STATISTICAL ISSUES IN THE ANALYSIS OF OUTCOMES IN CRITICAL CARE MEDICINE

A THESIS FOR THE DEGREE OF DOCTOR OF MEDICINE

JOHN LEITH MORAN

DEPARTMENT OF INTENSIVE CARE MEDICINE, THE QUEEN ELIZABETH HOSPITAL, 28 WOODVILLE ROAD, WOODVILLE SA 5011, AUSTRALIA

STATEMENT OF AUTHOR

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and , to the best of my knowledge and belief, contains no material previously published or written by any other person, except where due reference has been made in the text

I give my consent to this copy of my thesis , when deposited in the University Library, being available for loan and photocopying.

_____ _____

John L Moran

February , 2006

TABLE OF CONTENTS

INTRODUCTION

1.1 The focus of this thesis will be the nexus of statistical methods and clinical practice, as it applies to Critical Care Medicine and is reflected in the literature (for instance: Anaesthesia and Intensive Care (Anaesthesia and Intensive Care 2005) and Critical Care & Resuscitation (Critical Care and Resuscitation 2005) in Australia; and internationally: Critical Care Medicine (Critical Care Medicine 2005), Intensive Care Medicine (Intensive Care Medicine 2005), Chest (Chest 2005), American Journal of Respiratory and Critical Care Medicine (American Journal of Respiratory and Critical Care Medicine 2005) and Journal of the American Medical Association (JAMA 2005)).

1.2 Altman has documented the career of statistics in medical journals over a 20 year period and has lamented the general state of affairs (Altman 1982; Altman 1991b; Altman 1994; Altman 2000). The transfer of statistical techniques into medical literature is characterised by a significant lag-time (Altman *et al.* 1994b) and statistical input into medical research and publication, although "widely recommended …(is)… inconsistently obtained" (Altman *et al.* 2002), perhaps reflecting an undervaluation of statistical contributions to medicine, as articulated by one of the doyen's of biostatistics, Norman Breslow (Breslow 2003). The latter observed that, as opposed to the awarding of a Nobel Prize (in 2000) to econometricians Daniel McFadden and James Heckman for work on discrete choice models and selection bias, similar contributions to medicine by statisticians and epidemiologists have been, as yet, unrecognized.

1.3 Our comparators in statistical “critique” (Berk 2004; BROSS 1960) are drawn from analytic approaches, more than thirty years apart. First, the lucid contributions of Jerome Cornfield (Greenhouse 1982); in particular: the classic intervention (in 1959) into the tobacco smoking / lung cancer debate “Smoking and lung cancer: recent evidence and a discussion of some questions” (Cornfield *et al.* 1959); and “Further statistical analysis of the mortality findings” of the University Group Diabetes Program (Cornfield 1971), which was an elegant response to the controversy which raged (for some years (Kolata 1979)) over the discontinuance of tolbutamide and diet arm in that trial. The textual lucidity to which we refer was presumably a function of the literary background of Cornfield, as documented in the classic review by Salsburg of the rise of the modern statistical paradigm in the twentieth century (Salsburg 2001). Second, the muscular re-examination, or rather, dissection, by Freedman *et al* (Freedman *et al.* 2004) of the controversy surrounding breast cancer screening and its efficacy; being a detailed reading of the meta-analysis by Gotszche and Olsen (Gotzsche *et al.* 2000), who had questioned the role of mammography in breast cancer screening in terms of potential lives saved. Third, the subtle 1994 reappraisal by Petitti of the mortality treatment effect of patient “compliance” in randomized trials, as it related to both therapy and placebo groups in the Coronary Drug Project (The Coronary Drug Project Research Group 1981) and the Beta-blocker Heart Attack Trial (Byington 1984). The demonstration that the (cardiovascular) mortality reduction of compliance with placebo was of the same magnitude as that experienced by users of oestrogen replacement therapy, followed the publication of a quantitative assessment of the of the efficacy of oestrogen on coronary heart disease by Stampfer and Colditz, in which a relative risk of 0.56 (95% CI 0.5 –

0.61) was postulated (Stampfer *et al.* 1991). Petitti's review anticipated the null effects (of replacement oestrogen) demonstrated in the subsequent randomized trials of the Women's Health Initiative (The Women's Health Initiative Study Group 1998). These null effects caused extensive debate and some degree of angst in the epidemiological literature and the consequent death of observational epidemiology was rhetorically announced (Lawlor *et al.* 2004).

1.4 The thesis is divided into two parts:

1.4.1 First, a detailed expository analysis of various questions relating to the interpretation of the results of recent noteworthy trials in the medical and Critical Care literature. Initially we come to terms with the seemingly intractable *P*-value question which has regularly surfaced in the literature over the years. We also address the thorny but perennial parametric versus non-parametric test controversy. Next we look at the methodology of recent trials in Critical Care and find some problematic areas in terms of interim analyses and the reporting of results. These concerns are expanded into a detailed consideration of the issues surrounding group sequential and equivalence trials. The subsequent section analyses particular aspects of (i) effect size (ii) prognostic factors and responsiveness (iii) sample size, power and interpretation of trials and we conclude (iv) with a critique of various aspects of Critical Care practice, as it relates to certain key trials and overviews (meta-analyses) of these trials: the PROWESS trial of activated protein C in sepsis; hypothermia as therapy in cerebral injury; selective decontamination of the digestive tract; and nutrition as therapy.

1.4.2 Second, concrete focused analyses are performed on particular datasets and particular statistical techniques are subject to scrutiny. The first encompasses

multivariate analysis of phosphate metabolism in ICU patients; in particular, issues relating to regression to the mean, appropriate estimators (ordinary least squares or generalized linear models), model and variable selection, and missing data. The second looks at the analysis of cost data and explores the use of generalized linear models as appropriate estimators. The third introduces time-to-event analysis in and reviews the use of the Cox model and random effects estimators in a data set of patients with malignancies. The fourth is a in depth analysis of three aspects of meta-analysis as it applies in the Critical Care field: heterogeneity, publication bias and metaregression

1.5 In this endeavour, we are mindful of certain cautions regarding treatment effects: (i) it is reasonable to find odds ratio(s) below 0.6 “extremely surprising” (Speigelhalter *et al.* 2004) (ii) “If a result appears too good to be true, it probably is” (Yusuf 1997) and (iii) we may “require that data indicate an increased relative risk for a characteristic of at least 50 percent, on the assumption that an excess of this magnitude would not arise from extraneous factors alone” (Mantel *et al.* 1959). The latter proposition was first articulated in 1959 by Mantel and Haenszel, but needed to be reiterated (by Mantel) some thirty four years later (Mantel 1993). Finally, we endorse the admonition of Jerome Cornfield that “Any set of hospital or clinical data that is worth analysing at all is worth analysing properly” (Cornfield 1951).

1.6 The importance of statistical principles in both the interpretation and conduct of analysis would seem to be obvious and we must “grapple” with statistics in the same manner as Appleby urged with respect to health economics (Appleby 1987). To this extent, the evidence-based-medicine movement has mandated “critical appraisal”, which incorporates, to varying degree, statistical methods (Morris

2002b) and at least one prominent medical journal has recently welcomed papers "detailing important contributions in the design of studies or analysis of epidemiological data" (Dominici *et al.* 2004). Thus statistics is increasingly engaged with "front-line science" (Efron 2005) and these recent trends prefigure the overall thrust of the sections below.

2 P-VALUES, PRESENTATION OF RESULTS AND DATA ANALYSIS

2.1 P-VALUES (Moran *et al.* 2004b)

2.1.1 The status of *P*-values is a perennial one in the scientific literature. A recent paper in the medical literature by Sterne and Davey-Smith (Sterne *et al.* 2001a) has focused attention on some of the problems of interpretation of significance / hypothesis tests; in particular, the meaning and status of *P*-values associated with these tests. What were these concerns: that the division of results into "significant" and "non-significant" according to a *P*-value = 0.05 was arbitrary and not in accordance with the prescriptions of the founders of statistical inference; that the *P*-value is misinterpreted as the probability that the null hypothesis is true; that, as the absolute value of the *P*-value indexes the level of evidence against the null hypothesis, measures of effect should attract a *P*-value of 0.001, in preference to 0.05, where the evidence against the null hypothesis is not strong; Bayesian approaches to reporting of results may have advantage; and "significance" should not be a primary claim of the reporting of results, which should be accompanied by (90%) confidence intervals and interpreted in the context of the type of study and other available evidence.

2.1.2 As the authors observe, these matters are not new and have repeatedly surfaced in the literature of various scientific disciplines since the establishment of the "testing" paradigm in the 1920s and 1930s by AR Fisher and J Neyman & E Pearson (Gigerenzer *et al.* 1989). Two volumes, separated by almost 30 years and authored from within the behavioural science disciplines: "The significance test controversy – a reader" (Morrison *et al.* 1969) and "What if there were no significance tests" (Harlow *et al.* 1997), further attest to these controversies. In a provocatively entitled recent paper, "Two cheers for *P*-values", Stephen Senn (a

statistician), in a "limited defence of *P*-values", noted that "*P*-values are a practical success, but a critical failure. Scientists the world over use them, but scarcely a statistician can be found to defend them. Bayesians in particular find them ridiculous…." (Senn 2001). Nester, writing in 1996, suggested that "statisticians would be unwise to seek the limelight in any forthcoming 75th anniversary, centennial or tricentennial celebrations of hypothesis testing" (Nester 1996). Rindskopf asked why "Given the many attacks on it, null hypothesis testing…(was not) …dead" (Rindskopf 1997) and the demise of the *P*-value has been rhetorically reported (Evans *et al.* 1988). Within the medical literature similar sentiments have been expressed, as reflected in the titles of certain lead articles: "Confidence intervals rather than P values: estimation rather than hypothesis testing" (Gardner *et al.* 1986), "That confounded *P*-value" (Lang *et al.* 1998) and "Are all significant P values created equal?" (Browner *et al.* 1987). That this was not merely an academic question, was revealed by a decision of the Editor of the journal Epidemiology (Rothman 1998b): "When writing for Epidemiology, you can also enhance your prospects if you omit tests of statistical significance….we do not publish them at all" (Rothman 1998a). In psychology and the social sciences, the tone of discourse (against *P*-values) has at times been shrill, as noted by Nickerson in a recent exhaustive review (Nickerson 2000). In the biological (Johnson 1999) and econometric literature (Keuzenkamp *et al.* 1995) others have added to the chorus of complaint. How did this come about or is it all "a tempest in a tea pot" ? (Nickerson 2000)

2.1.3 History: the paradigm established

2.1.3.1 From our current (medical) perspective, "testing", *P*-values and Type I and II errors appear non-problematic; a "single, unified, uncontroversial means of statistical inference" (Hubbard *et al.* 2003), and the history of the development of statistical inference in the 20th century, a remote echo of current concerns (Goodman 1993). The state of statistics in the first decade of the 20th century has been described as "an unexplored archeological site" (Efron 1998) and the construction of the "testing" paradigm, first by Fisher and then Neyman-Pearson, was self-consciously defined with respect to the 19th century Bayesian dominance of "inverse probability". It was paradoxical that Gosset ("Student"), the inventor of the *t*-test in 1909, which initiated hypothesis testing (Lehman 1993), was a Bayesian (Senn 2002).

2.1.3.2 It is obviously difficult to relive the impact of the Fisherian revolution upon the statistical practice of the first decades of the 20th century, but we may be assured that it was fundamental, although the full significance of the first edition (in 1925) of the classic "Statistical methods for Research Workers" was , perhaps not surprisingly, "not immediately recognized"(Yates 1951). Fisher and Gosset in fact cooperated in calculating tables for the *t*-distribution presented (along with χ^2 and *z*-transformation) in the book (Keuzenkamp *et al.* 1995). The initial "common currency" of significant at 5% and 1% may well have been related to the fact that Fisher's tables (copied subsequently to many text books) were given for *P*-values of 0.01 and 0.05, partly in deference to the copyright limitations of the journal Biometrika, edited by Karl Pearson, the founder of the χ^2 test (Barnard 1990).

2.1.3.3 RA Fisher's position as the "founder" of modern statistics is presumably secure (Rao 1992); the comments of Savage (a Bayesian) attest to this: "It would be more economical to list the few statistical topics in which he displayed no interest than those in which he did" (Savage 1976). The Fisherian *significance* test, deriving from inductive inference, established a null hypothesis (H_0) and used discrepancies in the data to reject the null hypothesis: that is, H_0 posited a sample coming from a hypothetical (infinite) population with a known sampling distribution and H_0 was rejected if the sample estimate deviated from the mean of the sampling distribution by more than a specified criterion (the level of significance; formally, Pr ($x|H_0$)) (Hubbard *et al.* 2003). The *P*-value then, was the (tail area) probability of obtaining a result equal to or more extreme than what was actually observed (Senn 2002). However, for a P value of, say 0.05, it was *not* that the null hypothesis had a probability of (only) 5%. Under an assumption that the null hypothesis was true, it could *not* then be assumed that the P value was a "direct measure of the probability that the null hypothesis is false" (Goodman 1999). A P value of ≤ 0.05 on the null hypothesis indicated, according to Fisher, that: "Either an exceptionally rare chance has occurred or the theory is not true". It did not imply that the investigator accepted being deceived one in twenty occasions, rather it suggested what should be ignored: "all experiments in which significant results are not obtained". Fisher's further advice, oft quoted and ignored, was that "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty…or one in a hundred" (Hubbard *et al.* 2003). A subtle, but important point, was that the inference from the P value involved only one hypothesis and was partially based on unobserved data in the tail region (of the

sampling distribution); thus the "likelihood" of a hypothesis, deriving from the data , was not, at the same time, the "probability of being true" (Gibbons *et al.* 1975; Goodman 1993). That is, P values were not to be misinterpreted as posterior probabilities, a Bayesian proposition (De Groot 1973).

2.1.3.4 The statistical methodology of Neyman & Pearson (articulated primarily in two papers in 1928 and 1933 (Neyman *et al.* 1967)) sought to revise and improve upon Fisher's formulation, but from a deductive position, a paradigmatic difference, which from the Fisherian viewpoint of "rigorous inferences from the particular to the general", was a function of a certain mathematical "bias" of Neyman. Although somewhat distrustful of mathematicians ("Statistical methods for Research Workers" contained no formal mathematical proofs or lemmas), Fisher was in fact a Cambridge trained mathematician (Savage 1976). The Fisher-Neyman rivalry (Pearson later "distanced" himself from the Neyman-Pearson paradigm, more so when Neyman relocated to Berkley) was somewhat of a *cause-celebre* (Marks 2003), although the influence of Fisher's "Statistical methods for Research Workers" on the Neyman-Pearson enterprise was acknowledged by the latter (Lehman 1993), specifically the tabulation of the three distributions mentioned above. Neyman, for his part, charged Fisher with a (persistent) inability to operate with concepts (Neyman 1961); Fisher's circumlocutions were also a cause of irritation to those sympathetic to his view-point, such as Kempthorne: "The last sentence, particularly, leads me to the view that Fisher was talking on a plane barely understandable to the rest of humanity" (Kempthorne 1966). It was Neyman-Pearson methodology which formulated the now familiar two competing hypotheses paradigm, the null (H_0) and the alternate (H_A). This involved the probability of committing two kinds

of errors with respect to the null hypothesis; false rejection (Type I or α error) and false acceptance (Type II or β error). Power, defined as (1-β) or the P(rejecting H_0 | a particular alternative hypothesis), was introduced as a new and critical concept. Within the Fisherian schema, there was a notion of "sensitivity" in detecting departures from the null, but no formal concept of power (Johnstone 1986); although Barnard has argued otherwise (Barnard 1986). The α error was (in theory) prescribed *prior* to data collection and the focus was on minimizing β errors, subject to the bound upon α. What in effect was established were rules for making decisions between two hypotheses ("inductive behavior", although this behavioral aspect may only have been heuristic or hypothetical (Birnbaum 1977)), on the basis that in the "long run of experience, we shall not be too often wrong" (Hubbard *et al.* 2003). Thus for Neyman-Pearson, there was a tension between the control of long term error rates and judgment of the status of the individual experiment (Lehman 1993). The α and β error rates defined a rejection *region* for a test statistic; the significance level, α, was the "probability of a set of future outcomes", represented by the "tail" area of the null distribution. In principle, Neyman-Pearson theory also avoided an arbitrary element in Fisher's approach, the decision regarding the test statistic (Senn 2001). The Neyman-Pearson fundamental lemma guaranteed the existence of an optimal (uniformly) Most Powerful-α test (Pena *et al.* 1998); for a simple H_0 tested against a simple alternative H_A, the optimal test criterion was the likelihood ratio (LR) test (Barnard 1990; Neyman *et al.* 1933). The question of Fisher's approach to the alternative hypothesis is one of some difficulty: arguments that a small probability $p(E|h_o)$ of event *E* is "not enough *per se* to discredit the null

hypothesis h_0" have been "forcefully" advanced (Johnstone 1986). In particular, Berkson's oft cited paper from 1941 in which the question was posed: "If an event has occurred, the definitive question is not, "Is this an event which would be rare if H_0 is true?" but "Is there an alternative hypothesis under which the event would be relatively frequent?" If there is no plausible alternative at all, the rarity is quite irrelevant to a decision…" (Berkson 1947). Undoubtedly Fisher considered the 'alternative' as obligatory, as revealed in conversation with Kruskal and Savage in the 1950's, where the former recalls that "..Fisher agreed that, yes, naturally one had to think about distributions for the sample other than that of the hypothesis under test. And why were we making such a fuss about an elementary and trivial question" (Kruskal 1980).

2.1.3.5 That significance testing could provide statistical inference, rather than behavioural decisions, was the gulf that separated Fisher from (early) Neyman-Pearson. Some modification of the Neyman-Pearson position on this did occur, both internally; Pearson's description of a statistical test as a "means of learning" (Pearson 1955)and Neyman's subsequent equivocations on the matter of inference (Johnstone 1987); and externally, such as the position of Lehmann, in the classic 1959 volume "Testing statistical hypotheses", that in a hypothesis test the "information will be used for guidance…In such cases the emphasis is on the inference…" (Lehman 1959).

2.1.3.6 A data-based *P*-value is a random variable with a distribution, under the null hypothesis and for continuous test statistics, uniform over the interval [0,1], regardless of the size of the study (Hung *et al.* 1997; Schervish 1996). Under the alternative hypothesis, the distribution of the *P* value is skewed and is a function of both sample size ("a natural concept of power" (Berger *et al.*

1987a)) and the distribution of the test statistic that is used. *P*-values are therefore not α-error rates (Goodman 1999; Hubbard *et al.* 2003; Lehman 1993; Senn 2001), although both are tail-area probabilities under the null hypothesis. As Berger and Delampady note: "*P* values are not a repetitive error rate, at least in any real sense. A Neyman-Pearson error probability, α, has the actual frequentist interpretation that a long series of α level tests will reject no more than 100α% of true H_0, but the data dependent *P* values have no such interpretation" (Berger *et al.* 1987a). For a fixed, pre-specified α, the Neyman-Pearson decision rule "could be defined equivalently in terms of the *P*-value" (Hubbard *et al.* 2003) but what would be of interest was the fact that $P < \alpha$, not the specific value of P. It is ironic that standard applied statistical practice lies easily with an amalgam of the two methods, although from an alternate perspective, this amalgam may be considered as one of statistics "greater triumphs" (Carlton 2003). Tests of significance, as reported in journals, would appear to follow Neyman-Pearson "formally" but Fisher "philosophically" and practically (Johnstone 1986).

2.1.3.7 In the debate about the utility of *P*-values versus confidence intervals (CI), it is often forgotten that CI were introduced by Neyman in 1937 (Neyman 1937) and were considered by Neyman, and commentators, as integral to the overall theory of hypothesis testing (Neyman 1977), which embodied the (frequentist) theory of repeated sampling (an anathema to Fisher (Fisher 1955)). Thus for a 95% CI of a parameter θ, the interpretation is that in (an infinite number of) repetitions of a study, an exact proportion (95%) of all such intervals would enclose θ. Once the data has been collected and a single 95% CI has been calculated, the probability that θ lies within this CI is now 0 or 1. That is, a

95% CI is not equivalent to a 95% probability interval (which has a Bayesian interpretation) (Goodman 1994; Kupper 1998; Macdonald 2002; Young *et al.* 1997). Besides the theory of testing, the second great divide between Neyman and Fisher was that of confidence versus fiducial intervals, the latter being based on fiducial probability, introduced by Fisher as an alternative to Bayesian posterior probabilities (Seidenfeld 1998). Fiducial probability had as its basis certain sufficient statistics (*F* and *t* statistics and correlations) which contained all the information in a sample relevant to population parameters (inference from sample to population). Although agreement can be demonstrated between CI and fiducial intervals; the classic paper establishing so called exact (or Clopper-Pearson) CI of the binomial was entitled "The use of confidence or fiducial limits illustrated in the case of the binomial" (Clopper *et al.* 1934); the latter has not stood the test of Neyman's withering attacks (Neyman 1941; Neyman 1956; Neyman 1961) nor time (Fisher's "biggest blunder" (Efron 1998)). It is of interest, however, to record that Fisher (as early as 1935) apparently "recognized that 'confidence intervals' are 'only another way of saying that, by a certain test of significance, some kinds of hypothetical possibilities are to be rejected, while others are not' " (Dempster 1998).

2.1.4 History: the paradigm revised

2.1.4.1 The literature in response to the Fisher and Neyman-Pearson paradigm is, not surprisingly, enormous in breadth and detail, both from within the statistical (Berger 2003; Birnbaum 1977; Hacking 1965; Kyburg Jr 1974; Spielman 1973; Spielman 1974) and applied scientific disciplines, as noted above. One

of the more systematic and useful developments is that associated with DR Cox, in which *P*-values are treated as "rough tools for inspecting data" (Salsburg 1998). The perspective, developed fully in the 1974 volume of Cox and Hinkley (Cox *et al.* 1974), is one of eclecticism, the central theme being that "it is fruitful to contemplate problems formulated in different depths of detail and to use different methods accordingly…the most primitive formulation is that for a pure significance test, where only the null hypothesis under test need be explicitly formulated, and the richest formulation is that for Bayesian decision analysis…" (Cox 1978). Null hypotheses are not viewed as undifferentiated species, rather, are divided into plausible (close to the truth) and dividing (divide the range of possibilities into qualitatively different types) hypotheses, which may also be further sub-divided and specified. Statistical strategies, the use of significance tests and the actual *P*-value level, are determined contingent upon these hypothesis specifications (Cox 1977; Cox 1982).

2.1.4.2 A significance test (measuring the consistency of the data with a null hypothesis) has the following form: a function $t = t(y)$ of the observations exists, such that, the larger $t(y)$, the greater the inconsistency of y (the observed vector of responses) with H_0. $T = t(y)$ is the test statistic (a random variable). If the distribution of T is known (when H_0 is true), then the level of significance $p_{obs} =$ pr $(T \geq t_{obs}: H_0)$. The result of such a test is a significance level (not a decision); p_{obs} is a "guide, and no more, to interpretation" (albeit the mathematical connection between p_{obs} and "critical regions of pre-assigned size", that is, Neyman-Pearson testing) (Cox *et al.* 1974; Cox 1977; Royall 1986). A similar approach was also recommended by Kempthorne

(Kempthorne 1976). Of interest, Cox, in a somewhat sympathetic response to the paper by Sterne and Davey-Smith (Sterne *et al.* 2001a), suggested that "To distinguish several types of hypotheses that might be tested helps to understand the issues" (Cox 2001).

2.1.5 The problem revisited

2.1.5.1 Where then do we stand? In a response to the debate over the paper by Sterne and Davey-Smith (Sterne *et al.* 2001a), Berger (Berger 2001) outlined potential problems with hypothesis testing and it is useful to consider some of these.

2.1.5.1.1 *P*-values are misunderstood: the frequent misrepresentation of *P*-values and hypothesis testing, especially in textbooks, has been repeatedly documented (Dracup 1995; Smith 2001). Such, as with other statistical misrepresentations, is not an argument for their abolition.

2.1.5.1.2 *P*-values as a measure of support: Sterne and Davey-Smith (Sterne *et al.* 2001a) suggest a graded level of evidence against the null hypothesis, indexed by the *P*-value (page 229); such scales date back to the 1970s (Royall 1986). A corollary to this is the so called α-postulate of Cornfield (rejected by him in favour of likelihood ratios), that "All hypotheses rejected at the same critical level have equal amounts of evidence against them" (Cornfield 2004). However the question of sample size for equal *P*-values needs to be considered; a number of commentators have argued that for, say a *P*-value of 0.05, there is stronger evidence against H_0 for a small sample than a large one (Bandt *et al.* 1972; Berkson 1947; Freeman 1993; Gibbons *et al.* 1975; Pratt 1961; Royall 1986). Schervish also demonstrated that "the

interpretation of a particular value on the scale of support, such as the popular .05, must vary with the hypothesis" and was "unable to construct a consistent interpretation of the *P*-value as anything similar to a measure of support for its hypothesis" (Schervish 1996).

2.1.5.1.3 *P*-values are associated with rigid cut-off values: a flexible (eclectic) attitude to tests and *P*-values, associated with the approach of Cox and Kempthorne, has been outlined above. As opposed to Berger, it would indeed appear reasonable that there "should be no sharp distinction made between cases having a *P*-value of say 4.9% and those having a *P*-value of 5.1%- a distinction forced by the language of confidence interval testing " (Kempthorne 1972). In the presence of a bewildering array of possible statistics, the advice of Kempthorne to "Look at it" (Kempthorne 1972) seems apposite; similar to the admonitions of the adherents of the likelihood principle (Perneger 2001) and the Bayesians.

2.1.5.1.4 *P*-values are the wrong measure of evidence: from the Bayesian perspective, P values overstate the evidence against the null hypothesis and other methods to adduce evidence (likelihood ratios) may be of more utility (Goodman *et al.* 1988). In a frequently cited paper by J.O. Berger and Sellke, it was shown (two sided testing a normal mean) that with a *P*-value of 0.05, the posterior probability of the null was at least 0.30 for any objective prior distribution (Berger *et al.* 1987b). However, in the one-sided setting, where the different geometry of H_0 and H_A was not operative, the discrepancy, Bayesian posterior probability versus *P*-value, was no longer evident (Casella *et al.* 1987b). Technically, this 'geometry' relates to the fixing of a (prior) probability mass on the null and varying it on the alternative; Casella and R.

L. Berger suggest that the discrepancy between *P*-values and $P(H_0|x)$ is a function of the large (50%) prior probability mass placed on H_0 by J.O. Berger and co-workers (Berger *et al.* 1987a; Berger *et al.* 1987b) and conclude that "there is agreement between *P*-values and Bayesian interval null calculations in the more typical situation in which small prior probability is assigned to H_0" (Casella *et al.* 1987a). As Casella and R. L. Berger note, "We would be surprised if most researchers would place even a 10% prior probability on H_0 ", in accord with the sentiments of Meehl, who maintained that the point null hypothesis is "…[quasi-] always false in biological and social science" (Meehl 1978). *P*-values and posterior probabilities are not necessarily in competition and any difference in conclusions reached does not "….by itself invalidate either measure" (Dracup 1995). Of interest, J.O. Berger and co-workers recently proposed calibrating *P*-values such that they may be interpreted in either a Bayesian fashion ($B(p) = -e.p \log(p)$, when $p < 1/e$) or a frequentist way ($\alpha(p) = (1+[-e.p \log(p)]^{-1})^{-1}$ (Sellke *et al.* 2004).

2.1.5.1.5 Goodman, again from a Bayesian perspective, calculated the replication probability (using an uninformative prior) of trials at a *P*-value of 0.05; this was found to be 50% and lower than "expected" (by non-statisticians) (Goodman 1992). Senn has subjected the import of this finding to close scrutiny (Senn 2001; Senn 2002). Firstly, from a Bayesian perspective, *P*-values are not unreasonable given an uninformative prior. However, the problem is that "…the 'uninformative' prior is rarely appropriate…it is not possible to survive as a Bayesian on uninformative priors…" (Senn 2002). Secondly, the requirement that a single significant *P*-value should entail near certainty that a second will follow, is deemed by Senn to be an undesirable

property: “Anticipated evidence is not evidence, nor do we want it to be. To expect that it is, is to make exactly the same mistake that physicians make in saying, ‘the result was not significant, p = 0.09, because the trial was too small’ “ (Senn 2001). This being said of the general problem of replicability, empirical studies have suggested that the *P*-value does provide “a continuous measure that has an orderly and monotonic mapping onto confidence in the replicability of a null hypothesis rejection “ (Greenwald *et al.* 1996) and statistically significant exact replication (SSER) may be a useful interpretative measure (Posavac 2002).

2.1.6 *P*-values and CI

2.1.6.1 There would appear to be considerable virtue in reporting both *P*-values and CI, on the basis that singular statements such as $P < 0.05$, or P = nil sig, convey little useful information, although for a $100(1-\alpha)\%$ CI, it must be remembered that any violation of the assumptions that effect the true vale of α (obviously) effect CI precision (May 2003). From the Bayesian perspective, Lindley has summarized the position thus: “significance tests, as inference procedures, are better replaced by estimation methods…it is better to quote a confidence or credible interval….estimation procedures provide more information…..Nevertheless there remain cases where significance test have an advantage…” (Lindley 1986). Numerous papers within the medical literature have attested to the utility of CI (Braitman 1991; Simon 1993; Thompson 1987); in the epidemiological literature, the *P*-value has been condemned as confounded, in that the information “mixed” in the *P*-value should be

separately reported: the size of the effect (estimated by the, say, the risk ratio) and the precision of the estimate (described by the SE or CI) (Lang *et al.* 1998). Poole has suggested that for epidemiological measures such as relative risk, the estimates least influenced by chance are those with narrow confidence intervals, not low *P*-values (Poole 2001). However, as pointed out by Feinstein, *P*-values and CI methods are essentially reciprocal and do not provide "an evaluation of substantive importance for the 'big' or 'small' magnitude of the observed distinction…they offer no guidance for the basic quantitative scientific appraisals that depend on purely descriptive rather than inferential boundaries.." (Feinstein 1998). Similarly, Poole has also suggested that CI "are usually taken as nothing more than tests of significance" and proposed that the complete *P*-value (that is, the graph of all possible *P*-values or CI) or likelihood function be used for the "main result of an epidemiological study" (Poole 1987).

2.1.6.2 An interesting case study of the interpretation of "CI without *P*-values" and a focal point for a lively exchange in the American Journal of Public Health in the mid 1980's on the role of *P*-values, was the response of Fleiss to a relatively small case-control study (Foxman *et al.* 1985) in which all 12 reported CI for summary odds ratios included 1 (extending from 0.4 to 17); yet the associations were variously described as 'strong' 'negative' and 'positive'. Fleiss asked the not unreasonable question "There is no gainsaying that tests of significance have been abused, but at least they have the virtue of providing explicit, pre-specifiable criteria for inferring that an association is real. This is not the case with confidence intervals, at least as far as the paper in question is concerned….I would appreciate learning just what criteria were employed to

conclude that the ….associations were ‘strong’, ‘negative’ and ‘positive’ “ (Fleiss 1986b). The authors acknowledged the small study size, and suggested that the reader should be “cautious in generalizing our results”, but even when not “statistically significant” point estimates may “significantly add to our understanding” (Foxman *et al.* 1986). The latter position would appear to distance itself considerably from the Fisherian requirement of “rigorous uncertainty” (Marks 2003).

2.1.6.3 The question of the “propriety’ of associations and / or claims of efficacy also resonates with the reporting of drug trials (Cutler *et al.* 1966); the case for confidence intervals as estimates of effect has been well argued (Borenstein 1994) and indeed, would appear to be non-controversial. In the presence of competing claims and professional enthusiasts, one can, with V.W. Berger, question the likelihood of being mislead (Berger 2001) and find *P*-values eminently applicable to control this probability (Senn 1993), not withstanding the utility of other (for example, Bayesian) approaches (Hughes 1993).

2.1.7 Overview

2.1.7.1 Efron summarized the major reasons why, as opposed to the Bayesian 19th century, Fisherian and Neyman-Pearson ideas have held sway in the 20th, although subsequently suggesting that the 21st century would see “a combination of Bayesian and frequentist ideas” (Efron 2005): ease of use, model building, division of labour (parts of a complicated problem may be addressed separately) and objectivity (Efron 1986). Thus, despite the belief that

P-values are dead and buried (by some journals), we would agree with Fleiss that significance tests are “alive and well” (Fleiss 1986c).

2.2 NORMALITY OF DATA, PARAMETRIC AND NON-PARAMETRIC TESTS (Moran *et al.* 2002d)

2.2.1 Although the debate over the propriety of standard deviation (SD) versus standard error (SE) (Streiner 2000) for the presentation of results (Bartko 1985; Brown 1982) would appear to have concluded, the same cannot be said for the question of parametric versus non-parametric tests for the analysis of non-normally distributed data and / or small unequal data-sets. Conventional wisdom (Coyle 1996; Lumley *et al.* 2002) suggests that, when comparing two independent groups in the presence of one or both of the above conditions, the *t*-test may be unreliable and the Wilcoxon-Mann-Whitney (WMW) test is therefore preferable (Barber *et al.* 2000b; Ludbrook 1996; Murray 1996). That is, type I and II error rates are affected by violation of underlying test assumptions; type I errors are “liberal”, resulting in spurious rejection of the null hypothesis, and power rates are depressed resulting in undetected effects (Wilcox *et al.* 1998). Modifications of the *t-test* do exist, for both inequality of variance (Welch 1937) and skewness (Johnson 1978) (similarly for the WMW test (Fligner *et al.* 1981)), and these have been implemented variously in statistical packages. The important point to note is that such recommendations tend also to be detail non-specific, to the extent that they do not address the question of the degree of non-normality or “how small”. That such a

conventional strategy of preference for the WMW test for “non-normal data” analysis may lead to strikingly different conclusions was demonstrated by Barber and Thompson using a cost data-set (Barber *et al.* 2000b): a p value for the WMW test of 0.011 and for the *t-test*, 0.71. That one would select the *P*-value according to the underlying hypothesis is unconscionable, but presumably, not unknown, in the same manner as the presentation of the standard error (SE) instead of standard deviation (SD) to make the data look “better”, as above. Moreover, such a “test dredging” approach, involving multiple standard tests, increases the Type I error beyond the nominal 5% level (Gans 1984).

2.2.2 When considering the use of statistical tests such as the *t-test* and the WMW test, an important and frequently overlooked assumption is that of independence of observations. Assuming independence is highly likely to be reasonable in the clinical trial context, but this is not necessarily so for epidemiological studies, where for example, patients may be clustered by disease state or other criteria. As Cox and Hinkley observed some time ago, “The main emphasis in distribution-free tests is on avoiding assumptions about distributional form. In many applications, however, the most critical assumptions are those of independence.” (Cox *et al.* 1974). It is important to realize that the WMW test gives no more protection against false-positive inference than the *t*-test (Ludbrook 1996); the WMW test, in its normal approximation, is in fact equivalent to the *t-test* on the ranks of the original variables (Conover *et al.* 1981; Zimmerman *et al.* 1992). More-over, the interpretation of the null hypothesis being tested with the WMW test is not easily described (Lumley *et al.* 2002); conventional interpretation (including examples provided by statistical packages (Bergmann *et al.* 2000)) would have it that the null hypothesis is one

of equal group medians, but such is not the case; rather it is a test for equality of group mean ranks (Ludbrook 2001). Because the WMW test is a test of both location and shape, its interpretation as a test of medians is valid when the only distributional difference is a shift in location (Hart 2001); that is, for the MWM test to be "distribution free", under the null hypothesis of equal medians, the two populations being compared are assumed to be continuous and have the same shape (Fligner *et al.* 1981). The (Mood) median test has low power in small samples and is not recommended (Freidkin *et al.* 2000). That differences in spread may be as important as (putative) differences in medians is often overlooked. Furthermore, there are different outcomes from WMW test dependent upon the statistical package; these differences relate to the handling of ties, the use of the continuity correction and the use of the asymptotic approximation versus the exact permutation distribution. The latter form of the WMW test would appear to be the preferred option (Bergmann *et al.* 2000).

2.2.3 To overcome some of these problems, data transformation (log, square root and reciprocal) to achieve (approximate) normality is often used, but such transformations result in comparisons of geometric (for log transformation (Keene 1995)) and harmonic (for reciprocal transformation) means and statistical inference in comparing these means cannot be equated with the test of arithmetic means, unless (for geometric means, at least) the variances on the log-scale are equal (Barber *et al.* 2000a; Millns *et al.* 1995; Zhou *et al.* 1999). Due note of the potential loss of power (ranging from 2 to 10%) in analyzing transformed data must also be undertaken (Kingman *et al.* 1994). Back transformation (to the original scale) may also be problematic for "differences" following square root or reciprocal transformations; with logarithmic

transformation, the antilog of a mean log difference gives the ratio of geometric means (Briggs *et al.* 1998). Although frequently used in environmental and chemical research, the geometric mean has not been without its critics as an appropriate data summary statistic (Parkhurst 1998). Moreover the geometric mean is a biased estimator of the arithmetic mean and the latter statistic may be appropriate in considering such variables as costs and their surrogates; that is, where consideration of total costs is of importance (total costs = average costs x no of patients). The skewness (Benjamini *et al.* 1996) of the distributions of costs and length of stay (Weissman 1997) may mandate, for descriptive purposes, the reporting of summary statistics such as mean and standard deviation, median and range (see for example, in the critical care literature, Esteban *et al* (Esteban *et al.* 2002)). However, such does not negate the appropriateness of the arithmetic mean for statistical inference (Barber *et al.* 1998).

2.2.4 The above being said, what is known of the performance of the two tests under varying conditions? Lumley *et al* (Lumley *et al.* 2002), reviewing a number of studies of *t*-test performance under conditions of non-normality, with sample sizes ranging from as low as 3 to greater than 80, found the performance to be acceptable in terms of Type I and II errors; kurtosis (DeCarlo 1997) having less impact than severe skewness (the effect of positive skewness actually results in the sampling distribution of the *t* statistic becoming negatively skewed (Sutton 1993)). This also applied to extreme distributional situations of "floor effects" or "discrete mass at zero"; that is, when up to 50% of subjects record zero for the measured variable (Sawilowsky *et al.* 1992b; Sullivan *et al.* 1992). It also is often forgotten that at sample sizes of 25 to 30 and above, by virtue of the

Central Limit Theorem, the sampling distribution of *t* is effectively normal (Boneau 1960; Stonehouse *et al.* 1998). With respect to comparisons with the WMW test, results have been variable, dependent upon the experimental set-up; in particular, the use of "mathematically convenient" distributions or "real life" data sets from different disciplines, including psychology and education (Sawilowsky *et al.* 1992a). Skovlund and Fenstad (Skovlund *et al.* 2001), in a simulation study with sample sizes ranging from 5 to 15 and using combinations of equal and unequal variances and sample sizes, and distributions as normal, heavy tailed and skewed, reported that the *t-test* (and the Welch variant, for unequal variances) again had acceptable performance for most of the combinations, except for severely skewed distributions with unequal sample size and unequal variances. The WMW test was shown to be very sensitive to unequal variances (that is, deviations from a pure shift model) and was not recommended for any combination of data characteristics when this condition was present. Under these circumstances, the Welch *t-test* variant and / or data transformation was recommended. Similar results were noted by Zimmerman (Zimmerman 1998), who varied both normality and homogeneity of variance; in particular, the Type I error probabilities of the WMW test were more severely distorted with heavy tailed densities, unequal variances and sizes, with larger variances associated with the smaller sample size (n varying from 15 to 40). Bridge and Sawilowsky (Bridge *et al.* 1999), investigating the power (ability to detect a false null hypothesis) of the two tests in multimodal, mass at zero and extreme asymmetry distributions, with small n, found a comparative power advantage for the MWM test, which was substantial in some instances; supporting the previous study of Zimmermann and Zumbo (Zimmerman *et al.*

1992). However, these comparisons involved a location shift only and were not subject to multiple violations of assumptions, which, as Zimmerman observed (Zimmerman 1998), can produce “anomalous effects not observed in separate violations”.

2.2.5 The *t-test* would thus appear to be surprisingly robust to violation of assumptions, and any advantage of the MWM test would appear to be in situations of extreme skewness of underlying distributions, albeit such advantage may be compromised by variance non-homogeneity of the two groups being compared. Paradoxically, the WMW test becomes far less robust with increase in sample sizes (Stonehouse *et al.* 1998; Sutton 1993). However, alternatives to these tests are available (Zhou *et al.* 2001a). On the basis that biomedical research usually involves small samples and proceeds via randomization of a non-random sample rather than random sampling, and thus the randomization, not the population model applies, Ludbrook has argued for the use of permutation tests (Ludbrook 1994; Ludbrook *et al.* 1998). With appropriate software (Stat Xact 4 1999), permutation tests (Good.P. 2000) have become a feasible option and noted to have some advantage (Cohen *et al.* 1991). A second approach is the non-parametric bootstrap (Efron *et al.* 1993), in which an empirical estimate of the sampling distribution of the statistic in question is obtained by repeated sampling (for example, 1000 times) with replacement from the observe data. A number of recent papers have used bootstrap techniques in the analysis of skewed data (Barber *et al.* 2000a; Briggs *et al.* 1998; Desgagne *et al.* 1998; Rascati *et al.* 2001), although caution has been expressed about the robustness of the boostrap in these circumstances (O'Hagan *et al.* 2003). The lack of widespread use of these two alternative approaches to the *t-test* and the

WMW test may reflect the previously described lag-time of diffusion of statistical techniques into the medical literature (Altman *et al.* 1994b).

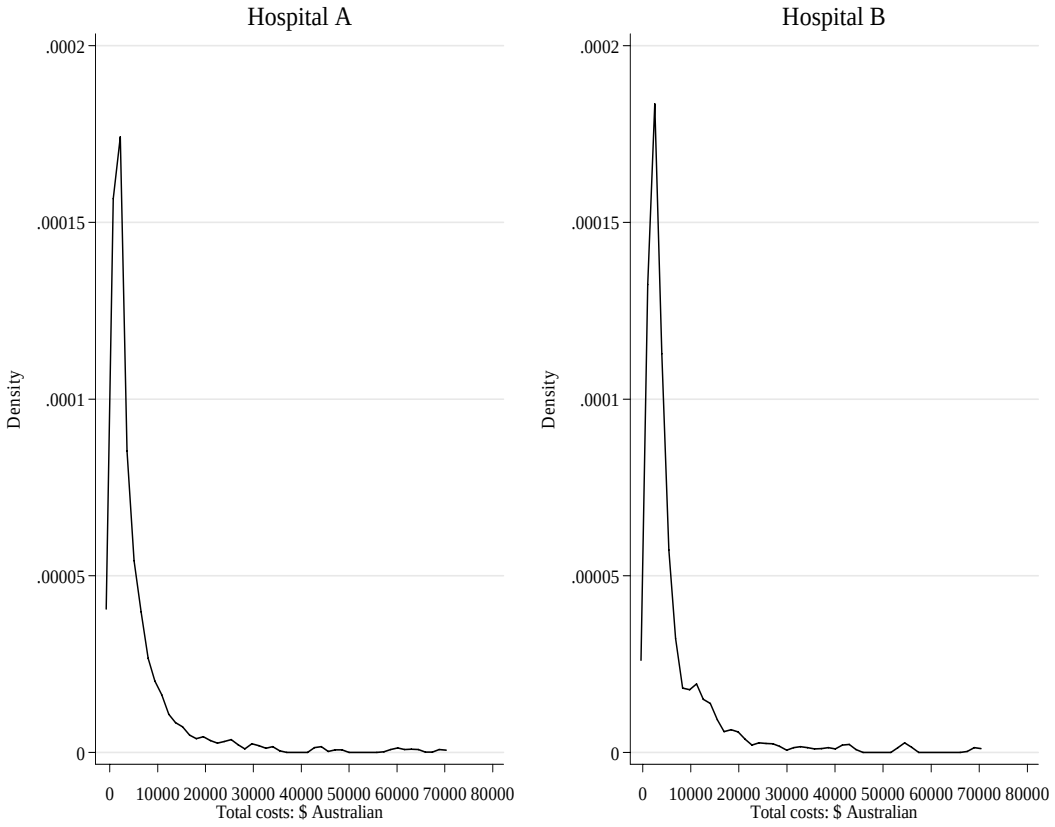
2.2.6 The points raised above are illustrated by a consideration of patient ICU-episode cost data (in $AUS), for two different hospitals, previously reported (Moran *et al.* 2001d). A tabular summary of the patient costs and a kernel density plot of costs for the two hospitals considered are below in Table 2.2.6 and Figure 2.2.6, respectively.

Table 2.2.6. Total costs ($AUS) for ICU admissions

Hospital	n =	mean	median	SD	min	max
A	410	5463	2478	8767	242	69327
B	244	6366	3155	9113	605	69426
Total	654	5800	2804	8901	242	69426

n, ICU patient number

Figure 2.2.6. Kernel density estimates of total costs, by hospital



2.2.6.1 The costs (both total and for each hospital) demonstrated significant kurtosis (p = 0.001) and skewness (p = 0.001), albeit there was no variance inequality between the hospitals (p = 0.50). Log transformation, in this case, did not effect normality (Shapiro-Wilk test, p = 0.0001) and served only to exacerbate variance disparity (p = 0.006). The difference in mean costs between the hospitals via the *t-test* was non-significant with a p = 0.21; the WMW test suggested a difference between hospital costs at the 0.0001 level. Log transformation of costs resulted in a significant *t-test* (p = 0.0004), but as noted

above, such refers to a comparison between geometric means and not a test of the "original" null hypothesis, of equality of arithmetic means, a point reiterated by Zhou *et al* (Zhou *et al.* 1997). Using the bias–corrected and accelerated (BCa) bootstrap method (Carpenter *et al.* 2000), no significant difference was noted between mean costs for the hospitals (95% CI of the difference: -$337 to +$2552); similarly, the two-sided p value for the (exact) permutation two-sample test, using the raw data as scores, was 0.11. It would appear, therefore, that there was no evidence of a difference between the (mean) ICU costs of the two hospitals.

