = 1 TO 10: y=y4z%-y4s*LOG10(k%) : IF y > y4max% THEN GOSUB
     1100
920  NEXT
930  FOR k% = 10 TO 70 STEP 10: y=y4z%-y4s*LOG10(k%) : IF y > y4max%
     THEN GOSUB 1100
940  NEXT
950  RETURN
960  REM drawing axes and scales for graph 2
970  x=x2z%+z1*x2s : GOSUB 1040: REM drawing vert axes
980  FOR j%=z1*10+1 TO zh*10-1 : z =j% / 10: x=x2z%+z*x2s : jrem%=j% MOD 10
     : IF jrem% = 0 THEN GOSUB 1040
990  NEXT j %
1000 x=x2z%+zh*x2s : GOSUB 1040: REM drawing vert axes
1010 x=x1z% : GOSUB 1040: x=x1z%+x1s/2 : GOSUB 1040: x=x1z%+x1s : GOSUB
     1040
1020 y=y2z% : GOSUB 1100: y=y2z%-y2s/2 : GOSUB 1100: y=y2z%-y2s : GOSUB
     1100 : REM drawing horiz axis
1030 RETURN
1040 REM drawing vertical axes for graph 2
1050 LEB 1 x, y2z%, x, y2z%-y2s
1060 RETURN
1070 REM drawing vert axes for graph 4
1080 LEB 1 x, y4z%, x, y4max%
1090 RETURN
1100 REM drawing horiz axes
1110 LEB 1 x2z%+z1*x2s, y, x2z%+zh*x2s, y
1120 LEB 1 x1z%, y, x1z%+x1s, y
1130 RETURN
1140 REM reading data
1150 DATA .5, .5398, .5793, .6179, .6554, .6915, .7257, .7580, .7881, .8159, .8413
1160 DATA .8643, .8849, .9032, .9192, .9332, .9452, .9554, .9641, .9713, .9772
1170 DATA .9821, .9861, .9893, .9918, .99379, .99534, .99653, .99744, .99813, .99865
1180 DATA .99903, .99931, .99952, .99966, .99977, .99984, .99989, .99993, .99995,
     .99997
1190 DATA .99998, .99999 ,.99999 ,.99999 ,1,1,1,1,1,1
1200 DATA 1,1,1,1,1,1,1,1,1,1
1210 DATA .3989, .3970, .3910, .3814, .3683, .3521, .3332, .3123, .2897, .2661, .2420
1220 DATA .2179, .1942, .1714, .1497, .1295, .1109, .0940, .0790, .0656, .0540
1230 DATA .0440, .0355, .0283, .0224, .0175, .0136, .0104, .00792, .00595, .00443
1240 DATA .00327, .00238, .00172, .00123, .00087, .00061, .00043, .00029, .00020,
     .00013
1250 DATA .00009, .00006, .00004, .00002, .00002, .00001, .00001, 0,0,0
1260 DATA 0,0,0,0,0,0,0,0,0,0
1270 nj %=60 : FOR j %=O TO nj % : READ j (60+j %) : j (60-j %)=1 j (60+j %) :
     NEXT j %
1280 FOR j%=0 TO nj% : READ jd(60+j%) : jd(60-j%)=jd(60+j%) : NEXT j%
1290 RETURN
1300 REM drawing graphs of cum freq distributions = graphs 2
1310 FOR b%=1 TO ncat%-1 : REM displaying the cum f dist for each category
1320 xprev=j(z1*10+60) : yprev=j((zl-zshift(b%))*10+60)
1330 FOR z10% = z1*10+1 TO zh*10 :z=z10%/10 : xnew j(z10%+60) : ynew=j
     (z10%-zshift(b%)* 10+60)
```

```
1340  LEB 1 x1z%+xprev*x1s, y2z%-yprev*y2s, x1z%+xnew*x1s, y2z%-ynew*y2s
1350  xprev=xnew : yprev=ynew : NEXT z10% : NEXT b%
1360  FOR b%=1 TO ncat%-1 : REM displaying the cum f dist for each category
1370  FOR z10% = z1*10 TO zh*10-1 :z=z10%/10 : j%=z10%-zshift(b%)*10+60 :
      LEB 1 x2z%+z*x2s, y2z% j(j%)*y2s, x2z%+(z+0.1)*x2s, y2z% j(j%+1)*y2s :
      NEXT
1380  NEXT b%
1390  RETURN
1400  REM calculating n patients if only 2 categories
1410  j%=z10%+60
1420  n2 = 4 * j(j%) * (1-j(j%))
1430  RETURN
1440  REM calculating n patients for the experimental categories
1450  f(0)=0 : f(ncat%)=1
1460  FOR b% = 1 TO ncat%-1 : j%=z10%-zshift(b%)*10+60 : f(b%) j(j%) : fd(b%)
      jd(j%) :NEXT b%
1470  F1=0 : FOR b%=1 TO ncat%-1 : F1 = Fl + fd(b%) * (f(b%+l) - f(b%-1)) :
      NEXT
1480  F2=0: FOR b%=1 TO ncat%-1 : F2 = F2 + f(b%) * f(b%+1) * (f(b%+1) -
      f(b%)) : NEXT
1490  n = 4 * F2 * jd(z10%+60) ^ 2 / F1 ^ 2 : REM the jd(z10%+60) rescales to a
      linear scale of response so that delta is in terms of response rather than z
1500  REM PRINT z; j%; f(1), fd(1), F1, F2, n
1510  RETURN
```

APPENDIX H: THE EFFICIENCY OF THE MANN-WHITNEY TEST RELATIVE TO THE CHI SQUARED TEST USING A DICHOTOMOUS CLASSIFICATION OF CONTINUOUS DATA

Consider the lateral shift model of section 2.6 when the number of categories k tends to infinity, the width of each category w tends to zero, and the distribution becomes continuous rather than categorical. The notation is changed so that a category is denoted by the efficacy E at which the lower boundary function has a value of 50% (as shown in Fig. H.1) and the width is denoted by δE rather than by w.