2.2.7 The conclusion to be drawn is that formulaic application of statistical tests is inappropriate; careful consideration of the null hypothesis being tested is needed to guide statistical inference.

3 CLINICAL TRIALS: DESIGN AND CONDUCT (Moran *et al.* 2001b; Moran *et al.* 2003c)

3.1 As opposed to other disciplines in medicine where defined therapy for acute disease process has been established (for example, in cardiology, fibrinolytic therapy for acute myocardial infarction (Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. 1994)), until recently “definitive” therapy in Critical Care has been difficult to establish. However, four trials reporting positive outcomes for interventions in critically-ill patients have recently been reported (Bernard *et al.* 2001; Rivers *et al.* 2001; The ARDS Network Authors for the ARDS Network 2000; van den Berghe *et al.* 2001) and were received with some interest by practitioners. All but one were large scale studies and used group sequential trial methods to allow the possibility of early study termination.

3.1.1 In the Prowess trial (Bernard *et al.* 2001) "O'Brien-Fleming spending function according to the method of Lan & DeMets", with initial enrolment of up to 2280 patients and two planned interim analyses at 760 and 1520 patients. In the ARDS Network trial (The ARDS Network Authors for the ARDS Network 2000) “interim analyses…after each successive group of approximately 200 patients. Stopping boundaries (with two-sided α level of 0.05) were designed to allow early termination of the study if the use of lower tidal volumes was found to be either efficacious or ineffective” No initial enrolment estimate was provided. Patient recruitment was, in fact, prematurely ceased in both studies due to a positive treatment effect, after the second and fourth interim analyses, respectively. In the Prowess trial, primary statistical analysis was based upon the Cochran-Mantel-Haenszel test; in the ARDS Network trial, primary analysis

was based upon the 180-day cumulative mortality. Both trials demonstrated modest treatment effects (6.5% and 9% absolute risk reduction) and both had similar lower 95% limits for the risk reduction (2.2 and 2.5% respectively). The Early Goal-Directed Therapy Collaborative Group trial (Rivers *et al.* 2001), performed two interim analyses (at one third and two thirds of the total patient enrolment), using the "alpha spending function of Demets and Land". The primary outcome of in-hospital mortality was reported as 30.5% in the treatment group (early goal-directed therapy, *n*= 130) and 46.5% in the standard therapy group (n = 133). In the "Intensive Insulin Therapy in Critically Ill Patients" trial (van den Berghe *et al.* 2001), a total of 1548 patients were enrolled with a primary outcome of intensive care mortality, which was 4.6% in the intensive-treatment group and 8.0% in the conventional-treatment group. The trial used the "Lan and DeMets method" to adjust for interim analyses and the trial was stopped at the fourth such analysis for efficacy (of the intensive-treatment group).

3.1.2 However, the physician who considers changing practice based upon the actual magnitude of these risk reductions must be circumspect, to the extent that these estimates are overly optimistic. Although little recognised in the clinical literature, it has been demonstrated for some time that early trial stopping based on group sequential designs results in inflated estimates of treatment effect and inappropriately narrow and incorrectly centred confidence intervals. Such bias(es) are also a function of the number of interim analyses (Pinheiro *et al.* 1997; Pocock *et al.* 1989; Stewart *et al.* 1996). As Demets and Lan observed, "Naïve estimates are biased after a sequentially designed trial has been completed, and appropriate adjustments for unbiased point estimates involve

parameters whose values are typically unknown." (DeMets *et al.* 1995). Hughes and Pocock, addressed the question of exaggerated magnitude of the treatment effect and proposed a Bayesian solution (Hughes *et al.* 1988; Pocock *et al.* 1990), Emerson and Fleming (Emerson *et al.* 1990) and Pinheiro and DeMets (Pinheiro *et al.* 1997) reported methods for estimating and reducing the bias of treatment differences. Emerson (Emerson 1993) and Kim (Kim 1989) both derived unbiased estimators following a group sequential trial and Liu *et al*. investigated appropriate adjustments for secondary hypotheses to control inflated Type I error and reduced power of conventional likelihood-based testing procedures (Liu *et al.* 2000). Two recent software packages EaST (Cytel Corporation 2003) and S+SEQTrial (Insightful Corporation 2002) provide facility for use of these unbiased estimators ((Kim 1989) and (Emerson *et al.* 1990) respectively); other packages also available are PEST (MPS Research Unit 2002) and a public domain routine (Reboussin *et al.* 2000) .

3.1.3 The above being said, it is of interest to note that, of the four above trials, only the "Intensive Insulin Therapy in Critically Ill Patients" trial (van den Berghe *et al.* 2001) reported adjustment of the final treatment estimate for sequential analysis of the trial; from an apparent risk reduction of 42% (95% CI: 22%-62%) to 32% (computed as the median unbiased estimate (Emerson 2000); 95% CI: 2%-55%). Group sequential methods obviously have an established place in trial methodology (Pocock 1992) and a review of randomised (non Critical Care) trials reported in the New England Journal of Medicine over a period of January 2000 to March 2001, revealed a total of 16 randomised trials conducting interim analyses. Of these, 10 used formal O'Brien-Fleming stopping rules; one used Pocock stopping rules and two mentioned "formal" stopping boundaries. A

positive treatment effect was established in 10 trials, all but one (using no formal stopping rules) being stopped early. Trial statistical methodology statements were variously detailed, but none specifically identified group sequential methods as a potential source of bias and no "positive" trial adjusted (downward) the primary outcome estimate(s). A detailed review of these methods is therefore appropriate.

3.2 DESIGN AND MONITORING OF CLINICAL TRIALS

3.2.1 The pre-eminence of the randomized clinical trial to assess medical interventions has been firmly established, although there has been some recent disquiet (Britton *et al.* 1998; Ioannidis *et al.* 2001). It would seem appropriate then, to review some aspects of trial design and termination in the light of new developments, theoretical and practical. By way of introduction, we highlight some details from a trial in the critically ill which was prematurely stopped.

3.2.2 The ALVEOLI trial (ARDS Network 2002) compared two ventilatory strategies in patients with acute lung injury and acute respiratory distress syndrome (ALI / ARDS): (a) higher end-expiratory lung volume / lower FIO_2 (HEELV / $LFIO_2$) versus (b) lower end-expiratory lung volume / higher FIO_2 (LEELV / $HFIO_2$), the latter ventilatory strategy being identical to that of the treatment group in the initial ARDS Network trial (The ARDS Network Authors for the ARDS Network 2000). The trial was stopped (The NHLBI ARDS Clinical Trials Network 2004) for lack of efficacy in early 2002, based upon a protocol specified futility stopping boundary (ARDS Network 1999). The trialists had planned to enroll a maximum of 750 patients (power 89%) over 2-3 years

assuming a 10% mortality difference between the LEELV / $HFIO_2$ group (mortality 28%) and the HEELV / $LFIO_2$ group (mortality 18%) with 89% power to detect such a difference. Two interim analyses were planned at 250 and 500 subjects. The trial stopping criteria for (i) efficacy, were O'Brien-Fleming boundaries (O'Brien *et al.* 1979) with one-sided p = 0.025, and (ii) futility were, at the first interim analysis, a mortality in the HEELV / $LFIO_2$ group greater than that observed in the LEELV / $HFIO_2$ group and , at the second interim analysis, a mortality in the HEELV / $LFIO_2$ group not at least 2% better than the LEELV / $HFIO_2$ group. If there were no true difference in mortality between the study groups, the chances of stopping at the first and second interim analyses were 50% and 24% respectively. An "informal" statement of the statistical requirements of ARDS Network trials has been provided by one of the co-authors of the initial ARDS Network trial (Schoenfeld 2001), who noted that the clinical consequences of such trial designs would be to reduce the cost of long-term drug development (by early stopping of futile trials) and guard against harm to patients. This would seem eminently suitable for trials in critically patients, as opposed to say, cardiovascular / cancer trials, where long term outcomes are of substantive interest (Peto *et al.* 1976), albeit recent recommendations for critically ill patients enrolled in trials to be followed for $\geq$ 90 days (Cohen *et al.* 2001).

3.2.3 The paradigm encompassing the randomized clinical trial has developed over a relatively short period of time (Friedman *et al.* 1998; Meldrum 2000). Notable watersheds were:

3.2.3.1 the impact of randomized allocation (Pocock 1979). Although randomization was introduced into agricultural science by RA Fisher in 1926, it was not

formally adopted into medical trials until the 1930's by Amberson and then, in the 1940's with the work of Hill . It has been suggested that the early history of experimental design may have been different if RA Fisher (Box 1976) had not been initially employed in agricultural research, where experiments are essentially non-sequential in nature (Armitage 1991).

3.2.3.2 Sequential analysis. The sequential nature of medical and industrial experimentation leads to the accumulation of data over time and the formulation of appropriate statistical approaches to sequential or interim analysis of such data, arising from the early work of Wald in the industrial sphere (McPherson 1990; Wald 1947), has been crucial (Armitage 1991; DeMets *et al.* 1984b; DeMets 1984).

3.2.3.3 Data Monitoring Committees (DMCs). The progressive integration of DMCs into the trial scenario has complemented the developing statistical approaches, to the extent that decisions regarding the termination or otherwise of trials are the result of a constellation of factors, statistical and other (Califf *et al.* 2001; Ellenberg *et al.* 2002; Greenberg Report 1988). This is exemplified by subsequent published reports of the deliberations of the DMCs in two early pivotal randomized controlled trials in the 1970's, the University Group Diabetes Project (UGDP) (Kolata 1979) and the Coronary Drug Project (CDP) (The Coronary Drug Project Research Group 1981) and, of more interest to critical care practitioners, the 1987 Department of Veterans Affairs Cooperative study of steroid therapy for systemic sepsis (Peduzzi 1991).

3.2.3.4 An understanding of "futility", as it applies to clinical trials, requires consideration of both sequential analysis and design in clinical trials and the role of the DMC in the context of the question: when to stop a clinical trial?

(DeMets *et al.* 1999; Pocock 1992). Some attention has been directed to these questions in the Intensive Care literature, but the focus has been more general (Hebert *et al.* 2002).

3.2.4 Definitions

3.2.4.1 If there are two groups of patients (*n* in each arm) in a trial, allocated to treatments A and B, then the interest is in testing the null hypothesis of no treatment difference H_0: $\mu_A = \mu_B$, where μ = the mean response (continuous or binary), against the (two-sided) alternative $\mu_A \neq \mu_B$, with a Type I and Type II error probability of α and β respectively (power being defined as [1-β]). These error probabilities are defined at a particular value of $(\mu_A - \mu_B) = \pm \delta$, where δ = treatment difference or effect. The standardized *Z* statistic can be used to assesses this difference and it can be further shown that.

$Z \sim N((\mu_A - \mu_B)\sqrt{(n/(2\sigma^2))}, 1)$, where *n* is total number, σ^2 = variance (and 1= SD of a standardized variable)

the necessary (fixed) sample size (per arm) is

$n_f(\alpha, \beta, \delta, \sigma^2) = (\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta))^2 2\sigma^2 / \delta^2$, where Φ is the standard normal cumulative distribution function, α is the Type I error and β is the Type II error (Jennison *et al.* 2000b). We would be further interested in specific rules or tests to stop the trial at an early stage for (a) efficacy, or in a worst case scenario, effect reversal and (b) futility, which we will define as the conditional probability that a clinical trial will fail to demonstrate the superiority of a new therapy given the accumulated data and projected sample size of the study.

3.2.5 Controversy

3.2.5.1 The question of one- or two-sided (tailed) tests in clinical trials has engendered, perhaps not surprisingly, much controversy. One-sided tests have greater power for rejecting the null hypothesis when the one-sided alternative applies, but for alternatives on the opposite side, there is less sensitivity. The larger sample size required of a two-sided test ("modest" increase only for some(Moye *et al.* 2002)) for the same power as a one-sided test in the same direction, is obviously offset by the power for the alternative in the opposite direction.

For a one-sided test assessing H_0: μ_A-μ_B = δ = 0 vs H_s: $\mu_A - \mu_B = \delta > 0$, the power is:

$1-\Phi(Z_{1-\alpha} - (\delta/\sigma_d))$, where σ_d is the pooled variance

The two-sided power is:

$[1-\Phi(Z_{1-(\alpha/2)} - (\delta/\sigma_d)\} + \Phi\{Z_{\alpha/2} - (\delta/\sigma_d))]$

3.2.5.2 Arguments for one-sided tests (at both the 0.05 and 0.025 level) in mainly drug-placebo trials, where the alternative hypothesis being tested is one-sided, have been advanced (Fisher 1991; Knottnerus *et al.* 2001; Overall 1990), but also vigorously contested (Fleiss 1988; Moye *et al.* 2002), especially in the light of the experiences of The Cardiac Arrhythmia Suppression Trial (CAST) trial. The CAST trial, in the belief that antiarrhythmics were (only) beneficial, was designed as a one-sided trial, albeit at the conservative 0.025 level, to assess "beneficial or .. no beneficial effect….(it)..was not designed to prove that an antiarrhythmic drug could cause harm" (Cast Investigators 1989), which in fact occurred, much to the surprise of the investigators. Currently, in line with conservative regulatory requirements (for example, the FDA), most trials are conducted with two-sided tests.

3.2.6 Sequential analysis

3.2.6.1 The effect of repeated "looks" at accumulating data, in terms of inflation of the Type I error has been known for some time (Armitage *et al.* 1969), but needs to be re-iterated. For a pre-defined number of data inspections (say, 5) a nominal *p* value level of ≤ 0.0159 must be obtained in any of the 5 tests for the results to be "truly" significant at ≤ 0.05 (McPherson 1974). The same may be said somewhat differently: a trial will continue if the test statistic (Fleming *et al.* 1984a) does not exceed some critical value (for example, the standardized *Z* statistic with an α level = 0.05, is 1.96*)*. If this critical value was used for each inspection of accumulating trial data (interim analyses), the probability of stopping would be 0.05 for the first test, 0.14 for the 5^{th} and 0.19 for the 10^{th} (DeMets *et al.* 1984b).

3.2.6.2 That this is a real concern was convincingly demonstrated by the profile of the *z* statistic over the course of the Coronary Drug Project Trial: the magnitude of the statistic fluctuated markedly between efficacy ($z < -2$) and null effect ($z \equiv 0$), but the final mortality curves (clofibrate- versus placebo-treated patients) were almost identical (Friedman *et al.* 1998; The Coronary Drug Project Research Group 1981). The multiple and diverse patient monitoring requirements in clinical trials (detection of treatment trends and toxicities, minimization of patient numbers, ethical concerns) mandate interim analyses of trial data and a sophisticated statistical analytic apparatus, incorporating at its core the above insights, has been developed.

3.2.7 Sequential design

3.2.7.1 Classical sequential designs, introduced by Wald (Wald 1947), used the likelihood ratio statistic (as opposed to the more familiar standardized *Z*

statistic), but could not prescribe exact sample sizes, merely that the experiment would stop ("open plan") (DeMets *et al.* 1984b; Friedman *et al.* 1998). The extension of "fully" sequential designs to the medical arena, initially by Armitage (Armitage 1975) ("closed plan"), required effective continual assessment of patients (or rather, patient pairs) and the test statistic (the size of which reflected the treatment effect magnitude) used to estimate this treatment effect was recalculated after each new outcome (hence increased as the trial progresses) and compared with certain criteria to control Type I error (DeMets 1998). Continual patient assessment is obviously burdensome and for the most part impracticable, except in single institution studies (Ware *et al.* 1985), and fully sequential designs have seen limited application. A notable exception was the relatively small (N = 196) MADIT trial (implanted defibrillator for high risk ventricular arrhythmias) which used weekly assessments after the first 10 deaths (Moss *et al.* 1996). The test statistic (log-rank (Mantel 1966)) was used to terminate the trial when it "crossed" one of the preset termination boundaries (efficacy (+ve value of the log-rank statistic), inefficacy (-ve value of the log-rank statistic) or null effect (values close to zero)) as determined by a two-sided triangular sequential design, proposed by Whitehead (Whitehead 1992). Similarly, the trial of prolonged methylprednisolone in unresolving acute respiratory distress syndrome (Meduri *et al.* 1998), a 4 institution single city study, used a one-sided triangular test at the 0.05 level (power 0.95) to demonstrate the superiority of drug to placebo. The boundaries for the triangular test are the score test statistic $S_k = Z_k\sqrt{I_k}$, where Z is the standardized statistic and I is the accrued information (see below).

3.2.8 Group sequential design

3.2.8.1 The seeming impracticality of the fully sequential design led to the development, initially by Pocock (Pocock 1977) and O'Brien & Fleming (O'Brien *et al.* 1979), of group sequential designs where data is analysed at intervals; the number of analyses being a function of factors such as anticipated sample size and recruitment rates, but in practice there appears little to be gained from more than 5 interim analyses (McPherson 1982; McPherson 1990). A maximum number (K) of patient blocks is determined: for example, in a 2 armed trial that would enroll 400 patients total (200 in each arm) according to a fixed sample scenario, it may be decided to have 4 groups of 100 patients and interim analyses (of the appropriate outcome(s)) would then occur at 100, 200 300 and 400 patients. The standardized statistic Z_k is computed over the k blocks as data accumulates; rejection of H_0 and trial termination occurs if $| Z_k |$ > critical value (c), there being a sequence of critical values (c_1,…., c_K) for each particular design; if $| Z_k | < c$, continuation of the trial occurs. Type I and II error probabilities are preserved under repeated testing.

3.2.8.2 The approach adopted by Pocock (Pocock 1977) was essentially one of a repeated significance test with a constant nominal significance level α', such that, for a given number of analyses (of *equal* patient number), the overall trial significance level will be α. So, for K = 5 analyses and α = 0.05 (two-sided), Z'_K = 2.413. The total sample size (assuming no stopping) needed for Pocock boundaries is in excess of the fixed sample design and is a function of a constant, R_p, which is defined for particular values of the parameters K, α and β; formally, R_p(K,α,β). That is, total $N = (N_{\text{fixed sample}} \times R_p)$. At a power (1-β) =

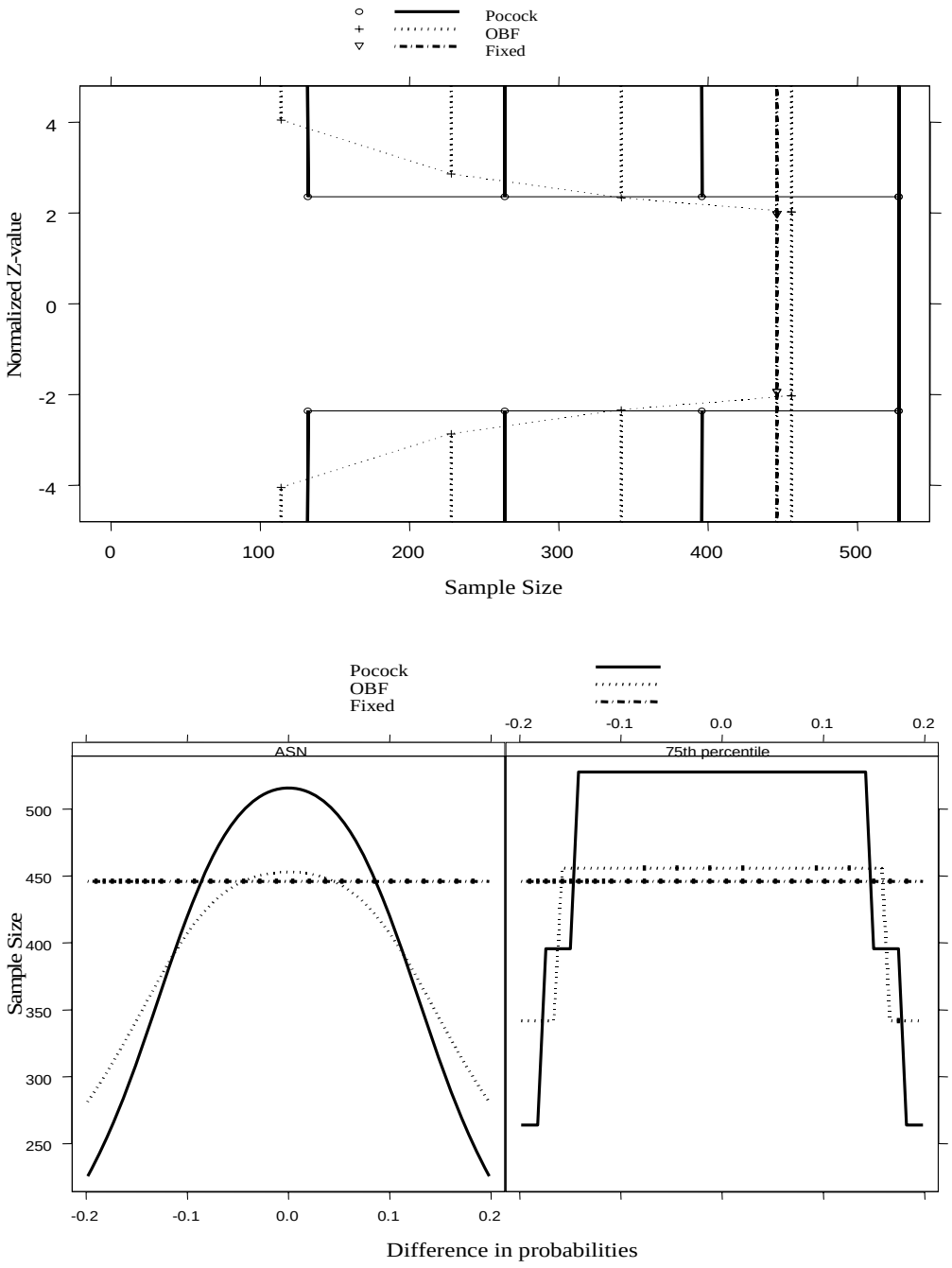
0.8 and (two-sided) α = 0.05, R_p = 1, for K = 1; R_p = 1.137 for K = 3 and R_p = 1.187 for K = 5 (Jennison *et al.* 2000b).

3.2.8.3 The O'Brien-Fleming (O'Brien *et al.* 1979) boundaries are such that the nominal significance level (for rejection of H_0) increases with data accumulation. Thus rejection is difficult early in the trial and becomes progressively easier; the critical value of the last test (K) is approximately the same as if a single test were done. The standardized statistic is computed as: $c_k = C_B(K, \alpha) \sqrt{(K/k)}$, where the constant $C_B(K, \alpha)$ ensures overall Type I error probability. Total sample size requirements for O'Brien-Fleming boundaries are less than for Pocock, but still somewhat greater (again, by a function R_B) than for a fixed sample design, again assuming no early stopping: at a power (1-β) = 0.8 and (two-sided) α = 0.05, R_B = 1, for K = 1; R_B = 1.007 for K = 3 and R_B = 1.015 for K = 5 (Jennison *et al.* 2000b). For both the Pocock and O'Brien-Fleming designs, given that there is a possibility to stop (increased at early analyses for the Pocock compared with the O'Brien-Fleming design), the expected number of patients (average sample number, ASN = ∑(probability of significance at jth test)×(N_j)), where N is the total sample size, is less than that of a fixed sample design, especially for large treatment effects (δ), although ASN for the Pocock design is uniformly greater than for O'Brien-Fleming.

3.2.8.4 We see the classic configuration of both these designs, using the normalized Z statistic scale, in Figure 3.2.8.4, for a two sided trial. The sequential design, 4 interim analyses and stopping only for the alternative hypothesis, has been applied to a trial postulating a reduction in mortality from 50% to 35% (-15%) with α = 0.05 and power (1-β) = 0.9. The critical boundary (Z) values of the two designs at each analysis are indicated by small circles (Pocock) and small

plus (O'Brien-Fleming) at the end of the vertical lines on the graph. Although the boundaries are discontinuous, they have been joined by horizontal solid lines (for the Pocock design) and sloping dotted lines for the O’Brien design to aid in visualization. Efficacy (a reduction in mortality) would be accompanied by a negative value of the *Z* statistic and the trial would (potentially) stop at any of the interim analyses

Figure 3.2.8.4 O'Brien-Fleming and Pocock boundaries and ASN for a two-sided trial: stopping for the alternative hypothesis only



Upper panel. Pocock and O'Brien-Fleming group sequential boundaries for the designed trial. Horizontal axis: sample size. Vertical axis: normalized Z statistic. Interim analyses are indicated by vertical lines (solid, Pocock design; dot, O'Brien-Fleming) and critical boundary values of Z for each interim analysis are indicated by the small circle (Pocock) and small plus (O'Brien-Fleming). Boundaries are discontinuous, but have been joined to aid in visualization. Vertical dash-dot line indicates the sample size for a fixed design of same power. OBF = O'Brien-Fleming. Fixed = fixed sample size for same power.

Lower panel.. Left insert: ASN is seen for both the Pocock and O'Brien-Fleming designs. Vertical axis: sample size; Horizontal axis: sample mean scale of difference in probabilities of the treatment effect (efficacy associated with –ve probabilities). Right insert: Sample efficiency for Pocock and O'Brien-Fleming designs at 75th percentile of the sample size distribution (over many possible values of the true treatment effect)

if the Z statistic were less than these values. Thus the (trial) continuation regions are the "white-space" areas outside the boundaries. As seen in the upper panel, stopping is possible earlier with the Pocock design but at the cost of inflation of the sample size. The sample size for a fixed design of same power is indicated by a vertical dash-dot line. In the lower panel, the ASN is seen to be less for O'Brien-Fleming design (left insert), and the 75th percentile of the sample size distribution (over many possible values of the true treatment effect) is also less (right insert).

3.2.8.5 Wang and Tsiatis (Wang *et al.* 1987) proposed a "power" family (Jennison *et al.* 2000b) of two-sided tests, indexed by a parameter Δ that determines the shape of the continuation region; $\Delta = 1$ is equivalent to the fixed sample size scenario, high values of Δ entail higher probabilities of early stopping, and low (0.3, 0.4) ensure minimal ASN. The standardized statistic is computed as: $c_k = C_{WT}(K, \alpha, \Delta) (k/K)^{\Delta-1/2}$ and the Pocock and O'Brien-Fleming designs are special cases. If $\Delta = 0.5$, the Pocock design results; $\Delta = 0$ gives the O'Brien-Fleming design. Unequal group sizes may be also accommodated, with the O'Brien-Fleming and Wang and Tsiatis (with low values of Δ) designs being robust to early group size variation. For trials comparing outcomes with survival analytic methods, the log-rank or a generalization of the Wilcoxon test are appropriate statistics, although interim analyses are done after certain numbers of events have occurred, rather than patients enrolled; again unequal event number may be tolerated (DeMets *et al.* 1985).

3.2.8.6 An alternative less formal approach had been introduced by Haybittle (Haybittle 1971) and Peto (Peto *et al.* 1976) who suggested conservative criteria at each interim analyses ($z = 3.09$: that is, for k analyses, setting $\alpha_1 = \alpha_2$

.. = $\alpha_{k-1} \equiv 0.001$). At the final ($k^{th}$) analysis, $\alpha_k = 0.05$ and the experimental error is nearly 0.05 (Fleming *et al.* 1984b); that is, providing the number of looks is not great, a fixed sample analysis can be undertaken at the final stage with no allowance for interim analyses (Jennison *et al.* 2000b). Any inflation of the α error may be avoided by application of the Bonferroni inequality to the final stage (Proschan 1999), such that, for example, after 5 interim analyses (at $p \leq 0.001$), the final p value, for a one-sided test $\alpha = 0.025$, would be 0.02 (*z* boundary of 2.05). A formal application of this strategy using the exact distribution of the test statistic has also been described (Fleming *et al.* 1984b). A conceptual problem with this type of rule is that a *z* value of 2.9 at the *k*-1 analysis does not suggest early termination, but is striking at the k^{th} analysis (DeMets *et al.* 1984b). Haybittle-Peto boundaries were used in the European Myocardial Infarct Amiodarone (EMAIT) Trial (Julian *et al.* 1997) and was used for the Australian and New Zealand Intensive Care Society (ANZICS) sponsored SAFE (saline versus albumin fluid evaluation) study (The SAFE Study Investigators 2004).

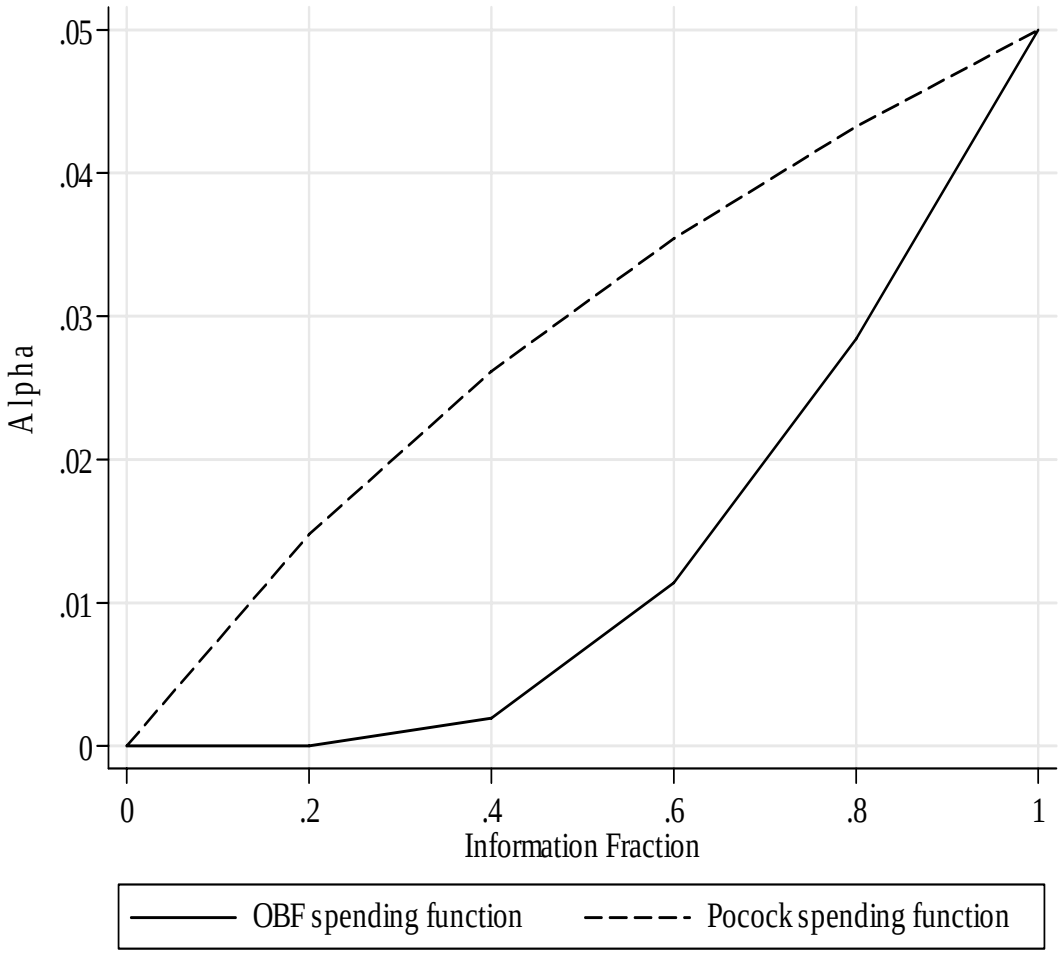
3.2.8.7 The nominal requirement of group sequential methods to specify equal numbers of patients or events (for survival analyses) in advance was formally overcome by the α spending function of Lan & DeMets (DeMets *et al.* 1995; Lan *et al.* 1983), which produced flexible discrete boundaries by the specification of an increasing function $\alpha(t^*)$ which characterizes the rate at which the α error is "spent" over the interim analyses. The exact number and / or timing of interim analyses need not be specified in advance, although the total sample size does. For any calendar time *t*, a certain fraction of the total information (where "information" may be defined as the total number of

observed patients or events accrued during a trial; see also Figure 2 in (Gillen *et al.* 2003)) t^* is observed ($0< t^* <1$, is given as the ratio of the inverse of the variance of the test statistic at any interim analysis and the final analysis). At trial beginning, $t^* = 0$ and $\alpha(t^*) = 0$; at trial end $t^* = 1$ and $\alpha(t^*) = \alpha$. Information fractions (of patients or events) may be variously defined (for example, 0.2, 0.4, 0.6, 0.8, 1.0) and critical test statistic values can be appropriately computed, using the techniques of Armitage *et al* (Armitage *et al.* 1969). The specification of "information" may be problematic and trials have been categorized as:

3.2.8.7.1 maximal information: maximal information is prescribed and trial termination is at either boundary crossing or maximum information; or maximal duration: maximum trial duration is pre-specified and (maximum) information is estimated, based upon trial design; current information is estimated on empiric grounds (current interim analysis) or on calendar time fractions (DeMets *et al.* 1995; Spiessens *et al.* 2000).

3.2.8.7.2 The approximate O'Brien-Fleming spending function is defined as $\alpha_1(t^*) = 2-2\Phi(Z_{\alpha/2} / \sqrt{t^*})$ and the Pocock, $\alpha_2(t^*) = \alpha \times \ln(1 + (e-1)t^*)$; other spending functions have also been described (Friedman *et al.* 1998). The difference in the rate of "spending" of the α (Type I error) over information fractions (0 through 1) for the two different functions, $\alpha_1(t^*)$ and $\alpha_2(t^*)$, is seen clearly in Figure 3.8.7.2 for a total (two sided) $\alpha = 0.05$. As mentioned above, the PROWESS study (Efficacy and Safety of Recombinant Human Activated Protein C for Severe Sepsis) utilized "the O'Brien-Fleming spending function according to the method of Lan and DeMets" (Bernard *et al.* 2001).

Figure 3.8.7.2. O'Brien-Fleming (OBF) and Pocock alpha spending functions

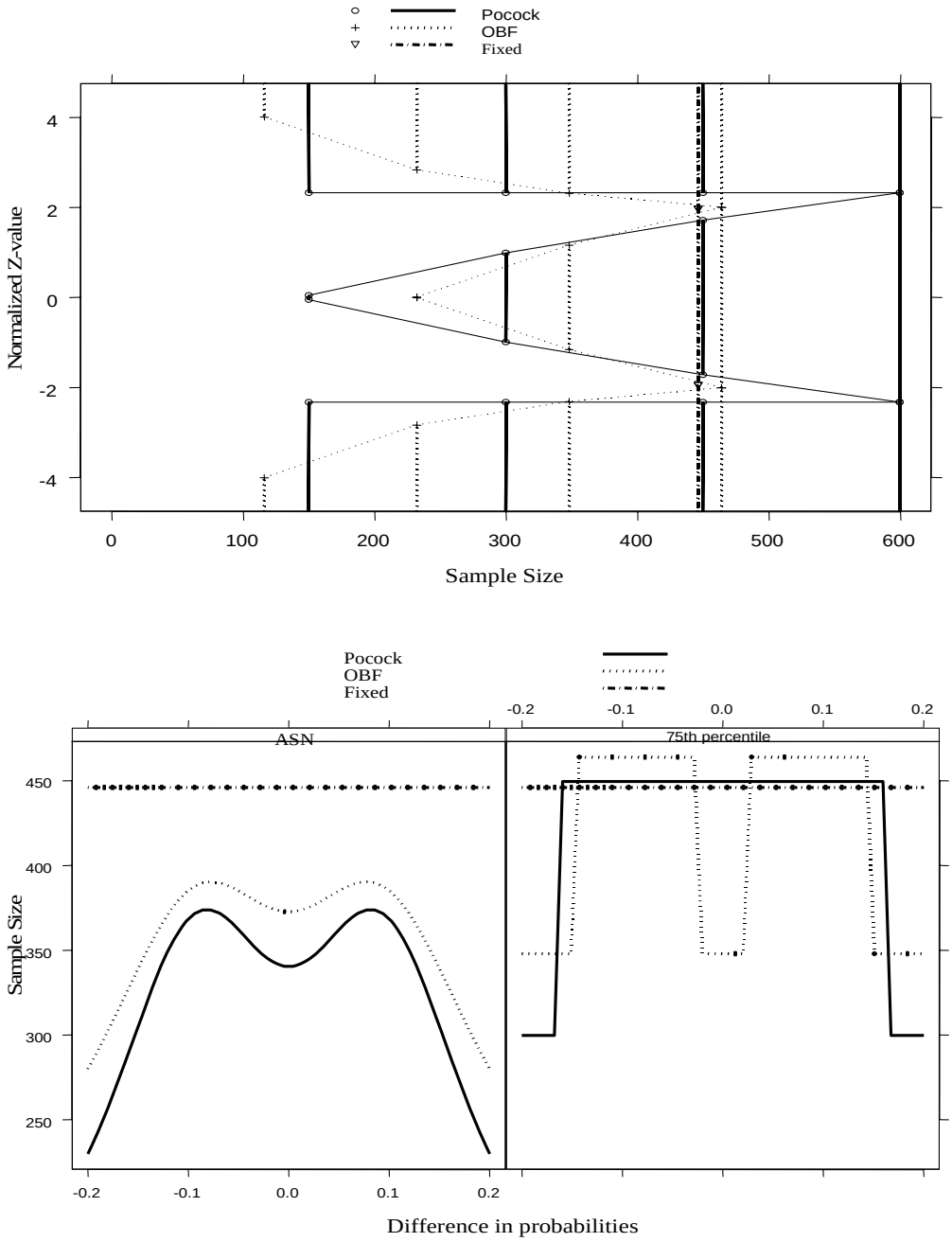


Vertical axis: Type I error (α). Horizontal axis: information fraction

3.2.8.8 In classical group sequential designs, originally described as two-sided tests, if the null hypothesis was in fact not rejected, then patient enrollment would occur until the final (K^{th}) analysis with probability of at least 1-α to satisfy Type I error constraints. Further refinement of group sequential design saw the introduction of one-sided tests, both one and two sided tests for early stopping under the null (Emerson *et al.* 1989; Jennison *et al.* 2000b; Overall *et al.* 1993; Pampallona *et al.* 1994; Pampallona *et al.* 2001; Spiessens *et al.* 2000) and

asymmetric boundaries (Califf *et al.* 2001; DeMets *et al.* 1999; Friedman *et al.* 1998). Figure 3.2.8.8 displays, in the upper panel, a two-sided design with stopping for both the null and the alternative for the same trial conditions as above in Figure 1, using both the O'Brien-Fleming and Pocock boundaries. The stopping boundaries for the null are seen as the so-called "inner-wedge" (Jennison *et al.* 2000b); stopping for the null occurs earlier (potentially) for the Pocock design. This being so, it is perhaps not surprising to see in the lower panel, that the ASN for both the Pocock and O'Brien-Fleming design is less than the fixed sample size and that the ASN for the Pocock is the less than the O'Brien-Fleming (left insert). The 75^{th} percentile of the sample distribution is also greater for the O'Brien-Fleming design (right insert).

Figure 3.2.8.8. O'Brien-Fleming and Pocock boundaries for a two-sided trial: stopping for the alternative and null hypotheses



Upper panel. Pocock and O'Brien-Fleming (OBF) group sequential boundaries for the designed trial. Horizontal axis: sample size. Vertical axis: normalized Z statistic. Interim analyses are indicated by vertical lines (solid, Pocock design; dot, O'Brien-Fleming) and critical boundary values of Z for each interim analysis are indicated by the small circle (Pocock) and small plus sign (O'Brien-Fleming). Boundaries are discontinuous, but have been joined to aid in visualization. Fixed sample size for equal power ("Fixed") is seen by the vertical dash-dot line. Null boundaries are seen as the "inner wedge".

Lower panel. Left insert: ASN for Pocock (solid line) and O'Brien-Fleming (OBF = dotted line) and fixed sample size seen as horizontal dash-dot line. Right insert: sample size for the 75th percentile of the sample size distribution (over many possible values of the true treatment effect).

3.2.9 Stochastic curtailment

3.2.9.1 The progressive review of trial data may be analyzed for a trend (positive, negative or null) by estimating what is termed the Conditional Power or the probability that, given current information, the trial will yield at endpoint a "significant" result (Betensky R.A. 2000). Two well known instances of this, in addition to the ALVEOLI trial , mentioned above (The NHLBI ARDS Clinical Trials Network 2004), were the Beta-Blocker Heart Attack Trial (BHAT), for positive effect (DeMets *et al.* 1984a), and a trial of prophylactic barbiturate coma in head injury for the null (Choi *et al.* 1985; Ward *et al.* 1985). If the conditional power under the *alternate* hypothesis is say, ≤ 0.15, given the information at the interim analysis (for a two sided test, minimum recommended $t = 0.64$), then consideration should be given to terminating the trial for futility (Davis *et al.* 1994; Ware *et al.* 1985). This stopping may result in a loss of power, but the loss is not substantial. If a trial is designed to have a (unconditional) power of 90%, the Type II error (β) = 0.1; if with curtailment, the (conditional) power is computed as 0.15 and the trial is stopped, the overall ("true") Type II error probability has found to be $\beta / \gamma = 0.1/(1-0.15) = 0.12$, (where γ =1- conditional power) (Lan *et al.* 1982). Similarly, the overall Type I error is given by α / γ. Curtailment boundaries may be generated and it is of interest that the O'Brien-Fleming and the 50% Conditional Power boundaries are coincident. In the presence of low unconditional (that is, at initial design stage) power and, for time-to-event outcomes with non-proportional hazards, aggressive futility monitoring and consequent early stopping may be difficult to sustain (Freidlin *et al.* 2002). The place of stochastic curtailment in group sequential clinical trials has been recently re-evaluated and found to be

problematic (Emerson *et al.* 2005b). A second approach to stochastic curtailment, Predictive futility, has a Bayesian flavor (Emerson *et al.* 2005a); a prior distribution for the treatment effect is specified and, given the data, the posterior distribution is then computed (Spiegelhalter *et al.* 1986).

3.2.10 Inference

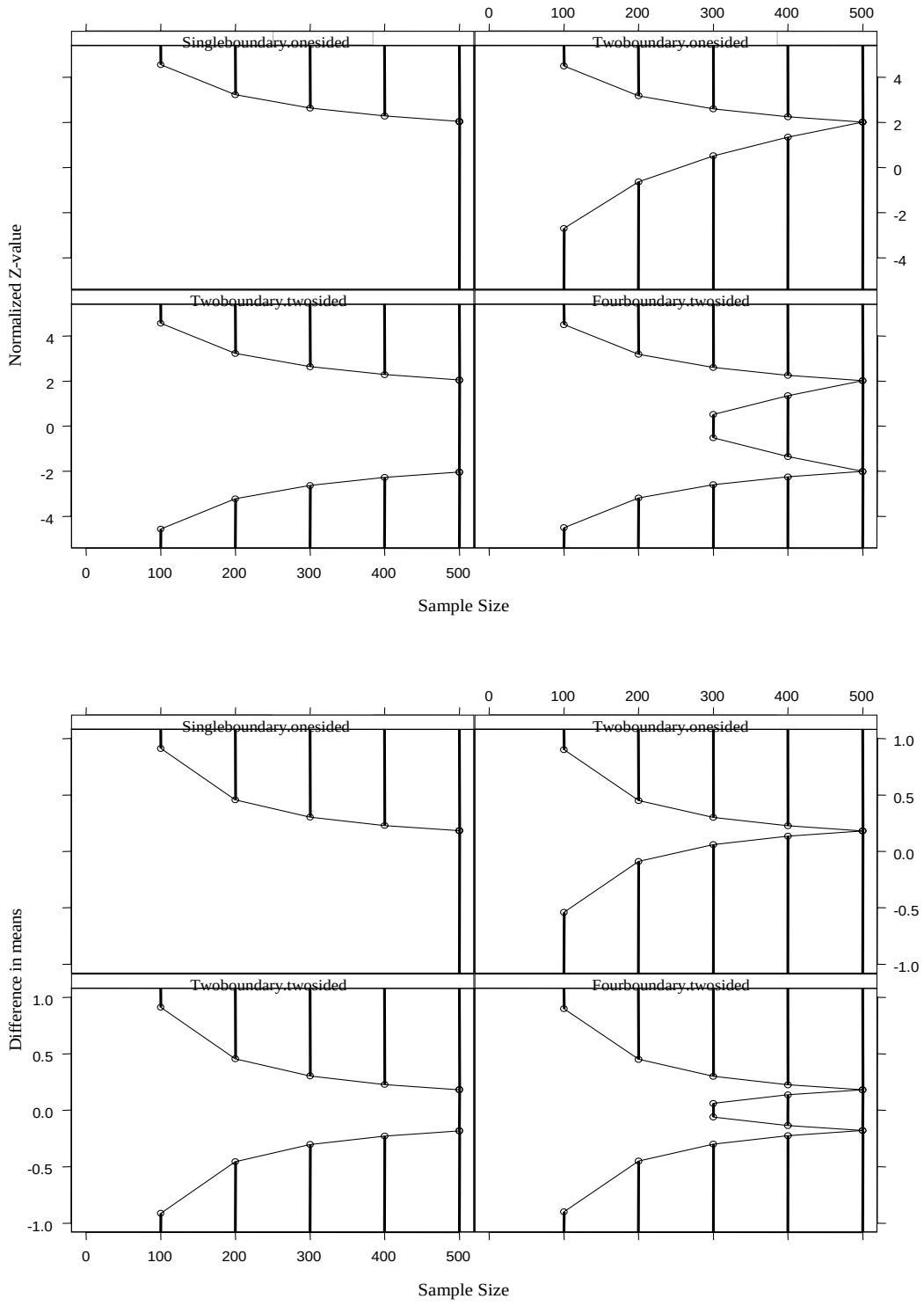
3.2.10.1 It is well known that early stopping of a trial introduces bias into the usual point estimates and confidence intervals of the various maximum likelihood estimators of treatment effect (for example, risk ratio, odds ratio). A number of revised estimators have been described and the bias adjusted mean would appear to have advantageous properties (Emerson *et al.* 1990). Suffice it to say that the reason for this bias is that the sampling distribution of the estimator, under the conditions of interim analysis, has an asymmetric jagged profile, quite unlike the smooth profile of standard distributions (Emerson 2000; Jennison *et al.* 2000b; Pinheiro *et al.* 1997). We have commented upon both this, and more importantly, the lack of use of these adjusted estimators in trial reports appearing in medical journals, above.

3.2.11 A unified approach

3.2.11.1 The above seeming plethora of group sequential designs was simplified with the introduction of the “Unified Family” by Kittelson and Emerson (Kittelson *et al.* 1999), resulting in a hybrid design incorporating aspects of both equivalence and superiority test designs. The particular insight was that the various methods involving seemingly different boundary functions (standardized *Z* statistic, partial sum statistic (used in the triangular test), alpha spending function, maximum likelihood estimate of treatment effect (sample mean scale), stochastic curtailment (conditional and predictive power)) are

transformations of each other (Emerson *et al.* 2000; Insightful Corporation 2002). The unified family, described initially on the sample mean scale which is invariant to hypothesis shifts, includes the above *formal* designs and allows a continuum between the four basic boundary designs, as seen in Figure 3.2.11.1, which we generate by adapting code given in the S+SeqTrial2 User's guide (Insightful Corporation 2002). The top panel uses the normalized Z scale and the bottom, the sample mean scale; again negative vales of the Z statistic and negative mean differences are associated with efficacy; O,Brien-Fleming boundaries are used. In clock-wise direction (repeated in each panel): a single boundary design for a one-sided test (early stopping against the null, or non rejection of the null); two boundary design for a one-sided test (early stopping for or against the null); four boundary design for two-sided hypothesis (early stopping for or against the null); and two boundary design for a two-sided test (early stopping against the null, but decision for the null only at final analysis). This innovation allows assessment of a spectrum of designs and "connection" between distinct families.

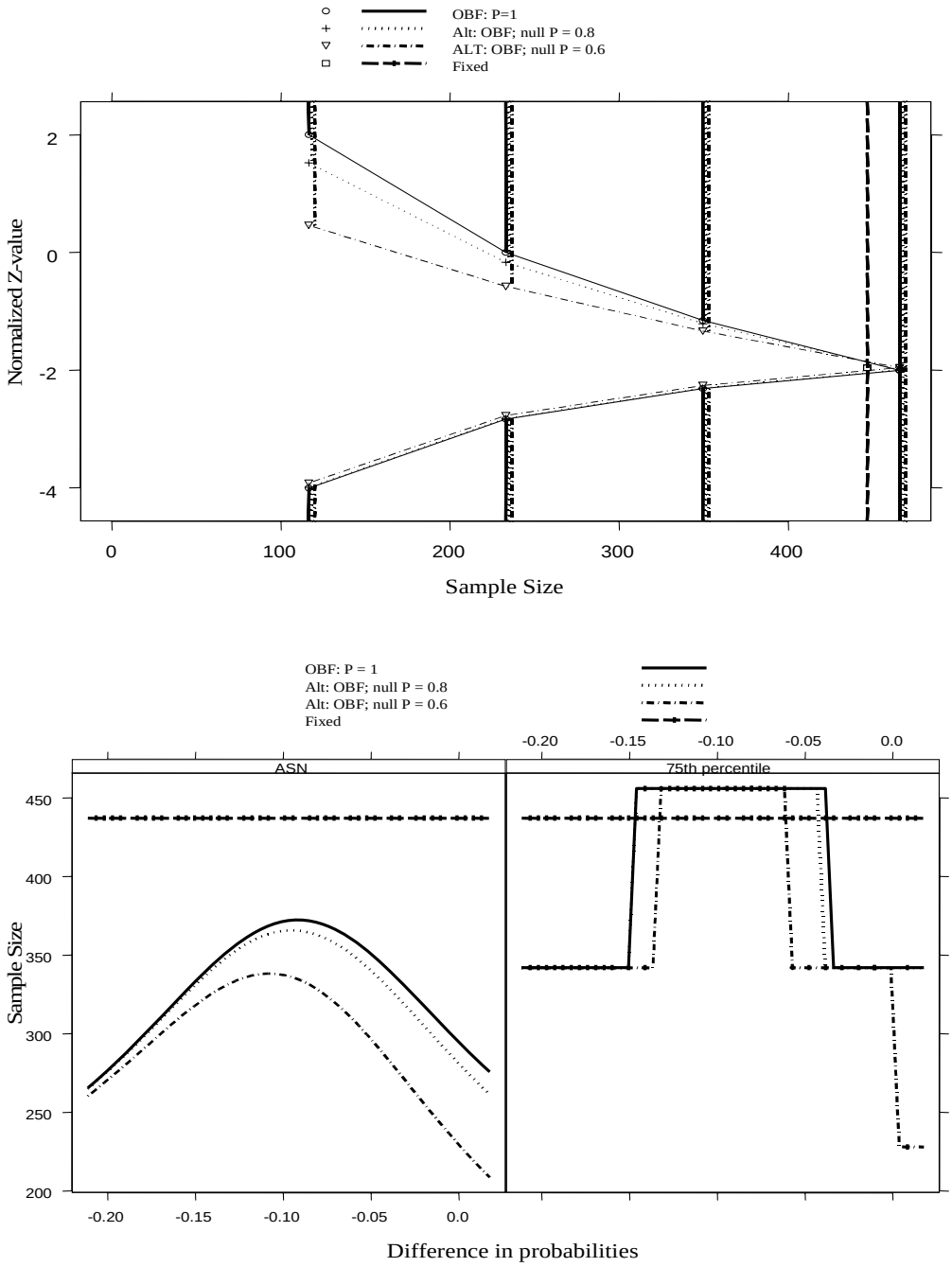
Figure 3.2.11.1. Four basic designs of the “unified family”



Top panel: O’Brien-Fleming boundaries for the 4 basic boundary designs (see Text). Horizontal axis: sample size. Vertical axis: normalized Z statistic. Bottom panel: Same design using the sample mean scale. Horizontal axis: sample size. Vertical axis: Difference in means. Efficacy is associated with negative values of the Z statistic and difference in means

3.2.11.2 Figure 3.2.11.2 shows the same trial as considered as in Figures 3.2.8.4 and 3.2.8.8, but with a one-sided test and early stopping for both the null and alternate hypothesis. The null boundaries (upper panel) have been progressively changed to allow earlier stopping (for futility) by manipulation of the Δ parameter in the Wang-Tsiatis power family, from a standard O'Brien-Fleming boundary through boundaries approaching the Pocock. That this is potentially advantageous is seen in the lower panel where ASN for the various designs is plotted; the design "Alt: OBF; null P = 0.6" which has a Δ value in the Wang & Tsiatis "power series" approaching the Pocock design has obvious reduced sample size as the probability difference approaches the null.

Figure 3.2.11.2. One-sided test with early stopping possible for both the null and alternate hypothesis.



Upper panel: Horizontal axis: sample size. Vertical axis: normalized Z statistic. Efficacy boundaries are the lower (O,Brien-Fleming) boundaries, which are co-incident. Null boundaries are the upper boundaries which are three in total: O,Brien-Fleming P = 1 (solid line), "null P = 0.8" (dotted line) and "null P = 0.6" (dash-dot line). The "P" values are specific to the implementation in S+SeqTrial2, but indicate progressive movement away from O,Brien-Fleming boundaries (P = 1) to Pocock boundaries (P = 0.5). OBF = O'Brien-Fleming. Alt = Alternative hypothesis (efficacy in a one-sided test). Vertical long-dashed line indicates the fixed-sample ("Fixed") size for equal power. Bottom panel. Left insert: ASN for three one-sided designs in the Upper panel. O,Brien-Fleming P = 1 (solid line), "null P = 0.8" (dotted line) and "null P = 0.6" (dash-dot line). Right insert: sample size for the 75th percentile of the sample size distribution (over many possible values of the true treatment effect). Vertical axis: sample size. Horizontal axis: difference in probabilities (null effect indicated by low or zero difference).

3.2.12 It was mentioned above (Stochastic curtailment 3.2.11) that the O'Brien-Fleming design and the 50% Conditional Power boundaries were coincident. Using the facility of the Unified Family, we are able to change scale and see this for the initial O'Brien-Fleming design for the two-sided trial illustrated in Figure 3.2.10.5. The lower and upper boundaries (there is no stopping for the null) are "a" and "d" respectively (Table 3.2.13.4) conditional probability (*computed at the particular boundary*) is 0.5 that at the last analysis the estimated treatment effect would correspond to an *opposite* decision.

Table 3.2.12. Stopping boundaries: Conditional Probability scale

	a	b	c	d
Time 1 (N= 114)	0.5	NA	NA	0.5
Time 2 (N= 228)	0.5	NA	NA	0.5
Time 3 (N= 342)	0.5	NA	NA	0.5
Time 4 (N= 456)	0.5	0.5	0.5	0.5

3.2.12.1 Therefore stopping occurs if the conditional power is < 50%, a rather elevated level. However, if we assume that the treatment effect is the *current best estimate* (and here we follow the discussion in the S+SeqTrial2 Manual, Chapter 6), we can revisit the boundaries on this scale, as seen in Table 3.2.13.5.We now see, looking at the boundaries "a" and "b" for the early analyses, that the probability of a reverse decision at the final analysis is small, which is consistent with the known early conservatism of the O/Brien-Fleming design.

Table 3.2.13.5 Stopping boundaries: Conditional Probability scale, current best estimate

a	b	c	d		
Time 1 (N= 114)	0.0000	NA	NA	0.0000	
Time 2 (N= 228)	0.0021	NA	NA	0.0021	
Time 3 (N= 342)	0.0886	NA	NA	0.0886	
Time 4 (N= 456)	0.5000	0.5	0.5	0.5000	

3.2.12.2 Software implementation. The implementation of sequential designs has been discussed above and a formal review is extant in the literature (Horton *et al.* 2001).

3.2.13 Conclusions

3.2.13.1 The specific statistical approaches described above are an aid to the "real-world" concerns of trial conduct (Delgado-Herrera *et al.* 2003). The termination of trials, for benefit, harm or the null, is a complex procedure, as has been attested to numerous times. The ramifications of these decisions may have a life of their own, as witnessed by the continuing debate of the decisions of the DMC in the UGDP trial for over 10 years (Kolata 1979). The factors recommending the adoption of a particular method of interim analysis perhaps reduce to the specifications of the individual trial, but some general determinants can be offered: the type of patient (critically ill or otherwise) and disease (acute or chronic); the nature of the intervention (life saving or equivalence testing); the extent of follow-up (short term 28 day mortality or long term observation); the need to define the "history" of the treated disease and the secondary end-points / toxicities; the estimate of the treatment effect (Wheatley *et al.* 2003); and any pressing requirement for early recognition of toxicity or null effect (Emerson 1995). For the ARDS Network at least, stopping for the null effect would appear to have assumed a priority. Current

statistical developments in the design of randomized clinical trials allow a large degree of flexibility which will aid in the tailoring of appropriate designs for the individual clinical trial.

3.3 EQUIVALENCE TRIALS (Moran *et al.* 2003d)

3.3.1 The demonstration of treatment efficacy by a prospective randomized placebo controlled trial has, as noted above, become established in medical practice. In situations where effective therapy already exists, the introduction of newer therapeutic agents using placebo controlled trials is controversial (Djulbegovic *et al.* 2001; Rothman *et al.* 1994; Temple 1996) and comparisons with "standard therapy" are frequently undertaken using so-called equivalence or non-inferiority trials (Ellenberg *et al.* 2000; Temple *et al.* 2000). The use of such trials obviously presupposes the established efficacy of a therapy, but the formulation of a "standard therapy" in the critical care setting has been a somewhat difficult enterprise, as opposed to, say, the practice of cardiology. For instance, the Fibrinolytic Therapy Trialists' Collaborative Group reviewed reports of fibrinolytic and standard therapy for myocardial infarction (ST segment elevation and/or bundle-branch block with randomization within 6 hours of symptom onset) in 58600 patients and demonstrated an overall absolute 35-day mortality reduction of -1.84% (95% CI: -2.34% to -1.35%) (Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. 1994). Thus fibrinolytic therapy, in particular streptokinase, has become a standard therapy and it is "no longer ethical to withhold ...(such therapy)..from patients…"(White 1998).

3.3.2 The usual (placebo) controlled trial is a superiority trial, where the aim is to rule out treatment equality by rejection of the null hypothesis that the two treatments

are the same. However, the converse proposition does not hold; that the failure to reject the null hypothesis (the “negative” clinical trial) establishes equivalence (Altman *et al.* 1995; Freiman *et al.* 1978; Kim *et al.* 1999; Spriet *et al.* 1979). An illustration of this was the report of a clinical trial comparing trimethoprim-sulphamethoxazole and pentamidine in the treatment of *Pneumocystis carinii* pneumonia. Forty patients were enrolled and no difference was seen in 21-day mortality rates ($p = 0.18$) or other indices of improvement or of toxicities (Wharton *et al.* 1986). The trialists concluded that the two study treatment arms were “probably of equal effectiveness”. As Polis and Blackwelder noted, apropos the question of study sample size and β error, “With additional patients, this study may have contributed more toward the resolution of this issue… (therapy of *P. carinii* pneumonia)… . However, it does not demonstrate that trimethoprim-sulphamethoxazole and pentamidine are equally effective; failure to show a significant difference …..is not at all the same as showing equivalence” (Polis *et al.* 1987). The equivalence trial reverses the logic of the superiority trial; the null hypothesis is instead that of a *specified difference* (δ) between the experimental therapy and an active control (Blackwelder 1982; D'Agostino *et al.* 2003; Makuch *et al.* 1986). Thus if μ is the “true” treatment difference (experimental vs control therapy and μ is positive when the experimental is superior to standard therapy) (Aras 2001; Hwang *et al.* 1999)

3.3.2.1 in a superiority trial the test is $\mu = 0$ vs $\mu \neq 0$ at the 5% level (or rather, $\mu \leq 0$ vs $\mu \geq 0$ at the 2.5% level)

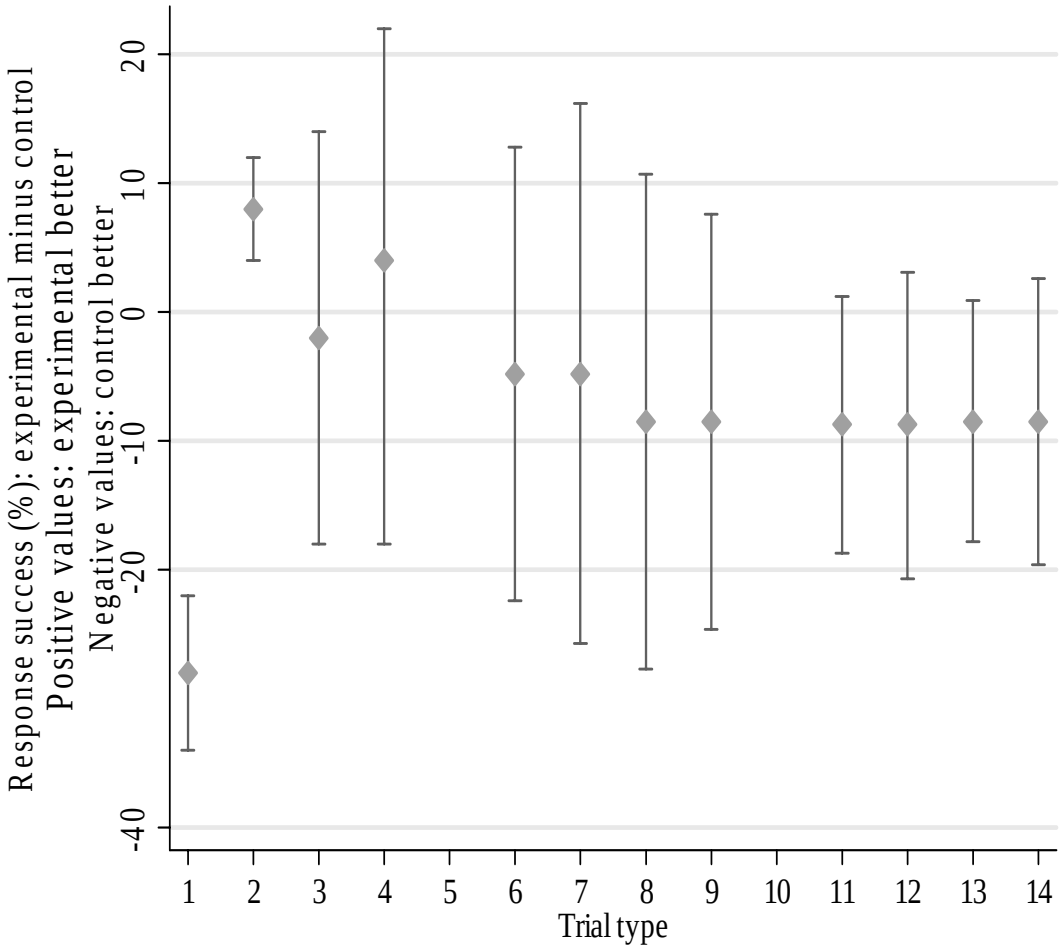
3.3.2.2 in an equivalence trial, the purpose is to demonstrate minimal differences (experimental therapy vs standard) in either direction. Therefore, $\mu \leq -\delta$ or $\mu \geq \delta$ is tested (two-sided) against $-\delta < \mu < \delta$ Alternatively, a pair of one-sided

hypotheses are tested: H_1 $\mu \leq -\delta$ vs $\mu > -\delta$ and H_2 $\mu \geq \delta$ vs $\mu < \delta$ (both one-sided hypotheses need to be rejected). True equivalence trials are usually bio-equivalence trials (Williams *et al.* 2002).

3.3.2.3 in a non-inferiority trial, the purpose is to demonstrate that the experimental therapy is not substantially worse than active-control. Therefore, $\mu \leq -\delta$ is tested against $\mu > -\delta$. The test is one-sided at α significance level (usually 0.05). Such testing may be subject to the known limitations of hypothesis testing in general (Makuch *et al.* 1978; Makuch *et al.* 1989). Alternatively, a 100(1-2α) percent two-sided confidence interval for the treatment difference is computed and if the lower bound of the CI is > -δ, non-inferiority can be claimed. Whether this be at the 90% or 95% is a point of some dispute, although recent regulatory recommendations suggest $\alpha = 0.025$ for one-sided testing of non-inferiority (Hauschke 2001). It is also noted that the strategy of using a two sided 90% CI for a one-sided 5% test assumes that the 90% CI is equal-tailed (each end of the interval excludes 5%) (Hauck *et al.* 1999). In the clinical literature equivalence is often used synonymously with non-inferiority and, unless specified otherwise, this review will conform to this practice.

3.3.2.4 In the Figure 3.3.2.4, trials 1-4, shows the above as hypothetical trials with point estimates and CI (95% for trials 1 & 2 and 90% CI for trials 3 & 4). In trial 1, placebo vs standard drug, the lower 95% CI approximates, in this scenario, the value of δ which is set at 20%. Trial 2 shows a successful superiority trial with lower 95% CI above zero. Trial 3 is an equivalence trial showing upper and lower 90% CI within $\pm\delta$. In trial 4, a non-inferiority trial, the lower 90% CI is > -δ (but the upper 95% CI is > +δ).

Figure 3.3.2.4. Graphic display of point estimates and CI of various trials



Vertical axis: absolute risk response success (%) as (experimental - standard), such that +ve values reflect efficacy of experimental drug and -ve values, efficacy of standard drug. Diamonds: point estimate with 90% CI, unless indicated. Horizontal axis: trial type. Trials 1-4, hypothetical examples: 1. Placebo / standard drug 2. Superiority trial (successful, 95% CI) 3. Equivalence trial (successful) 4. Non-inferiority trial (successful). Trials 6-9, Phillips *et al* (reference 65): 6. PP analysis 7. PP analysis with 95% CI .8 ITT analysis 9. ITT analysis, 95% CI. Trials 11-14 Rex *et al* (reference 67): 11. PP analysis 12. PP analysis with 95% CI 13. ITT analysis 14. ITT analysis 95% CI

3.3.3 Issues in equivalence trials (Jones *et al.* 1996; Kirshner 1991; Makuch *et al.* 1989; Windeler *et al.* 1996). The assumptions made when conducting equivalence trials are those of

3.3.3.1 *assay sensitivity*: that is that the active-control would have been superior to a placebo if such had been employed in the current trial. That is, an equivalence trial requires the consideration of "…information external to the trial." (McAlister *et al.* 2001)

3.3.3.2 *sensitivity to drug effects*: the ability of well designed trials to reliably demonstrate active-control drug effect (with respect to placebo). If the above two assumptions have not been met, then the interpretation of equivalence trials can be problematic. This was forcefully demonstrated by Tramer *et al* (Tramer *et al.* 1998) in their recent review of anti-emetics; in particular, the efficacy of ondansetron. They concluded that where no gold standard treatment existed and event rates (in this case, of emesis) varied widely "…trial designs without placebo controls are unlikely to yield sensible results"

3.3.3.3 *constancy assumption*, that the historical treatment difference is preserved in the current trial. This may be difficult to sustain given changes in medical practice and the effect of different patient populations.

3.3.3.4 Intention-to-treat (ITT) vs per-protocol (PP) analysis: In superiority trials ITT is the preferred analysis as compared with PP (Lewis *et al.* 1993). Such is not the case with equivalence trials where PP analysis is the more conservative and ITT tends to make treatment arms appear similar (D'Agostino *et al.* 2003; Hauck *et al.* 1999; Wiens 2001), although this will depend upon the pattern of patient drop out and treatment assignment. Both types of analyses should be

presented, but this strategy must take into consideration the reduced patient number in a PP analysis when initial sample size calculations are made.

3.3.3.5 Biocreep: whereby a slightly inferior treatment becomes the active control for the next generation of equivalence trials and active controls become little different from placebos (D'Agostino *et al.* 2003; Hauck *et al.* 1999; Shlaes *et al.* 2002).

3.3.3.6 Conduct of the trial: poor trial conduct in an equivalence trial will widen CI of the observed treatment effect and make the declaration of equivalence more difficult (Chuang-Stein 1999; Hauck *et al.* 1999; Jones *et al.* 1996) , whereas in a superiority trial there will be a tendency to a null result which may be mistakenly claimed as indicating equivalence.

3.3.3.7 Sample size requirements for equivalence studies are variably increased above similar superiority trials; on average about 10% (Djulbegovic *et al.* 2001). Formulas for such calculations are provided in numerous articles (Hwang *et al.* 1999; Jones *et al.* 1996; Makuch *et al.* 1980; Makuch *et al.* 1986) and specialized software is available (Elashoff 2003; Hintze 2002).

3.3.3.8 The determination of δ (or non-inferiority margin): this may be formally defined as the largest acceptable clinical difference in treatment efficacy (experimental therapy vs standard) or, in the reverse, as a difference in patient status with an effect size $\leq \delta$ that is non-detectable (Aras 2001). Thus, for example, δ is different from (and usually smaller (Makuch *et al.* 1978)) the difference in proportions ($\pi_1 - \pi_2$) or means between two treatments used in routine sample size calculations for superiority trials (Hatala *et al.* 1999; Makuch *et al.* 1986). Recommendations for the calculation of δ have been numerous and are found, not surprisingly, in the biopharmaceutical literature

(Hauschke 2001; Holmgren 1999; Ng 1993; Ng 1995), but the regulatory literature has been somewhat circumspect in prescribing δ a priori (D'Agostino *et al.* 2003; Wiens 2002). From a statistical perspective, δ has been defined as a certain fraction of (a) the treatment effect of control drug vs placebo (for example, 0.2 – 0.5) or (b) the lower 95% CI of this treatment effect (for example, 0.5) derived from a meta-analysis or large trial (Blackwelder 2002; Hauck *et al.* 1999; Ng 1993; Ng 2001). In the anti-infective drug testing domain, the Food and Drug Administration (FDA) in the USA had informally (1992) provided a so called step-down function of δ that reflected the observed response rates in the equivalence study. For response rates (one or both arms) of at least 90%, ≥ 80% but < 90%, and ≥70% but < 80%, δ was suggested to be 10, 15 and 20% respectively (Wiens *et al.* 2001). The recent removal of this step-down function (2001) and the use of a more conservative δ (unofficially 10% (D'Agostino *et al.* 2003)) has provoked comment regarding the unavoidable and large increment in trial size consequent upon this decision (Shlaes *et al.* 2002).

3.3.3.8.1 In cardiology, where standard care in the treatment of acute myocardial infarction has been effectively established, equivalence trials have used streptokinase as the standard and δ has been set at a much lower level. In the INJECT trial (Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. 1994), which compared reteplase and streptokinase using 35 day mortality as end-point, equivalence was established if the (upper) CI of mortality difference excluded the possibility that reteplase mortality was > 1% worse than streptokinase mortality (Hampton 1996) (note here and subsequently, the algebraic reversal when a positive treatment difference (μ, above) indicates

worse outcome) . The difference was in fact 0.5% with two-sided 90% CI for the difference: -1.7% to 0.71% (two-sided 95% CI: -1.96% to 0.98%). A different perspective may be taken of the attempt to infer equivalence from a (very) large negative superiority trial (Ng 1995). GUSTO III (GUSTO III Investigators 1997) was designed as a superiority trial (15059 patients enrolled) to detect a 20% difference in 30-day post myocardial infarction mortality, comparing double-bolus reteplase relative to an accelerated infusion of atelpase. The mortality for the reteplase arm was 7.47% and that of alteplase 7.24, an absolute mortality difference of 0.23% (two-sided 95% CI: -0.65% to 1.11%) This would, as the trialists noted “exceed a definition of equivalence requiring a difference of less than 1%” (1997; Fleming 2000), but they did note, parenthetically, that the INJECT trial (Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. 1994) had used 90% CI to establish equivalence. On this basis, equivalence would have been established in GUSTO III (two-sided 90% CI: -0.51% to 0.98%).

3.3.3.8.2 What is an appropriate equivalence margin is further illustrated by a consideration of the COBALT equivalence trial (COBALT Investigators 1997), where 7169 patients were enrolled to compare 30-day post myocardial infarction mortality, comparing weight adjusted accelerated alteplase with double bolus alteplase. On the basis of a 0.4% lower 95% CI for the absolute difference of accelerate infusion of alteplase vs streptokinase in the GUSTO I trial (GUSTO I Investigators 1993), equivalence was defined in the COBALT trial if the upper boundary of a one-sided 95% CI of the difference in mortality did not exceed 0.4%. The absolute mortality difference of 0.44% with two-sided 90% CI: -0.57 to 1.49% thus failed to sustain equivalence.

The differences in these approaches provoked editorial comment by Ware and Antmann (Ware *et al.* 1997) who also noted the consequences of the calculation of sample size when based upon the assumption of unequal mortality rates in the two arms. In the COBALT trial, it was assumed that 30-day mortality rates would be 6.3% in accelerated alteplase and 5.4% with double bolus alteplase and the trialists calculated an initial equivalence sample size of 4029 per group (4039 by our calculations using the software package PASS 2002 (Hintze 2002)). However, the sample size required for the approximately equal rates of 7.5% (the range mortality of the two arms reported in the COBALT trial) was identified by Ware and Antman as approximately 50000 in each group and the power of the COBALT study as effectively 0.16 (our calculations: 53634 in each group and power 0.158). Ware and Antman further suggested a absolute difference of 1.5% as a reasonable compromise for equivalence studies of this type, which, with 80% power, would require 3832 in each arm, a not impossible task it would appear for cardiology trials.

3.3.3.9 Testing for noninferiority and superiority: within the same trial it is possible to test sequentially for non-inferiority and superiority (Dunnett *et al.* 1996), although there are inherent problems in this strategy (Chuang-Stein 2001; Wiens 2001). At the least, trial methodology statements must pre-specify these analyses; the direction of testing should be (a) initial demonstration of non-inferiority and (b) subsequent testing for superiority (the reverse is problematic in interpretation, if superiority is shown in ITT analysis, but non-inferiority is not demonstrated in PP analysis); type I error must be preserved and the potential problems of different/unequal patient populations (PP analysis for

non-inferiority and ITT analysis for superiority) must be addressed, with for example, imputation of missing values (Hauck *et al.* 1999). Declaring (formal) non-inferiority when the primary trial methodology of superiority has been unsuccessful "…should be looked upon with healthy skepticism" ...(Wiens 2001); that is, it has the status of a post-hoc analysis.

3.3.4 Overviews of equivalence trials

3.3.4.1 Two recent papers have assessed the performance of trials where clinical or therapeutic equivalence have been claimed. Greene *et al* (Greene *et al.* 2000b) studied 88 reports (1992 to 1996) claiming equivalence (in title or abstract); δ was formally set in only 23% of reports, in 67% equivalence was declared after a failed test of superiority, the sample size was calculated in advance in only 33%, with 25% of reports having $n \leq 20$ per group. Of interest, δ ranged from 0 to 76% for proportionate differences. McCalister *et al* (McAlister *et al.* 2001), reviewing 4 recent hypertensive trials which appeared to show equivalence between treatment arms, found a lack of fulfillment in all of these trials of the 6 additional features that were defined as distinguishing superiority from equivalence trials; that: the active control was previously shown to be effective, similarity of (current) patients and outcome variables to those of original trials, optimal application of regimens, appropriate analysis, pre-specified δ and adequate sample size.

3.3.5 Critical care implications

3.3.5.1 Large trials with small treatment effect margins.

3.3.5.1.1 In this context, it is if interest to look again at the protocols of the ANZICS Clinical Trials Group SAFE trial (Finfer *et al.* 2003) which compared saline and albumin resuscitation. A total of 7000 patients were to be enrolled to

detect a ≥ 3% absolute mortality difference between treatment groups based upon an assumed 15% control mortality and β error 0.1. The magnitude of this mortality difference was derived from the lower 95% CI of the estimated treatment effect (albumin versus non-albumin use) from the 1998 Cochrane Injury Group Albumin Reviewers paper (Cochrane Injuries Group Albumin Reviewers 1998). As the SAFE trial (a superiority trial by its methodology description) was of large size with a relatively small treatment difference targeted, a question may be posed: if the null hypothesis of no treatment effect is not rejected, are we able to conclude equivalence between the two regimens. Despite the cautions above, some support for this scenario is provided by Ng who argues that "If the sample size is such that the …β error …at some δ is sufficiently small (eg < 0.05) then we can conclude that the two treatments are δ-equivalent" (Ng 1995). However, two restrictions are seen to apply to this proposition: the small β error and a robust δ. From the 1998 Cochrane paper, the lower 95% CI of treatment effect with a random effects estimator was 0.016 (our calculations are based upon the "metan" routine (Bradburn *et al.* 1998), using Stata™ software (Stata Statistical Software 2003)) . Thus possible values of δ would be 3% (lower 95% CI of fixed effects meta-analysis), 0.016 (lower 95% CI of random effects meta-analysis), 0.015 (50% of lower 95% CI of fixed effects meta-analysis). Total sample sizes under these scenarios are variably in excess of that of the SAFE trial.

3.3.5.2 Treatment of mycotic infections

3.3.5.2.1 Therapy for mycotic infections in the ICU has been recently transformed by the introduction of a group of drugs which have been marketed as

"equivalent" to the enduring yardstick, amphotericin B (Kam *et al.* 2002). Since its introduction in 1957, amphotericin B has been the standard (and until recently, the only) antifungal therapy (Gallis *et al.* 1990), despite never being compared with placebo and only two randomized trials of its action reported before the more recent comparisons with fluconazole (Rex *et al.* 2001). The effect of no specific therapy for candidaemia is difficult to estimate, but the prospective observational study (n=427) of Nguyen *et al* (Nguyen *et al.* 1995) suggests mortality rates of 27% with and 74% without antifungal therapy for candidal infections.

3.3.5.3 Amphotericin B and fluconazole.

3.3.5.3.1 There are 5 randomized trials in the literature comparing amphotericin B and fluconazole (Abele-Horn *et al.* 1996; Anaissie *et al.* 1996; Kujath *et al.* 1993; Phillips *et al.* 1997; Rex *et al.* 1994); in 2, amphotericin B was combined with flucytosine (Phillips *et al.* 1997; Rex *et al.* 2003), and in 2 there were intention to treat and "efficacy" groups (Phillips *et al.* 1997; Rex *et al.* 1994). Study sizes ranged from 40 to 237, for a total of 619 patients, a modest number, the implications of which, for the conduct of further comparative trials in mycotic infections, has been commented upon (Rex *et al.* 2001). For all of these studies, the primary evaluable end-point was responsiveness to therapy (variously defined), not mortality and in the "evaluable" groups responsiveness varied from 57% to 78% (mean, 68% for amphotericin and 61% for fluconazole) with overall mortality ranging from 12% to 38% (mean mortality 28% and no difference between the two drugs). The Phillips *et al* 1997 study was formally conducted with an "equivalence" protocol: $\delta = 0.2$, one-sided $\alpha = 0.05$ with an assumed response rate of 70% and requiring a

total of 148 patients (Phillips *et al.* 1997). Recruitment was limited to 106 patients total: point estimates and CIs for PP ("efficacy") and ITT analyses are seen in Figure 3.3.2.4 as trials 6-9. Lower borders (90% and 95%) of CI for both PP and ITT analyses are > -δ and the Westlake version (Westlake 1976) of the two one-sided hypothesis tests for equivalence also fails (p = 0.08 & 0.12 (Goldstein 1994)). The largest of these trials, by Rex *et al* in 1994 (Rex *et al.* 1994) recruited 237 patients but the methodology statement, although specifying a null hypothesis "…that the difference between the proportions of patients with favourable responses in the two groups would be less than 20 percent." is not clear as to this being an equivalence trial. This being said, equivalence can be established for both the PP and ITT analyses at (see Figure 3.3.2.4, trials 11-14), with p values for the two one-sided hypothesis tests for equivalence at 0.03 and 0.02 respectively, albeit the 95% lower border for the PP analysis was marginal. Neither trial was able to demonstrate equivalence / non-inferiority at δ = 0.1. The other three trials (Abele-Horn *et al.* 1996; Anaissie *et al.* 1996; Kujath *et al.* 1993) used superiority methodology and inferred "equivalence" from failure to demonstrate a difference. Pooling the studies and using quantitative meta-analytic techniques (Bradburn *et al.* 1998), for the PP analysis the treatment effect , fluconazole versus amphotericin, was – 6%, 95% CI: -13.8% to 1.9% (p = 0.14) and for the ITT analysis, -8.2%, 95% CI: -16.7% to 0.4% (p = 0.06). Although the pooled effect fail to demonstrate a definite treatment advantage for amphotericin, what can be inferred from the CI of the treatment effect is that a δ of 0.2 (20%) appears to be too large.

3.3.5.3.2 Four of the trials (Abele-Horn *et al.* 1996; Anaissie *et al.* 1996; Phillips *et al.* 1997; Rex *et al.* 1994) reported various toxic effects of therapy which were more common in the amphotericin group. Surprisingly, other than in the Annaissie *et al* trial (Anaissie *et al.* 1996), where indices of amphotericin induced renal toxicity returned to baseline level (at final review) in 73%, there was no systematic evaluation of the consequences of renal toxicity nor reporting of toxicity surrogates such hospital length of stay. This is important as it is problematic to claim "equivalence" from a randomized trial of efficacy and suggest "superiority" using arguments about toxicity (or other non-formally assessed end-points) (Djulbegovic *et al.* 2001). Although not germane to this thesis, other areas of concern in the interpretation of these trials are patient population (neutropenic versus non-neutropenic), the number and the appropriate management of catheter associated infections (79% in the Rex trial (Rex *et al.* 1994)), high versus low dose amphotericin (< 500mg total vs > 500mg (Nguyen *et al.* 1995)), especially in the context of a large number of catheter associated infections, and the omission of pre-emptive management of known toxic side-effects.

3.3.5.4 Amphotericin versus amphotericin modifications

3.3.5.4.1 The "high" incidence of toxicity, especially renal (Wingard *et al.* 1999), associated with amphotericin B has prompted re-assessments of the first-line position of this drug in mycotic infections (Ostrosky-Zeichner *et al.* 2003; Rex *et al.* 1999). Two recent large equivalence trials of Caspofungin (total n= 687, $\delta = 0.1$) (Mora-Duarte *et al.* 2002) and liposomal amphotericin B (total n= 239, $\delta = 0.2$) (Walsh *et al.* 1999) versus amphotericin B have suggested comparable efficacy of these two drugs with respect to amphotericin B and a

reduction in toxicities. A similarly large superiority trial of Amphotericin B colloidal-dispersion versus the parent compound concluded “comparable efficacy” (White *et al.* 1998), but of interest noted increased non-renal infusion toxicities in the colloidal-dispersion amphotericin, a point re-iterated by Winston *et al* in correspondence over the liposomal amphotericin B trial (Winston *et al.* 1999). What is pertinent in this series of trials is the cost of the group of amphotericin comparators; that is the trade-off of the extra costs of toxicities (especially renal) versus drug costs. Cagnoni *et al* (Cagnoni *et al.* 2000)undertook a pharmaco-economic analysis of the liposomal amphotericin B trial above and found that the hospital acquisition cost of the drug was critical in determining the break-even point, but it is noted that only 60% of patients were evaluated and “costs” were inferred from billing data, a strategy which has been criticized (Gyldmark 1995). O’Connell *et al* (O'Connell *et al.* 2002) undertook a similar cost-effectiveness study of the use of liposomal amphotericin B in 2002 and estimated the cost per additional life saved to be £ 23,819 (≡ $AUS 57, 574).

3.3.6 Conclusions

3.3.6.1 Claims in the literature as to the demonstration of “equivalence” must be subjected to careful scrutiny. The particular methodology of equivalence trials is of critical importance with respect to the conclusions that may be inferred, especially as these trials require the concurrent assessment of appropriate “external information”. Extension of conclusions beyond the “equivalence” hypothesis must also be formally assessed.

4 CLINICAL TRIALS IN CRITICAL CARE: INTERPRETATION

4.1 General comments (Moran *et al.* 2001a). The accompanying editorial (Evans 2001) to the Early Goal-Directed Therapy Collaborative Group and "Intensive Insulin Therapy in Critically Ill Patients" trials (Rivers *et al.* 2001; van den Berghe *et al.* 2001) pointed to problems of interpretation of unblinded and single-institution trials and it is convenient to begin a review of "interpretation" of clinical trials in Critical Care at this point.

4.1.1 The impact of "unblindedness": we may assess the effects of this on treatment efficacy by use of the Mann-Whitney statistic, as an alternative to the more traditional estimates such as risk ratio or risk difference. The Mann-Whitney statistic (Colditz *et al.* 1988; Colditz *et al.* 1989) estimates the probability (0 to 1.0) that a randomly selected patient given an innovative therapy will respond better than a randomly selected patient given "standard" treatment; that is, for a value of the statistic of, say, 0.6, the interpretation is that there is a 60% probability that the (next) randomly selected patient (or, more correctly, one of a pair of patients) on therapy will improve compared with no therapy. The empirical studies above have suggested that, in the presence of non-blinded randomised studies, the statistic be decreased by a value 0.1 or 0.15. In the current context, two of the ICU trials reviewed above (The ARDS Network Authors for the ARDS Network 2000; van den Berghe *et al.* 2001) were non-blinded and a third (Rivers *et al.* 2001) was partially blinded. Table 4.1.1 shows the Mann-Whitney statistic for the 4 trials mentioned above (Section 3.1. (Bernard *et al.* 2001; Rivers *et al.* 2001; The ARDS Network Authors for the ARDS Network 2000; van den Berghe *et al.* 2001) and the (positive, but non-blinded) trial of Amato *et al* (Amato *et al.* 1998), which reported a protective-

ventilation strategy in ARDS. What is apparent is the increased unadjusted probability of better response in the smallest study (Amato *et al.* 1998)) and the non-blinded studies compared with the large blinded PROWESS study (Bernard *et al.* 2001); but, after "adjustment", the response probabilities, at least for the larger 4 studies, are reasonably comparable. This comparability of responsiveness may say something of future expectations of therapeutic success in the ICU.

Table 4.1.1

Study	Blinded	Total n	MW stat	MW stat, adj
Amato	no	53	0.67 a	0.52 to 0.57
ARDS network	no	861	0.55 a	0.40 to 0.45
PROWESS	yes	1708	0.53 a	0.53
Van den Berghe	no	1548	0.52 h	0.37 to 0.42
Rivers *et al*	no	263	0.58 h	0.43 to 0.48

Total *n*= total study patient number. MW stat = Mann-Whitney statistic
MW stat, adj = adjusted Mann-Whitney statistic. a = related to 28 day mortality.
h = related to hospital mortality

4.1.2 Single versus multi-institutional trials: that the results of a single-institution trial may reflect, uniquely, the local structure of care has been previously attested to (Burns *et al.* 1991), but the interpretations(s) of the results of multi-institutional trials may also be a cause for some controversy, if we are to believe the exchanges in current mailing lists. What we are concerned about here is the recurrent debate on the implications of heterogeneity (Louis 1991), at the patient (Ioannidis *et al.* 1997; Rothwell 1995) institutional (Horwitz *et al.* 1996), or analytic level (Britton *et al.* 1999; Simon 1980). That is, will "dissimilar" patient and site characteristics and numbers vitiate the results of a trial which reports an overall treatment-effect. These concerns may be subsumed under the notion of "effect reversal" and was the object of the Horwitz *et al* study, above (Horwitz

et al. 1996), where a significant difference (using the Gail-Simon test for qualitative interaction) was found between 10 “divergent (survival under placebo was better) and 21 “dominant” (results agreeing with the trial as a whole) centres. As Senn and Harrell subsequently pointed out (Senn *et al.* 1997; Senn *et al.* 1998a), such a post-hoc exercise is problematic and a Galbraith plot of the centre effects revealed all lying within ± 2 SE. Thus “chance was an adequate explanation” for these findings; that is the probability of “effect reversal” may be considered a function of the number of centres in a multi-centre trial. In particular, for 80% power, the probability of at least one effect “reversal “ is ≥ 50% with ≥ 6 trial centres (Senn 1997). However, the problems of multi-institutional trials are not uniquely different from those of the single institution and the preponderance of multi-institutional trials may force us, rather, to seek solutions to these problems and not ignore them. In particular, we note recent recommendations for individualisation of patient therapy (Dans *et al.* 1998; Glasziou *et al.* 1995) and debates on the appropriate form of site-weighting and site-treatment interactions (the so-called type II vs type III model) (Gallo 2000; Kallen 1997; Schwemer 2000), the use of random-effects approaches to model site-effects (Fleiss 1986a; Senn 1998) and Bayesian methods as an alternative to traditional ANOVA analysis (Gould 1998). These questions are, of course, pertinent to another paradigm, that of meta-analysis and the close comparisons, at least at the analytic level, between meta-analysis and multi-centred trials has been commented upon (DeMets 1987; Senn 2000b). This close relation has seen changes in clinical trials mandated by meta-analysis (Chalmers *et al.* 1996) and the meta-analysis of individual patient data has been suggested as being at the top of the hierarchy of strength of evidence concerning

efficacy of treatment (Olkin 1995). Two other aspects of trial methodology deserve our further attention.

4.1.3 The problem of "optimal" cut points in the evaluation of (and implementation in subsequent trials) prognostic factors. There has been extensive discussion of the problems of this approach in the cancer (Hilsenbeck *et al.* 1992; Altman 1991a; Altman *et al.* 1994c) and statistical literature (Lausen *et al.* 1996; Morgan *et al.* 1986); the problems are those of increased Type I error, inflated estimates of effect and increased variance of these estimates and decrease in the efficiency of analysis. Appropriate adjustment of *P*-values may be made for this form of exploratory analysis (Hilsenbeck *et al.* 1996). That this may effect the outcome of trials may be indicated by recent research into the effect of anti-tumor necrosis antibodies (MAK 195F) in sepsis and septic shock. In a preliminary assessment of safety and efficacy of MAK 195F (Reinhart *et al.* 1996), the prognostic value of IL-6 concentrations of greater than 1000 pg/ml was established by (retrospective) cut-point analysis with no adjustment of the *P*-value. Although in the subsequent randomised placebo-controlled RAMSES study (Reinhart *et al.* 2001) of MAK 195F, a difference in mortality rates was observed between patient groups with IL-6 levels above and below 1000 pg/ml (40% vs 50%, $p < 0.001$), this did not translate into a mortality difference between treatment and placebo groups with IL-6 levels > 1000 pg/ml ($p = 0.36$). Similar prognostic cut-point classification occurred in the recent investigation of cortisol levels and cortisol response to corticotrophin in septic shock (Annane *et al.* 2000). The subsequent report of improved mortality in septic shock (Annane *et al.* 2002) with supplementary cortisol deserves further comment.

4.1.3.1 The trial was based upon a dichotomy of “responsiveness” to corticotrophin (primary end-point, 28 day survival in adrenal non-responders) with non-responders having a 63% vs 53% in responders, a non-significant difference ($p = 0.1$). Significance of the mortality difference was achieved using adjusted Cox regression (covariates: baseline cortisol, cortisol response, McCabe classification, LOD score, lactate and PaO_2/FIO_2 ratio); this may be problematic as discussion below will demonstrate. Furthermore, sample size was calculated using a one-sided formulation (the authors had “no interest in formally demonstrating a hypothetical deleterious effect of corticosteroids”) and postulated a 95% mortality rate (referenced to studies in calendar years 1983 and 1991) in cortisol non-responders, which was considerably greater than the 82% mortality found in the “worst” sub-group of the previous prospective cohort study (Annane *et al.* 2000). In addition, the final mortality estimates appeared not to have been adjusted for the two interim analyses.