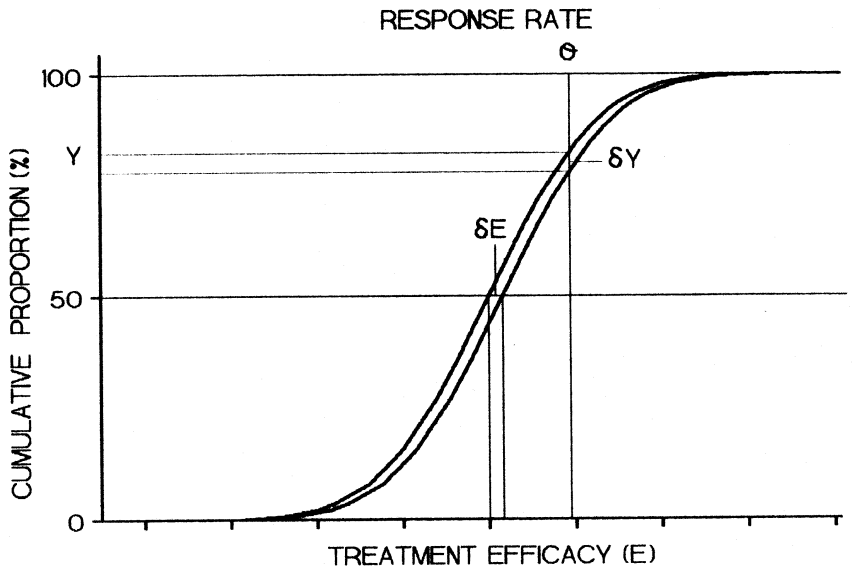


Fig. H.1. Definition of terms.

Then when the response rate is θ (and the efficacy is $\Phi^{-1}(\theta)$)

$$Y = \Phi\{\Phi^{-1}(\theta) - E\}$$

$$Y - \delta Y = \Phi\{\Phi^{-1}(\theta) - (E + \delta E)\}$$

and the proportion of the population contained in category E (denoted δY) is

$$\Phi\{\Phi^{-1}(\theta) - E\} - \Phi\{\Phi^{-1}(\theta) - (E + \delta E)\}$$
$$= \varphi\{\Phi^{-1}(\theta) - E\} \times \delta E \text{ when } \delta E \text{ becomes small,}$$

and thus the distribution of the population amongst the categories δE wide is an asymptotically normal continuous distribution, centred on $E = \Phi^{-1}(\theta)$ and with unit variance.

The asymptotic relative efficiency for two statistical tests can be calculated (Pratt and Gibbons, 1981) by

$$\lim_{n \to \infty} \frac{\{\mu_1'(\theta)\}^2 / \sigma_1^2(\theta)}{\{\mu_2'(\theta)\}^2 / \sigma_2^2(\theta)} \qquad \text{(H.1)}$$

where

μ_1' is the derivative of the first test statistic,

σ_1^2 is its variance, and

μ_2' and σ_2^2 are the same quantities for the second test

The derivatives are normally taken with respect to the index variable (here θ) but here they will be taken with respect to the variable E. This results in no change in expression H.1.

Consider first the Mann-Whitney test applied to continuous normally distributed data, Randles and Wolfe (1979) give

$$(\mu')^2 / \sigma^2 = 3n \{ \int_{-\infty}^{\infty} f^2(E)\, dE \}^2$$

For a normally distributed variable E with unit variance,

$$\{ \int_{-\infty}^{\infty} f^2(E)\, dE \}^2 = 1/(4\pi)$$

Therefore $(\mu')^2 / \sigma^2 = 3n / 4\pi$

Consider now that the continuous normally distributed data is divided into a dichotomy of total response versus non-response, and analyzed by a Chi squared test. This analysis is equivalent to a test of two proportions (Armitage & Berry, 1987) and thus $(\mu')^2 / \sigma^2$ can be derived from consideration of the difference in two proportions and its variance. Hence

$$\begin{aligned} \mu' &= {}^{d}/_{dx}\,\theta \\ &= {}^{d}/_{dx}\,\Phi(E) \\ &= \varphi(E) \\ &= \varphi\{\Phi^{-1}(\theta)\} \end{aligned}$$

For the difference between two proportions θ and $\theta + \delta\theta$, the variance is given by (Armitage & Berry, 1987)

$$\sigma^2 = \frac{\theta(1 - \theta)}{n / 2} + \frac{(\theta + \delta\theta)(1 - \theta - \delta\theta)}{n / 2}$$

$$= 4\,\theta(1 - \theta) / n \quad \text{when } \delta\theta \text{ is small.}$$

Thus substituting in expression H.1, and denoting the efficiency of the Mann-Whitney test of ungrouped continuous data relative to the Chi squared test of 2 categories by e_{MW,χ^2}

$$e_{MW,\chi^2} = \frac{3n / 4\pi}{\varphi^2\{\Phi^{-1}(\theta)\} \;/\; \{4\,\theta(1 - \theta) / n\}}$$

$$= \frac{3\,\theta(1 - \theta)}{\pi\;\varphi^2\{\Phi^{-1}(\theta)\}}$$

This result is used in section 2.6 and Appendix G.