4.1.4 The problem of “responsiveness” in assessing therapeutic interventions: this rather subtle problem was first given prominence in the cancer literature. It involves the responsiveness of patients to, say, chemotherapy in terms of a “remission” or rate of favourable response, being interpreted as evidence that the effect of treatment was to prolong overall survival. At the most, this is trivially true (patients who survive longer have a better outcome), but when differences in survival time (responders vs non-responders) are subjected to formal test (for example, the log rank test), bias in favour of responders may occur because these patients are being counted at risk of failure (death) before the time of response. As response to treatment and survival are both outcome variables, the use of testing to compare responder and non-responder survival distributions

merely serves to confirm or deny the association between response and survival; it does not necessarily invoke a causal pathway (Weiss *et al.* 1983; Morgan 1988; Anderson *et al.* 1983). We may think of a number of reasons why this scenario may occur: early deaths are counted as non-responders, by definition response involves a "guarantee time" before the response occurs, "response" may be a surrogate for the selection a particular patient subset not previously identified at initial randomization (that is the distribution of frailties (Vaupel *et al.* 1985; Yashin *et al.* 1995), an "unobservable" prognostic index, between responders and non-responders, differs) and the failure to appropriately operationalize the notion of response.

4.1.5 The translation of these analytic principles to the critically-ill context may not be straight forward, given the obvious different time scales and problems of definition of response. For instance, in the observational studies and trials looking at the effectiveness of increases in oxygen delivery in septic patients (Heyland *et al.* 1996), responsiveness was usually defined as an increase in cardiac output, but this may be an insensitive criteria. Such responsiveness was located in particular patient subsets in the trials addressing the question of goal-oriented therapy (Yu *et al.* 1998; Gattinoni *et al.* 1995). This may not be surprising, but it serves to caution us in our interpretations of what a response actually means (did it select out a group of patients with "better" (or worse) prognosis) and how it should be appropriately analysed. We may also speculate that where no improvement in mortality has been observed, but there are improvements in other end-points (rates of infection, length of stay), the same mechanisms may be operative (Bower *et al.* 1995).

4.2 Sample size, power and interpretation (Moran *et al.* 2004c)

4.2.1 Two recent editorials (Cooper 2004; Morgan 2004) have highlighted particular aspects of significant clinical trials in Critical Care: the French Pulmonary Artery Catheter Study Group's trial of pulmonary artery catheters (PAC) in septic shock and ARDS (Richard *et al.* 2003) and the SAFE trial of saline and albumin resuscitation in ICU patients (The SAFE Study Investigators 2004). An apposite comment was made regarding the SAFE trial that "Clinician's interpretation of this … [trial] …will undoubtedly be influenced by previous convictions" (Cooper 2004). Such "convictions" should properly include expectations regarding trial conduct; in particular, appropriate sample size and the correct interpretation of results when 'sufficient' trial sample size is not achieved. It is therefore instructive to consider, from these perspectives, the above two studies (Richard *et al.* 2003; The SAFE Study Investigators 2004) and a third, Transfusion Requirements in Critical Care (Hebert *et al.* 1999), which has been repeatedly cited in follow-up transfusion studies, such as the recently reported CRIT study of anaemia and blood transfusions in the critically ill (Corwin *et al.* 2004).

4.2.2 In the French Pulmonary Artery Catheter Study Group trial (Richard *et al.* 2003), 676 patients were randomised to receive a PAC (n = 335) or not (n = 341), with a primary end point of 28 day mortality. Initial sample size (with one interim analysis at 500 enrolled patients) was estimated at 1100 ($\alpha = 0.05$, $\beta = 0.1$), based upon an anticipated 10% mortality difference (35% vs 45%, for a global mortality of 40%, with "balanced group mortalities of 35% and 45%"). The treatment estimate was a risk ratio (RR) = 0.97 with 95% CI 0.86-1.10 and P = 0.67. At study conclusion (cessation was at 30 months by the Data Safety & Monitoring Board due to slow recruitment), the power to detect a 10% mortality

difference was 78% and, as the authors noted, underpowered also to detect the postulated 5% absolute mortality difference (≡ odds ratio of 1.24) of the previous Connors *et al* observational study of the mortality effect of PAC (Connors, Jr. *et al.* 1996). The observed RR (0.97) in the French Pulmonary Artery Catheter Study Group trial equated to an absolute risk difference (RD) of -1.6% (95% CI: -9% to 5.8%). The appropriate question is: what are we to make of this effect estimate when the final sample size and power is reduced?

4.2.3 One stratagem is to assess the treatment effect with respect to the estimated post-hoc power; with a total sample size of 680 and a single interim analysis (at 45% of total sample size), the power to detect 5% and 10% mortality difference was 26% and 76% respectively (by our calculations, using the S+SeqTrial2 module running under S-Plus® V 6.2 software, with O,Brien-Fleming stopping boundaries (Insightful Corporation 2002; O'Brien *et al.* 1979). However, inference from retrospective power has been properly criticized (Goodman *et al.* 1994; Hoenig *et al.* 2001; Zumbo *et al.* 1998) and an alternative approach, adopted in the French Pulmonary Artery Catheter Study Group trial, is to consider inference from the observed difference, as outlined by Hauck and Anderson (Hauck *et al.* 1986). The latter employed an equivalence testing approach to quantify (with the generation of appropriate P values) "…what was actually determined from the study….a possible outcome of the equivalence testing approach is the conclusion at the 5 per cent level that two…proportions …do not differ by more than some specified amount" (Hauck *et al.* 1986). Using both *90%* confidence intervals and two simultaneous one-sided (*t*) tests (the TOST procedure) for the specified difference (Rogers *et al.* 1993), it can be shown (we use the Stata™ module "equipi" (Goldstein 1994) and the "Analysis

of proportions" module in NCSS, release 2004 (NCSS 2004)) that "equivalence" is achieved for a threshold 28-day mortality difference between the two treatment groups of 7.8%, in agreement with the estimate reported by the French Pulmonary Artery Catheter Study Group authors. Therefore "we can conclude at an α risk of 5% that the absolute difference in mortality rate between the 2 groups is no more than 7.8%" (Richard *et al.* 2003).

4.2.4 The Transfusion Requirements in Critical Care trial, conducted by Hebert *et al* (Hebert *et al.* 1999) and published in 1999, has been pivotal in determining Critical Care physician attitudes to transfusion; in particular, that a restrictive red blood cell transfusion strategy (in this case, haemoglobin concentrations maintained at 7.0 to 9.0 g per decilitre) was "equivalent" to a liberal strategy (haemoglobin concentrations maintained at 10.0 to 12.0 g per decilitre). The primary outcome of the trial was "death from all causes in the 30 days after randomization". The trial enrolled 838 patients with 418 randomized to restrictive and 420 to liberal transfusion strategies. The 30-day mortalities of 18.7 and 23.3 percent respectively, were described in the published report as "similar" (P = 0.11) and the conclusion was that a "…restrictive strategy of red-cell transfusions is at least as effective as and possibly superior to a liberal transfusion strategy in critically ill patients". As opposed to the French Pulmonary Artery Catheter Study Group trial, Hebert *et al*, as outlined in the published methods, conducted an "equivalency trial" (Hebert *et al.* 1999). As previously shown (see 3.3, above), the null hypotheses in superiority and equivalence trials are reversed: in a superiority trial, the null hypothesis (H_0) is that the treatments have equal effects and in an equivalence trial, H_0 is that there is a specified difference (Δ) . Retention of H_0 in a superiority trial does not

establish equivalence (the null hypothesis of no difference is not "proved"); rejection of H_0 (and acceptance of the alternative hypothesis, H_a) in an equivalence trial establishes that the treatments do not differ by more than the specified Δ. Testing for equivalence (or non-inferiority) and superiority within the same trial is possible, but trial methodology statements must pre-specify this and there must be initial demonstration of equivalence.

4.2.5 Estimated sample size in the Transfusion Requirements in Critical Care trial showed progressive re-adjustments over time and the final statement was that the (recalculated) sample size of 1620 "…allowed us to rule out an absolute difference in the 30-day mortality rate of 5.5 percent…". However, due to poor recruitment, the study was terminated at 838 patient enrolments. Neither the study authors nor commentators formally canvassed the consequences of this early termination in terms of trial end-points. Applying the same methods as above, for 30-day mortality at an α risk of 5%, the absolute difference in mortality rate between the 2 groups is calculated to be no more than 9.3%, and for hospital mortality 10.9%. Hospital mortality, albeit a secondary end point, was a focal–point of discussion in both the trial report (Hebert *et al.* 1999) and the accompanying editorial (Ely *et al.* 1999), the latter describing the higher in-hospital mortality associated with liberal transfusion practices as "striking". However, the "significance" of the in-hospital mortality difference was marginal ($P = 0.05$, rounded: on a battery of 8 tests provided by NCSS software, P was always ≥ 0.051) and no adjustments were made for multiple testing. Combined testing for both equivalence and superiority yielded 'significant' results (for in-hospital mortality difference) only at $\Delta = \pm 10.9\%$. Thus the trial goal of an equivalence margin of 5.5% between restrictive and liberal red blood cell

transfusion regimens was *not* achieved and was demonstrated only at the 9-11% level.

4.2.6 The publishing of the results of the SAFE trial (The SAFE Study Investigators 2004) was welcomed on a number of fronts (Cook 2004; Cooper 2004), none the least of which was the ability to conduct large trials (Jennison *et al.* 2000a) in the critically ill over a relatively short period of time without insurmountable enrolment difficulties. A trial sample size of 7000 (with two interim analyses at 2333 (33%) and 4666 (67%) patients, using Haybittle-Peto boundaries (Jennison *et al.* 2000b) and 3.2 above) provided a 90% power to detect a 3% absolute mortality difference between the two treatment groups from an estimated baseline mortality rate of 15%. The hypothesis tested (H_0) was that "when 4 percent albumin is compared with 0.9 percent sodium chloride (normal saline) for intravascular-fluid resuscitation in patients in the ICU, there is no difference in the 28-day rate of death from any cause"; a superiority hypothesis, which was not rejected: RR = 0.99 (95% CI: 0.91 - 1.09) corresponding to an absolute RD of 0.07% (95% CI: -2% to 1.8%). The conclusion drawn was that there was "…evidence that albumin and saline should be considered clinically equivalent for intravascular volume resuscitation in a heterogeneous population of patients in the ICU". However, the accompanying editorial noted that the overall treatment effect "suggests equivalence, although proof of equivalence would require a different sample-size calculation" (Cook 2004). This editorial claim is of some importance: first, because a negative (superiority) trial cannot assert the null hypothesis "proved" (see above); second, no formal demonstration of equivalence was proposed in the published trial report (The SAFE Study

Investigators 2004) nor in the earlier methods paper (Finfer *et al.* 2003). What are we to think?

4.2.7 Nominal fixed sample size for a 3% difference in mortality outcomes would vary between 5500 (mortality reduction 15% to 12%) to 6000 (global mortality of 15% based upon "balanced group mortalities" (see 4.2.2, above) of 16.5% and 13.5%). The influence of the interim analysis with Haybittle-Peto stopping rules is to increment the nominal sample size by a function $R_{H\text{-}P}$ which is defined by the number of analyses or groups of observations (K), α and β, such that total $N = (N_{\text{fixed sample}} \times R_{H\text{-}P})$. For 3 analyses with two-sided $\alpha = 0.05$ and $\beta = 0.1$, $R_{H\text{-}P} = 1.007$ (Jennison *et al.* 2000b), a seemingly small increment. For the classical Pocock and O'Brien-Fleming designs, R is 1.15 and 1.016 respectively; these constants refer to the maximum sample size, not the average sample size (ASN), which, given the possibility of stopping early for these latter two designs, is less than the fixed sample size (Jennison *et al.* 2000b). As Jennison and Turnbull note: "Although Haybittle-Peto tests do not attain the maximum possible reductions in expected sample size, this is not always the key issue…"; and it may be that where large sample sizes are needed "…the investigator's objective is really to gather as much information as possible on all aspects of treatment, and there is little incentive for early stopping apart from the ethical need to cease randomizing patients to a clearly inferior treatment" (Jennison *et al.* 2000b). This perspective is contrasted to that operative in the recently reported ALVEOLI trial (high versus low PEEP in ARDS), where a 10% mortality reduction (28% to 18%) was sought: "Asymmetric stopping boundaries (with a two-sided (alpha) = 0.05) were designed to allow early termination of the trial if the use of higher PEEP was found to reduce mortality or if there was a low

probability that the trial could demonstrate a lower mortality rate in the higher-PEEP group than in the lower-PEEP group (futility stopping rule)" (The NHLBI ARDS Clinical Trials Network 2004).

4.2.8 For a sample size of 7000 and a simulation based (90%) two-sided CI approach for the difference (3%) in proportions based upon an equivalence hypothesis, the power at basal mortalities of 15% (projected mortality) and 21% (the actual mortality of the SAFE study) is computed to be 93% and 87% respectively (Elashoff 2002) (similar results were generated from "PASS"). As noted, H_0 in the SAFE trial was not rejected and using the above method of "inference from the observed difference" (Hauck *et al.* 1986), at an α risk of 5% the absolute difference in mortality rate between the 2 groups is no more than 1.7%. The 3% difference in mortality rates targeted in the SAFE trial was based on the "approximate minimal effect suggested by the lower confidence interval in the Cochrane Injury Group Albumin Reviewers Paper" (ANZICS Clinical Trials Group and Institute for International Health SAFE Study Investigators 2003), which used a fixed effect estimator to calculate the pooled difference, but this lower confidence interval was 1.6% for the random effect estimate. The choice of meta-analytic estimator is contentious, independent of the demonstration of heterogeneity (Normand 1999). An appropriate Δ has been defined as a fraction (0.2 – 0.5) of either the treatment effect, control drug versus placebo or of the lower 95% of this treatment effect, derived from a large trial or meta-analysis (Blackwelder 2002) and Section 3.3.4.6, above. Although these margins (which operationally may be defined as minimal clinically important differences) may seem small, from the perspective of a therapy applied to a "diverse population of critically ill patients" they translate into a noteworthy absolute number of

(potential) deaths: for the 16 units in the trial enrolling 7000 patients over 20 months, 210 potential deaths at the 3% mortality treatment difference and 119 at a 1.7% difference. It would seem that a precise definition of these minimally clinically important differences will "…undoubtedly be influenced by previous convictions" (Cooper 2004).

4.2.9 What can be construed from the above: first and obvious, that achieving planned sample size is critical for the assessment of trial reports (Moore *et al.* 1998). Second, both the French Pulmonary Artery Catheter Study Group trial and the Transfusion Requirements in Critical Care trial failed to establish their primary end-points and treatment recommendations based on these end-points must be circumspect. Third, the "equivalence" of 0.9% saline and 4 percent albumin for intravascular-fluid resuscitation in patients in the ICU would appear to be located at the 1.7 % absolute risk-difference level. Fourth, minimal clinically important differences for critically-ill patient categories need to be established.

4.2.9.1 The interpretation of treatment effects in trial reports is never that simple; a consideration of "inference from the observed difference" is an aide to the perplexed clinician.

5 CRITIQUE OF TRIALS IN CRITICAL CARE

5.1 Controversies with sepsis trials (Moran *et al.* 2002b)

5.1.1 The recent publication in the New England Journal of Medicine of a number of articles (Manns *et al.* 2002a; Siegel 2002c; Warren *et al.* 2002b; Wenzel 2002) looking again at the question of the efficacy of recombinant activated protein C, Xigris ®, Eli Lilly (APC) in severe sepsis, as originally reported in the PROWESS trial (Bernard *et al.* 2001), and the response of the primary investigators of the latter trial (Ely *et al.* 2002e), has some resonance with the "second look" saga (Ely *et al.* 2002d; Quezado *et al.* 1994; Schulman *et al.* 1991; Warren *et al.* 1992; Wenzel 1992) that surrounded the publication of the original HA-1A monoclonal antibody study (Ziegler *et al.* 1991) and the corresponding response of the HA-1A trialists (Ziegler *et al.* 1992b). This debate is even more ironic when one considers that some of the "critics" of both trials are the same and the recommendation for licensing of the now withdrawn HA-1A apparently occurred with unanimous approval by the United States Food and Drug Administration (FDA) advisory panel (Ziegler *et al.* 1992a); whereas the FDA Anti-Infectious Drugs Advisory Committee is reported to have been split 10 to 10 as to the safety and efficacy of APC (Warren *et al.* 2002d). The purpose of two of the papers above (Siegel 2002b; Warren *et al.* 2002e) was to alert readers to some of the potentially contradictory trial detail amassed in the licensing process of APC and the effect that this could have on the overall interpretation of the original report (Bernard *et al.* 2001); in particular, study protocol amendments, change in APC master cell line, sub-group efficacy and

adverse events (bleeding). For the clinician, the practical interpretation of a single reported trial (Peck *et al.* 2002) has become both onerous and uncertain; particularly as publicly available electronic material, in this case, FDA reports and submissions (Eli Lilly and Company 2001; FDA/CBER 2001b; FDA/CBER 2001a; United States Federal Drug and Food Administration 2001a; United States Federal Drug and Food Administration 2001b), is not routinely accessible and runs into hundreds of pages.

5.1.2 With respect to the PROWESS trial, the response of some recent editorialists (Hinds 2001; Hoth *et al.* 2001) has been, perhaps not surprisingly, cautious; provoking one commentator to lament the lack of excitement about the trial (Morris 2002a). The contrasting results of the PROWESS (positive treatment effect) and the antithrombin III (ATIII) KyberSept (null effect) sepsis trials have been reviewed (Crowther *et al.* 2001b), as have the putative pathophysiological mechanisms (Opal 2001b), in particular the role of heparin, both as an anti-sepsis agent (Davidson *et al.* 2002) and in terms of adverse effects (Crowther *et al.* 2001a). Of note was the difference in mortality rate of the placebo group in each trial: KyberSept 38.7% and PROWESS 31.3%, $p = 0.001$ (Fisher exact test), suggesting difference in illness severity on enrollment (unfortunately the PROWESS study used APACHE II and the KyberSept study used SAPS II as severity instruments). Both studies exemplified the limiting requirements of controlled clinical trials methodology, particularly those of patient selection, and such limitations have been the subject of formal review (Britton *et al.* 1999). Patient selection is particularly pertinent in consideration of the treatment effect of the now licensed APC product. Although the exclusion criteria were provided in Appendix 2 of the original report (Bernard *et al.* 2001), what was not

mentioned was the substantial protocol amendment at 720 (of 1690) enrolled patients plus a change in APC master cell production line. The sponsor detailed these amendment changes (10 points) in the FDA submission (Eli Lilly and Company 2001) and the trialists explained them in terms of enrolling ".. patients with a high likelihood of dying from severe sepsis and a low likelihood of dying from other causes…." and excluding "…patients in whom life support might be curtailed during the 28-day study period…" (see also Table 1 in their response (Ely *et al.* 2002c)). Surprisingly there was no change in placebo group mortality subsequent to the amendment (30% before and 31% afterwards), but there was a change in the efficacy of APC from null to active (relative risk (RR) 0.94, 95% CI 0.75-1.17, p = 0.57 to RR 0.71, 95% CI 0.57-0.87, p = 0.001; interaction test, p = 0.08) and such drew detailed comment from Warren *et al*, Siegel and the FDA (Siegel 2002a). No convincing evidence was adduced for a role of either protocol amendment or the concomitant new master cell bank in this efficacy change, but such cannot be excluded.

5.1.3 The sponsor noted that the "95% relative risk confidence intervals for the original and amendment results both include the overall relative risk estimate for the trial of 0.806, and …there was considerable overlap in the relative risk confidence intervals of the two subpopulations" (Eli Lilly and Company 2001). However, the inferential strategy of using overlapping CI is biased to the extent that it rejects the null hypothesis less often (that is, it is a conservative measure) (Schenker *et al.* 2001). The sponsor strategy for demonstrating consistency of effect across trial sub-groups, overlap of 95% CI of the various sub-group treatment effects (approximately 70) with the overall study relative risk estimate (0.806), produced different estimates of treatment consistency compared with

that adopted by the FDA and by Warren *et al*, who argued for severity dependence of treatment. As shown in Table 1 in Warren *et al* (Warren *et al.* 2002c) the first two APACHE II quartiles had no significant treatment effects, using the more familiar definition of the null effect as 95% CI spanning 1. A number of issues are raised here about the sponsor's analytic strategy, which, as Ely *et al* (Ely *et al.* 2002b) mentioned, resulted in "only one .. not meet[ing] the trial definition of subgroup consistency." First, it is problematic using the point estimate of 0.806 as a null hypothesis in that this is based on the same data as the subgroups being tested; second, how likely is it with multiple testing that one should obtain fewer false positives than statistically expected (at $p = 0.05$); third, the correspondence between confidence intervals and hypothesis tests (that is, point estimate as null hypothesis) does not always hold, depending on the calculations of the standard errors; fourth, the method of overlap of CI is sensitive to the study design (it will be deficient in a randomized trial context when standard errors of the two populations are similar) and the correlation between point estimates (with a positive correlation, likely in a trial context, the overlap method loses power) (Schenker *et al.* 2001).

5.1.4 Apropos of consistency of effect, it was of more than passing interest to note that in the sponsor submission (Eli Lilly and Company 2001), the first APACHE II quartile group was effectively deemed discordant (unadjusted RR point estimate of 1.25 with lower 95%CI was 0.78) and a lengthy analysis was undertaken in an attempt to explain this, presumably on the basis that all other APACHE II quartile point estimates were <1, which argued, in some sense, against consistent sub-group effects. Paradoxically, no such analysis was undertaken to explain a RR of < 0.5 with upper 95% CI < 0.806 (treatment

benefit greater than the entire population) in the 1st IL-6 quartile group. The sponsor further suggested that the "..observed variability in relative risk estimates.. [of treatment effect]…appears predominantly due to changes in the mix of investigators actively enrolling patients during the course of the trial…"; 20 sites enrolled patients only under the original protocol version, 45 investigative sites only under the amended version of the protocol (227 patients). The treatment-by-protocol interaction for the 99 sites enrolling under both protocols was non-significant at p = 0.5 (Eli Lilly and Company 2001). Evidence of a treatment-by-site interaction effect was present at p = 0.08 (over the 160 sites) (RM87179 2001). Of interest was the lack of treatment effect (95% CI of RR spanning 1) in the regional areas of both Europe and Australasia (Amos 2001). Such effects must, of course, be tempered by the known problems of post-hoc subgroup analysis (Freemantle 2001).

5.1.5 The original publication (Bernard *et al.* 2001) and the sponsor submission to the FDA Anti-Infective Drugs Advisory Committee (Eli Lilly and Company 2001) argued for a consistent treatment effect of APC across subgroups, but the FDA licensed the product for treatment of adult patients with severe sepsis and an APACHE II score of 25 or more (3rd and 4th quartiles, see above), the latter requirement still being effectively an untested hypothesis. Using simulation studies in an economic evaluation, Manns *et al* (Manns *et al.* 2002b) gave further credence to an APACHE II treatment threshold in terms of cost effectiveness, but although the mean APACHE II score of their cohort (787 patients) was 21, only 40 patients were sampled for formal chart review to confirm a diagnosis of severe sepsis, suggesting the potential for insensitivity of sepsis diagnosis. At this stage the requirement for a second blinded efficacy trial

has not been mandated for APC as opposed to the HA-1A case (McCloskey *et al.* 1994) where severe doubts about the activity of the monoclonal antibody were expressed (Quezado *et al.* 1993). Ely *et al* (Ely *et al.* 2002a), commenting upon the difficulties inherent in the use of an APACHE II treatment threshold, referred to the complementary evidence from organ dysfunction scores where treatment-induced RR showed a progressive fall from 0.92 through 0.6 with organ systems dysfunction of 1 to 5 (Eli Lilly and Company 2001). However, the 95% CI of the RRs all spanned 1, suggesting that organ dysfunction was an insensitive means of assessing risk-related treatment effects and therefore an unsatisfactory surrogate. Given that one of the reasons proffered for the differential efficacy between APC (some) and ATIII (none) in sepsis was the superior anti-inflammatory of APC (Opal 2001a), it was surprising then that the only significant and consistent organ system (time) change (either 28 day mean-time average or time-windowed, days 1-4, 1-7 & 1-14), treatment versus placebo, was in the cardiovascular system. Using Kaplan-Meier estimates with non-surviving patients censored, time to resolution of both cardiovascular and respiratory organ dysfunction was significantly reduced with APC ($p = 0.009$) (Eli Lilly and Company 2001). This was also reflected in a significant treatment decrease in vasopressor (20.1 versus 18.8, $p = 0.014$) and ventilator free days (14.3 versus 13.2, $p = 0.05$), although, paradoxically, not in SIRS-, ICU-, or hospital-free days.

5.1.6 However, the strategy of censoring deaths with Kaplan-Meier estimates depends upon the assumption of “non-informative censoring” and this has been questioned in acute illness (Clark *et al.* 1997). The study methods for dealing with missing data were also problematic: for imputation, last observation

available was carried forward (LOCF), an ad hoc method known to be based upon unrealistic assumptions (Siddiqui *et al.* 1998); for informative drop-outs (deaths), a non-surviving patient received an organ dysfunction score of 4 (worst score) for the day of death and for every day thereafter until Study Day 28. The latter scenario, where there is no joint modeling of the marker level and drop-out process, has been shown to yield biased estimates (Touloumi *et al.* 2002). With respect to the FDA assessment of the efficacy of APC (FDA/CBER 2001a), the following groups were identified where treatment effect predominated : 3rd and 4th APACHE II quartiles, laboratory evidence of disseminated intravenous coagulation, not on heparin, age > 50 years, shock and ≥ 2 organ system dysfunctions.

5.1.7 Where then does this leave the clinician? Obviously treatment efficacy has retreated from a uniform effect (Bernard *et al.* 2001), the absolute magnitude of which may not be as great as initially reported to one located in sub-groups, albeit some pre-defined and intuitively obvious. The conclusion must be with Warren *et al* (Warren *et al.* 2002a), that APC is not be a standard of care in severe sepsis and, despite some evidence for cost effectiveness if tailored to an APACHE II threshold, a confirmatory trial seems mandated.

5.2 Cerebral injury and the role of induced hypothermia (Moran *et al.* 2002a)

5.2.1 In the last few years a series of randomised controlled trials have reported the use of therapeutic mild hypothermia (33^0C to 35^0C) in two groups of patients: (a) post cardiac arrest (The Hypothermia after Cardiac Arrest Study Group 2002; Bernard *et al.* 2002) and (b) acute traumatic brain injury (Shiozaki *et al.*

2001; Marion *et al.* 1997; Jiang *et al.* 2000; Clifton *et al.* 2001c; Clifton *et al.* 1993). Therapeutic hypothermia was reported to produce improved outcomes in group (a) but not uniformly in group (b); rather the "definitive" large-scale trial, 392 patients (Clifton *et al.* 2001c), in group (b) was unable to demonstrate benefit. Editorial response to these trials, not surprisingly, differed, from endorsement in the case of post cardiac arrest (Safar *et al.* 2002) to analysis of potential reasons for lack of response (Narayan 2001) in the case of traumatic injury. What are the factors responsible for these disparities?

5.2.2 All trialists referred to the multifactorial mechanisms which may be responsible for the protective effect of hypothermia; the non-effect of hypothermia in traumatic cerebral injury was been explained in terms of potential differences in these mechanisms with respect to initiating injury (Curfman 2002). Outcomes in the trials were reported as deaths or scales of (cerebral) performance. Only one study (Bernard *et al.* 2002) made specific reference to difficulties in such assessments (Coste *et al.* 1995) and no studies reported inter-rater reliability studies of outcome assessment. Despite these potential problems, for all trials it was possible, within this review, to classify outcome into "poor" (as severe disability, both awake or unconscious, with institutionalisation or death) and "good" (normal or moderate disability) categories. Given this, it is noted that the normo-thermic (control) patients in both groups (a) and (b) had almost identical "poor" cerebral outcome percentages, at 67% and 60% respectively, suggesting, at least, common functional outcomes. Across the trials, cerebral outcomes were assessed at different times, from hospital discharge through 12 months post discharge. Evidence from one trial (Marion *et al.* 1997), where outcomes were reported at 3, 6 and 12 months, suggests that the proportion of

"bad" outcomes may not have materially changed over this time (p=0.09, Fisher exact test).

5.2.3 Time to treatment. The time to reach target hypothermia (however defined) has been thought to be a critical therapeutic factor, more so in explaining the lack of efficacy of hypothermia in traumatic cerebral injury (Safar *et al.* 2001; Narayan 2001; Clifton 2001) . Over all the trials considered above (The Hypothermia after Cardiac Arrest Study Group 2002; Shiozaki *et al.* 2001; Marion *et al.* 1997; Jiang *et al.* 2000; Clifton *et al.* 2001c; Clifton *et al.* 1993; Bernard *et al.* 2002), this target time varied from 2 to 15 hours (mean, 8.6 hours). Using meta-analytic methods (meta-regression, with a restricted maximum likelihood estimator (Sharp 1998)), no linear relationship was demonstrated between log odds of poor cerebral outcome and time to target hypothermia in the 5 cerebral trauma trials, p = 0.07; with the outlier (Clifton *et al.* 2001b) removed (time to target =15 hours), p = 0.4. Similarly, no effect was demonstrated using individual patient data in the largest trial to date in cerebral trauma (Clifton *et al.* 2001a; Clifton *et al.* 2001c).

5.2.4 Centre effect. Heterogeneity between participating centres was suggested as a reason for non-effect of hypothermia in cerebral trauma (Safar *et al.* 2001) and Clifton *et al* (Clifton *et al.* 2001a) further reported on "inter-centre variance in treatment effect and outcomes" from the original trial (Clifton *et al.* 2001c). Such centre variation was thought to have reduced the "overall sensitivity" of the trial and treatment effect reversal was demonstrated for some centres. A considerable literature has accumulated on these issues (Agresti *et al.* 2000; Ioannidis *et al.* 1998b), but as pointed out by Senn and Harrell (Senn *et al.* 1997) and Peto (Peto 1982; Peto *et al.* 1995), such heterogeneity (including effect

reversal) is to be expected, purely by chance, and at a certain level, may always be demonstrable if the number of centres is at least 8. From the clinical perspective, heterogeneity with respect to such variables may not be a defect, rather a strength (Marshall 2001; Peto *et al.* 1995).

5.2.5 Baseline variables. All trials under consideration reported baseline variables between treatment groups; at face value, this would seem to be a reasonable procedure. However, two points need to be made regarding this practice: with appropriate randomisation, by definition, differences are due to chance and, randomisation into treatment groups without bias does not necessarily lead to groups having "similar" baseline characteristics (Altman 1985; Altman *et al.* 1990; Senn 1995). It may be difficult, in fact, to demonstrate "no" difference between groups, depending upon the number of covariates recorded and comparisons made and protocols for describing just which covariates to tabulate (Enas *et al.* 1990). Moreover, variable tabulations are essentially marginal summaries and provide no insight into the joint distribution of prognostic factors in treatment groups (Piantadosi 1990).

5.2.5.1 In the Clifton *et al* trial (Clifton *et al.* 2001a) of traumatic cerebral injury, specific mention was made of the unbalanced assignment of age groups and patients with spontaneous hypothermia (a previously unknown prognostic factor) to treatment groups (Ginsberg 2002). Bernard *et al* (Bernard *et al.* 2002) noted differences in percentage of males and bystander performed cardio-pulmonary resuscitation between the two treatment groups. Again, this may have been by chance, or due to the atypical randomisation protocol which prescribed odd / even day allocation; circadian and seasonal variations in acute cardiovascular disease have been noted (Evans *et al.* 2000; Arntz *et al.* 2000).

5.2.5.2 However, the question of balance between treatment groups relates not to the validity but to the efficiency of statistical inference (Begg 1990; Senn 1994a). Thus, concerns expressed that such “unbalance” may have affected trial outcome (Safar *et al.* 2001) are misplaced, to the extent that statistical analysis is not invalidated, rather the ability to derive conclusion from the results is affected.

5.2.6 Treatment effect: adjusted versus unadjusted estimates. Three studies also reported statistically adjusted treatment effect estimates on the basis of either (perceived) baseline differences between treatment groups or covariates related to outcome:

5.2.6.1 on the basis of “slightly less severe “ traumatic cerebral injuries in computer tomographic (CT) class (the patients being stratified on allocation to a Glasgow coma score (GCS) of 3-4 or 5-7), Marion *et al* (Marion *et al.* 1997) adjusted risk ratios (presumably equated with odds ratios) for treatment effect using initial Glasgow coma score and CT class only. Over all trial patients, unadjusted (and significant) risk ratios for beneficial treatment effect at 3, 6 and 12 months post hospital discharge moved towards the null and became non-significant . For patients with a Glasgow Coma score (CGS) of 5-7, adjusted risk ratios at 3 and 6 months (but not at 12 months) maintained statistical significance.

5.2.6.2 In the Bernard *et al* trial (Bernard *et al.* 2002), the unadjusted estimate of “good” outcome, in the odds ratio metric was 2.65 (95% CI: 1.02-6.88; p = 0.046) and in the risk ratio metric 1.84(0.97-3.49). That this estimate had border line statistical significance is noted, especially as the study protocol details an unspecified and apparently unblinded interim analysis at 62 eligible

patients. The statistical consequences of such data inspections (especially if the interim treatment difference was examined) are known to inflate the probability of a type I error (Armitage *et al.* 1985; Gould 2001). That is, for 1 or 2 pre-planned interim data inspections, the nominal significance level required to achieve a "true" level of 0.05 (the type I error rate), are 0.03 and 0.021 respectively (McPherson 1974). These figures refer to normally distributed data; recent investigations in the sample size re-estimation (SSR) literature suggest that calculations for binary data are comparable (Shih 1995; Shun *et al.* 2001). The adjusted (odds ratio) estimate via logistic regression, including only two covariates, age and time from collapse to return of spontaneous circulation, was 5.25 (95% CI: 1.47-18.76; p = 0.011).

5.2.6.3 In the Hypothermia after Cardiac Arrest Study Group trial (The Hypothermia after Cardiac Arrest Study Group 2002), baseline differences were thought to be due to "random variation", but the risk ratio of six month "good" cerebral outcome "changed only minimally" with the addition of all recorded baseline covariates (risk ratios 1.40(95% CI: 1.08-1.81) vs 1.47(1.09-1.82) respectively). The unadjusted odds ratio for "good" outcome was 1.89(1.17-3.05).

5.2.6.4 That both the adjusted treatment estimate (odds ratio 5.25, equivalent to 2.4 as risk ratio) and its upper 95% CI (odds ratio 18.76, equivalent to 8.57 as risk ratio) in the Bernard *et al* trial (Bernard *et al.* 2002) was substantially greater than that of the numerically larger Hypothermia after Cardiac Arrest Study Group trial (The Hypothermia after Cardiac Arrest Study Group 2002) is cause for some comment. The small size of the trial, with relatively few events, is certainly one explanation, to the extent that treatment effects tend to be inflated

in small positive trials (Stern *et al.* 2001). Paucity of data / events may be surmounted by the use of exact inference or appropriate estimators (Tomz *et al.* 1999). In non-linear regression models with randomised studies, covariate adjustment tends to move the treatment effect estimate away from null and to have a variable, unpredictable effect upon the variance of that estimate, as opposed to the increase in precision seen with linear models (Beach *et al.* 1989; Buyse 1989; Ford *et al.* 1995; Robinson *et al.* 1991; Tomz *et al.* 1999). On the other hand, omission of balanced covariates, leads to bias in estimates of effect (Chastang *et al.* 1988; Hauck *et al.* 1991; Robinson *et al.* 1991). Both trials (The Hypothermia after Cardiac Arrest Study Group 2002; Bernard *et al.* 2002) also appeared to use post-hoc data-driven adjustments, in that the particular limited set of conditioning covariates selected were not pre-specified in methodology statements, as has been recommended (Senn 1989a; Senn 1994a; Senn 2000a); such approaches are also known to overestimate treatment effects (Hauck *et al.* 1998; Raab *et al.* 2000). It is also uncertain if model selection was undertaken to derive adjusted estimates, such a strategy being problematic in randomised trials. This is not to suggest that one should necessarily condition on "all" covariates (Tukey 1991), rather that covariates specified in the design phase are included in the model (Greene *et al.* 2000a); if covariates are orthogonal to treatment, then standard errors of treatment effects will also be minimised.

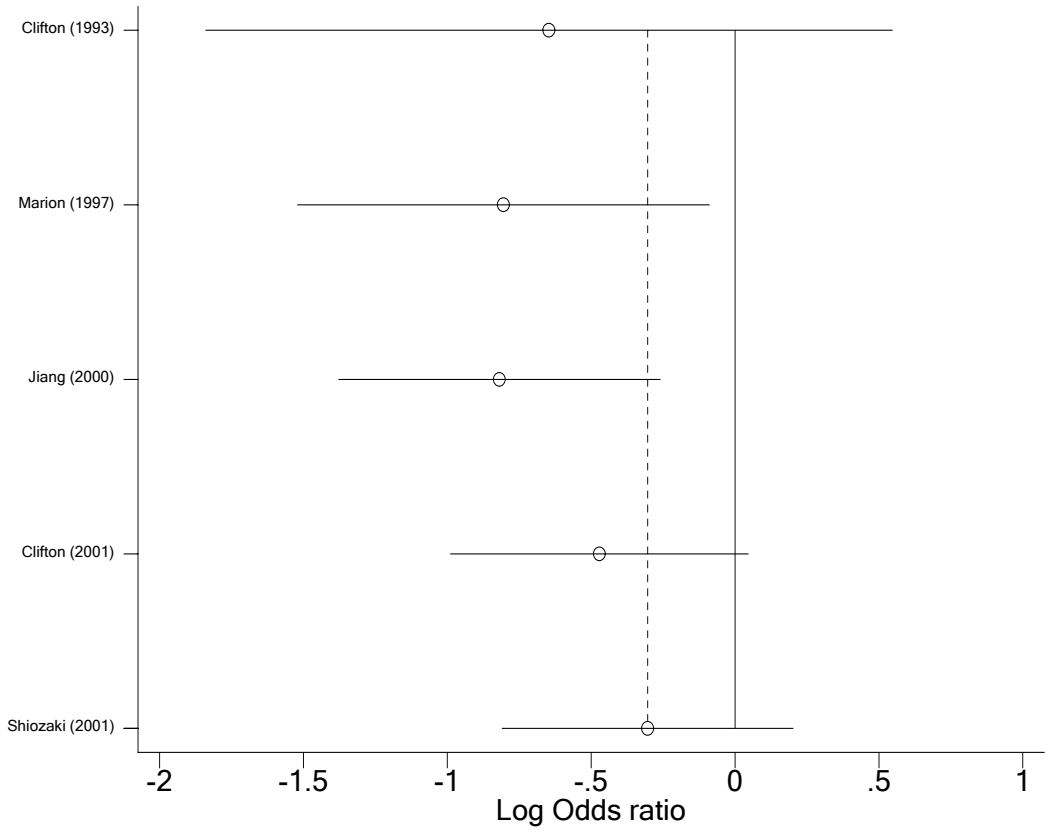
5.2.7 Effect measures. Measures of effect were considered by the Hypothermia after Cardiac Arrest Study Group trial (The Hypothermia after Cardiac Arrest Study Group 2002); in particular, risk ratio estimates were derived from odds ratio using a transformation after Zhang *et al* (Zhang *et al.* 1998), although it has been

pointed out that this transformation method is potentially biased in the presences of unmeasured confounders (McNutt *et al.* 2003). That odds ratios do not approximate risk ratios, except with rare outcomes, and are often misinterpreted as risk ratios needs to be reiterated (Laupacis *et al.* 1988; Sinclair *et al.* 1994); the latter are in fact available "directly" from certain multivariable regression estimators (Wacholder 1986). The advantages of the various metrics (or scales) of treatment effect have been extensively discussed (Deeks *et al.* 2001).The odds ratio has superior statistical properties and is the key parameter in the linear logistic model. The log odds scale is unbounded in both directions, but is numerically greater than the risk ratio when underlying event rates are frequent. Risk ratio is sensitive to small counts and is bounded above in a manner dependent on the control group risk. Although more intuitive as a measure than odds ratio, it has the peculiarity that the relative risk of dying is unrelated to that of surviving (Senn *et al.* 1998b). Risk difference is immediately intuitive and expresses the consequences of no therapy (unlike both odds and risk ratios), but is constrained from –1 to 1 and suffers from potential bias with varying time to follow-up which may also impact on the derivation of number needed to treat (NNT) (Walter 2001). The Hypothermia after Cardiac Arrest Study Group provided estimates of NNT to prevent one unfavourable neurological outcome (6; 95% CI 4-25), although the NNT has not been without its critics (Hutton 2000).

5.2.8 Specific focus, traumatic brain injury. As revealed by the title of the editorial accompanying the Clifton *et al* paper, "a good idea proved ineffective" (Narayan 2001), the efficacy of hypothermia in traumatic brain injury is not established. The total number of patients enrolled into the 5 trials was 673 and the primary

cerebral outcome, as good versus poor, was only available at differing times post hospital discharge: 3 months (2 trials), 6 months (2 trials) and 12 months (1 trial). If we assume that these outcomes do not significantly change from 3 through 12 months (see above), then the pooled risk ratio (random effects estimator) for poor outcome was 0.88 (95% CI: 0.71-1.09; p = 0.23). Cumulative estimates (Sterne 1998) are seen in Figure 5.2.8 in the log odds metric; further large positive outcome trials would appear to be necessary to establish significant efficacy.

Figure 5.2.8. Cumulative estimates for treatment effect of induced hypothermia in traumatic brain injury



Cumulative meta-analytic estimate of therapeutic hypothermia on "poor" outcome in cerebral trauma. Vertical axis; trials in year order. Horizontal axis: treatment effect, log odds ratio. Solid vertical line; null effect (+ve log odds ratio, "poor" outcome; -ve log odds ratio, "good" outcome). Dashed vertical line; point estimate of final treatment effect. Horizontal lines; 95% CI of cumulative treatment effect for trials. Circles; point estimates of cumulative treatment effect.

5.2.9 Specific focus, post cardiac-arrest.

5.2.9.1 The two trials in post cardiac arrest patients yielded apparent positive treatment effects. Primary cerebral outcome was not assessed beyond hospital discharge for the Bernard *et al* trial (Bernard *et al.* 2002), but there was no difference in the cerebral performance status in the subset of patients of the

patients actually discharged alive (good versus poor outcome: 21 / 0 in the hypothermia group and 9 / 2 in the normothermia group; p = 0.11, Fisher exact). In the Hypothermia after Cardiac Arrest Study Group trial (The Hypothermia after Cardiac Arrest Study Group 2002) a similar result was obtained for patients discharge out of hospital and assessed at 6 months (good versus poor outcome: 64 / 20 in the hypothermia group and 42 / 24 in the normothermia group; p = 0.11, Fisher exact). The same results are obtained if we discount for deaths occurring after discharge (6 and 7 respectively; p = 0.22). Hospital deaths were characterised only in the Bernard trial (Bernard *et al.* 2002) and equal numbers of patients died of cardiac failure in the two treatment groups (5 in the hypothermia and 4 in the normothermia group). No evidence was adduced that hypothermia was cardioprotective, in fact the tendency to hyperglycaemia in the hypothermia group of Bernard *et al* (Bernard *et al.* 2002) may be inimical (Capes *et al.* 2000). In the Hypothermia after Cardiac Arrest Study Group trial (The Hypothermia after Cardiac Arrest Study Group 2002), the hospital death rate was improved with hypothermia (p = 0.03, Fisher exact), but we are surprisingly given no information as to the causes of death.

5.2.9.2 The protective effect then is apparently manifesting itself by preventing “early” catastrophic non-cardiac (presumably cerebral) deaths, but with no impact upon less severe forms of cerebral injury. Such would appear to be counter-intuitive, although there may be a confounding effect of small numbers and insensitivity of cerebral performance scales alluded to above. The impact of non-blinding of the studies exaggerating treatment effects, a point raised by the Hypothermia after Cardiac Arrest Study Group (The Hypothermia after

Cardiac Arrest Study Group 2002), must also be considered. Furthermore, it is also apparent that in this trial up to 25% of recorded temperatures in the normo-thermic group were >= 38^0C (Figure 1 in (The Hypothermia after Cardiac Arrest Study Group 2002)) for a number of hours. In the absence of protocol statements as to what constituted "normo-thermia", this may have been a source of bias in terms of eventual cerebral outcomes.

5.2.10 The question of what constituted appropriate effect size also needs clarification. Both trials reported preliminary studies where outcome rates were initially investigated; such studies are unfortunately known to have wide confidence intervals for control rates and may be unrepresentative (Wittes *et al.* 1990):

5.2.10.1 Bernard *et al* (Bernard *et al.* 1997), in a pilot trial over the years 1993 to 1996, reported in1997 a "good" outcome with therapeutic hypothermia in 50% (11 of 22) of patients compared with "poor" outcome in 14% (3 of 22) of historical controls (1991-1993) treated with standard measures. On the basis of this study, the initial sample size was set for a treatment effect of 36% at a power of 0.8; 31 patients in each group. This would appear to be a substantive treatment effect, given the known problems of estimating such rates using non-concurrent controls. The change over time in control rates may also be problematic and was evident in the prospective study (Bernard *et al.* 2002), where the rate of good outcome in the normo-thermic group had almost doubled to 26.5%. To prevent under-estimation of the sample size based upon an unrepresentative estimate of control rates (see above) , Gould (Gould 2001) has recommended using the 75^{th} percentile of the confidence distribution of the (population) variance. In this case, the 95% CI of the assumed underlying control rate (14%) is 3% to 35% and the 75^{th} percentile would correspond to a

control rate of 24%, much closer to the observed trial control rate of 26.5%. Under this scenario, the sample size to achieve a treatment group "good" outcome rate of 50% (corresponding to a treatment effect now of 24%) would increase to 61 in each group.

5.2.10.2 Evidence from four recent "positive" trials of therapy in the critically ill considered above (Bernard *et al.* 2001; Rivers *et al.* 2001; The ARDS Network Authors for the ARDS Network 2000; van den Berghe *et al.* 2001), suggests a treatment difference of 3.4% to 16% (average 8.7%), albeit for mortality. In the Bernard *et al* (Bernard *et al.* 2002) trial, the risk difference was 22.4% (95% CI: 13.2% to 43.4%) and the observed post-hoc power for the trial was 42%. However, such calculations are known to be methodologically flawed (Hoenig *et al.* 2001; Zumbo *et al.* 1998); what can be estimated is the precision of the (observed) difference for any given power and "clinically important difference", after Goodman and Berlin (Goodman *et al.* 1994). For a total sample size of 62 , power of 0.8 and treatment effect of 36%, the predicted precision is ± 25% and for a hypothesised 0% difference between treatments, the observed treatment effect in the trial (22.4%) is not excluded; that is, it is within the ± 25% precision bounds. This uncertainty of effect is consistent with the p value of 0.06 for primary outcome by the Fisher exact test which is recommended in the current context of randomisation inference (Ludbrook *et al.* 1994). On the basis of these analyses, plus the effect of the unspecified interim analysis (see above), the status of the "observed" treatment effect is questioned.

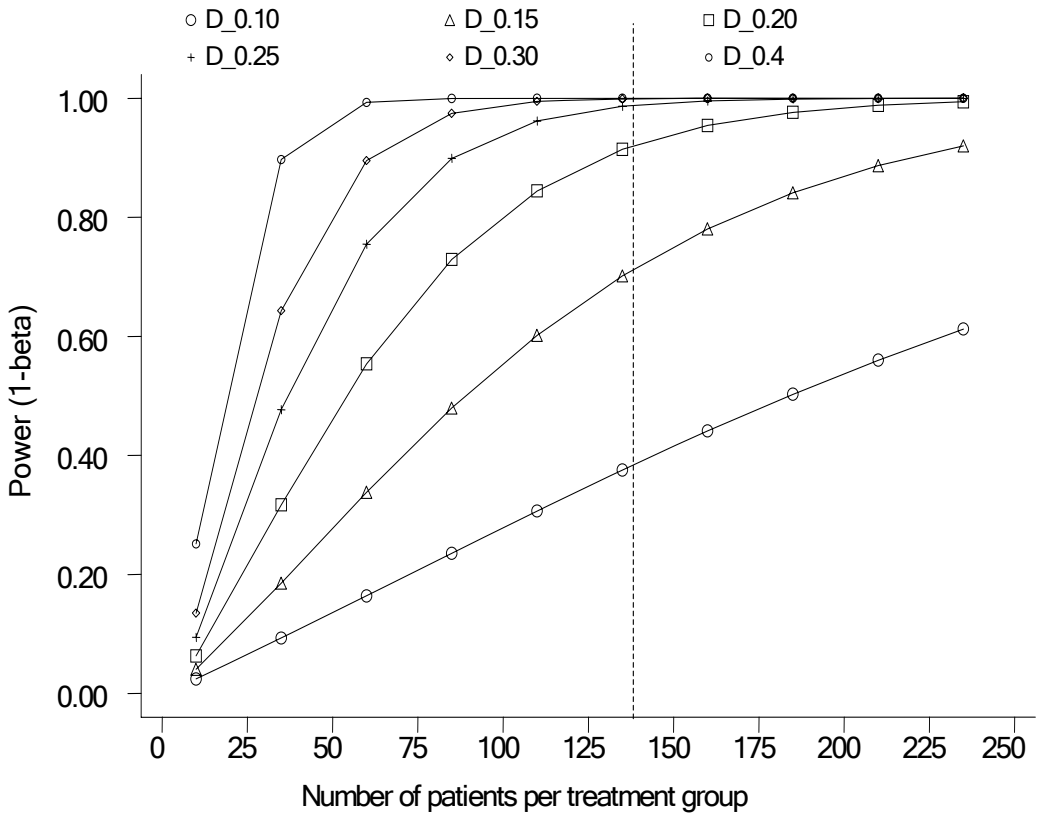
5.2.10.3 The initial pilot study the Hypothermia after Cardiac Arrest Study group (Holzer *et al.* 1997) reported a good outcome with hypothermia, at six months

review, in 52% (14 of 27) of patients versus 26% in unspecified historical controls. On this basis the trialists planned to enrol a total of 500 patients. In the prospective trial methodology statements (The Hypothermia after Cardiac Arrest Study Group 2002), no sample or effect size was provided and enrolment was stopped at a total of 275 patients for ad-hoc reasons (lack of funding). The control rate for "good" outcome was also noted to have "improved" to 39%. No indication of interim analyses was given, perhaps unusual in multicentre trial planning to enrol 500 patients and using innovative therapy. It is useful then to look at power curves for the estimated effect size(s) implied by a total sample size of 500. This is seen in Figure 5.2.10.3 , where power is plotted against sample size (per group) by differences between two groups of 0.1, 0.15, 0.2, 0.25, 0.3 and 0.4, with one group fixed at a base proportion of 0.26, the historical control rate for "good" outcome. For a total trial size of 450 patients, 80% power is established for an effect size of approximately 12-13%. For the prospective trial (2), with total patient number of 275, power was 0.8 for a reduced effect size of 17-18% (again, we use this latter estimate for pedagogical reasons only). Thus some caution is needed in understanding what effect size would be appropriate for the hypothermic intervention after cardiac arrest and the performance of both trials in respect of this.

5.2.11 That clinical trials such as reviewed are very difficult to perform is attested to by the long recruitment times and patient exclusions. All trials recorded a low and acceptable incidence of side effects of therapy. With respect to therapeutic hypothermia in cerebral trauma, at this juncture no advantage is demonstrable. In the case of post cardiac arrest, a treatment effect was claimed in both trials,

but the methodological concerns raised above would counsel caution towards these claims and not mandate hypothermia as a current standard of care. Future studies should report separate analyses for both deaths and scaled cerebral outcomes; the latter should be analysed as ordinal data.

Figure 5.2.10.3. Power curves for estimated effect size



Power vs Sample size (per group) by differences between two groups; one group fixed at a proportion of 0.26.
Vertical axis; power as (1-B). Horizontal axis: number of patients per treatment arm
Large circles: line of effect for a treatment difference (D_0.1) of 0.1
Large diamond: line of effect for a treatment difference (D_0.15) of 0.15
Squares: line of effect for a treatment difference (D_0.2) of 0.2
Plus sign: line of effect for a treatment difference (D_0.25) of 0.25
Small diamond: line of effect for a treatment difference (D_0.3) of 0.3
Small circle: line of effect for a treatment difference (D_0.4) of 0.4

5.3 The role of selective decontamination of the digestive tract (Moran *et al.* 2003b)

5.3.1 It is nearly 20 years since the first paper describing the implementation of selective decontamination of the digestive tract (SDD) (Stoutenbeek *et al.* 1984) and neither the enthusiasm for testing the hypothesis of SDD efficacy (de Jonge *et al.* 2003b) nor the attendant controversy over this technique (van Saene *et al.* 2003; Kollef 2003; Bonten *et al.* 2003) appears to have waned. The recent publication of the largest (positive outcome) prospective randomized trial of SDD in the literature by de Jonge *et al* (de Jonge *et al.* 2003b) and the claims, primarily based upon two recent meta-analyses (Nathens *et al.* 1999a; D'Amico *et al.* 1998a), that (i) SDD reduces mortality and (ii) "the main reason for SDD not being widely used is the primacy of opinion over evidence" (van Saene *et al.* 2003) would appear to warrant close scrutiny. This requirement for scrutiny appears all the more justified as the recent critiques of SDD (Webb 2000; Kollef 2003; Bonten *et al.* 2003) seemingly accept the reduction in rates of mortality and respiratory tract infection, but base their repudiation of SDD upon arguments regarding real or anticipated rates of antibiotic resistant bacteria, a concern also reiterated by the editorial (Vincent 2003) accompanying the current de Jonge *et al* study (de Jonge *et al.* 2003b).

5.3.2 Meta-analysis overview:

5.3.2.1 Nathens & Marshall 1999 (Nathens *et al.* 1999a). This meta-analysis described a number of patient groups in which SDD had been compared with no prophylaxis; critically ill surgical patients, critically ill medical patients, transplantation, major elective surgery, thermal injury and acute pancreatitis, but consideration will only be given to the first two groups: "surgical", comprising 11 individual studies and "medical" with 10 studies. The "surgical"

group was defined when ≥ 75% of subjects evaluated has been admitted following trauma or surgery (although the Cochrane Database (Liberati *et al.* 2003) identified the Unertl *et al* trial (Unertl *et al.* 1987) as having 52% medical "admission diagnosis") and the "medical" when < 25% met this criterion. Studies using both topical and systemic component SDD and topical component only SDD were combined. No further justification was given for the characterization of these groups and no primary trial results were reported. Nathens & Marshall claimed that SDD effectively reduced (ICU) mortality in "surgical" (odds ratio (OR): 0.70; 95% CI: 0.52-0.93) as opposed to "medical" patients, where no mortality effect was apparent (OR: 0.91; 95% CI: 0.71-1.18). Within the "surgical" group only topical and systemic component SDD was effective (OR: 0.60; 95% CI: 0.41 to 0.88).Of the nosocomial infections identified, SDD effectively reduced the rate of pneumonia and urinary tract infection in both "surgical" and "medical" patients, but bacteraemia was reduced in the "surgical" group only. Of interest, SDD had no impact upon wound infection.

5.3.2.2 D'Amico *et al* 1998 (D'Amico *et al.* 1998a). This meta-analysis identified 34 individual studies or sub-studies and was originally reported in the Cochrane Database in 1997 and updated there in July 2001, although no new studies were identified as of that date (Nathens *et al.* 1999b). In addition to background trial information provided in the Cochrane Database, the grouped data used in the standard meta-analytic analyses was intention-to-treat (provided by contact with original authors) not per-protocol as reported in published original trials, a notable difference from that presumably used by Nathens and Marshall. The underlying strategy of analysis was to consider

studies (n = 17) comparing topical and systemic component SDD with no prophylaxis separately from topical component only SDD (n = 10); a further subgroup of seven trials was also identified where systemic antibiotics had been given to the control group. A significant reduction of (ICU) mortality was apparent only for topical and systemic component SDD (17 studies: OR: 0.80, 95% CI: 0.69 to 0.93; p = 0.007), but interestingly not when topical and systemic component SDD was compared with systemic prophylaxis was given to the control groups (OR = 0.98; 95% CI: 0.73 to 1.32). Individual patient data (IPD) was also available from 25 of 33 trials. When analysis was performed using IPD (Nathens *et al.* 1999c), the treatment effect of topical and systemic component SDD was consistent across the three patient groups identified on admission diagnosis; medical, surgical and trauma, but could not be identified in the (small number of) patients in each group with APACHE II scores > 30. No substantive difference was noted in the results of grouped versus IPD analyses.

5.3.2.2.1 Only nosocomial respiratory infections were considered by D'Amico *et al* and a significant protective effect of both topical and systemic component (OR = 0.35, 95% CI: 0.29 to 0.41) and topical component SDD (OR = 0.56; 95% CI: 0.46 to 0.68), but not when topical and systemic component SDD was compared with systemic component only SDD (OR = 0.81; 95% CI: 0.61 to 1.08).

5.3.3 Critique

5.3.3.1 The studies identified by Nathens and Marshall (Nathens *et al.* 1999a) were the same as those used by D'Amico *et al* with the exception of that of Godard *et al* (Godard *et al.* 1990), a non-randomized trial not included in the D'Amico *et al*

meta-analysis (D'Amico *et al.* 1998a). The utilization of intention-to-treat data versus per-protocol published trial data overcomes potential bias due to under-reporting of patient numbers, a problem identified by D'Amico *et al* (D'Amico *et al.* 1998a) and, in a meta-analysis on immuno-nutrition, by Beale *et al* (Beale *et al.* 1999). Using the intention-to-treat data provided by D'Amico *et al*, there was no significant mortality effect of SDD in the "surgical" group of Nathens and Marshall, with (OR: 0.79; 95% CI: 0.62 to 1.02, p = 0.07) or without (OR: 0.79; 95% CI: 0.611 to 1.02; p = 0.07) the Unertl *et al* trial data ((Unertl *et al.* 1987),see above). Quantitative analysis was performed by the "metan" routine (Bradburn *et al.* 1998), using Stata™ statistical package (D'Amico *et al.* 1998b). This finding is consistent with the results presented by D'Amico *et al* (D'Amico *et al.* 1998a), using individual patient data.

5.3.3.2 Primary mortality end-point of the SDD trials considered in both meta-analyses was invariably that in the ICU. Given recent recommendations for prolonged follow-up in clinical sepsis trials (Cohen *et al.* 2001), the translation of this to improved hospital (Azoulay *et al.* 2003) or post hospital mortality is uncertain. More importantly, in both meta-analyses, only the use of combined topical and systemic therapy in SDD was advantageous with respect to mortality and, as indicated by D'Amico *et al* (D'Amico *et al.* 1998a) and discussed in three recent over-views of SDD (Silvestri *et al.* 2000; Kollef 2003; Bonten *et al.* 2003), the parenteral broad spectrum antibiotic (usually cefotaxime) would appear to be the critical factor in SDD. This can be demonstrated by combining the data from D'Amico *et al* (D'Amico *et al.* 1998a), albeit only in a "suggestive" (non-weighted) analysis:

5.3.3.2.1 24 studies reported the use of topical plus systemic SDD, with a mortality of 23.3% (539 deaths in 2313 patients), as opposed to a mortality of 18.6% in the sub-group of studies (n = 7, see above) where systemic antibiotics were administered to the control group (130 deaths in 699 patients); a significant difference (p = 0.009; two-sided Fisher's exact test)

5.3.3.2.2 29 studies reported no prophylaxis, with a mortality of 28.62% (705 deaths in 2463 patients), as opposed to the mortality of 18.6%, where systemic antibiotics only were administered to the control group; p = 0.001.

5.3.4 Although D'Amico *et al* considered that a trial comparison of combined topical and systemic therapy with systemic therapy alone would be "a logical next step" (D'Amico *et al.* 1998a), Silvestri *et al* (Silvestri *et al.* 2000) resiled from this position to claim that "…the efficacy of SDD, rather than the addition of cefotaxime, determines outcome….". The latter authors cited evidence from a study of traumatic and medical head injury (Ewig *et al.* 1999) where "previous (short-term) antibiotics" were a risk factor (OR 0.2; 95% CI: 0.05 to 0.86) for Gram-negative enteric bacilli and Pseudomonas species colonization within 24 hours of ICU admission. However, "short-term" was defined in the study as "any dose of any antibiotic prior to first sampling" and the CI of the estimate were wide, as were other CI in the study, up to an OR of 128 suggesting unstable or implausible estimates, presumably due to the small sample size (n = 48) and the multiple statistical comparisons.

5.3.5 The mechanism for the SDD protective effect has not been well characterized. The relationship between respiratory infections and (subsequent) ICU death has been described by SDD protagonists as "weak" (Silvestri *et al.* 2000; D'Amico *et al.* 1998a). Kollef (Kollef 2003) suggested that "Trauma and surgical patients

have previously been shown to benefit from the use of systemic antibiotic prophylaxis, including reduced rates of nosocomial infection and improved hospital survival", but the two references offered in proof (Lizan-Garcia *et al.* 1997; Classen *et al.* 1992) refer only to prevention of surgical wound infection by antibiotic prophylaxis and contained no analysis of mortality. Silvestri *et al* (Silvestri *et al.* 2000) noted experimental and cardiovascular by-pass patient evidence of reduction in gut endotoxin and its absorption leading to "recovery of systemic immunity". Nathens & Marshall (Nathens *et al.* 1999a) reported a reduction of bacteraemia, as opposed to other infections, pneumonia and urinary tract, in their "surgical" group compared with the "medical", but offered no pathophysiological explanation as to why this should be; furthermore, the definitions of the two categories of patients would appear to be ad hoc and problematic from an inference point of view. Thus there appears to be a paradox within the SDD paradigm; topical component SDD appears to impact upon nosocomial respiratory infection, which has no direct impact upon mortality, as systemic prophylaxis alone reduces mortality, but has no effect upon nosocomial respiratory infection.

5.3.6 The reduction of nosocomial infection by SDD would have logically been expected to be translated into favourable differences for SDD in proxy outcome measures, such as mechanical ventilation time and / or ICU length of stay. A previous meta-analysis, by Heyland *et al* (Heyland *et al.* 1994), found no difference in ICU length of stay; 15.5 days versus 17.0 days, $p = 0.48$. In their "surgical" group, Nathens & Marshall (Nathens *et al.* 1999a) found a reduction in ICU length of stay (8 studies only; 16.9±13 vs 15.2±12.5 days, *t-test*, $p < 0.05$), which may be expressed as a weighted mean difference (WND): -1.8 days

(favouring SDD); 95% CI: -3.4 to -0.2 days; p = 0.03. However mechanical ventilation time was not reduced by SDD in the "surgical" group (WMD: -0.8 days; 95% CI: -2.1 to +0.4 days, p = 0.2) and in the larger D'Amico *et al* (D'Amico *et al.* 1998a) analysis of topical and systemic SDD vs no prophylaxis, where the mean percentage of surgical patients was 25%, neither ICU length of stay (WMD: -0.9 days; 95%CI: -2.4 to + 0.5 days; p = 0.21) nor mechanical ventilation time (WMD: -1.5 days, 95% CI: -3.0 to +0.1 days; p = 0.06) demonstrated reduction in the SDD arm. However, as noted by Beale *et al* (Beale *et al.* 1999), differences in mortality rates and increased early death rates when overall mortality rates are the same, may confound the interpretation of time-dependent variables such as ventilator days and length of stay; and addressed this potential problem by the use of individual patient data in their analysis and censored for non-survivors. Such an analytic approach has not been repeated within the SDD paradigm, despite individual patient data being available. In individual trials, contradictory results have been found for the impact of SDD upon ICU length of stay for survivors; no influence in two studies (Rocha *et al.* 1992; Korinek *et al.* 1993) and favourable in one (Sanchez *et al.* 1998).

5.3.7 The impact of the costs of SDD has been addressed in a minority of papers; both surveys of SDD practice from SDD protagonists (Silvestri *et al.* 2000; van Saene *et al.* 2003) cite the both the low costs of all non-patent SDD drugs and four studies (Stoutenbeek *et al.* 1996; Sanchez *et al.* 1998; Rocha *et al.* 1992; Korinek *et al.* 1993) which have demonstrated reduced costs per survivor with SDD. However, studies (Verwaest *et al.* 1997; Krueger *et al.* 2002; Gastinne *et*

al. 1992) may be cited which show an increase in costs with SDD and a comparison of these two cohorts is instructive (Table 5.3.7).

Table 5.3.7. Demographic comparisons for costing studies and SDD

Study control	N	Year	AP II	CPn rate	LOS: (%)	LOS: SDD	MVti_SDD control	MVti_
Increased costs								
Sanchez-Garcia	271	1998	26.6	43	16.6	19.9	13.4	16.9
Roacha	151	1992	15.6	46	25.4	26.6	14.2	14.6
Stoutenbeek	91	1996	N/A	19	13.1	16.8	N/A	N/A
Korinek	191	1993	15.6	39	25.4	26.6	14.2	14.6
Decreased costs								
Gastinne	465	1992	13.5	15	18	19	N/A	N/A
Kreuger	660	2002	20.3	10	10	10	4.9	6.4
Verwaest	578	1997	18	22	19.6	18.9	N/A	N/A
Comparator								
Esteban	5183	2002	20			11.2		5.9

N, total study number. Year, year of publication. AP II, APACHE II score. CPn rate, control arm rate of pneumonia. LOS: SDD, mean ICU length of stay for SDD patients (days). LOS: control, mean ICU length of stay for control patients (days). MVti_SDD, mean mechanical ventilation time for SDD patients (days). MVti_control, mean mechanical ventilation time for control patients (days).

5.3.7.1 The studies associated with decreased costs for SDD displayed higher control rates of pneumonia (compared also with the overall control rate of 32% in the D'Amico meta-analysis (D'Amico *et al.* 1998a)) and longer ICU length of stay and mechanical ventilation time than the studies with increased SDD costs. Comparison with the large prospective international cohort study of ventilated ICU patients of Esteban *et al* (Esteban *et al.* 2002) is also informative; mean ICU length of stay and mechanical ventilation time is better approximated to the comparator patients of Esteban *et al* by those studies showing increased SDD costs. As no difference in ICU length of stay and mechanical ventilation time has been demonstrated between SDD and control patients (see above), any cost difference must be assumed to reflect costs of both SDD and the

diagnosis and treatment of (acquired) infections. Thus costs and cost differentials are not fixed by the use per se of SDD, but are functions of variables such as the underlying control rate of pneumonia (and other infections) and length of ventilation.

5.3.8 Uncertainty has also been expressed regarding two fundamental aspects of SDD. Firstly, the individual constituents of the "package", whether this be the topical and systemic antibiotics or topical alone (Kollef 2003). Bonten *et al* further note that a "head-to-head comparison of the complete SDD package vs oropharyngeal decontamination in a randomized fashion has never been performed" (Bonten *et al.* 2003). Secondly, the diagnosis of pneumonia has been clinically based with a variable and uncontrolled use of protected specimens; the latter technique may vary the study diagnosis sensitivity. More disquieting was the reported inverse relationship between the methodological quality score of the SDD studies and the rates of pneumonia (but not mortality), suggesting overly optimistic estimates of SDD benefit (van Nieuwenhoven *et al.* 2001) with respect to nosocomial infection.

5.3.9 De Jonge *et al* study (de Jonge *et al.* 2003b)

5.3.9.1 What then may be said of the latest large study of SDD? Firstly, this was an un-blinded study which showed OR for ICU mortality of 0.59 (95% CI: 0.42 to 0.82) and hospital mortality of 0.71 (95% CI: 0.53 to 0.94). The authors noted the magnitude of improvement in outcome compared with the D'Amico *et al* (D'Amico *et al.* 1998a) meta-analysis (OR for ICU mortality, 0.8, 95% CI: 0.69 to 0.93) and suggested that various modifications of their SDD regimen may have been responsible. Discordance between meta-analyses and (subsequent) large trials are well described (Ioannidis *et al.* 1998a; LeLorier *et*

al. 1997), with overall correlation of treatment effects varying between -0.12 to 0.76 and for primary end-points, 0.50 to 0.76. Un-blinded studies are known to yield exaggerated treatment effects (Schulz *et al.* 1995) and this factor cannot be excluded, given that the two ICUs where the study was performed were co-located and shared staff. Paradoxically and almost counter-intuitively, in the Cochrane Database update of the D'Amico *et al* meta-analysis (Liberati *et al.* 2003), there was no effect of SDD (topical plus systemic) in un-blinded studies (total number of patients = 2568; OR 0.90, 95% CI: 0.74 to 1.08; p = 0.2) compared with double-blind (total number of patients = 1013; OR 0.63, 95% CI: 0.48 to 0.83; p = 0.0009).

5.3.9.2 However, the calculation of the trial sample size for mortality needs further comment. Sample size for what appeared to be the "dominant" primary endpoint, anticipated incidence of colonization with certain prescribed resistant bacteria, yielded "at least 503 patients….in each group" (de Jonge *et al.* 2003b). With this sample size, the 95% CI for the effect on mortality, about a target OR of 0.8 (the OR point estimate of the D'Amico *et al* meta-analysis (D'Amico *et al.* 1998a), above) and 25% mortality in the control group (less than the control mortality of D'Amico *et al* at 28.2%), extends from 0.60 to 1.07; that is, a significant mortality effect cannot be declared. By way of clarification, as the OR metric is not transparent (Sinclair *et al.* 1994), the above trial set-up corresponds to proportions (π) of 25% in the control group (π_2) and 21% in the treatment group (π_1), as OR = $[\pi_2(1-\pi_1)] / [\pi_1(1-\pi_2)]$. A sample size of 875 per group would give a 95% two-sided CI of 0.64 to 1.00 about an OR of 0.8 (computations are based upon the statistical package nQuery Advisor®); thus > 1750 patients in total would be needed to detect a

significant mortality effect at the 0.8 OR level. Following a suggestion of Flather *et al* (Flather *et al.* 1997), the power for the D'Amico *et al* meta-analysis (D'Amico *et al.* 1998a), calculated by conventional means, is 0.79. The OR observed in the de Jonge *et al* study of ICU mortality was 0.59 (95% CI: 0.42 to 0.82) and a sample size of 420 per group would give 95% CI about an OR of 0.42 to 0.83 (de Jonge *et al.* 2003a). However, these are post-hoc calculations which are known to be flawed from a methodological viewpoint. Thus it is perhaps surprising that the observed mortality treatment effect of a trial that was under-powered for one of its primary endpoints was substantially greater than that of a large meta-analysis.

5.3.9.3 The perspective of de Jonge at al (de Jonge *et al.* 2003b) was presumably to conduct their trial based upon all cause mortality. However, some degree of uncertainty would appear to exist regarding the pathophysiological basis this of SDD (see, above). Silvestri *et al* (Silvestri *et al.* 2000), assuming an "association between pneumonia and mortality", a baseline mortality of 30% in a mixed ICU population and 27% of deaths in ICU being "directly attributable to pneumonia", calculated that 2000 to 3000 patients were needed to detect a 10% to 20% mortality reduction; the presumption being that SDD reduced mortality (primarily) via reduction in antecedent respiratory infection. If the mortality effects of de Jonge *et al* trial (de Jonge *et al.* 2003b) are accepted, an alternative estimate of treatment efficacy may be exploited; the Mann-Whitney statistic, which estimates the probability (0 to 1.0) that a randomly selected patient given an innovative therapy will respond better than a randomly selected patient given "standard" treatment (see above). For both ICU and hospital mortality, the Mann-Whitney statistic is 0.54, the interpretation being

that there is a 54 % probability that the (next) randomly selected patient (or, more correctly, one of a pair of patients) on therapy will improve compared with no therapy. In the context of unblinded trials, it is recommended that the Mann-Whitney statistic be reduced by 0.11, which would suggest a less than 50% overall probability of improvement for SDD.

5.3.9.4 With respect to sub-groups, for ICU mortality SDD improved (statistically significant) outcome only in the "urgent surgery" group ($p = 0.02$), whereas for hospital mortality, SDD did not improve outcome in any ("medical" group, $p = 0.07$), albeit the overall point estimates for all groups demonstrated a favourable impact of SDD. Again, no information regarding potential pathophysiologic mechanisms underlying the observed treatment effect (incidence of bacteraemia, respiratory infection) were offered and although ICU length of stay was decreased by SDD, mechanical ventilation time was not reported. The primary end-points of the study were three: acquired colonization by any resistant strain ($p = 0.001$), ICU ($p = 0.002$) and hospital mortality ($p = 0.02$). These end-points were reported simultaneously without statistical adjustment; this is problematic (Chi 1998; Sankoh *et al.* 2003). A number of adjustment procedures are available; the most well-known and simple (but conservative), the Bonferroni, would yield adjusted *P* values (above) of 0.003, 0.006 and 0.06 respectively.

5.3.9.5 Conclusions.

5.3.9.5.1 A number of substantive paradoxes and contradictions have been exhibited within the SDD paradigm. The accompanying editorial to the de Jonge *et al* trial (Vincent 2003) posed the question: "SDD: for everyone, everywhere?" and replied that SDD "worked" (reduced mortality) but was less certain about

its application to “all environments”. An alternate question may be: “What is the future of SDD”, to which a reasoned response may be: “Still uncertain”.

5.4 The role of nutrition as therapeutic intervention (Moran *et al.* 2002c)

5.4.1 A large body of recent literature has directed itself to the impact of nutrition, both enteral and parenteral, on various outcomes in hospitalised patients, including the critically ill (ASPEN Board of Directors and the Clinical Guidelines Task Force 2002; Heyland 2000a). Perhaps since the definitive large scale trial (The Veterans Affairs Total Parenteral Nutrition Cooperative Study Group 1991), which helped define its role some 24 years after the initial description (Dudrick *et al.* 1968), the star of parenteral nutrition (TPN) is waning (Heyland 2000b; Heyland *et al.* 2001a) and enteral nutrition (EN) has displaced parenteral as the preferred form of nutrition in the critically ill (Maynard *et al.* 1991). However, recent reviews have questioned the role and value of nutrition itself (Heyland 2000b; Koretz 1994) and it is perhaps now apposite to consider the treatment effect of nutrition.

5.4.2 What could we mean by nutrition as therapy? This could involve a search for evidence that prescription of “nutrition” to hospitalised patients was associated with improvement in certain patient variables; measured (before / after) biochemical variables, the normalisation of indices of immune function, changes in nitrogen balance or increase in patient muscle strength (Braga *et al.* 1995; Furst 2000; Wagenmakers 2001). Alternatively, evidence could be sought for improvement in key outcome variables, such as mortality, complications, length of stay and cost. With the benefit of the lessons of evidence based medicine, it is

the latter set of outcomes with which we are now most comfortable to demonstrate efficacy of therapy and the first list of patient variables are properly considered surrogate-end points for the latter (Leung 2001; De Gruttola *et al.* 1997).

5.4.2.1 A valid surrogate must not be merely correlated with a true clinical outcome, rather, the effect of the intervention on the surrogate end-point must *predict* the effect on the clinical outcome (Fleming *et al.* 1996). The classic statistical definition of a surrogate end-point was provided by Prentice (Prentice 1989): "..a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based upon the true endpoint". This is a quite restrictive definition and, paradoxically, may be not be able to be "meaningfully tested" (Begg *et al.* 2000). Freedman *et al* (Freedman *et al.* 1992) provided alternative criteria in the form of the proportion explained; the ratio of the treatment effects on the primary end-point, with and without adjustment of the surrogate. A high proportion equated with a useful surrogate. However, this measure has been shown to have considerable variability (Buyse *et al.* 1998; Bycott *et al.* 1998) and, with the realisation that any proper demonstration of the efficacy of a surrogate requires large patient numbers and strong treatment effects, recent attention has been directed to meta-analytic methods for establishing criteria (Gail *et al.* 2000). The current position is one of uncertainty about the establishment of exclusively statistical criteria for (pure) surrogate studies (Leung 2001) and cautions have been expressed regarding such attempts (Gotzsche *et al.* 1996; Fleming *et al.* 1996). That the recent clinical review of surrogate end-points by Bucher *et al* (Bucher *et al.*

1999) offered no precise (quantifiable) criteria for the “resolution of …(this)… scenario” and concluded that “treatment recommendations based on surrogate outcome effects can never be strong” is illustrative of the above concerns. Therefore, further use of such surrogate end-points with respect to “nutrition as therapy” will be for illustrative purposes only. How then to assess the question?

5.4.3 A number of recent reviews have highlighted particular problem areas within the nutrition paradigm; the discourse has been either at the level of the individual study (Buchman 2001; Koretz 1995; Lipman 1998; MacFie 2000) or a qualitative overview (“gestalt”) of the same (Zaloga 1998; Zaloga 1999).The procedure here will be to focus on the assessment of the efficacy of nutrition at the meta-analytic level, looking at the aggregate results (Braunschweig *et al.* 2001) of randomised controlled trials (RCTs). That such a strategy may obfuscate differences between patient subsets must be acknowledged; however, subset or sensitivity analysis, even when protocol specified, is at best, hypothesis generating. Moreover, the reduced size of such subsets with respect to the overall meta-analysis introduces further uncertainty into treatment estimates, although this is rarely commented upon and sub-analyses defined apriori are implicitly given the same status as more powerful aggregated estimates. For instance, in the meta-analysis of early EN in gastro-intestinal surgery (Lewis *et al.* 2001), anastomotic dehiscence rates were the main outcome measure. These rates were reported in only 8 of the 11 studies; in 2 studies the anastomosis was proximal to the site of enteral feeding and in 6 where the anastomosis was distal to the feeding site. Yet the conclusion of the

meta-analysts was that there was “little evidence” that the site of anastomosis relative to the EN feeding site was of importance. Thus, in the context of paucity of data, observed risks may not adequately reflect (true) underlying risk and estimates may be inefficient or biased. This is especially important when considering the significance (p value) of the treatment effect in meta-analyses and patient subsets; the usual variance estimator in meta-analyses is biased (Li *et al.* 1994) and the “standard” test procedure, model choice between fixed and random effects dependent upon diagnosis of heterogeneity, is anti-conservative (type I error rates of nearly 10% if heterogeneity is present (Knapp *et al.* 2000). Estimation techniques which may offer advantage in this context are restricted maximum likelihood estimation (Bockenhoff *et al.* 2000; Thompson *et al.* 1997), which produces conservative standard errors (that is type I error rates close to nominal levels) and full Bayesian analysis (Smith *et al.* 1995) where, in the process of producing posterior estimates of parameters, “strength is borrowed” from larger studies to inform smaller.

5.4.4 Table 5.4.4 shows nutritional effect estimates in 11 meta-analyses (Al Omran *et al.* 2002; Avenell *et al.* 2000; Braunschweig *et al.* 2001; Ferreira *et al.* 2000; Heyland *et al.* 1998; Heyland *et al.* 2001a; Heyland *et al.* 2001b; Lewis *et al.* 2001; Marik *et al.* 2001; Yanagawa *et al.* 2000; Zachos *et al.* 2001). The later meta-analysis of immuno-nutrition by Heyland *et al* (Heyland *et al.* 2001b) is used in preference to a consideration, separately, of the similar theme meta-analyses of Beale *et al* (Beale *et al.* 1999) and Heys *et al* (Heys *et al.* 1999). With respect to nutritional intervention in “general” patients (Table 1, meta-analyses number 1-3), there was no effect on mortality, nor a treatment effect of enteral nutrition on remission in Crohn’s disease, nor on various outcomes in

stable COPD and elderly patients with hip fracture. In the “acutely / critically ill” patient (Table 1, meta-analyses 4-11), no mortality effect was demonstrable from specific nutritional intervention (immuno-nutrition as EN, EN or TPN), timed delivery of EN (early vs late) nor where EN was compared with TPN. Infectious complications were reduced with early delivery of EN, immuno-nutrition and EN versus TPN, but there was significant heterogeneity also demonstrated. No statistical significant advantage of nutritional modality was evident with respect to non-infectious complications, but both early EN (versus late) and immuno-nutrition were associated with a significant reduction of length of stay (LOS), albeit there was again, heterogeneity of effect. It is thus seen that in nearly 10000 patients and over 100 randomised controlled trials, the only consistent aggregate effect was a decrease in infectious complications and length of stay with the provision of enteral nutrition, but, in a minority of trials. Although these two benefits are analytically consistent, what was obvious was a systematic under-reporting of non-mortality outcomes across all meta-analyses (see Table 5.4.4, “effect (study no)”). That this has not been commented upon by the meta-analysts is cause for concern, as the implications of missing data (Piggott 1994) are those of bias and inefficiency in estimation. This problem has been termed “within-study selective reporting of subgroups” (Hutton *et al.* 2000; Hahn *et al.* 2000b). As with publication bias, the most likely explanation of this selective reporting is that of non-significance. When the “missingness” has been addressed by appropriate imputation techniques (Hahn *et al.* 2000b), the initial estimate of effect was seen to be exaggerated. In the absence of covariates an alternative strategy could be meta-regression of treatment effect against baseline risk, as a surrogate for patient severity of illness. However , the uncritical use of

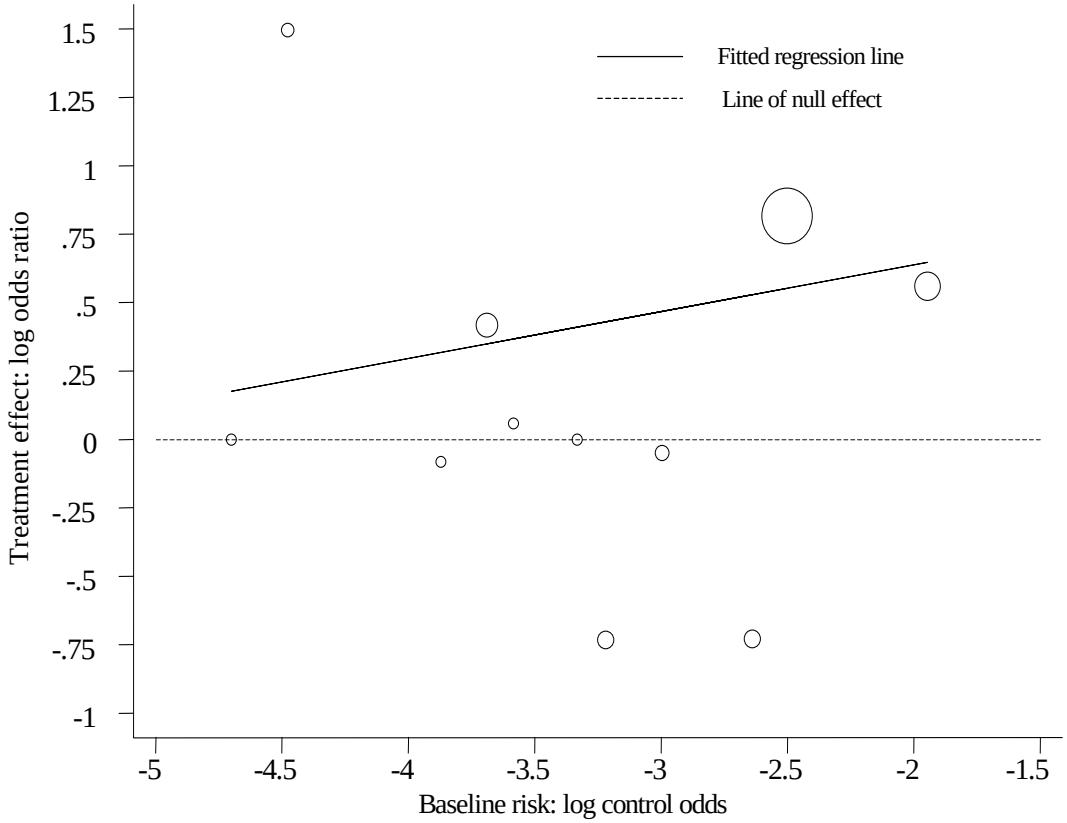
weighted linear regression (Sun *et al.* 1996b), by ignoring regression to the mean, is associated with biased estimates and full Bayesian analysis, producing empirical estimates of the true posterior parameters, is recommended (Sharp *et al.* 2000). Using such a Bayesian approach (Moran *et al.* 2001c), with an uninformative (uniform) prior, a significant relationship was demonstrated (on the log odds scale) between treatment effect and control arm mortality in the Heys *et al* (Heys *et al.* 1999) meta-analysis of immuno-nutrient supplements, as demonstrated in Figure 5.4.4. The fitted regression line shows an increase (> 0) in log odds ratio from low to high baseline risk; that is, as baseline risk (and presumably patient severity of illness) increases the adverse effect on (mortality) outcome produced by immuno-nutrient supplements also increases.

Table 5.4.4

Metan no	Meta-analysis: specific	Ref no	Year	RCT	Patient no	Mortality effect (study no)	Infections effect (study no)	Complications effect (study no)	LOS (days) effect (study no)
1	EN in Crohn's disease remission	38	2002	9	298				
	(a) elemental vs non-elemental diets					no difference: OR 1.15(0.64-2.08)			
	(b) EN vs steroids			4	253	EN less effective OR 0.3(0.17-0.52)			
2	Nutritional supplements in stable COPD	37	2002	9	277	no effect on lung function, exercise tolerance, anthropometric measurements			
3	Nutritional supplements in hip fracture in elderly	35	2002	15	1054	no strong evidence of effect over multiple nutrient supplement types			
4	TPN vs EN:	36	2001	27	1829				
	(a) tube feeding			20	1033	no effect (9) RR 0.96(0.55-1.66)	decreased (15) RR 0.6(0.56-0.79)	*nutritional* no effect (10)* RR 1.36(0.96-1.83) *other* no effect (9) RR 0.92(0.55-1.65)	
	(b) standard care			7	796	no effect (6) RR 1.14(0.69-1.88)	decreased (7)* RR 0.77(0.65-0.91)	*other* no effect (7) RR 0.87(0.74-1.03)	
5	EN: early vs late acutely-ill patients	43	2001	15	753	no effect (6) RR 0.74(0.37-1.48)	decreased (12) RR 0.45(0.3-0.66)	no effect RR 0.82(0.56-1.19)	decreased (12) 2.2(0.81-3.63)
6	EN: early vs late post GIS surgery	42	2001	11	837	no effect (6) RR 0.48(0.18-1.29)	decreased (8) RR 0.72(0.53-0.98)	no effect RR 0.53(0.26-1.08)	decreased (11)* 0.84(0.36-1.33)
7	Immuno-nutrition	40	2001	22	2419	no effect (22) RR 1.10(0.93-1.31)	decreased (18)* RR 0.66(0.54-0.80)	NA	decreased (17)* 3.3(1.025-5.63)
8	TPN vs standard therapy in critically-ill	6	1998	26	2211	no effect (26) RR 1.03(0.81-1.31)	NA	no effect (22)* RR 0.84(0.64-1.09)	no effect (9)* 1.55 (-1.25 to 4.36)
9	TPN vs standard therapy in surgical patients	41	2000	27	2907	no effect (27) RR 0.97(0.76-1.24)	NA	no effect (2)* RR 0.81(0.65-1.01)	no effect (8) 0.65 (-1.82 to 3.11)
10	TPN vs EN: head injury	44	2002						
	(a) early vs late			6	257	no effect (6) RR 0.71(0.43-1.17)	NA	NA	NA
	(b) parenteral vs enteral			5	207	no effect (5) RR 0.66(0.41-1.07)	NA	NA	NA
11	EN vs TPN in pancreatitis	39	2002	2	70	no effect	NA	NA	NA

Ref no = reference number,; Year = year of publication of meta-analysis; RCT = number of randomised controlled trials considered in meta-analysis; Patient no = total patient number in meta-analysis; Mortality = effect of nutritional intervention on mortality, Infections = infection rate recorded in meta-analysis, Complications = complications (non-infective) recorded in meta-analysis analysis; (i) nutritional = complications specific to provision of nutrition, (ii) other = complications not specific to provision of nutrition, effect(study no) = effect of nutritional intervention (number of RCT upon which estimate is based), LOS = length of stay (hospital) in days, RR = risk ratio (point estimate(95% CI)), * = significant heterogeneity recorded in meta-analysis, EN = enteral nutrition, TPN = total parenteral nutrition, COPD = chronic obstructive pulmonary disease, GIS = gastro-intestinal, NA = not available.

Figure 5.4.4. Treatment effect versus baseline risk



Treatment effect (log odds ratio) versus baseline risk (log control odds) for immuno-nutrient supplements using full Bayesian regression analysis. Vertical axis, log odds ratio; +ve log odds, adverse effect; -ve log odds, beneficial effect. Horizontal axis, log control odds; increase in risk as log control odds becomes less negative. Horizontal dotted line, line of null effect (log of odds ratio of 1)
Solid line, regression fit. Circles, individual studies; size of circles inversely proportional to inverse of log odds ratio (large circles correspond to large studies)

5.4.5 Nutritional efficacy was demonstrated in some patient subsets: major complications were significantly lower in malnourished patients with TPN versus standard therapy (Heyland *et al.* 1998), in studies with lower methods scores and when TPN was initiated pre-operatively (Heyland *et al.* 2001a); early versus late EN was associated with a substantive reduction in LOS in trauma / head injured / burn patients (Marik *et al.* 2001); immuno-nutrition formulas with

high (versus low) arginine content appeared to be of advantage with respect to infectious complications and LOS (Heyland *et al.* 2001b); immuno-nutrition reduced infectious complications and LOS in elective surgery versus critically ill patients; in the latter group, LOS was reduced by immuno-nutrition, but, in the absence of a reduction in infections, this was thought to be dependent upon the trend to increased mortality in studies using low arginine EN (Heyland *et al.* 2001b). The status of these sub-analyses has been commented on above; a more consistent approach would be to use (multivariate) meta-regression (Berlin *et al.* 1994). Calendar time dependence of treatment effect, such as produced by dichotomising the trial time span within a meta-analysis (Braunschweig *et al.* 2001; Heyland *et al.* 1998; Heyland *et al.* 2001a), may also be better understood using cumulative meta-analysis (Lau *et al.* 1992). Similarly, the frequent use of trial quality scales (Braunschweig *et al.* 2001; Heyland *et al.* 1998; Heyland *et al.* 2001b; Heyland *et al.* 2001a) to interpret effect estimates ("poorer" quality associated with inflated estimates) presupposes a linear and additive notion of "quality", which notion may be inconsistent (Greenland *et al.* 2001; Juni *et al.* 1999). Finally, given the frequently noted low power of tests for heterogeneity (Fleiss 1986a), the persistent use of a strict 0.05 level of p value (instead of 0.1) for the diagnosis of heterogeneity (Marik *et al.* 2001) will increase the type I error rate and is unjustifiable.

5.4.6 The above cautions on the (over-)interpretation of meta-analytic reviews of nutrition has resonance with the 1995 critique of Koretz on nutrition in the ICU (Koretz 1995), although the promise of "therapeutic foods" would not seem to have fully materialised. What can be said at this juncture is that: (i) there is no evidence that mortality is affected by specific nutritional intervention other than

the obvious intervention to prevent actual starvation (ii) early EN, compared with delayed EN or TPN, is associated with a decrease in the incidence of infectious complications and LOS, but this advantage is substantially confounded by heterogeneity of effect and the consequences of missing data (iii) enthusiasm for the aggressive provision of EN (or TPN) as a therapeutic modality is misplaced. Two recent meta-analyses have further defined the role of EN and TPN in Critical Care practice (Peter *et al.* 2005; Simpsom *et al.* 2005).

6 MULTIVARIABLE ANALYSIS OF PHOSPHATE METABOLISM IN CRITICALLY ILL PATIENTS WITH RESPIRATORY FAILURE

6.1 Introduction: Profound disturbances of inorganic phosphate metabolism have been observed during the course of respiratory illness associated with acute respiratory failure (Fisher *et al.* 1983). In the critically ill ICU patient, variable acute changes have been reported; both admission hyper-phosphataemia and acute hypophosphataemia associated with mechanical ventilation (Laaban *et al.* 1990). A prospective study was thus undertaken to investigate phosphate metabolism in patients over the first 24 hours after admission to the ICU; in particular, the change (delta) of plasma phosphate and red blood cell 2,3-diphosphoglycerate concentration (as an index of intracellular phosphate shift) over this 24 hour period and concomitant renal phosphate handling. The patients investigated were those with: (i) acute respiratory failure associated with a primary respiratory system diagnosis (ii) cardiogenic acute pulmonary oedema and (iii) non-respiratory illness as a primary admission diagnosis. Multivariable linear regression was used to establish quantitative relationships between key physiological predictor variables of delta phosphate. Such a statistical strategy would normally be considered relatively non-controversial. However, some pertinent questions may be highlighted in the course of analysis: what is the appropriate analytic form of change (Harrell Jr 1997; Kaiser 1989); what are the effects of missing data (Little 1992); is model variable selection consistent (Derksen *et al.* 1992) and are there advantages to other statistical models compared with the usual form of ordinary least squares linear regression (OLS) (Hardin *et al.* 2001)? That some answers to the above questions involve relatively

unfamiliar statistical techniques, may reflect the lag phenomenon of statistical method transfer into the medical literature, referred to above (Altman *et al.* 1994b) and we note recent similar contributions in the biostatistical literature (Carlin *et al.* 2001).

6.2 Methods:

6.2.1 Full details of patient definitions and biochemical analytic techniques have been reported (Moran *et al.* 2002e) and a brief summary is provided here: patients were enrolled on admission to the TQEH ICU with diagnoses of: (i) acute respiratory failure with an intrinsic respiratory system diagnosis (Group I) (ii) cardiogenic acute pulmonary oedema (Group II) (iii) patients admitted to the ICU and acting as controls (Group III), with primary diagnoses other than respiratory failure and/or cardiogenic pulmonary oedema. Plasma phosphate reference range was 0.8-1.45 mmol/L; hypophosphataemia was classified as mild (plasma concentration 0.61-0.8 mmol/L), moderate (0.32-0.6 mmol/L) and severe (< 0.32 mmol/L). Hyperphosphataemia was defined as a plasma concentration exceeded 1.45 mmol/l. Patients were observed over a 24 hour period and intravenous fluids as 0.45% or 0.9% saline only were prescribed for the first 24 hours in the ICU, to avoid hypophosphataemia due to concomitant infusion of glucose-containing fluids . No phosphate or magnesium-containing fluids were given during this time period, but potassium was supplemented in maintenance fluids to maintain plasma potassium ≥ 4.0 mmol/L. At study entry (ICU-admission, T_0) and at 24 hours post ICU admission (T_{24}) arterial blood specimens were taken for measurement of (i) plasma biochemical variables (ii) arterial blood gases (iii) plasma lactate (iv) intact parathyroid hormone (v) red blood cell 2,3-diphosphoglycerate (2,3-DPG); concentration was determined

after the method described by Luzzato; reference range 9.4-16.9 umol/g of haemoglobin. At T_0 a random urine sample was taken followed by a 24 hour urine collection; both specimens were assayed for concentrations of sodium, potassium, creatinine, calcium, magnesium and phosphate. Between 1000 to 1200 hours on the day after admission (24 hour clock), a two hour urine collection and arterial blood sample at mid-point were taken. Time from admission to beginning of two hour collection was recorded (time delay of collection for real threshold phosphate).

6.2.2 Urine and plasma concentrations of creatinine and phosphate urine volume were measured.

6.2.2.1 Clearance of creatinine and phosphate were calculated according to the formula:

$$\text{Clearance(ml/min)}=\left(\frac{\text{Urine concentration }(\text{mmol/l})}{\text{plasma concentration}}\right)\text{x}\left(\text{urine volume (ml)}\right)/120\,(\text{min})$$

6.2.2.2 Renal threshold phosphate concentration (RTP) was calculated using the Walton and Bijvoet nomogram (Walton *et al.* 1975).

6.2.2.3 Fractional excretion of phosphate (FE_{PO4}) was calculated according to the formula: $FE_{PO4} = [(U/P)_{PO4}/(U/P)_{Cr}] \times 100$, where U and P represent concentrations in urine and plasma; PO_4=phosphate and Cr = creatinine (Zarich *et al.* 1985).

6.2.3 An APACHE II score was recorded over the first 24 hours. Patient details, pertinent to the study, were separately recorded: diagnosis, age, sex, mechanical ventilation and prescription of diuretic, (parenteral) corticosteroids, aminophylline and sympathomimetic agents.

6.3 Statistical methodology:

6.3.1 The appropriate analytic form for delta phosphate was investigated by regression of candidate variables:

change (T_{24}-T_0 phosphate)

percentage change ([T_{24}-T_0 phosphate] / T_0 phosphate*100)

ratio (T_{24} phosphate / T_0 phosphate)

log ratio (log [T_{24} phosphate / T_0 phosphate] = [log(T_{24} phosphate) – log(T_0 phosphate)])

against both *initial phosphate* (T_0 phosphate) and *average phosphate* ([T_0 phosphate +T_{24} phosphate]/2), as suggested by Harrell (Harrell Jr 1997), to find a dependent variable not related to initial values (T_0 phosphate). Normality was assessed by kernel density estimation and specific tests (Shapiro-Wilk) (Fox 1990).

6.3.2 To establish the propriety of change vs percentage change, Kaiser's R (Kaiser 1989) was also calculated:

$$R = \left(\overline{X_g}\right)^2 \sum_{ij}\left[\left(P_{ij} - \overline{P_i}\right)/100\right]^2 / \sum_{ij}\left(C_{ij} - \overline{C_i}\right)^2$$

with $\overline{X}_g$ the geometric mean of T_0 phosphate and $\overline{C}_i$ and $\overline{P}_i$ the arithmetic means of the *change* and *percentage change* in group *i*, respectively. If R >1, *change* is suggested as the appropriate form of delta phosphate; if R <1, *percentage change*.

6.3.3 The effect of regression to the mean on the β coefficient of the of the *change* vs *initial phosphate* relationship, was computed after Blomqvist's method: $\hat{\beta} = [\hat{\beta}' + (1-\hat{\rho})]/\hat{\rho}$, where $\hat{\beta}'$ is the regression slope of *change* on *initial phosphate*, as derived from the data-set and $\hat{\rho} = 1 - \hat{\delta}^2 / s_x^2$, where s_x^2 is the

observed variance of measured initial (T_0) values and $\hat{\delta}^2$ is an estimate from an "external" source of within-subject variance (Blomqvist *et al.* 1978; Hayes 1988). Estimates of $\hat{\delta}^2$ were obtained from computerised records of patients admitted to the TQEH ICU; duplicate values of phosphate were obtained at admission from patients admitted from 1998-1999, where the difference between time of admission and time of duplicate observation was < 3 hours.

6.3.4 Predictor variables were defined by a backward selection process (Mantel 1970) from the full model (19 variables) using Akaike information criterion ($\mathrm{AIC} = -2\ln\hat{L}(M_k) + 2P$ (Lindsey *et al.* 1998), where $\hat{L}(M_k)$ is the likelihood of the model and *P* is the number of parameters in the model), corresponding to a nominal p value of 0.157 for stepwise selection (Sauerbrei 1999). Consistency of variable selection was investigated by using bootstrap (1000 samples, (Carpenter *et al.* 2000)) of the backward selection with p=0.157, using a user written bootstrap technique in the statistical program Stata™ and recording the variables selected with a frequency > 50% in the OLS regression (Chen *et al.* 1985; Mick *et al.* 1994; Sauerbrei 1999). Multi-colinearity (variance inflation factors [VIF] < 10 and condition number < 15 (Chatterjee *et al.* 2000)), interactions (adjusted for multiple comparisons after Holm (Holm 1979)) and predictor non-linearity were also investigated (Weesie 1999). Where non-linearity was evident, covariate effect was modelled using fractional polynomials (Sauerbrei *et al.* 1999). Specific attention was paid to the selection problems with highly correlated ($r \geq 0.5$) variables (Fried *et al.* 2001): (arterial) pH and $PaC0_2$; fractional excretion of phosphate; 24 hour urine phosphate excretion; and renal threshold phosphate. Models assessed were: OLS and a generalized linear model (GLM) with the log link (Hardin *et al.* 2001). GLMs

are empirical transforms of the classical linear (Gaussian) regression model and are distinguished from OLS by particular model, rather than data, transformations: specifically, a response distribution of one of the exponential family of distributions (normal, Poisson, gamma, binomial, inverse Gaussian) and a (monotonic) link function (identity, logarithmic, square root, logistic, power) which relates the mean of the response to a scale on which the model effects combine additively (Myers *et al.* 1997).

6.3.5 Models were compared using estimates of fit (Hardin *et al.* 2001; Long *et al.* 2000; Sheiner *et al.* 1981):

$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$, where RSS= residual sum of squares and TSS is total sum of squares

adjusted R^2, $R_a^2 = 1 - \left(1 - R^2\right)\left(n - c\right)/\left(n - k\right)$, where n is observation number, k is number of parameters (including constant) and c =1 if a constant in model, =0 otherwise

AIC, as defined above (smaller values indicating preferred models)

adjusted likelihood-ratio index, $R_{alri}^2 = 1 - \frac{L\left(M_\beta\right) - k}{L\left(M_\alpha\right)}$, where $L\left(M_\beta\right)$ is the likelihood of the model with intercept and predictors, $L\left(M_\alpha\right)$ is the likelihood of the model with intercept only and k is the number of predictors.

6.3.6 Missing values: a set of key predictor variables in addition to the candidate variables for delta phosphate were defined and any missing value pattern was identified (Horton *et al.* 2001). The data set was tested for “missingness” being completely at random (MCAR; that is, the missing values were a random sub sample of the entire data-set) (Little 1988a), utilising the missing value module

provided in Systat® statistical software (Schafer 2001a). Subsequently, a multiple imputation approach after Schafer was adopted (Schafer 2001b). Firstly, a "complete" data set of key predictors of delta phosphate was generated via the expectation-maximisation (EM) algorithm (Dempster *et al.* 1977) and this data set was used in variable and model selection (see iv, above). Secondly, *k* (=10) imputed data sets were created using the technique of data-augmentation for a final multivariable model with appropriate point estimates (average of the k estimates) and standard errors (square root of the average within data set and across data set variances, multiplied by a bias correction factor because $k < \infty$) . Comparisons between (i) the set of mean values of the initial, EM and multiple imputed data sets and (ii) parameter point estimates and standard errors of the OLS and GLM models for the initial and multiple imputed data sets were also undertaken.

6.4 Results:

6.4.1 Fifty seven patients, 32 males and 25 females, of mean±SD age 67±12 years and Apache II score 22±6, of whom 32 were ventilated, were enrolled into the study. In Group I, patient diagnoses (n=) were: exacerbation of COPD with acute respiratory failure, 20; pneumonia, 10; status asthmaticus, 2. Group II comprised 10 patients with acute pulmonary oedema; no patient had a diagnosis of myocardial infarct based upon 12 lead electrocardiogram and creatine kinase estimation. In Group III diagnoses were: postoperative abdominal aortic aneurysm, 3; gastrointestinal perforation, 3; septic shock, 4; post cardiac arrest, 3; self-ingestion, 1 and subarachnoid haemorrhage, 1. Further details of the 3 diagnostic groups are given in Table 6.4.1. Groups were comparable for age, sex, ventilatory status and Apache II scores, but differed in frequency of

prescription of therapeutic agents: parenteral diuretic (frusemide), corticosteroids (intravenous hydrocortisone), aminophylline and sympathomimetics (intravenous adrenaline and salbutamol used only).

Table 6.4.1. Patient diagnostic groups

	Group I (Respiratory)	Group II (Cardiac)	Group III (Mixed)	*p*
Number (n =)	32	10	15	
Age (years)	67±11	67±12	68±16	NS
APACHE II score	20±5	24±3	24±9	NS
Sex (M / F)	18/14	6/4	8/7	NS
Ventilated	18/14	6/4	11/4	NS
Diuretic	9/23	10/0	4/11	0.0001
Steroid	23/9	1/9	3/12	0.0001
Aminophylline	26/6	3/7	1/14	0.0000
Sympatho-mimetics	4/28	4/6	5/10	0.0002

Values for age and Apache II score are given as mean±SD. M =male, F = female. For variables Ventilated, Diuretic, Steroid, Aminophylline and Sympatho-mimetics; numbers indicate “yes” / “no”.

6.4.2 Initial (T_0) values for variables in the three groups are shown in Table 6.4.2a. At T_0, 3 patients (one in each Group) had mild hypophosphataemia, 25 (17 in Group I and 8 in Group III) had normal levels and 29 (14 in Group I, 9 in Group II and 6 in Group III) were hyperphosphataemic. All groups showed a 24 hour decrement of plasma phosphate (non-significant in Group III). Red blood cell 2,3-diphosphoglycerate concentrations showed 24 hour increments in Groups I and II, but a slight decrease in Group III; over all groups, the change was non-significant. Arterial blood gas variables showed initial acidaemia in all groups and hypercapnia in Group I; all variables tended to normalise over the 24 hour observation period. Variables for which intergroup differences existed were: plasma lactate (T_0), parathyroid hormone (T_0 and T_{24}-T_0 value) and glucose (T_0 and T_{24}-T_0 value). Statistically significant differences (T_{24}-T_0) were found for

the following variables: phosphate, pH, $PaCO_2$, bicarbonate, lactate and glucose (Table 6.4.2b). At T_{24}, 15 patients (26%) had hypophosphataemia; 13 in Group I and one patient each in Groups II and III. Of these 15 patients, 9 had mild, 4 had moderate and 2, both Group I, had severe hypophosphataemia.

6.4.3 Renal threshold phosphate. Twenty four hour urinary phosphate and sodium excretion showed no significant difference between groups, whether expressed as absolute amount (mmol/24 hours) or relative to urine volume and/or creatinine excretion per 24 hours (data not shown). Two-way analysis of variance revealed no diagnostic-group / frusemide prescription effect with respect to 24 hour urine sodium excretion. Time delay of collection before assessment of renal threshold phosphate was 16(6) hours and no relationship was demonstrated between renal threshold phosphate and this time delay ($p = 0.11$; no evidence of non-linear relationship). Renal threshold phosphate concentration showed a reduction below reference range in Groups I and II (0.65 ± 0.29 and 0.57 ± 0.29, respectively); values were significantly lower for both groups compared with Group III (1.09 ± 0.50, $p=0.001$). Of the 15 patients hypophosphataemic at T_{24}, 13 had renal threshold phosphate concentrations < 0.8 mmol/L (lower limit normal range), compared with 20 of 40 patients who were normo- or hyperphosphataemic (Fisher exact, $p= 0.01$).

Table 6.4.2a. Initial (T_0) values (mean±SD) for group variables.

Variable	Group I	Group II	Group III
Phosphate (0.8-1.45 umol/l)	1.45±0.66	2.06±0.81	1.42±0.66
2,3-DPG (9.4-16.9*u*mol/gHb)	13.5±.85	10.85±2.68	14.67±3.81
pH (7.36-7.44)	7.31±0.14	7.23±0.12	7.36±0.17
$PaCO_2$ (35-45mmHg)	62±32$^{\Delta}$	47±10	37±11
Bicarbonate (22-30mmol/L)	28±8$^{\Delta}$	19±4	22±4
PaO_2/FIO_2	227±116	143±76	256±150
Lactate (0.5-2.0mmol/L)	2.4±1.5*	6.5±5.3*	3.7±3.2
Parathyroid hormone (1.0-7.0*p*mol/L)	5.7±4.4**	20.7±19.4$^{**\#\#}$	4.2±2.0$^{\#\#}$
Glucose (3.0-5.0mmol/L)	11.4±5.1*	20.2±11.7$^{*\#}$	10.6±4.7$^{\#}$
Calcium⊕ (2.10-2.60mmol/L)	2.31±0.16	2.24±0.14	2.39±0.21
Creatinine (0.05-0.12mmol/L)	0.093±0.036*	0.201±0.151$^{*\#}$	0.115±0.054$^{\#}$
Urine phosphate (mmol/L)	21.3±21.4	14.6±6.5	14.7±10.2
Urine sodium (mmol/l)	71±41	85±32	66±46
24hour urine phosphate (16-48mmol/24h)	22.5±14.0	22.7±10.6	19.9±16.6
24hour urine sodium (mmol/24hr)	130±103	218±183	117±129
RTP (n=55) (0.8-1.35mmol/L)	0.65±0.29	0.57±0.29	1.09±0.50$^{\Delta\Delta}$
FE_{PO4} (6-20%)	27±22	35±23	19±13

(***** or **#** , p<0.01 and ****** or **##**, p<0.001) indicate significant differences in pairwise comparisons (across groups); (**Δ**, p<0.01and **ΔΔ** , p<0.001) indicate single group difference with respect to two other groups. ⊕ Calcium levels corrected for plasma albumin by 0.02 mmol/L per g of albumin (to a value of 40g/L). 2,3-DPG, Red blood cell 2,3-diphosphoglycerate. FE_{PO4}, fractional excretion of phosphate. RTP, renal threshold phosphate concentration. No adjustments were made for multiple comparisons.

Table 6.4.2b Initial (T_0) and 24 hour (T_{24}) values (mean±SD) for variables; all patients.

Variable	T_0	T_{24}	*p*
Phosphate (0.8-1.45 umol/l)	1.55±0.71	1.00±3.0	<0.0001
Red cell 2,3-diphosphoglycerate (9.4-16.9*u*mol/gHb)	13.5±3.3	14.1±3.0	NS
pH (7.36-7.44)	7.31±0.15	7.42±0.07	<0.0001
$PaCO_2$ (35-45mmHg)	53±7	44±15	0.008
Bicarbonate (22-30mmol/L)	25±8	27±7	<0.0001
PaO_2/FIO_2	222±122	220±111	NS
Lactate (0.5-2.0mmol/L)	3.4±3.1	1.8±1.0	<0.001
Parathyroid hormone (1.0-7.0*p*mol/L)	7.8±10.1	7.3±7.05	NS
Glucose (3.0-5.0mmol/L)	12.7±7.4	8.3±2.3	<0.0001
Calcium ⊕ (2.10-2.60mmol/L)	2.32±0.18	2.31±0.19	NS
Creatinine (0.05-0.12mmol/L)	0.12±0.08	0.108±0.08	NS
T_0 urine phosphate (mmol/L)	18.3±17.1		
T_0 urine sodium (mmol/L)	73±41		
24hr urine phosphate (16-48mmol/24h)	21.8±14.0		
24hr urine phosphate/ mmol creatinine (1.8-2.7 mmol/mmolCr/24h)	2.8±1.7		
24h urine sodium (mmol/24h)	142±129		
RTP (0.8-1.35mmol/L)	0.76±0.39		
FE_{PO4} (6-20%)	26±20		

Calcium levels corrected for plasma albumin by 0.02 mmol/L per g of albumin (to a value of 40g/L). RTP, renal threshold phosphate concentration. 2,3-diphosphoglycerate, Red blood cell 2,3-diphosphoglycerate. FE_{PO4}, fractional excretion of phosphate . *P* refers to significance of *t-test* for T_{24}-T_0. No adjustments were made for multiple comparisons.

6.4.4 Delta phosphate: Phosphate change was significantly related to both initial (phosphate change = 0.673-0.787*T_0PO_4; R^2= 0.72, p= 0.0001) and average phosphate (phosphate change = 0.522-0.836*average phosphate; R^2 = 0.35, p= 0.0001). Kaiser's R, computed at 0.73, favoured percentage change as a predictor. Percentage change, ratio and log-ratio demonstrated a significant relation to T_0 phosphate, but no relation (p >0.12 for all) to average phosphate suggesting an independence from initial values for this variable. However, only log-ratio phosphate was normally distributed and was thus adopted as the appropriate form of Δ phosphate.

6.4.5 The confounding effect of regression to the mean was quantified by the use of Blomqvist's adjustment of the regression relationship change vs T_0 phosphate . The initial relation was: phosphate change=0.673-0.787(T_0PO_4); p=0.000, R^2=0.72, RMSE=0.35. s_x^2(T_0 variance) was 0.51 and $\hat{\delta}^2$ (within-subject "external" variance) was 0.243 yielding a corrected $\hat{\beta}$ of -0.593, a 25% reduction in effect.

6.4.6 Predictive models. Using log ratio T_{24}/T_0 phosphate as the form of delta phosphate (the dependent variable), the significant predictors initially identified by backward selection were delta 2,3-diphosphoglycerate, delta pH and renal threshold phosphate concentration. No effect of time delay of collection before renal threshold measurement was evident (p=0.58). Further exploration, guided by theoretical considerations, identified a significant (p=0.003) interaction between the prescription of aminophylline and renal threshold phosphate concentration No multi-colinearity was present. By bootstrap of the backward selection process, the variables selected (frequency > 50%) were delta 2,3-diphosphoglycerate, delta pH, renal threshold phosphate concentration and

aminophylline; this occurred in both the initial and complete (EM) data sets. Parameter estimates with standard errors (SE) and p values and indices of model fit (R^2, adjusted R^2, AIC and R^2_{alri}) for the various models are seen in Table 6.4.6. Non-linearity of covariate effect was not demonstrated.

Table 6.4.6. Parameter and SE estimates with performance indices of multivariable models

Model Parameter (SE)	OLS	p	GLM: log link	p	GLM: eform
del 23dpg	0.043(0.024)	0.07	0.057(0.03)	0.06	1.059(0.032)
del pH	-2.01(0.344)	0.0001	-2.715(0.598)	0.0001	0.066(0.04)
RTP	0.965(0.30)	0.003	0.740(0.364)	0.04	2.10(0.763)
aminophylline	0.836(0.261)	0.003	0.611(0.351)	0.08	1.841(0.646)
amin_rtp	-0.99(0.349)	0.007	-0.771(0.408)	0.05	0.462(0.188)
R^2	0.65		0.6		
R^2 adj	0.6		0.55		
AIC	32.2		25.4		
R^2_{alri}	0.52		0.5		

del, difference (T_{24}-T_0). 23dpg, , Red blood cell 2,3-diphosphoglycerate. RTP, renal threshold phosphate concentration. amin_rtp, interaction between aminophylline & RTP. OLS, ordinary least squares. GLM, generalized linear model. eform, exponentiated form.

6.4.7 Missing data: The initial variables considered in the full data set were 19 in number and missing values occurred in 10 (all continuous variables); the frequency of missing values (per variable) ranging from 2 to 17%. The *P*-value for Little's MCAR test (Little 1988b) was 0.8, indicating that the missing data was in fact a random sub-sample of the data-set. Model parameters using multiple imputation are seen in Table 6.4.7; parameter estimates and standard errors were consistent with those of the original data-set. Because of missing values in the initial data set, observation number (n) for the multivariable models was reduced to 46 compared with 57 for the complete (EM) and multiple imputation data sets. No difference existed between the set of means of

(continuous) key predictor variables of the initial data set and (i) the complete EM data set (Hotelling's T-squared test) and (ii) the k = 10 multiply imputed data sets (*t-test*, data not shown). Data summaries and figures are therefore reported from the initial data set.

Table 6.4.7: Parameter and SE estimates of multiple imputation data sets (k=10).

Model Parameter (SE)	OLS	p	GLM: log link	p
del 23dpg	0.047(0.021)	0.03	0.052(0.024)	0.03
del pH	-2.048(0.308)	0.0001	-2.720(0.483)	0.001
RTP	0.914(0.266)	0.001	0.656(0.296)	0.03
aminophylline	0.687(0.229)	0.003	0.481(0.286)	0.08
amin_rtp	-0.847(0.181)	0.004	-0.65(0.335)	0.05

del, difference (T_{24}-T_0). 23dpg, , Red blood cell 2,3-diphosphoglycerate. RTP, renal threshold phosphate concentration. amin_rtp, interaction between aminophylline & RTP. OLS, ordinary least squares. GLM, generalized linear model.

6.5 Discussion

6.5.1 Plasma phosphate and respiratory illness:

Decrease in plasma phosphate over the initial treatment period of acute respiratory failure has been previously noted, especially with mechanical ventilation and was evident in the current study, being significant (compared with T_0 values) in Groups I and II only. Delta pH was a significant independent variable in this study (see below) and mechanical ventilation, postulated as a “key” factor in previous studies, is thus a surrogate for the (induced) change of delta pH (and delta $PaCO_2$, with which delta pH was obviously highly correlated; rho = -0.84, p=0.0001). The effects of other potentially important predictors must also be considered: (i) pre-admission (treatment) variables; although these may have affected phosphate metabolism, there was no group difference in time-zero urine phosphate or sodium excretion

(expressed absolutely, Table 2, or relative to creatinine excretion, data not shown) (ii) extracellular volume change; possible group differences were not reflected in any urinary excretion index nor in the effect of administered diuretics (iii) mechanical ventilation; the effect of different modes of ventilatory support was not explored (iv) gastro-intestinal phosphate losses, presumed small, were not measured (v) the effect of endogenous stress hormones such as insulin and adrenaline was not quantified. Potentially dissimilar patients were also combined in Group I, but this was thought reasonable as there were no differences in the set of mean values between the patients with exacerbation of COPD and pneumonia (Hotelling's $T^2 = 64.37$, $p < 0.89$). A detailed consideration of phosphate metabolism (not the focus of this thesis) proceeding from these results is found in Moran *et al* (Moran *et al.* 2002e).

6.5.2 Statistical considerations

6.5.2.1 Dependent variable(s): a considerable literature has addressed the appropriate form “change” (Gill *et al.* 1985; Harrell Jr 1997; Hayes 1988; Oldham 1962). Initial analysis was not unreasonably directed at finding a measure of change that was independent of the initial value (T_0 phosphate). Plots of candidate measures against both initial and average phosphate, as recommended by various authors, were used to overcome the intrinsic mathematical relationship involved in analysing raw change variables; in particular, the simple difference (T_{24}-T_0). In this study, the measure adopted as the dependent variable was the log ratio (log T_{24} / T_0 phosphate) which, being normally distributed was considered the most appropriate dependent variable for linear regression. As Harrell (Harrell Jr 1997) has noted, the above endeavour may be difficult

because of regression to the mean; a significant relation of the dependent variable to initial phosphate, but not to the average, may suggest such regression to the mean. Hayes, reviewing measures to adjust for the potential effects of regression to the mean, found Blomqvist's adjustment the most robust (Hayes 1988). In the current study the "effect" of regression to the mean was to inflate by 25% the coefficient for the regression relation, phosphate change versus initial phosphate. An alternate analysis could have considered T_{24} phosphate as the dependent regression variable, with T_0 phosphate as baseline, in an ANCOVA analysis (Frison *et al.* 1992). Such a strategy would, however, have yielded the same analytic results as the use of phosphate change as the dependent variable (Senn 1994b). Baseline corrections have a significant impact when the correlation, "baseline" vs "current value", is > 0.5 (Senn 1989b). In the current study, the correlation T_{24}:T_0 phosphate was 0.26.

6.5.2.2 Missing data: in the presence of missing values, multivariable analysis is usually accompanied by complete-case analysis. That is, only complete observations are considered across variables, resulting in a decrease in the total *n* and potential bias and / or loss of efficiency in estimation. Recent recommendations on the conduct of multivariable analysis have been unusually silent on appropriate statistical procedures to deal with missing values (Concato *et al.* 1993). Few studies actually comment upon adjustments for missing data; straightforward strategies such as normal value replacement or mean substitution have been criticised (Harrell, Jr. 2001; Schafer *et al.* 1998). We would presume that missing values in the data-sets reported in the (medical) literature have been a ubiquitous phenomenon, but only in the last 10 years has formal attention been drawn in this literature to the problem, despite

statistical techniques being available for nearly 20 years (Rubin 1996). There has been a recent emphasis on (i) the uncertainty aspect of imputation, to the extent that analysis ignoring such uncertainty will generate inappropriately small standard errors and p values and rates of Type I error higher than nominal levels, and (ii) the notion that missing values are a source of variability to be averaged over. A maximum likelihood approach to missing value imputation, using the EM algorithm, is more convenient (involving only one data-set), but it is a deterministic (non-random) methodology subject to the cautions above. In multiple imputation, each missing value is replaced by a set of m > 1 plausible values drawn from their predictive distribution. The efficiency of an estimate based upon m imputations is approximately $\left(1+\frac{\gamma}{m}\right)^{-1}$ where γ is the fraction of missing information for the quantity being estimated (Schafer *et al.* 1998). The Little test for MCAR (albeit a test with relatively low power) was non-significant in this data set, suggesting that there would be minimal bias in the parameter estimates from conventional analysis and such was generally found (Table 6.3.6.7). Thus effective reduction of observation number to n = 46 in variable selection and model construction in the initial data set appeared to relatively robust to “missingness”. That this may not be the case in other data sets is, of course, an empiric question.

6.5.2.3 Competing models: The three variables which were demonstrated to predict log ratio phosphate were consistent from the physiologic viewpoint. The full model (all predictors considered) had an R^2 of 0.68, but an adjusted R^2 of only 0.36, suggesting that a more parsimonious model was appropriate. Considerable controversy has attended the question of selection of variables in

statistical models using “automatic” techniques (Derksen *et al.* 1992; Harrell, Jr. 2001); formal procedures have been established for incorporating the bootstrap into variable selection (Shao 1996). In the current paper, we used bootstrap of the selection process and a heuristic rule as confirmatory test for the initial AIC-guided backward selection.

6.5.2.4 Parameter interpretation: the OLS model used a log-transformed dependent variable (log ratio T_{24}/T_0 phosphate); fitted values and parameter estimates were in terms of the log response (the geometric mean). Natural log differences (log [T_{24} phosphate / T_0 phosphate] = [log(T_{24} phosphate) – log(T_0 phosphate)]) correspond to fractional differences on the original scale; that is, the expected value will change by $100(\exp(\beta_j)-1)\%$ for each 1-unit change in the independent variable x_j (β being the appropriate regression coefficient), regardless of whether x_j is continuous or dichotomous (Cole 2000; Zhou *et al.* 2001b). Appropriate back-transformation (Duan’s smearing estimate) avoids bias in predicted values from either a “simple” exponentiation or a naïve back transformation (exponential transform + 0.5*[MSE]; where MSE = mean squared error of estimation) (Duan 1983; Manning 1998); as seen in the comparison of the descriptive statistics of the various forms (original data and model predictions) of the phosphate ratio in Table 6.5.2.4.

Table 6.5.2.4. Summary statistics of T_{24} / T_0 ratio and predictions from OLS and GLM equations

	PO_4 T_{24}/T_0 Ratio	OLS exp trans	OLS naïve transform	OLS Duan smear est	GLM loglink
n	46	46	46	46	46
mean	0.78	0.74	0.78	0.77	0.77
SD	0.42	0.28	0.29	0.29	0.31
min	0.18	0.26	0.28	0.27	0.14
p25	0.48	0.51	0.54	0.54	0.56
p50	0.75	0.76	0.8	0.79	0.75
p75	0.93	0.93	0.98	0.97	0.94
max	2.44	1.23	1.3	1.29	1.47

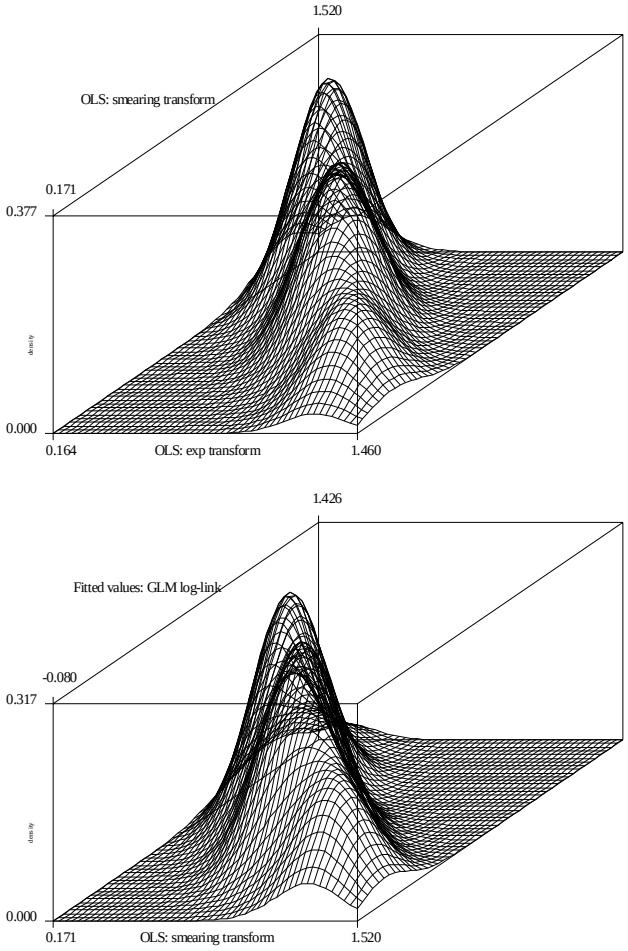
n = number of observations; mean = mean; SD = standard deviation; min = minimum value; p25 = 25th percentile; p50 = median; p75 = 75th percentile, max = maximum observation. PO_4, phosphate. . OLS, ordinary least squares. OLS exp transform, OLS with exponential back-transformation. OLS Duan smear est, OLS with Duan's smearing estimate. GLM, generalized linear model.

This comparison is further highlighted using bivariate kernel density and wire-frame plots of the predictions: OLS with exponential and smearing transforms and OLS with smearing transform and GLM log link, as seen in Figure 6.5.2.4.

A particular advantage of the GLM log-link model is that it provides estimates of the (exponential) conditional mean function (ratio T_{24}/T_0 phosphate). The GLM log-link logs the predictor ($x\beta$), rather than the response, to linearize the relationship between response and predictors. That is, if we consider a transformation (say, logarithmic) *g*, then the expectation (*E*) of classic linear model has the form: $E\left\{g\left(Y_i\right)\right\} = a + x\beta$, whereas the GLM has the form: $g\left\{E\left(Y\right)\right\} = a + x\beta$ (Firth 1991). Thus parameters are equal to the logs of arithmetic means and their ratios (the ratios being for either groups defined by discrete predictors, in this case "aminophylline, yes / no", or changes in response to a unit increase in a continuous predictor). The original arithmetic means and ratios are given by the exponential form (Table 6.4.6.; GLM: eform). In this sense, the GLM log-link model provides more convenient and intuitive estimates

than the traditional log-transformed OLS. Other GLM models could also be considered: gamma and inverse Gaussian families with an identity or log link (Hardin *et al.* 2001).

Figure 6.5.2.4: Bivariate kernel density and wire-frameplot



Upper graph: plot of predictions of OLS with smearing back-transform versus predictions of OLS with exponential (=exp) back-transform. Lower graph: plot of fitted values from GLM model with log-link versus OLS with smearing back-transformation

6.5.2.5 Model uncertainty: the traditional approach to modelling would appear to assume (by omission) that, although parameter estimates are imbued with a degree of uncertainty (95% CIs), there exists fixity in terms of the “final” model structure; that this is not the case has been argued at various levels

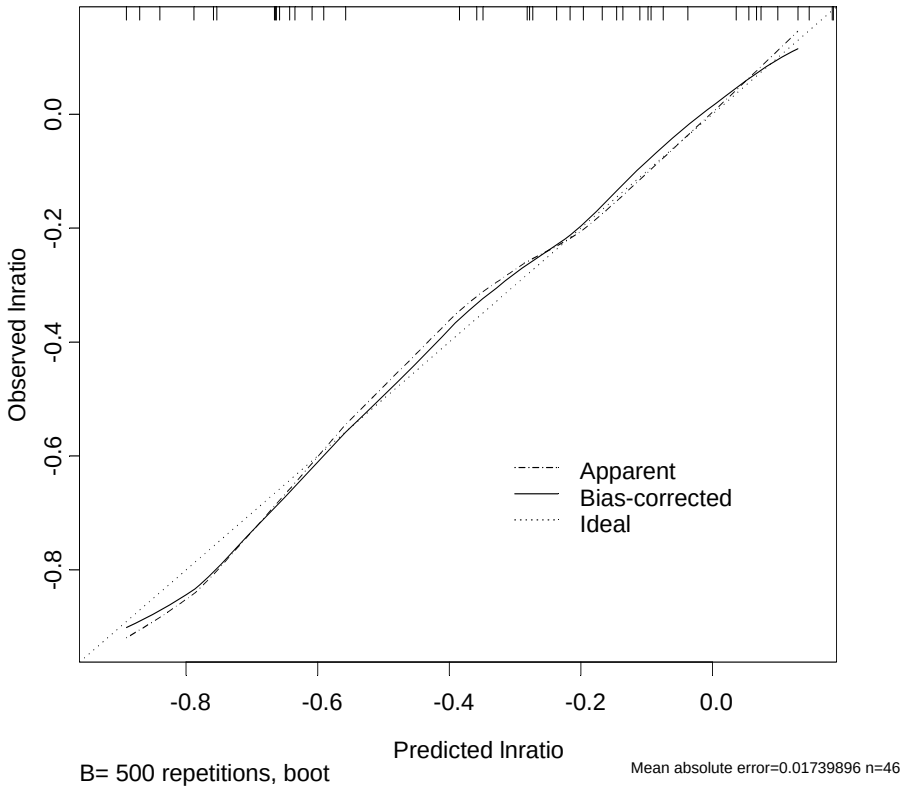
(Buckland *et al.* 1997; Chatfield 1995; Hastie *et al.* 2001). With regard to analysis of the current data set: the relative consistency of variable selection between the original and complete (EM) data sets and parameter estimates between the initial and multiple imputation data-sets is noted. The uncertainty aspects of the modelling enterprise were extended by:

6.5.2.5.1 using the bootstrap method (500 bootstrap samples) and functions from Harrell's S-Plus Design library (Harrell Jr 2005), validation (estimates of model fit) and calibration (plot of observed versus predicted) of the OLS model (see Table 6.4.6) were further undertaken. The data-set size prohibited the more familiar training and (set-aside) validation sets, although a critique of this particular strategy has been formulated (Hirsch 1991). The original R^2 and adjusted-R^2 were 0.65 and 0.6 respectively (Table 4a). The "optimism" of prediction (in "new" data sets; in this case, averaged over 500 bootstrap samples) was estimated at 0.1 and hence realistic estimates of model fit were correspondingly less, by that amount ("shrinkage"). Of note is that these were unconditional estimates; conditional estimates would have require sampling from quantities such as residuals, but under most conditions, conditional and unconditional estimators are similar. A more conservative estimate of optimism of prediction is the "0.632" bootstrap (Harrell, Jr. 2001; Hastie *et al.* 2001) , which, for each observation, uses predictions from bootstrap samples not containing that observation (the average number of distinct observations in each bootstrap sample is approximately 0.632*N); here the optimism was estimated at 0.15. Figure 6.5.2.5.1 shows the contrast of the apparent and bias-corrected agreement of observed versus predicted log ratio

T_{24}/T_0 phosphate, with respect to line-of-identity. The fit was adjudged as reasonable.

6.5.2.5.2 uncertainty in model selection may be explicitly formulated by estimates of posterior model probability for ***all*** possible models, using Bayesian Model Averaging; the “bicreg” function of Raftery, running under S-Plus statistical software was used (Hoeting *et al.* 1999; Raftery *et al.* 1996). The total posterior probability of the “top” 10 models (as assessed by posterior model probability) was computed at only 45% of total probability, suggesting that considerable model uncertainty existed, even for the current relatively small data-set.

Figure 6.5.2.5.1. Observed versus predicted log ratio T_{24}/T_0 phosphate



Observed versus predicted log Phosphate ratio(lnratio). Line of identity is shown as (……); apparent relationship is shown as (. _ . .) and bootstrap (500x) bias-corrected relationship as (_____).

6.6 Conclusions

6.6.1 Modern statistical techniques, for example the bootstrap, data-augmentation and generalized linear models, are able to supplement traditional approaches to multivariable prediction; the effect of regression to the mean may be quantified, missing values are able to be managed appropriately, modelling of skewed variables may be more effectively accomplished and model uncertainty, in its various forms, may be quantified.

7 ANALYSIS OF COST DATA

7.1 Introduction: Medical cost data are usually right skewed with variability increasing as the mean costs increases. The traditional model for cost prediction has been multivariable ordinary least squares regression (OLS) (Sznajder *et al.* 1998; Diehr *et al.* 1999; Becker *et al.* 1995), with or without initial transformation, usually logarithmic, of the dependent cost variable (Manning 1998). As previously noted by Chhikara and Folks (Chhikara *et al.* 1989), the use of transformations suggested by the data still leaves the problem of interpretation of the results; analysis on transformed scales does not "provide inferences about population mean costs which are of primary interest" (Barber *et al.* 2004). Thus "simple" logarithmic transformation has attendant problems in terms of both the appropriate back transformation into the original scale (that is, dollars) (Duan 1983) and the interpretation of regression coefficients, as noted above (Cole 2000). Recently, generalized linear models (GLM; see section 6.3.4, above), have also been introduced into the analysis of cost data (Austin *et al.* 2003; Manning *et al.* 2001; Kilian *et al.* 2002; Buntin *et al.* 2004; Diehr *et al.* 1999; Blough *et al.* 2000b). It has been suggested that health care expenditure and use-data frequently have a log-normal or gamma distribution (Manning *et al.* 2001) and the studies using GLM for cost analysis have focussed on the gamma response distribution. However, the shape of the inverse Gaussian distribution, with a high initial peak and long right tail (Chhikara *et al.* 1989), may recommend its use for cost data.

7.1.1 Our purpose then was to compare the performance of OLS and various GLMs (specific combinations of distribution (family) and link) in the analysis of individual patient costs derived from a "ground-up" ICU utilisation study and to

answer the question: do GLMs, in particular a GLM using the inverse Gaussian distribution response distribution, have particular advantage when analysing medical cost data? Performance was adjudged using established indices (mean absolute error (MAE), root mean square error (RMSE) and various coefficients of determination (R^2)) and graphical residual analysis (Diehr *et al.* 1999; Hardin *et al.* 2001).

7.2 Data sources and settings

7.2.1 Cost data for ICU patient stay, including all related management activity, but excluding costs associated with provision of services external to the ICU, was generated from a nine month study (in 1991) in three South Australian adult ICUs; an in detail analysis of this data has been recently reported (Moran *et al.* 2004a).

7.2.2 Data collection: in each ICU, dedicated unit data collectors recorded daily activity and utilisation. The specific utilisation elements were: drugs; procedural; costs for pathology, radiology, physiotherapy, nursing staff and medical staff costs; overhead costs derived using the Yale DRG costing methodology (Fetter *et al.* 1980; Moran *et al.* 2004a), and allocated to patients on the basis of ICU length of stay; and residual costs. Total costs (1991 Australian $) were computed as the sum of various cost fractions: (i) medication and procedural (ii) nursing, physiotherapy and medical (iii) radiology and pathology (iv) overhead and other; individual (patient) day costs were not available for analysis.

7.2.2.1 Additional patient data recorded included: Demographics: age, gender, ethnicity, comorbidities consistent with the APACHE III algorithm (Knaus *et al.* 1991); ICU stay variables: patient source, admission diagnosis and principal

physiological system dysfunction on admission, ventilatory status, cardio-respiratory (heart and respiratory rate, systolic and diastolic blood pressure), arterial blood gas (pH, PaO_2, $PaCO_2$) and biochemical variables such that an APACHE III score could be computed, ICU length of stay and outcome; hospital stay variables: treating hospital, DRG, hospital length of stay and outcome. Categorical variables were score as 0/1, 0/1/2 as indicated. The previous analysis (Moran *et al.* 2004a), using ordinary least squares with untransformed cost data, had, on the basis of a significant Chow test (p = 0.0001; (Chow 1960)) , considered survivors and non-survivors separately. For the purposes of the current analysis: only ICU survivors were considered; potential predictor variables were drawn from demographic and ICU admission day data only; two extreme (cost) outliers (ICU costs > $ AUS100000) and a single case with incomplete first day data were not considered.

7.3 Statistical methodology

7.3.1 Variables are reported as mean(SD) unless otherwise indicated; Stata® statistical software was used (Stata Statistical Software 2003). Probability plots (*P-P*) were initially used to compare the cost distribution with hypothesised distributions (normal, lognormal, gamma and inverse Gaussian). If $x_1, x_2, \ldots x_n$ is the ordered sample (size n) from a distribution with location and scale parameters α and β and F is the cdf , the *P-P* plots $Z_i = F\left(\left[X_i - \hat{u}\right]\hat{\sigma}\right)$ against p_i, where $\hat{u}$ and $\hat{\sigma}$ are estimators of location and scale respectively and p_i are plotting positions (Gan *et al.* 1990).

7.3.2 Multivariable models to predict total costs were as follows: OLS; OLS with log transformation of costs and back-transformations of log-costs as (i) simple exponential (ii) “naïve”, that is the exponential of (predicted costs +

$0.5*(RMSE)^2$), where RMSE = square root of the mean square error of the OLS equation (iii) Duan's smearing estimate and (iv) heteroscedastic retransformation (Manning *et al.* 2001; Kilian *et al.* 2002); GLM with Gaussian family & log link; GLM with gamma family & log link; GLM with inverse Gaussian family & log link; (Hardin *et al.* 2001). Variable selection from a full model used the Akaike information criterion (AIC= -2(*L*) + 2(c+p+1)), where *L* is the log-likelihood, c is the number of model covariates and p is the number of model-specific ancillary parameters (Lindsey *et al.* 1998). Non-linearity of covariate effect was investigated by using (parametric) fractional polynomials and all first order interactions were explored. Model performance was variously assessed:

7.3.3 quantitative predictive indices

7.3.3.1 mean absolute error (MAE) as mean of absolute difference between observed and predicted cost

7.3.3.2 root mean square error (RMSE) as predicted cost minus observed, square of the difference, mean of the squared difference and square root of this value

7.3.3.3 correlation (Pearson, *rho*) of cost and predicted cost (Zheng *et al.* 2000) with 95% bootstrap (BCa) confidence intervals using 1000 repetitions (Carpenter *et al.* 2000)

7.3.3.4 squared correlation (R^2)

7.3.3.5 a "pseudo-R^2" statistic from the GLM literature, the Ben-Akiva and Lerman adjusted likelihood-ratio index = $(1-(L(M_\beta-k) / L(M_\alpha))$, where M_β is the log-likelihood of model with intercept and predictors, *k* is the number of model parameters M_α is the log-likelihood of model with intercept only (Hardin *et al.* 2001)

7.3.3.6 Lin's concordance correlation coefficient (*rho_c*, to be distinguished from Pearson's correlation coefficient) of cost and predicted cost, with 95% BCa CIs. As noted by commentators, correlation (*rho*) implies data for two variables Y_1 and Y_2 lying on a line $Y_1 = \alpha + \beta Y_2$ and large values of *rho* may occur with $\alpha \neq 0$ and / or $\beta \neq 1$ (that is, in the absence of "perfect" agreement between Y_1 and Y_2). *rho_c* assesses the agreement between two paired sets of measurements by measuring the variation from the 45^o line of identity (Zheng 2000).

7.3.3.7 For the OLS models, formal tests for heteroscedasticity (non constant variance(Greene 2000)) were performed; Breusch-Pagan / Cook-Weisberg, Szroeter and a likelihood ratio test for groupwise heteroscedasticity

7.3.3.8 graphical analysis using Anscombe residuals (Hardin *et al.* 2001)

7.3.3.8.1 residual versus fitted values plots, looking for even distribution of the residuals about the y = 0 line

7.3.3.8.2 standardised normal probability plots (*P-P* plots, focusing on centre of the distribution) and inverse normal quantile plots (Q-Q plots, emphasizing the tails of the distribution) of the residuals, looking for close approximation to the 45^o line of identity (Gan *et al.* 1991)

7.3.3.8.3 plots of residuals to assess residual heteroscedasticity; heteroscedasticity was adjudged by the degree of slope (away form the horizontal) of the lowess (locally weighted scatter plot smoothing (Cleveland 1979)) plot line relating the SD of the residuals to the mean values of grouped fitted values and grouped APACHE III scores looking for lack of trend

7.3.3.8.4 plots in 7.3.3.8.1-7.3.3.8.3 above were compared with those using deviance residuals

7.4 Results

7.4.1 The cohort consisted of 1098 patients of mean(SD) age 56(19.5) years and 41% were female. Further patient details are shown in Table 7.4.1.

Table 7.4.1. Patient demographics: mean(SD) or absolute numbers as indicated

Variable	
n	1098
Age: years	56(19.5)
APACHE III score	51(22.5)
Hospital (patient number)	
415	
257	
426	
Gender (female / male; n =)	447 / 651
Ventilated (n =)	552
Chronic dialysis (n =)	10
COPD (n =)	14
Hepatic failure (n =)	6
Metastatic carcinoma (n =)	26
ICU length of stay (days)	2(0.5-67)*
Hospital length of stay (days)	16(0.5-248)*

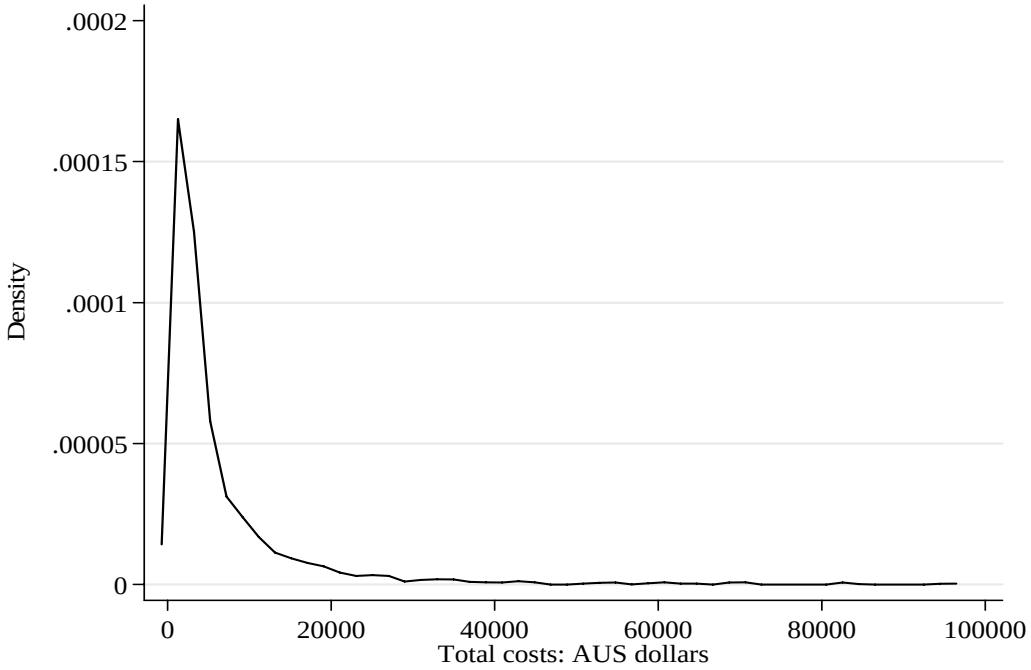
COPD; chronic obstructive pulmonary disease. *; median(range)

7.4.2 Total costs (1991 $AUS) were $6311(9689) with a range of $106 to $95602. The distribution as seen in Figure 7.4.2a, showed marked kurtosis and skewness (p=0.0001) and log transformation did not yield a normal distribution (Shapiro-Wilk W test, p = 0.0001), albeit the kurtosis was modified (p = 0.44). Figure 7.4.2b shows (i) in the upper panel, a probability (*P-P*) plot of gamma and inverse Gaussian distributions generated from the total cost data: in particular, for the (two parameter) gamma distribution, the shape parameter (*alpha*) = 0.953 and the scale parameter (*beta*) = 6604; and for the inverse Gaussian distribution, mean (*mu*) = 6311 and *lambda* = 2677, where variance is mu^3 / *lambda*. and (ii) in the lower panel, quantile-quantile (Q-Q) plots of the above two generated distributions against total costs. Total costs were better approximated (clustering of data points about the 45^o line of identity) by the inverse Gaussian

distribution; no routine transformation of costs (Buchner *et al.* 1990) yielded a normal distribution.

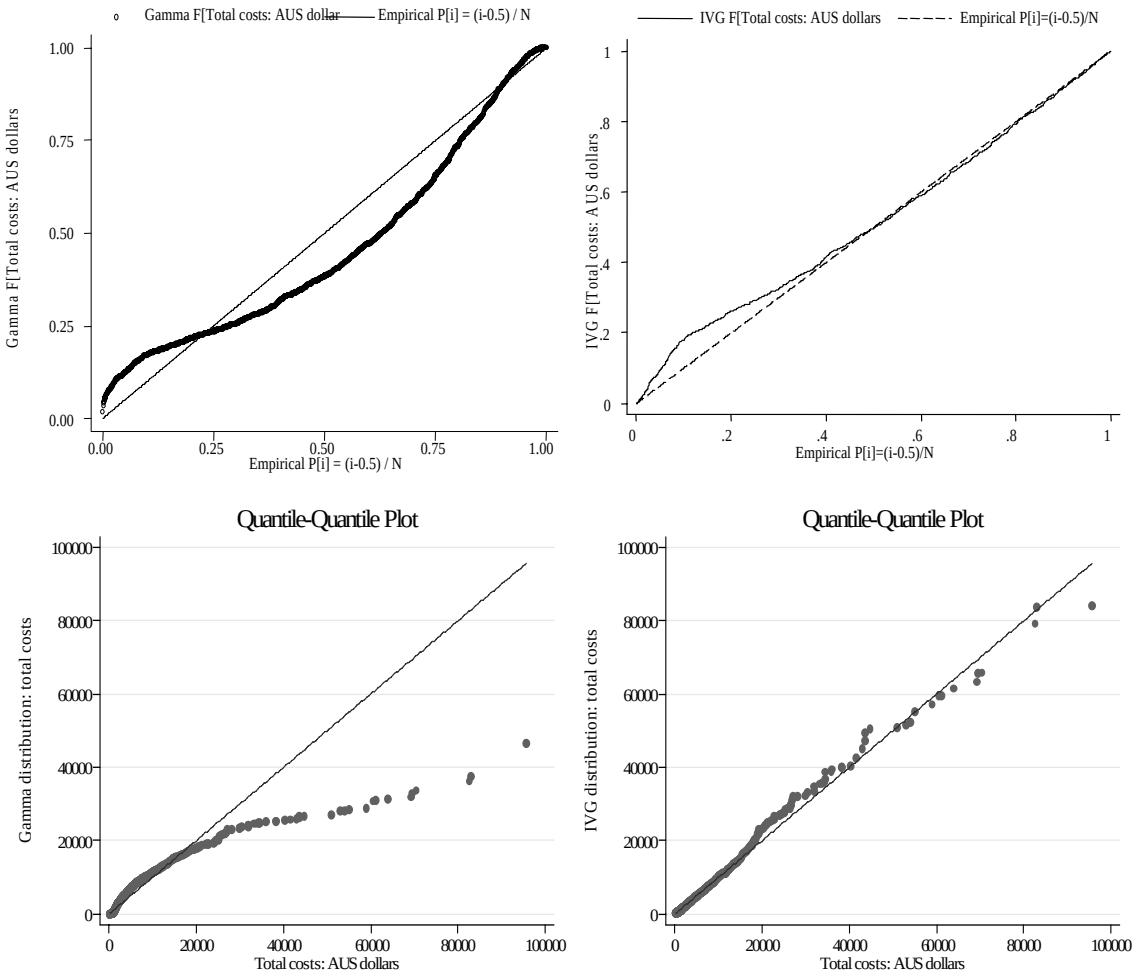
7.4.3 Model covariates and performance indices are seen in Table 7.4.3. Consistency of covariate selection for APACHE III score, ventilation and hospital source was demonstrated across all models, with COPD and chronic dialysis being the next most frequent selections. No significant interactions were demonstrated and continuous variables (APACHE III score and age) demonstrated consistent linear effects. Mean predicted costs, MAE and RMSE varied considerably across models (Table 2). Of note was the severe under-prediction of predicted costs by simple exponentiation in the OLS: log costs model and modest over prediction of mean costs by GLM: inverse Gaussian family & log link. The range of total raw costs was considerable, $106 to $95602; only three models had predicted costs > $35000 (maximum predicted costs $52152); OLS: log costs with heteroscedastic retransformation, GLM: Gaussian family & log link and GLM: inverse Gaussian family & log link. MAE and RMSE were minimal using OLS: log costs with heteroscedastic retransformation and GLM: Gaussian family & log link. Correlation (observed versus fitted costs) and R^2 were best with the GLM: Gaussian family & log link and OLS: log costs with back transformation. Lin's concordance correlation coefficient (observed versus fitted costs) suggested best performance with GLM: Gaussian family & log link, GLM: inverse Gaussian family & log link and OLS: log cost with heteroscedastic retransformation. Considerable variation in concordance was observed between the various back transformations of the OLS log cost model.

Figure 7.4.2a. Cost distribution (kernel density plot)



Kernel density plot of cost distribution. Vertical axis; density. Horizontal axis; Dollars

Figure 7.4.2b. Probability and quantile-quantile plots for gamma and inverse Gaussian cost distributions



Upper panel: probability plot (*P-P*) of two parameter gamma (left) and inverse Gaussian (right) distributions against costs. Vertical axis, (cumulative) probability, 0-1; horizontal axis, Hazen plotting position (= (i-0.5)/n, where i=rank and n=count, (Gan *et al.* 1990)). Lower panel: ordered quantile plots of distributions (gamma, left and inverse Gaussian, right) generated from total costs (vertical axis) against total costs (horizontal axis)

Table 7.4.3. Total and predicted costs ($ Australian) and model performance indices.

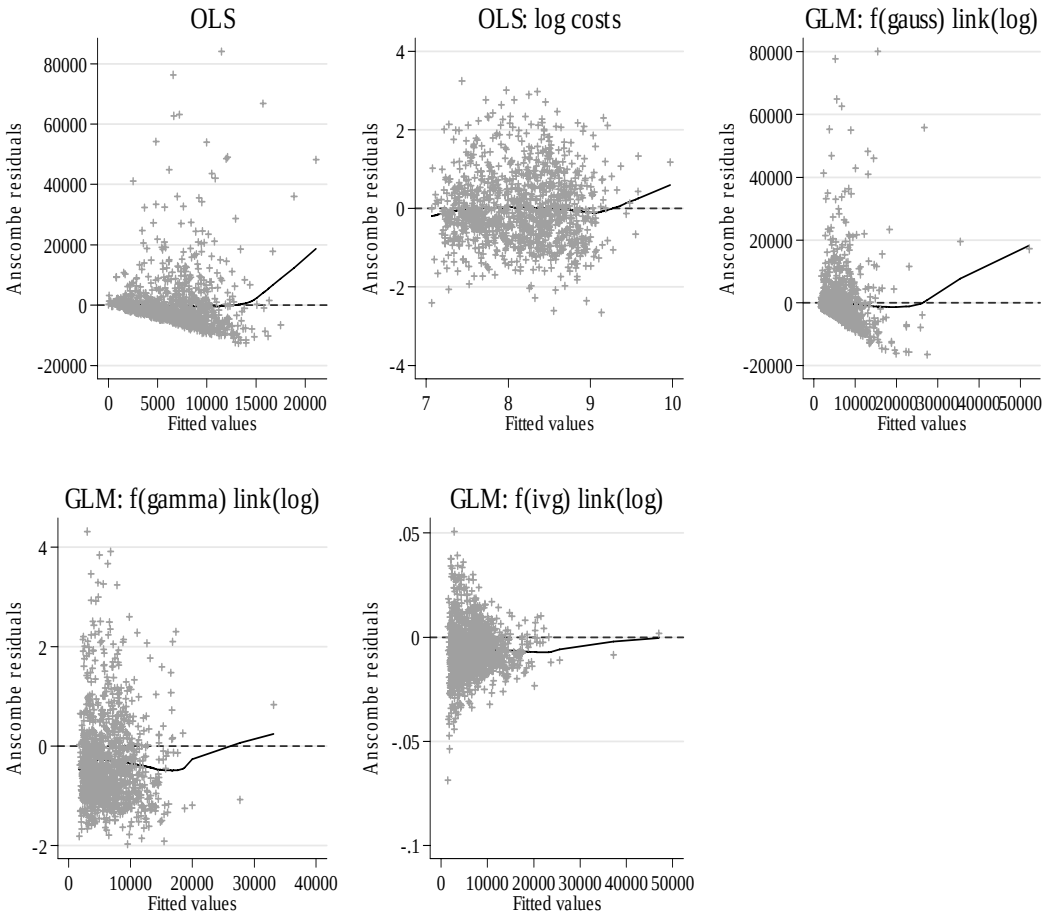
	Covariates	Mean	SD	MAE	RMSE	Corr (95% CI)	rho_c (95% CI)	R^2	BAL
Total costs		6311	9689						
OLS	APIII,metca age,vent,copd	6311	3313	4995	9101	0.342 (0.284-0.413)	0.209 (0.159-0.256)	0.117	0.006
OLS: log, exp	Hosp,AP3,age,vent copd,metca,hfail	3936	2131	4242	9420	0.369 (0.283-0.290)	0.146 (0.11-0.197)	0.136	0.106
OLS: log, naïve	Hosp,AP3,age,vent copd,metca,hfail	5902	3195	4753	9018	0.369 (0.283-0.290)	0.219 (0.167-0.295)	0.136	0.106
OLS: log, Duan	Hosp,AP3,age,vent copd,metca,hfail	6298	3410	4914	9002	0.369 (0.283-0.290)	0.231 (0.176-0.309)	0.136	0.106
OLS: log, het	Hosp,AP3,age,vent copd,metca,hfail	6037	3753	4780	8965	0.379 (0.295-0.499)	0.255 (0.189-0.363)	0.144	0.106
GLM: gausslog	Hosp,AP3,age,vent copd,metca,hfail	6121	4102	4798	8907	0.396 (0.304-0.517)	0.283 (0.202-0.402)	0.155	0.009
GLM: gamlog	Hosp,AP3,vent	6368	3540	4990	9077	0.349 (0.283-0.454)	0.225 (0.171-0.295)	0.122	0.014
GLM: ivglog	Hosp,AP3,vent,cdial	6805	4467	5198	9136	0.353 (0.280-0.468)	0.268 (0.202-0.369)	0.124	0.0004

SD, standard deviation. MAE; mean absolute error. RMSE; root mean square error. Corr; Pearson correlation with 95% bootstrap (BCa) CI. rho_c; Lin's concordance correlation coefficient with 95% bootstrap (BCa) CI. R^2; coefficient of determination. BAL; Ben-Akiva and Lerman adjusted likelihood ratio index.
Hosp; hospital source. AP3; APACHE III score. age; age in years. Vent; ventilation on ICU admission day. copd; history of chronic obstructive pulmonary disease. metca; evidence of metastatic carcinoma. hfail; hepatic failure. cdial; chronic dialysis.
OLS; ordinary least squares regression. OLS: log, exp; ordinary least squares regression using log transformed costs and exponential back transformation. OLS: log, naïve; ordinary least squares regression using log transformed costs and naive back transformation. OLS: log, Duan: ordinary least squares regression using log transformed costs and Duan's smearing back transformation. OLS: log, het: ordinary least squares regression using log transformed costs and heteroscedastic back transformation GLM: gausslog; generalized linear model with Gaussian family & log link. GLM: gamlog; generalized linear model with gamma family & log link. GLM: ivglog; generalized linear model with inverse Gaussian family & log link.

7.4.4 Overall model performance (systematic departure from model assumptions) was assessed by inspection of plots of residuals against fitted values (shown in Figure 7.4.4a) and standardised normal probability (shown in Figure 7.4.4b) and inverse normal quantile plots of residuals. Symmetrical distribution (residuals versus fitted values) and normality of residuals (probability and quantile plots), suggesting optimal model performance, was observed for OLS: log costs and

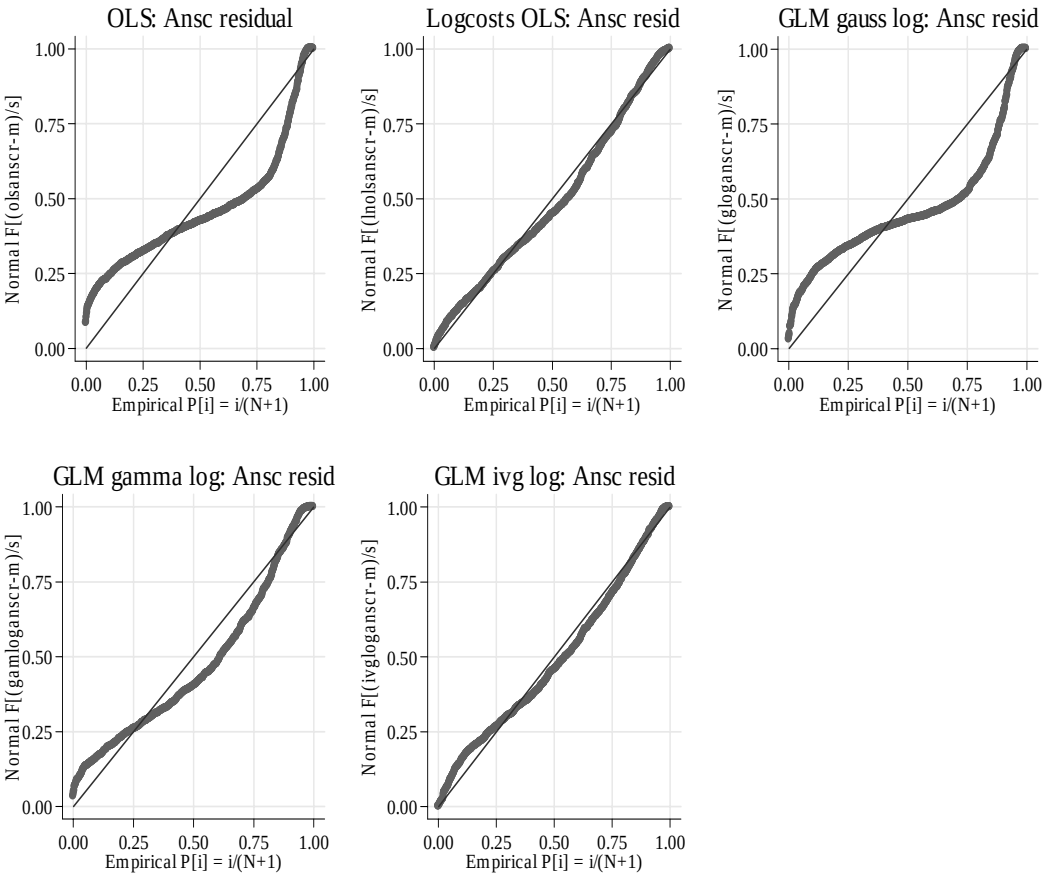
GLM: inverse Gaussian family & log link. The only models to reasonably satisfy homoscedasticity (constant variance assumption) were GLM: Gamma family & log link and OLS: log costs, although all models appeared suspect (Figure 7.4.4c). This being said, tests of heteroscedasticity identified significant overall ($p = 0.001$) and covariate specific (APACHE III score ($p = 0.001$), ventilation status ($p = 0.001$)) heteroscedasticity for both OLS and OLS: log costs. No differential diagnostic sensitivity in the plots between Anscombe and deviance residuals was noted.

Figure 7.4.4a. Plots of Anscombe residuals versus fitted values for various models



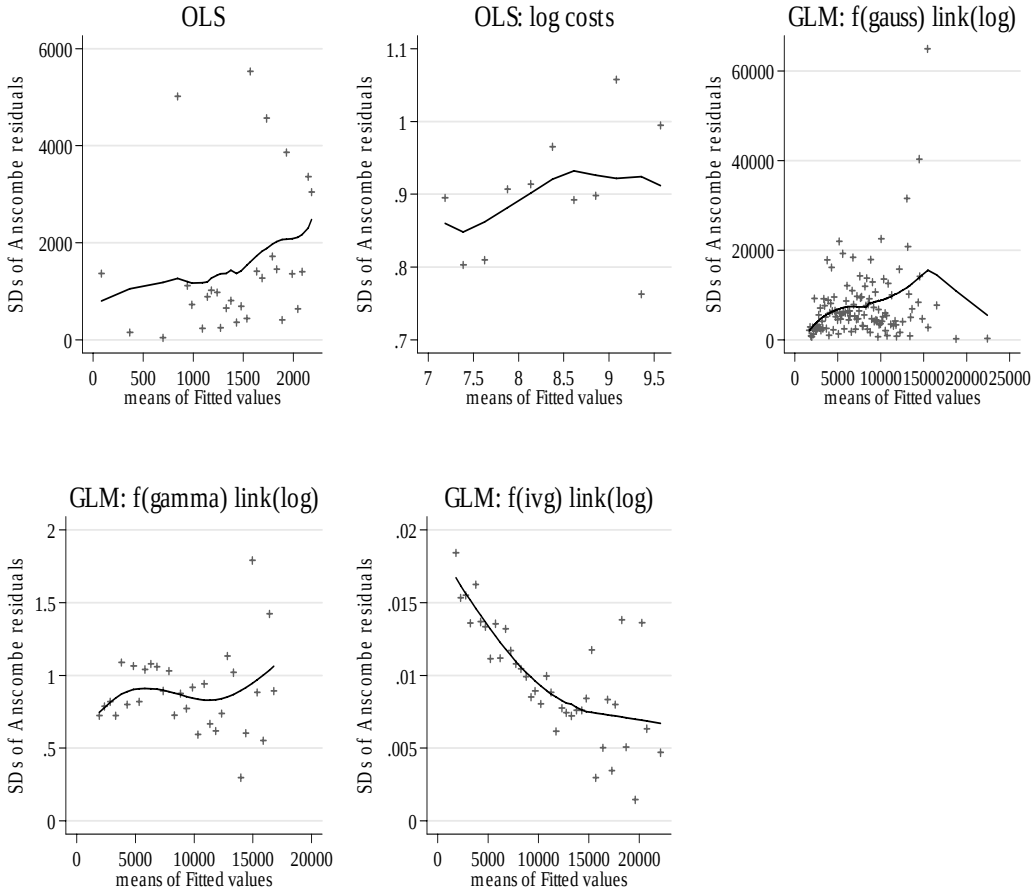
Plots of Anscombe residuals (vertical axis) against fitted values (horizontal axis) from regression models. Upper panel (left to right): OLS, OLS with log transformed costs, GLM Gaussian family & log link. Lower Panel (left to right): GLM gamma family & log link, GLM inverse Gaussian family & log link

Figure 7.4.4b. Standardized normal probability plot of Anscombe residuals (*P*-norm plot)



Standardised normal probability plot ("*P*-norm" plot) of regression models. Vertical axis: cumulative probability related to normal distribution), where is mean of data and is standard deviation; horizontal axis: plotting positions (Weibull, pi = i/(N+1). Upper panel (left to right): OLS, OLS with log transformed costs, GLM Gaussian family & log link. Lower Panel (left to right): GLM gamma family & log link, GLM inverse Gaussian family & log link.

Figure 7.4.4c. Lowess plots of SDs of Anscombe residuals versus means of grouped fitted values



Heteroscedasticity diagnostic plots using “Lowess” smoothing (26). Vertical axis: standard deviation of Anscombe residuals. Horizontal axis: means of grouped fitted values of regression models

Upper panel (left to right): OLS, OLS with log transformed costs, GLM Gaussian family & log link. Lower Panel (left to right): GLM gamma family & log link, GLM inverse Gaussian family & log link.

7.5 Discussion

7.5.1 The models considered above addressed the estimation of mean or total costs using particular covariate sets (conditional mean modelling (Manning *et al.* 2001)); formally, the estimation of $E(y \mid x)$. Although the dependent variable (y) was positively skewed, estimation of median costs was not considered, as this would have been less relevant to ICU administrative concerns, which focus on total costs = average costs × number of patients (Austin *et al.* 2003).

7.5.2 Distributions and transformations

7.5.2.1 The traditional model for skewed health data is one of logarithmic transformation of the dependent variable (Diehr *et al.* 1999). Such a transformation usually induces symmetry rather than normality into the cost variable; if the variance-mean relationship is a power (square) function, logarithmic transformation serves to stabilise variance (homoscedasticity). OLS with a logged dependent variable ($\log(y)$) is contingent upon a linear relation of mean $\log(y)$ to the covariates and the constancy of variance, not necessarily normality. This being said, inference is on the log-dollar scale (Blough *et al.* 2000a). Logarithmic transformation results in comparison of geometric means and inference in comparing such means cannot be equated with a test of arithmetic means unless log-scale variances (between groups) are equal (Briggs *et al.* 1998). Back transformation to the original scale of the dependent variable (in this case, dollars) is not simply a matter of exponentiation. As seen from Table 7.4.3, the concordance (rho_c) of total costs with predicted costs for the OLS log costs model is dependent upon the method of back transformation, with rho_c varying from 0.146 with simple exponentiation, 0.219 with “naïve” transformation, 0.231 with Duan’s

smearing estimate and 0.255 for heteroscedastic retransformation. For OLS with normally distributed homoscedastic errors, the recommended transformation is: exponential(fitted values + 0.5*(RMSE)^2). For non-normally distributed, homoscedastic errors, the smearing estimate ($\hat{\phi}$) has superior properties and is given by: $\hat{\phi} = \frac{1}{N}\sum_{i=1}^{n}\exp(\hat{\varepsilon}_i)$, where ε are the absolute residuals from the OLS regression (Rutten-van Molken *et al.* 1994); the predicted costs are then calculated as $E(y) = \hat{\phi}.\exp(X\beta)$, where $X\beta$ is the OLS linear predictor. In the current study, $\hat{\phi}$ = 1.60. For normally distributed heteroscedastic residuals: exponential(fitted values + 0.5*(log scale variance function, $v(x)$)), where $v(x)$ is the variance of the log-scale, obtained by regressing the squared residuals on the covariates (Kilian *et al.* 2002). Thus a seemingly "simple" log transformation entails a rather complex model dependent re-transformation process to recover costs in the original scale, without incurring bias.

7.5.2.2 Similarly, the interpretation of regression coefficients with log transform of the dependent variable is not facile: for the homoscedastic normal regression, the effect on the (untransformed) dependent variable is in terms of a percentage change =100×(exp(β)-1), where β is the (independent) variable regression coefficient, continuous or categorical (Cole 2000; Zhou *et al.* 2001b). For heteroscedastic regression, where the covariate (x_j) also appears in the variance model ($\sigma^2 = \exp(\gamma' x_j)$), covariate effect is somewhat more complex, as developed by Zhou *et al* (Zhou *et al.* 2001b), with different interpretations of unit change of dependent variable for categorical and continuous covariates.

7.5.2.3 The gamma distribution, most useful with positive responses (≥ 0) having a constant coefficient of variation, has also been suggested as an appropriate distribution with which to model costs (Blough *et al.* 1999). In a recent empirical investigation of costs generated from a randomised trial, the gamma distribution was found to be the most appropriate, based primarily upon analysis of residuals; no initial approximation of the cost distribution to the exponential family of distributions was provided (Barber *et al.* 2004). Such was not the case in the current study, where the total cost distribution was poorly approximated by the gamma distribution (Figure 7.3.3.2b). The inverse Gaussian distribution, with a high initial peak with rapid drop off and long right tail, would appear to adequately reflect cost and length of stay distributions, although little has been published on this (Chhikara *et al.* 1989). A previous paper, applying the inverse Gaussian distribution to length of stay, used the method of sample quantiles (agreement of fitted and observed distributions at specific quantiles) (Whitmore 1975), but did not model the length of stay. This being said, regression models appropriate for cost data are not necessarily optimal for length of stay prediction (Austin *et al.* 2003).

7.5.3 Generalized linear models

7.5.3.1 The generalized linear model synthesizes the general techniques used to analyse continuous and discrete data into a unified conceptual framework (Breslow 1996). Explanatory features are combined additively as in classical linear models. The properties of the response variable are matched by the particular distribution; the variance is a function of the mean ($\mathrm{var}(y \mid x) = \sigma^2 v(x)$), except for the normal distribution, where the mean and variance are independent; and the link function determines the appropriate

scale (Lane 2002). For example, in OLS with a (log) transformation (g), the expectation (E) is $E(g(Y_i)) = \alpha + x\beta$; for the GLM, the form of the expectation is $g(E(Y))= \alpha + x\beta$. That is, the GLM log-links the predictor ($x\beta$) rather than the response and parameters are equal to the logs of arithmetic means (continuous variables) and their ratios (categorical variables); thus parameters can be interpreted directly in a manner similar to odds ratios (Blough *et al.* 2000a). GLM are fitted by either maximum likelihood or iteratively re-weighted least squares and a key parameter is the deviance = 2log λ, where log λ = log-likelihood(full or "saturated" model) – log-likelihood(null or intercept only model). For the normal distribution model, the deviance is the residual sum of squares and hence the notion of R^2 (= 1-(residual sum of squares/total sum of squares)) may be interpreted as the familiar "percent variance explained". Although there are "pseudo-R^2" statistics for the GLM, the deviance for non-normal distributions is different from the residual sum of squares and the scalar values of these various statistics are not monotone transformations, as would apply to the normal linear model. Thus, the squared correlation (R^2) of models showed modest correlation (rho = 0.52, p = 0.1) with the Ben-Akiva and Lerman adjusted likelihood ratio index (Table 7.4.3), but poor concordance (rho_c = 0.07, p = 0.15).

7.5.4 Model performance

7.5.4.1 Overall predictive performance was low, as adjudged by R^2, but similar to that of Becker *et al* (Becker *et al.* 1995), who reported $R^2 = 0.13$ for a multivariable regression equation predicting costs after cardiac surgery and also limited the covariate recording period to ≤ 3 days post-operatively. There was no apparent advantage, in terms of R^2, of a "full" model (17 covariates, data not shown),

although the total patient number would have been sufficient (Green 1991). Covariate selection was not constrained to be constant between models and variation of the model covariate sets occurred, similar to that reported by Dudley *et al* (Dudley *et al.* 1993). Formal data trimming was not initially undertaken (Lee *et al.* 1998) and there may have been a tendency in the OLS and GLM: inverse Gaussian family & log link models to over fitting, as evidenced by a relatively low RMSE and high MAE (Manning *et al.* 2001). Across both quantitative indices and graphical assessment of model performance, OLS: log costs with heteroscedastic retransformation and GLM: inverse Gaussian family & log link seemed the preferred models. That the GLM model(s) had comparative performance compared with OLS: log costs is of obvious advantage, in that re-transformation is avoided and $E(y|x)$ or $\ln(E(y|x))$ is "directly" available (Manning *et al.* 2001). In terms of selection between GLMs, an assessment of the power function of the variance mean (= μ) relation has been proposed (Manning *et al.* 2001; Blough *et al.* 1999), using regression of the log of squared residuals ($\log(y_i - \hat{y}_i)^2$, where y_i = observed costs and $\hat{y}_i$ = fitted or predicted values) against the log of the fitted values in the raw-scale: $\ln\left(y_i - \hat{y}_i\right)^2 = \lambda_0 + \lambda_1 \ln\left(\hat{y}_i\right) + v_i$, the scalar quantity of the coefficient (λ_1) of the logged fitted values indicating the degree of this power relationship. For the GLM gamma family, $\lambda = 2$ (that is, variance = μ^2) and for the GLM inverse Gaussian family, $\lambda = 3$ (variance = μ^3). In the current data set, λ was calculated as 2.1, suggesting initial model preference for the gamma distribution; this was also reflected in model AIC values (normalised for *n*,

lower values being preferred), comparing across GLMs (Table 2): 21.04, 19.24 and 26.28 (Gaussian, gamma and inverse Gaussian family GLM, respectively).

7.5.5 Heteroscedasticity

7.5.5.1 The primary concern in this study was the prediction of total costs from ICU admission day data; that is, pragmatic rather than explanatory (Schwartz *et al.* 1967). Thus, unlike other studies (Kilian *et al.* 2002; Manning 1998; Manning *et al.* 2001), the effect of, for example, patient groupings (into hospitals) and covariate heteroscedasticity on precision of the β coefficients and the appropriate compensation for this via robust or bootstrapped variance estimates (Kilian *et al.* 2002), was not a focus of attention, although this would be an issue in assessing the relative importance of various covariates to cost determination. This being said, all models (including the "full" model, data not shown) demonstrated heteroscedasticity to some degree, with the GLM: gamma family & log link exhibiting least tendency (Figure 7.4.4c). A number of factors undoubtedly contributed to this: the skewness of the cost data, patient groupings and the non-normality of the two continuous predictors, APACHE III score and age. Standard transformations and quantile (n=4) categorisation of the latter two covariates did not resolve this heteroscedasticity.

7.6 Conclusions

7.6.1 GLMs offer an alternative to the standard OLS model for cost prediction. OLS with log transformation of the dependent cost variable must appropriately formulate the problem of back transformation to avoid predictive bias. GLM using the inverse Gaussian response distribution may be of advantage in the analysis of cost data.

8 OUTCOME OF PATIENTS ADMITTED TO ICU WITH HAEMATOLOGICAL AND SOLID MALIGNANCIES (Moran *et al.* 2005b)

8.1 Introduction: The role of the intensive care unit (ICU) in the care of critically ill patients with haematological and solid malignancies has been a matter of controversy (Rubenfeld *et al.* 1996). Recent reports suggest that the outcome of these patients may not have substantially improved over time (Hulme *et al.* 1999; Azoulay *et al.* 2000). The impact on survival of various patient covariates such as prior bone marrow transplantation (BMT; allogeneic or autologous), mechanical ventilation and multiple organ dysfunction and its treatment(s) has been variously assessed; divergences being presumably due to specific patient factors (solid tumor versus haematological malignancy, medical versus surgical) and illness severity, and causation (respiratory failure and / or shock, extent of associated neutropenia and proximity to chemotherapy).Using a previously defined methodology a review was undertaken of potential risk factors for time-to-mortality, censored at 30 days post-ICU-admission, of patients admitted to a single multidisciplinary adult ICU at a university teaching hospital, over a 10 year period 1989 to 1999. Primarily, we were concerned to evaluate any change in outcome over time (Milberg *et al.* 1995); and the effect of severity of illness (Knaus *et al.* 1985b), comorbidity burden (Pittet *et al.* 1993) and mechanical ventilation (Knaus 1989) on outcome. Secondarily, we sought evidence for latent patient heterogeneity about important covariates (cohort effect and mechanical ventilation) using random effects models (Hougaard 1995) and gauged the ability of the Cox regression to adequately model survival, given recent cautions that the Cox model may not be optimal in acute severe illness (Knaus *et al.* 1993).

8.2 Methods:

8.2.1 All ICU patient admissions with haematological or solid malignancies from 1989 to 1999, directly referred by the haematology-oncology unit, were identified from a computerised prospective database, incorporating the APACHE II scoring system. Case notes and ICU data sheets of the patients were subsequently reviewed to confirm diagnoses and to extract relevant study data. Access to these records was obtained under extant guidelines of the TQEH Ethics of Research Committee and informed consent was waived. The following data was recorded:

8.2.1.1 Premorbid – (defined as that available most proximate to the index admission) date of birth and gender; nature and initial time of malignancy diagnosis; Karnofsky score (Karnofsky *et al.* 1948). Information was obtained from hospital sources and referring medical officers.

8.2.1.2 During the hospital admission – cohort (early: admission October 1989- July 1994; late: August 1994-March 1999); admission status (index admission, repeat hospital and ICU admission or repeat ICU admission within a hospital admission); Charlson comorbidity score (CCI) (Charlson *et al.* 1987); ICU admission diagnosis; source (ward or emergency service); hospital admission to ICU admission time (lead time); recent surgery (within 7 days); time of last chemotherapy; performance of stem cell transplantation; prescription on ICU admission of steroid, granulocyte colony-stimulating factor [G-CSF] and antibiotics; leucocyte and platelet count on ICU admission, on days 1 (admission) through 8 following ICU admission; details of mechanical ventilation, ICU therapeutics, APACHE II score, organ failure (Knaus *et al.* 1985c) and sepsis status (notated as systemic inflammatory syndrome (SIRS),

sepsis, severe sepsis and septic shock) (Beal *et al.* 1994) on days 1 (admission) through 8 following ICU admission; ICU and hospital length of stay and outcome.

8.2.1.3 Follow up - case note and computerised information systems review and telephone contact with local medical officers.

8.2.1.4 Patient exclusions: To preserve an uniform cohort, only patients being directly cared for by the haematology-oncology unit were included in the review. Thus, for example, cancer patients being subjected to "routine" surgical resection and such patients referred from the surgical wards to the ICU for post-operative complications, were not considered.

8.3 Statistical methodology:

8.3.1 Time to mortality for the *index* admissions, right censored at day 30, was assessed using Kaplan-Meier and Cox model estimates. The Cox model was structured for (i) pre-morbid and ICU admission day variables and (ii) pre-morbid, admission day and time-varying covariates (including first degree lagged and differenced values), where these were recorded, over days 1 (admission) through 8 following ICU admission (Altman *et al.* 1994a). The parameterization of the Cox model as hazard ratio is not to be simply equated with risk ratio, in that hazard ratios (HR), as opposed to risk ratios, are rates. Time-varying covariates were identified as those having significant interactions ($p< 0.05$) of the (continuously time-varying) covariate with failure-times (time to death) over 30 days. Predictor variables were defined by a backward selection procedure from the full model of potential predictors using Akaike information criterion (Lindsey *et al.* 1998); initial bi-variable selection screening was not undertaken (Sun *et al.* 1996a). Attention was directed to the

question of model selection with correlated variables and the potential effect of multi-colinearity was carefully assessed. First order interactions were explored and non-linearity of covariate effect was also investigated by inspection of residual plots and parametric (fractional polynomials) and non-parametric (cubic smoothing splines) methods. Overall Cox model fit was assessed by residual plots and specific tests for goodness-of-fit; concordance (Harrell's C statistic, comparable with area under the receiver operator characteristic curve (ROC) in logistic regression) and non-proportionality (Harrell, Jr. *et al.* 1996; Hosmer Jr *et al.* 1999).

8.3.2 The Cox analysis for multiple record patient data (days 1 through 8) was extended to a random effects formulation using the Stata® module GLLAMM (Rabe-Hesketh *et al.* 2000), which fits generalised linear latent and mixed models. Mortality was modelled using a Poisson analysis with the default log link (the offset being the log of the interval lengths between failure times) and the exponentiated regression coefficients were interpreted as conditional (on the random effects) hazard ratios. The baseline log hazard was modelled via restricted cubic splines (Herndon *et al.* 1995). For instances observed within subjects (785 within 89 patients), random effects were modelled at the subject level and both the ventilation and cohort effect were also allowed to vary randomly between subjects. Adequacy of modelling the baseline hazard was assessed by comparing the parameter estimates of the Cox model and (fixed effects) Poisson regression model (with the internal knots (n=5) as variables to be estimated in the regression). Overall utility of the random effects approach was determined by the likelihood ratio test (random effects versus fixed effects Poisson regression). Frailty variance [θ], defined as the exponential of the

(overall) random intercept, was reported; values of $\theta > 1$ were interpreted as reflecting a larger than average hazard and for $\theta < 1$, the hazard was less than average (Hosmer Jr *et al.* 1999).

8.4 Results:

8.4.1 During the study period there were 108 admissions in 89 patients (13 admissions were repeat ICU admissions in a hospital admission and there were 6 repeat hospital and ICU admissions); 54% of admissions were in the early cohort. ICU admission diagnoses and patient variables are seen in Table 8.4.1. The most common diagnoses were sepsis with shock and non-specific acute respiratory failure; in four admissions surgery had occurred within 7 days. The patients were severely ill and over the first day in ICU, 50% had a diagnosis of septic shock. Graphical display of the time change (days 1 (admission) through 8) for 30 day survivors vs non-survivors for (i) APACHE II score, leucocyte and platelet count and plasma bilirubin and (ii) percentage of patients ventilated, inotrope dependent, experiencing cardiovascular, respiratory, renal, haematologic and neurologic failure (Knaus *et al.* 1985d), are seen in Figures 8.4.1a and 8.4.1b respectively. For these variables, the most apparent differences over time between 30 day survivors versus non-survivors were for the APACHE II score, ventilation and haematological and neurologic failure. For variables leucocyte and platelet count and plasma bilirubin, due to skewness of distribution, time points were plotted as median, inter-quartile range.

8.4.2 Overall, 46% of patients were ventilated in ICU and the mortality of these patients, with an APACHE II score 34 ± 8, was 73%. For the early versus late cohort, there was no difference in: APACHE II score (28 ± 8.4) versus 28 ± 10, p=0.9), the percentage of patients ventilated in (56% versus 44%, p=0.99) nor in

the unadjusted mortality (78% versus 67%, p=0.49). Median length of mechanical ventilation was 3.5 days (range, 0.5-26 days); median ICU and hospital length of stay were 3 (range, 0.5-41) and 20 (range, 0.5-141) days respectively. ICU and 30-day mortality were 39% (95% CI, 30%-52%) and 54% (95% CI: 45%-65%) respectively.

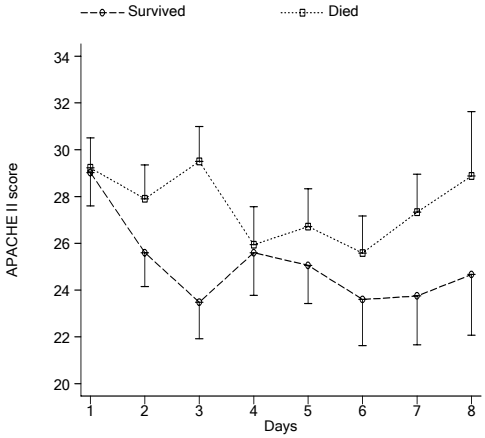
Table 8.4.1 Patient variables

ICU referral diagnosis (n, (%))	
Septic shock	28(25.93)
Acute respiratory failure (non-specific)	23(21.30)
Sepsis	22(20.37)
Pneumonia	13(12.04)
Hypovolaemic shock	8(7.40)
Tumor-lysis syndrome	6(5.56)
Cardiac arrest	3(2.78)
Pulmonary embolus	2(1.85)
Status-epilepticus	2(1.85)
Gastrointestinal perforation	1(0.93)
Pre-morbid	
Malignancy diagnosis (%)	
leukaemia (acute & chronic)	35
lymphoma (+ multiple myeloma)	38
solid tumour	27
Median time from diagnosis to ICU admission (months)	
Leukaemia	1
Lymphoma	8.7
Solid tumour	6.4
Karnofsky score (%)	57(22)
Charlson comorbidity index*	3(range 2-12)
ICU Admission	
Chemotherapy within 30 days (%)	57
Stem cell transplant (%)	20
Steroid on admission (%)	40
G-CSF on admission (%)	18
Age (years)	54.5(14)
Gender (% female)	43
Lead time (days) *	5(range 0-67)
APACHE II score	28(9)
WCC *#	1.2(0.1-45.8)
Platelet count *#	32(2-552)
Ventilated (%)	34
Inotropes / vasopressors (%)	50
SIRS (%)	99
Sepsis (%)	81.5
Severe sepsis (%)	66
Septic shock (%)	50

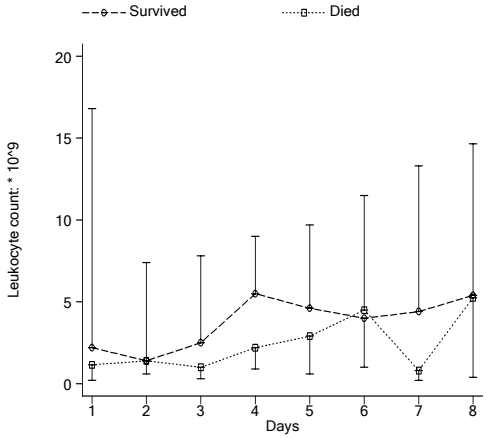
Median (range); #×10^9

Figure 8.4.1a. Time course of patient variables

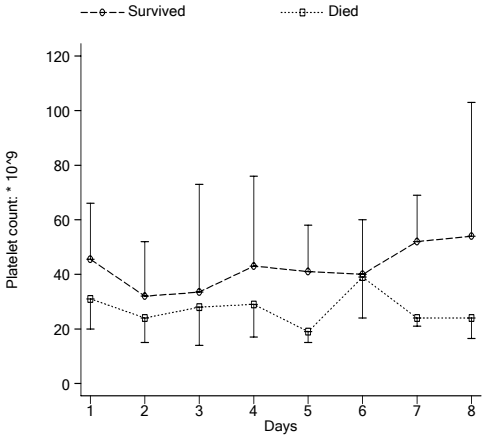
A



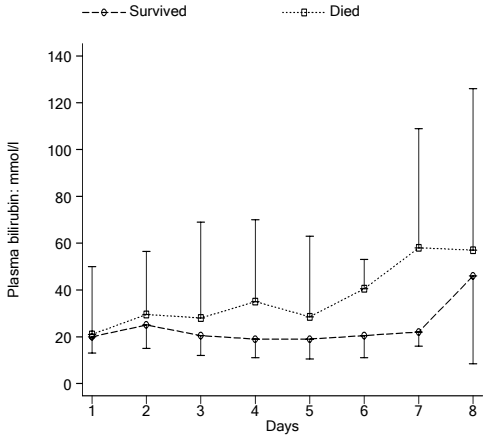
B



C

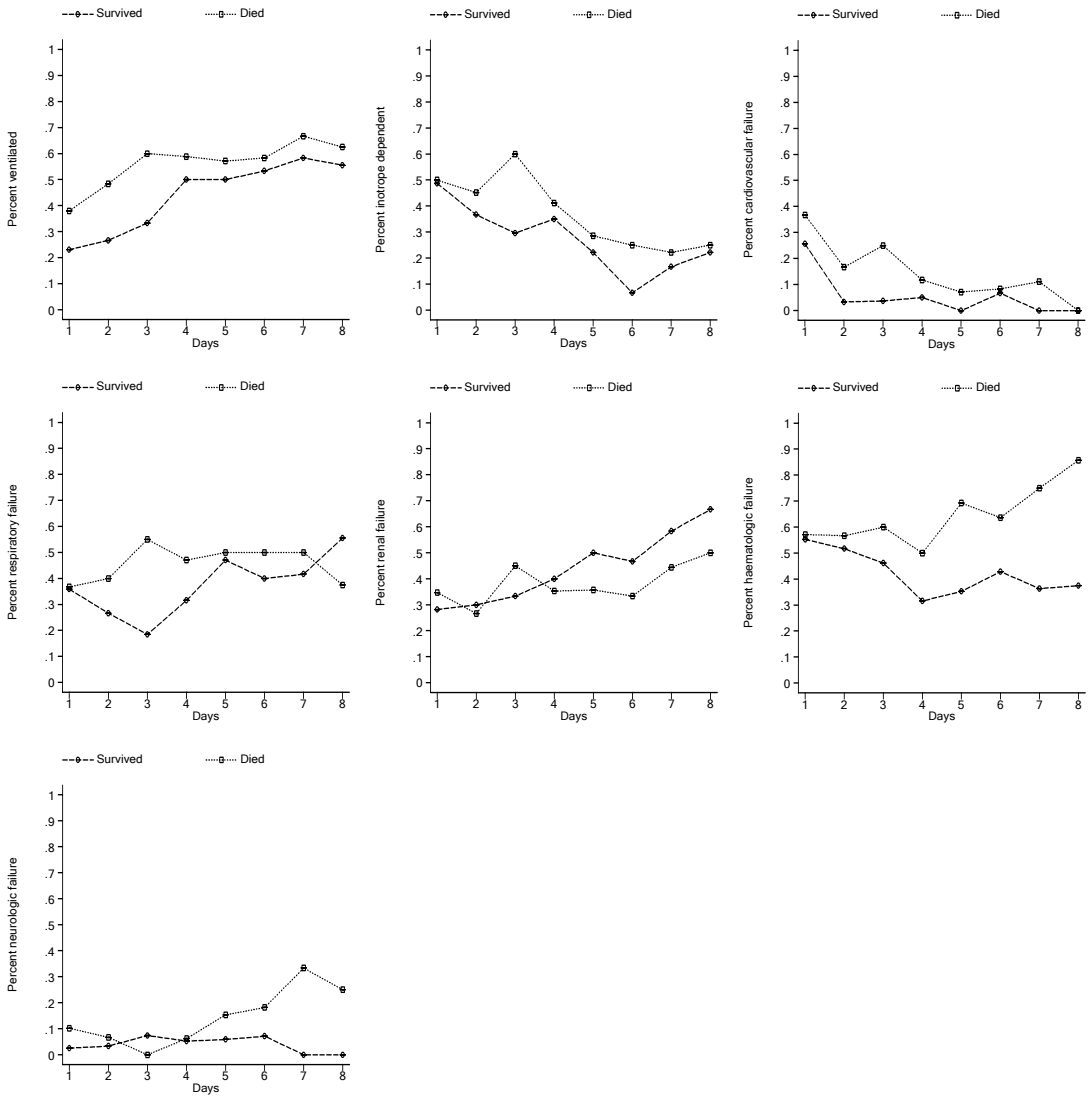


D



Time course of change of variables for survivors (dashed line) and non-survivors (dotted line) over the first 8 days of ICU stay. A. APACHE II score (as mean values with vertical bars as 95% CI) . B.Leucocyte count (*109/L). C. Platelet count (*109/L). D.Plasma bilirubin (mmol/L). For variables B-D values are median with vertical bars as inter-quartile range.

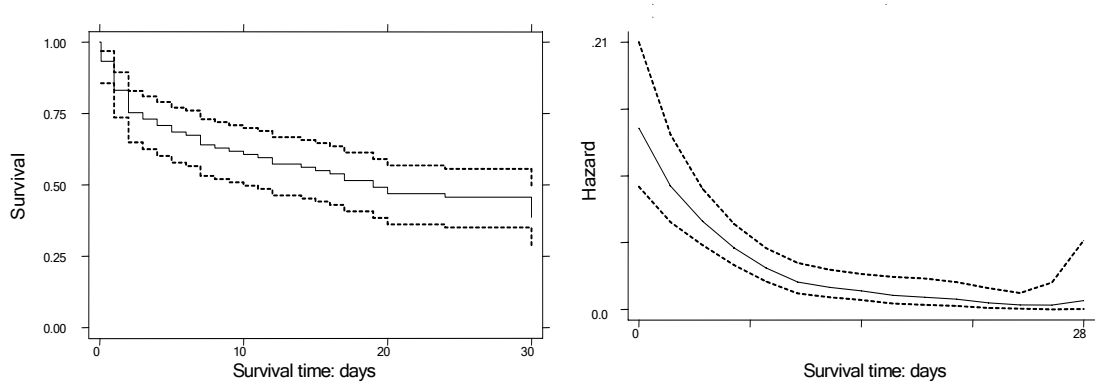
Figure 8.4.1b. Time course of patient variables



Time course of change of variables (as percentages) for survivors (dashed line) and non-survivors (dotted line) over the first 8 days of ICU stay. Upper panel, left to right: Ventilation, inotrope dependence and cardiovascular failure. Middle panel, left to right: Respiratory, renal and haematologic failure. Lower panel: Neurologic failure. Criteria for organ failures are as in Methodology section, above.

8.4.3 Kaplan-Meier estimates of 30 day survival probability and the corresponding smoothed hazard plot (Klein *et al.* 1997) are shown in Figure 8.3.3.3 .

Figure 8.43. Kaplan_Meier survival estimates



Left panel: Kaplan-Meier survival estimates with 95% point-wise CIs (dashed lines). Right panel: Smoothed hazard (vertical axis) with 95% CIs (dashed lines); horizontal axis; time(days)

8.4.4 Using admission day data only, Cox regression predictors of survival were: Charlson comorbidity index, lead time (in days) and mechanical ventilation (considered as a categorical variable). For the multiple-record data Cox model (using patient data over days 1-8 in ICU), predictors were: Charlson comorbidity index, lead time (in days), mechanical ventilation, APACHE II score and cohort effect (considered as a categorical variable; second versus first five year period); see Table 8.4.4.

Table 8.4.4 Cox model estimates (hazard ratio ± SE) for significant predictors

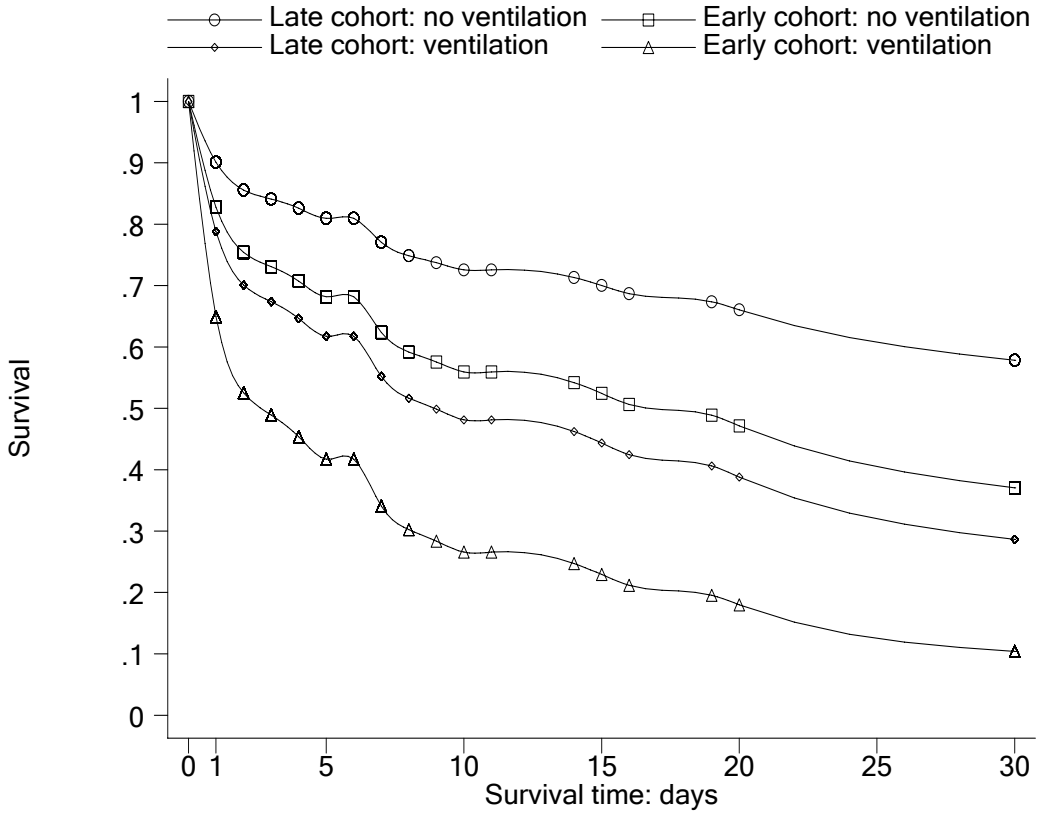
	COHORT (2^{nd} vs 1^{st} 5 year period)	CCI (range 2-12)	TIME to ICU admission (days)	APACHE II score	MECHANICAL VENTILATION
Admission day Data Hazard Ratio ± SE p value		1.15 ± 0.063 0.009	1.02 ± 0.012 0.05		3.21 ± 0.981 0.001
Admission to day 8 Data Hazard Ratio ± SE p value	0.62 ± 0.20 0.05	1.12 ± 0.056 0.02	1.02 ± 0.001 0.02	1.05 ± 0.023 0.02	2.59 ± 0.723 0.01

CCI; Charlson comorbidity index.

8.4.5 Multi-colinearity was not present and no significant interactions were evident despite careful consideration of the potential interactions between (i) lead time and cohort effect (ii) cohort effect and malignancy diagnosis (iii) mechanical ventilation, APACHE II score and malignancy diagnosis (iv) inotrope dependency and APACHE II score and (v) age at diagnosis and specific malignancy diagnosis. Non-linear covariate effect was also not demonstrated. The Cox model was well specified and the decile goodness-of-fit test demonstrated p >0.1 across all deciles; Harrell's C concordance statistic was 0.70 and the global proportionality test was non-significant at p = 0.16 (Day 1 model) and p = 0.9 (multiple patient record data model). Time-varying covariate effects were not demonstrated for the above covariates (including the first lag or difference of APACHE II score), nor were they demonstrated for other potential covariates, such as leukocyte and platelet count, plasma bilirubin and inotrope dependency. Graphical display of the survival probabilities for the 4 distinct categorical groups (multiple record patient data model), at covariate values of

APACHE II score = 26, CCI = 4 and lead time = 5 days, is seen in Figure 8.4.5 (back-projected to a common survival probability of 100% at day "0").

Figure 8.4.5. Cox model: survival probabilities



Cox model survival probabilities for the 4 distinct categorical groups (multiple record patient data model), at covariate values of: APACHE II score=26, CCI=4 and lead time=5 days (back-projected to a common survival probability of 100% at day "0"). Circles: Late cohort, no mechanical ventilation. Squares: Early cohort, no mechanical ventilation. Diamonds: Late cohort, mechanical ventilation. Triangles: Early cohort, mechanical ventilation.

8.4.6 Table 8.4.6 shows parameter estimates from the random effects approach: the Cox and fixed effects Poisson models yielded similar estimates for the predictor variables above (Charlson comorbidity index, lead time, mechanical ventilation, APACHE II score and cohort) signifying that the modelling of the baseline

hazard was sufficient. The random effects model parameters suggested an increased impact of both the cohort effect and mechanical ventilation. However, the likelihood ratio test comparing this model with the fixed effects Poisson model was non significant ($p > 0.5$), indicating no unmeasured subject heterogeneity.

Table 8.4.6. Comparison of Cox, fixed effect Poisson and random effects regression model estimates

Cox model	Poisson		GLLAMM	θ
Cohort (2nd vs 1st 5 year period)	0.62 ± 0.2	0.61 ± 0.19	0.54 ± 0.207	1.03
CCI	1.12 ± 0.056	1.12 ± 0.06	1.16 ± 0.082	
TIME to ICU admission (days)	1.02 ± 0.009	1.02 ± 0.01	1.02 ± 0.013	
APACHE II score	1.05 ± 0.023	1.05 ± 0.022	1.06 ± 0.025	
Mechanical ventilation	2.59 ± 0.723	2.61 ± 0.868	3.07 ± 1.344	1.02
Log Likelihood	-185.72	-148.5	-148.28	

Cox model; estimates as hazard ratios ± SE for multiple record patient data (days 1-8, cf Figure 5). Poisson; fixed effects Poisson model with baseline hazard modelled with restricted cubic spline (point estimates as incidence rate ratios ± SE). GLLAMM; random effects model (subject, mechanical ventilation and cohort), estimates as conditional hazard ratios ± SE). θ; frailty variance

8.5 Discussion:

8.5.1 Numerous articles have reviewed the outcome of patients suffering from haematological and solid tumours and have addressed the question of predictive variables at a descriptive (Tremblay *et al.* 1995; Huaringa *et al.* 2000) and modelling level (Sculier *et al.* 2000; Price *et al.* 1998). A dominant theme has been the poor outcome with hospital or 30 day mortalities ranging from 40 to 50% in general cancer patients (Azoulay *et al.* 2000), up to 80% or greater in ventilated, inotrope dependent patients undergoing bone marrow transplantation (Rubenfeld *et al.* 1996). The current study attests to a high mortality, revealed by the patient subsets displayed in Figure 8.4.5.

8.5.2 Modelling considerations:

8.5.2.1 The analytic approach was to contrast the more conventional premorbid / admission variable analysis with that of an "updated" covariate model, using patient data recorded over the first eight days of ICU admission. As noted above, no time varying covariate effect for the predictors could be demonstrated and thus, in the "updated" covariate model, the coefficients were interpreted as an "average" over all days for which failures occurred; that is the coefficients representing covariate effect were "time-invariant" (Altman *et al.* 1994a). Of interest, no prognostic effect of tumour type or leukocyte or platelet count (at least over the first 8 days) was evident, which would argue against any policy of optimism that depended upon recovery of hematological indices.

8.5.2.2 Neither the cohort effect nor the APACHE II score were predictors using the day 1 data set, which would therefore indicate advantage for the "updated" covariate model; the improvement of statistical efficiency by repeated subject observation (in the presence of information loss due to censoring), has been previously noted (Hogan *et al.* 1998). The (mortality) hazard over the 30 days showed a monotonic decline (Figure 8.3.3.3). This was different from the peaked (non-monotonic) hazard curve reported by Knaus and co-workers (Knaus *et al.* 1993) and suggesting a maximal hazard on (or before) ICU admission, indicated presumably by the high admission APACHE II score and the predictive importance of time to ICU admission. In the Knaus *et al* report, the impact of the physiology score of the APACHE III algorithm demonstrated a progressive time-related decrease and a log-normal accelerated failure time model was preferred. Such an effect was not found in this study, but it is noted that a non-monotonic (initial rise and fall) mortality hazard does not

necessarily imply that a proportional hazards model (Cox or otherwise) may not apply. Proportional-hazard formulations of the parametric log-normal form exist, but such are not available in standard software packages (Bruedel *et al.* 1995). The cohort effect in this study was modelled as a categorical variable as this approach had an intuitive interpretability and was a compromise between a long study period (10 years) and relatively small study numbers. There are problems with such a "cut-point" approach, as reviewed above those of increase in Type I error, over-estimation of effect at each of the cut-point levels and the conceptual problem of sudden marked changes in effect at the various levels. It was of interest however, that a post-hoc search for an "optimal cut-point" for the cohort effect (using the maximal chi-square of the generalised log-rank statistic via isotonic regression analysis (Putten 2002)) identified the same cut-point (July to August 1994) as was used in the study (data not shown).

8.5.3 Cohort effect: A cohort effect was demonstrated, with the second 5 year period having a better (risk adjusted) prognosis (see Table 8.4.4). That this was not an effect of a change in referral pattern or severity and / or type of illness was indicated by the non-significance of the interaction between the cohort effect and lead-time, APACHE II score and tumor diagnosis. There was also no difference in the percentage of censored patients, nor in the distribution of censored survival times, early versus late cohort ($p = 0.13$ & $p = 0.9$, respectively), which could have potentially explained this cohort effect. There would appear to be little comment in the specific haematological-oncological literature about ICU mortality improvement over time; two recent notes have suggested that this may not be the case (Martin *et al.* 1998; Hulme *et al.* 1999).

8.5.3.1 However, Azoulay *et al* (Azoulay *et al.* 2001) reported a single institution study with an improved survival in 105 patients of a 1996 to 1998 cohort compared with 132 in a 1990 to 1995 cohort (the overall 30 day mortality was 72.5%,). Using a nested cohort study, the authors proposed that the use of non-invasive mechanical ventilation (NIMV) was protective and thus an explanation of the cohort effect. There are potential problems in such a scenario: the interpretation of treatment effects (non-random allocation to NIMV) is known to be problematic in cohort studies (Copas *et al.* 1997), the positive effect of NIMV on mortality in randomised studies involving non-COPD patients has yet to be demonstrated (Brookes *et al.* 2001b) and a (significant) treatment-cohort interaction was not reported. Furthermore, an analysis of the matching of exposed (NIMV)-unexposed (no NIMV) patients yielded differences ($p < 0.1$) in 7 patient characteristics, of which 6 favoured the NIMV group in terms of their impact upon mortality, suggesting the potential for bias in the final odds ratio estimate of the effect of NIMV. The factor(s) responsible for the cohort effect in the current study were not immediately apparent and no statistically significant patient heterogeneity about the cohort effect was demonstrated (Table 3). Overall power considerations may also have been important, especially in the elucidation of important interactions, given the requirement for increased patient numbers (approximately 4×) to demonstrate any interactions as compared with the main-effects (Brookes *et al.* 2001a). The observed improvement in outcome with time may be consonant with the overall improvement in ICU outcomes noted by Azoulay *et al* (Azoulay *et al.* 2001) and Kress *et al* (Kress *et al.* 1999) and identified in other specific patient groups (Milberg *et al.* 1995).

8.5.4 Utility of severity of illness scoring: A number of papers have assessed the utility of severity scores and a general relationship between increase of severity of illness and mortality has been observed (Guiguet *et al.* 1998; Kress *et al.* 1999; Azoulay *et al.* 2000). However no study has specifically considered the statistical consequences of a lack of independence between organ failure categorisation and severity scores, or between organ failures per se. Moreover, as discussed below with respect to the effect of mechanical ventilation, organ failure (as a categorical variable) may also be considered as an outcome. As such a competing risks approach would more effectively encompass organ failure (Moeschberger *et al.* 1995) and in the presence of such multiple confounding dependencies, the current study used daily APACHE II scores.

8.5.5 Premorbid assessment: Although the APACHE II score incorporates an assessment of chronic health status, the predictive ability of other composite chronic health indices, Karnofsky score and Charlson comorbidity index, was also assessed. The Karnofsky score was not found to be predictive. Other referenced studies used univariate assessment of chronic health status, which is paradoxical as the multivariable Charlson index (Charlson *et al.* 1987), introduced in 1987, was validated in a cohort of patients with breast cancer. This index has also been found to be of predictive value in other patient groups (Fried *et al.* 2001).

8.5.6 Impact of mechanical ventilation: Most studies have found mechanical ventilation to be an independent adverse predictor of outcome, as with the present report. The comments above, regarding the relationship of overall mortality and severity of illness, are also pertinent to survival from mechanical ventilation; the recent studies of Azoulay *et al* (Azoulay *et al.* 2001) and Kress

et al (Kress *et al.* 1999) being illustrative here. In the former study, the median(inter-quartile range) SAPS II score of those ventilated was 44.5(36-59) with a 71% mortality rate; in the latter, ventilated patients had a SAPS II score of 51(41-62) and APACHE II score 22(16-27) with a mortality of 67%. In the current study, a 73% mortality with ventilated patients was seen with a median(inter-quartile range) APACHE II score of 35(27-40) and computed SAPS II score of 65(54-73). Thus claims for low mortality or changes of mortality must be interpreted against the background severity of illness.

8.5.6.1 Some cautions may apply to the interpretation of the effect of mechanical ventilation. There was significant statistical dependence between APACHE II score, mechanical ventilation and inotrope dependency (data not shown). Ventilation may also be considered as a surrogate for outcome, as in the paper by Crawford and Peterson (Crawford *et al.* 1992). Thus, within the same data sets, these outcomes are correlated and, therefore, it may be no surprise to find a significant effect of ventilation. The same would apply to any putative relation between outcome and prolonged ventilation and / or length of stay. The bias of estimators that adjust for a concomitant variable affected by treatment has been noted (Rosenbaum 1984; Copas *et al.* 1997). Marginal structural and g-estimation models may offer less biased estimation, but they are not pursued at this juncture (Robins *et al.* 1992; Robins *et al.* 2000; Fewell *et al.* 2004). This being said, no colinearity was demonstrated between APACHE II score, mechanical ventilation and inotrope dependency and there was no substantive change in the point estimates and standard errors of the hazard ratios of the Charlson comorbidity index, lead time, APACHE II score

and cohort when mechanical ventilation (considered as a categorical variable) was excluded from the estimation.

8.5.7 Random effects (frailty) model: The notion of frailty or individual (or group) heterogeneity and its extension to survival and event data has assumed some importance recently (Hougaard 1995). A frailty model is a random effects model for time variables where the random effect has a (latent) multiplicative effect on the hazard; that, is the hazard at time *t*, for an individual with frailty *Y*, is $h(t|Y=y) = yh_0(t)*\exp(x_j.b)$). Frailty addresses unexplained variability (in time to failure) in terms of omitted covariate(s) or measurement error; thus, if frailty is ignored, an underestimation of covariate effect will be observed (Henderson *et al.* 1999). The approach adopted in this paper was a generalization to the multivariate case (Pickles *et al.* 1994) in endeavour to identify patient heterogeneity with respect to both ventilation and cohort effect, where the some uncertainty existed about the adequacy of modelling of the effect. Although the random effects point estimates suggested a modification of the effect of the recorded significant covariates (Table 8.4.6), there was no statistically significant advantage (likelihood ratio test) between the random effects and the fixed effects Poisson model; that is, the traditional pooled analysis was sufficient.

8.6 Conclusions:

8.6.1 It is concluded that mortality outcomes were consonant with the severity of illness and that improved 30 day survival occurred over a 10 year period. Questions of excess mortality or appropriateness of care must be considered against the background of the level of severity of illness and cohort composition, including comorbidity burden. Analysis restricted to admission data alone may

be insensitive to particular covariate effects. Neither tumour type nor recovery of neutrophil or platelet count over first 8 days after admission was prognostic, but mechanical ventilation would appear to be an independent mortality determinant. Peak mortality hazard occurs proximate to ICU admission and declines monotonically thereafter.

9 METHODOLOGY IN META-ANALYSIS

9.1 Introduction: Since its introduction by Glass in 1976 (Glass 1976), meta-analysis has become an established review process within the medical and Critical Care literature (Egger *et al.* 2001). Some controversy, however, attends particular aspects of meta-analytic practice; these may be conveniently classified as (i) procedural (Edwards *et al.* 2002), relating to the individual studies of the particular meta-analysis: questions relating to search strategy, study quality, the impact of large versus small studies and the propriety of pooling the studies in the first place and (ii) statistical: these questions relate to the choice of the metric of the treatment effect in the case of binary outcomes, the diagnosis of and adjustment for both heterogeneity of treatment effects and publication bias, and the relation of the treatment effect to underlying demographic and patient characteristics (meta-regression).

9.1.1 In selected meta-analyses relating to Critical Care practice, the aim of this study was to investigate certain aspects of the meta-analytic process which were deemed central to the interpretation of any pooled effect estimate; in particular (i) heterogeneity: by the frequency of determination in individual reports; the frequency of “undiagnosed” heterogeneity based upon conventional (asymptotic) statistics; the effect of treatment effect metric upon the sampling distribution of the statistic used to diagnose heterogeneity and the overall “impact” of heterogeneity on the pooled estimate. Heterogeneity reflects clinical, methodological and / or statistical features of the component studies and the search for “predictors of between-study heterogeneity” may be the “primary value” of meta-analysis (Greenland 1994) (ii) publication bias: by its frequency

of determination within individual reports and the sensitivity of various methods of its determination. Publication bias may represent *the* major problem in meta-analysis (Dickersin *et al.* 1993) (iii) meta-regression. The positive relationship between severity of illness and treatment efficacy (improved efficacy in the more severely ill) is the dominant paradigm in the interpretation of therapeutic interventions in the critically ill (Knaus *et al.* 1996; Knaus *et al.* 1985a). In the absence of uniform reporting of patient severity of illness scores in individual studies (for example, APACHE II scores (Knaus *et al.* 1985b), where a linear relationship between score and mortality holds), control arm risk would appear an adequate surrogate for patient severity of illness (Eichacker *et al.* 2002; Sun *et al.* 1996b) and the sensitivity of different meta-regression methodologies in identifying the relationship between treatment effect and control arm risk was explored. As the "…most appropriate effect measure might be any of the three scales" (Warn *et al.* 2002), the explorations of heterogeneity, publication bias and metaregression of control arm risk covered all three metrics, odds ratio (OR), risk ratio (RR) and risk difference (RD).

9.2 Materials and methods

9.2.1 Study selection: Meta-analyses were identified via electronic (MEDLINE™) and hand journal search; the sample was restricted to meta-analyses in Critical Care practice reflecting major therapeutic concern. In this sense, the purpose was similar to recent studies (Gyldmark 1995; Steyerberg *et al.* 2000) where in-depth methodological analysis has been the primary focus, rather than the elucidation of a pooled estimate of effect across an exhaustive catalogue. Original references of the studies cited in the meta-analyses were accessed and where indicated, data was transcribed for analysis. Note was taken of the assessment and presence of

both heterogeneity and publication bias in the reports of each of the meta-analyses. All meta-analyses considered in this study were re-analysed, as indicated below.

9.3 Statistical methodology

9.3.1 Treatment effect: (mortality as outcome) was computed, using both fixed (Mantel-Haenszel) and random effects (DerSimonian-Laird) model estimators, as OR, RR and RD using the "metan" routine (Bradburn *et al.* 1998) and Stata® statistical software (Cochran 1954a).

9.3.2 Heterogeneity: was assessed for each of the metrics using (i) the *Q* statistic where weights are the reciprocal of the variance of the treatment effect (Cochran 1954b; DerSimonian *et al.* 1986) and (ii) the I^2 statistic, which summarizes the impact (rather than the extent, as given by *Q*) of heterogeneity (formally, the percentage of variability in point estimates due to heterogeneity rather than sampling error) as being "mild" if < 30%, and "notable" if >> 50% (Higgins *et al.* 2002). Homogeneity of OR was also separately determined using the Zelen exact test via StatXact-4® software (Stat Xact 4 1999). Empirical bootstrap distributions (Carpenter *et al.* 2000) of the *Q* statistic for each meta-analysis, for the metrics OR, RR and RD, were produced and graphically represented using kernel density plots (Fox 1990).

9.3.3 Publication bias: was assessed across all metrics by funnel plots (Light *et al.* 1984) and formal quantitative tests of publication bias; in particular, the adjusted rank correlation (Begg *et al.* 1994) and regression asymmetry (Begg *et al.* 1994) tests, and "trim and fill" methodology of Duval and Tweedie (Duval *et al.* 2000b), as implemented in the Stata® routines "metabias" and "metatrim" (Steichen 1998; Steichen 2000). Funnel plot axes were set to treatment effect

and standard error and significance was ascribed at p = 0.1 for the adjusted rank correlation and regression asymmetry tests (Sterne *et al.* 2001b). For the "metatrim" routine, the analysis was based on both random and fixed effects meta-analytic point estimates and the linear iterative trimming estimator was used. As analysis utilising the funnel plot with < 10 individual studies is not recommended (Sterne *et al.* 2000b), this threshold was used within the current study to select meta-analyses for reporting, in accordance also with a recent exemplar study using "trim-and-fill" methodology (Sutton *et al.* 2000). Evidence for publication bias via the "metatrim" routine was denoted as "some" if the number of filled studies was > 0 and "significant" ($p < 0.05$) if the number of filled studies was > 3 (Duval *et al.* 2000b; Sutton *et al.* 2000).

9.3.4 Metaregression (Berlin *et al.* 1994): of treatment effect (log OR, log RR and risk difference) versus control arm risk (log odds) was also undertaken using (i) weighted (inverse variance) least squares regression (WOLS) (ii) restricted maximum likelihood (REML), as implemented in the Stata® routine "metareg" (Sharp 1998) and (iii) Bayesian hierarchical regression with WINBUGS software (Speigelhalter *et al.* 2000), using Markov Chain Monte Carlo methodology with Gibbs sampling to obtain empirical estimates of the true posterior parameters, as implemented by Warn *et al* (Warn *et al.* 2002). The method was to model the joint distribution of the points of a L'Abbe plot of the "true" treatment and control group risks; the latter having independent uniform(0,1) priors, equivalent to treating them as "fixed effects". For τ, the following uniform priors were placed: risk difference (-1,1), log relative risk (0,2) and log odds ratio (0,2); for δ (the treatment effect): on the absolute risk scale, uniform (-1,1), on the relative risk scale, normal (0,10) and on the log

odds scale, normal (0,10). Median estimates of slope and 95% credible Bayesian intervals for each of the treatment metrics were computed using BUGS software, with burn-in of 5000 iterations and 10000 iterations being subsequently monitored and used to estimates posterior quantities. Convergence was assessed by the methods of Gelman and Rubin (Gelman *et al.* 1996).

9.4 Results

9.4.1 Fourteen meta-analyses were considered (Table 9.4.1): (1) effect of parenteral nutrition vs standard therapy (Heyland *et al.* 1998), (2) corticosteroid treatment for sepsis (Cronin *et al.* 1995), (3) efficacy of selective decontamination of the digestive tract (Selective Decontamination of the Digestive Tract Trailists' Collaborative Group 1993), (4) efficacy of maximizing oxygen delivery in sepsis (Heyland *et al.* 1996), (5) non-invasive ventilation in respiratory failure (Keenan *et al.* 1997), (6) human albumin administration in critically ill patients (Cochrane Injuries Group Albumin Reviewers 1998), (7) stress ulcer prophylaxis in critically ill patients; sucralfate versus antacids (Cook *et al.* 1996), (8) inflammatory therapies in sepsis (Zeni *et al.* 1997), (9) colloid or crystalloid solutions in critically ill patients (Schierhout *et al.* 1998), (10) enteral nutritional supplementation with key nutrients in patients with critical illness and cancer (Heys *et al.* 1999), (11) immuno-nutrition in the critically ill (Beale *et al.* 1999), (12) effect of early versus delayed enteral nutritional support (Zaloga 1999), (13) survival after human albumin administration, update re-analysis (Wilkes *et al.* 2001c), and (14) update on immuno-nutrition use in the critically ill (Heyland *et al.* 2001b). An additional meta-analysis on non-invasive ventilation in respiratory failure, using similar methodology, has been the subject of a further report and was not considered here (Peter *et al.* 2002). Table

9.4.1. shows details of the above meta-analyses in terms of patient numbers (per study and per meta-analysis), control and treatment arm mortalities; no differences were noted between the estimates in the current study and those of the primary references, albeit estimates of all metrics were not presented in these primary references. Three meta-analyses (Cronin *et al.* 1995; Heyland *et al.* 1996; Keenan *et al.* 1997) had less than 10 trials each; minimum to maximum patient number per trial showed considerable variation (11 to 808, median number 15) as did event rates (mortalities from 0.05 to 0.45). Heterogeneity was formally assessed in all but 2 meta-analyses and publication bias in only 4.

9.4.2 Table 9.4.2. displays, for each metric, the treatment effects and estimates of the "impact" of heterogeneity as the I^2 measure, expressed as percentages. Of the 14 meta-analyses, consistent treatment effect across the metrics, for both fixed effect and random effects, was seen in three (Cochrane Injuries Group Albumin Reviewers 1998; Keenan *et al.* 1997; Zeni *et al.* 1997) and non-uniform treatment effects (with $p<0.1$) in two (Heys *et al.* 1999; Schierhout *et al.* 1998). Heterogeneity was evident ($p < 0.1$) in two instances (Cronin *et al.* 1995; Heyland *et al.* 1996) for all treatment effect metrics and in one (Cook *et al.* 1996) for the risk difference metric. As diagnosed by the Zelen exact test, but not identified by standard analysis, heterogeneity was present in another three meta-analyses; at $p = 0.07$ (Heyland *et al.* 1998), $p = 0.001$ (Beale *et al.* 1999) and $p = 0.008$ (Heyland *et al.* 2001b). For the OR metric at least, (assuming that the Zelen test represented a "gold standard"), the *Q* test (threshold, $p <0.1$) had a sensitivity of 40% and a negative predictive value of 75%. The assessment of heterogeneity impact by the I^2 measure, across meta-analyses and metrics, was generally consistent with the inferences afforded by a combination of the *Q* and

Zelen tests; notably, the three meta-analyses (Beale *et al.* 1999; Heyland *et al.* 1998; Heyland *et al.* 2001b) identified by the Zelen, but not the *Q* test, had I^2 percentages > 0% <

Table 9.4.1. Overview of meta-analyses considered

Number	Meta-analytic Study types	Year	Trials	Min Pat.	Max Pat.	Total Pat	Control Mort.	Treat Mort.	Heterog Assessed	PB Assessed
1	Parenteral Nutrition[1]	1998	27	18	459	2259	0.14	0.12	yes	no
2	Steroids in sepsis[2]	1995	9	48	381	1297	0.44	0.45	yes	no
3	Selective gut decontamination[3]	1993	23	31	445	4142	0.32	0.29	yes	no
4	Oxygen supply in sepsis[4]	1996	7	51	808	1977	0.37	0.38	yes	no
5	Non-invasive ventilation[5]	1997	5	11	85	228	0.36	0.12	yes	no
6	Albumin supplementation[6]	1998	24	14	219	1204	0.13	0.19	yes	yes
7	Gastric bleeding prophylaxis[7]	1996	11	50	213	1348	0.24	0.22	yes	no
8	Anti-inflammatory agents in sepsis[8]	1997	18	29	971	6380	0.40	0.38	no	no
9	Colloid vs crystalloid[9]	1998	19	12	174	1315	0.20	0.21	yes	yes
10	Immunonutrients[10]	1999	11	28	279	1008	0.05	0.08	yes	no
11	Immunonutrition[11]	1999	12	22	390	1529	0.14	0.12	yes	yes
12	Early enteral nutrition[12]	1999	12	15	195	793	0.10	0.11	no	no
13	Albumin supplement_review[13]	2001	42	16	300	2958	0.18	0.21	yes	yes
14	Immunonutrition_review[14]	2001	22	20	390	2419	0.12	0.13	yes	no

Trials, number of trials in each meta-analysis. Year, year of publication. Min Pat., minimum number of patients per meta-analysis. Max Pat., maximum number of patients per meta-analysis. Total Pat., total number of patients per meta-analysis. Control Mort., control arm mortality. Treat Mort., treatment arm mortality. Mortality as mean values for each meta-analysis. Where individual studies have no events (deaths), these have not been considered in the calculations. Heterog Assess, was heterogeneity formally assessed in the meta-analysis. PB Assessed, was publication bias formally assessed in the meta-analysis.

Table 9.4.2. Effect measures of meta-analyses

Meta-analysis type	OR_{FE}	OR_{RE}	RR_{FE}	RR_{RE}	RD_{FE}	RD_{RE}	HET_{OR}	HET_{RR}	HET_{RD}
Parenteral Nutrition[1]	0.98	0.95	0.98	0.98	-0.002	0.000	14.0^{Z}	12.2	0.0
Steroids in sepsis[2]	1.13	1.03	1.07	1.02	0.026	0.012	72.8**	70**	77.7**
Selective gut decontamination[3]	0.9	0.91	0.93	0.93	-0.019	-0.016	0.0	0.0	0.0
Oxygen supply in sepsis[4]	0.9	0.94	0.94	0.96	-0.023	-0.008	71.6**	69.6**	73.1**
Non-invasive ventilation[5]	0.3**	0.32**	0.41**	0.45**	-0.193**	-0.187**	0.0	0.0	13.8
Albumin supplementation[6]	1.92**	1.84**	1.68**	1.46**	0.069**	0.046**	0.0	0.0	0.1
Gastric bleeding prophylaxis[7]	0.79	0.78#	0.85	0.82#	-0.034	-0.018	2.1	0.0	33*
Anti-inflammatory agents sepsis[8]	0.9*	0.9*	0.93*	0.93*	-0.025*	-0.024#	0.0	0.0	0.0
Colloid vs crystalloid[9]	1.3#	1.3#	1.19#	1.15	0.039#	0.017	0.0	0.0	0.0
Immunonutrients[10]	1.76*	1.75#	1.67*	1.66#	0.028#	0.009	0.0	0.0	0.0
Immunonutrition[11]	1.08	1.01	1.05	1	0.09	0.007	29.6^{Z}	27.9	18.2
Early Enteral nutrition[12]	1.09	1.1	1.08	1.06	0.006	-0.001	0.0	0.0	4.2
Albumin supplement_review[13]	1.14	1.1	1.12	1.06	0.018	0.020	0.0	0.0	13.4
Immunonutrition_review[14]	1.17	1.17	1.12	1.1	0.015	0.003	7.7^{Z}	6.5	0.0

OR, odds ratio. RR, risk ratio. RD, risk difference. FE, fixed effect. RE, random effects.
HET, heterogeneity. Numbers refer to I^2 statistic (as percentage) of Higgins & Thompson [57].
#; $0.05 < p < 0.1$. *; $0.01 < p < 0.05$. **; $0.001 < p < 0.01$
^{Z}, heterogeneity present by Zelen exact test

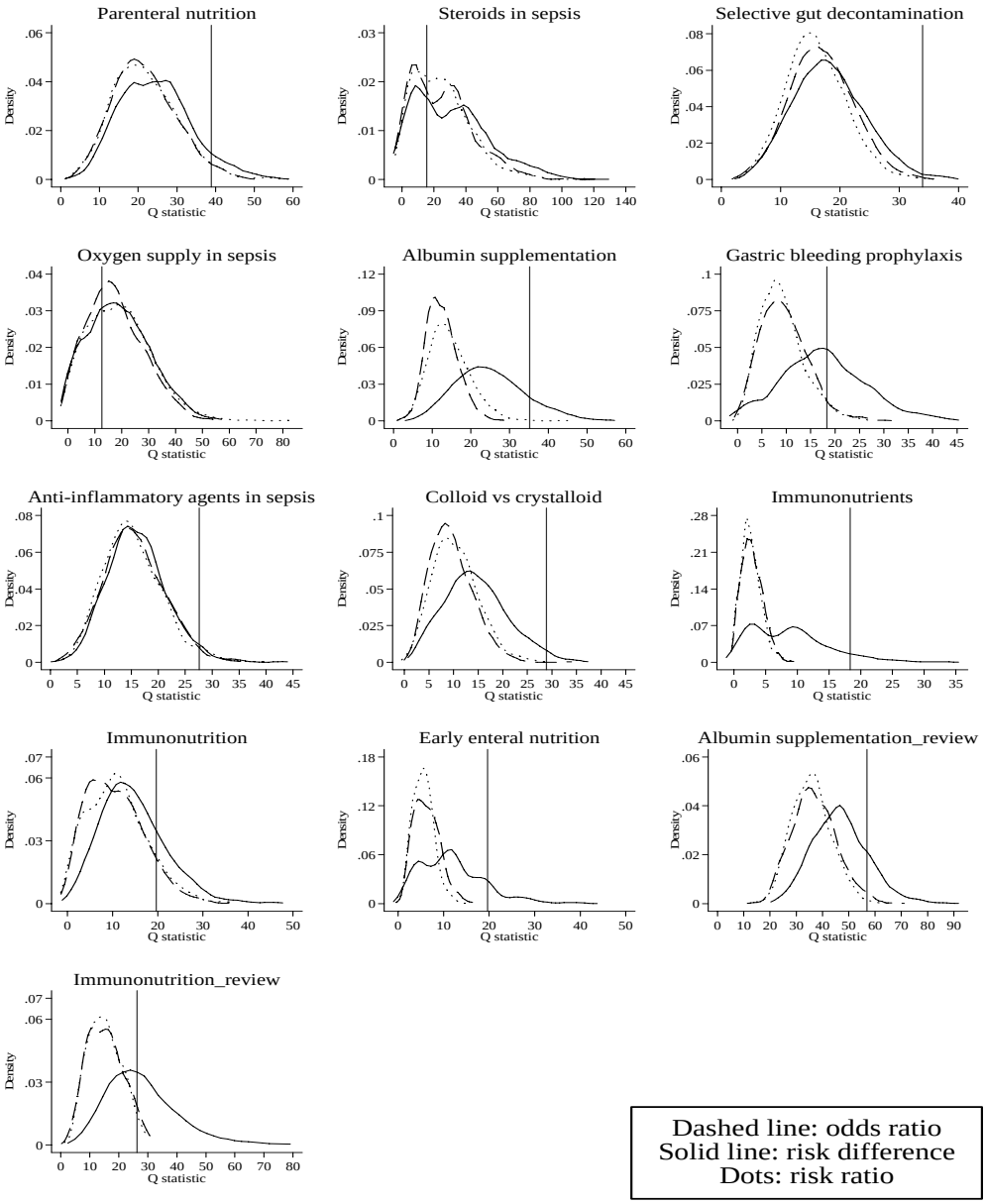
30% and the single meta-analysis (Cook *et al*. 1996) with heterogeneity at p = 0.03 for RD had an I2 = 49.1%.

9.4.3 Figure 9.4.3 shows kernel density plots of the bootstrap distribution of the *Q* statistic for the 13 meta-analyses (in separate panels) for the three treatment effect metrics: OR (dashed lines), RR (dots) and RD (solid line). Due to the small number of individual studies (n=5) within the non-invasive ventilation in respiratory failure meta-analysis (Keenan *et al.* 1997), no bootstrap distribution was generated. For each panel a vertical line indicates the value of the *Q* statistic for a p value of 0.05. Overall, the bootstrap distribution of the risk difference metric was displaced to the right, closer to the "significant" value of the *Q* statistic. The distributions for OR and RR tend to approximate each other. Formal testing of the distributions over the 13 meta-analyses by the Kruskall-Wallis test revealed a significant difference between the RD distribution and that of both the OR and RR (p = 0.01, adjusted for multiple comparisons) and no difference between OR and RR distributions (p = 0.11).

9.4.4 Funnel plot asymmetry was diagnosed variably between the studies and not uniformly across metrics (Table 9.4.4), noting that three meta-analyses (Cronin *et al.* 1995; Heyland *et al.* 1996; Keenan *et al.* 1997) were excluded from consideration because of individual study numbers < 10. Publication bias was suggested by (i) the regression symmetry test in the OR metric for one meta-analysis (p = 0.0004 (Cook *et al.* 1996)), in the RR metric for three (p = 0.04 (Heyland *et al.* 1998), p = 0.01 (Cook *et al.* 1996), p = 0.03 (Wilkes *et al.* 2001c)) and in one for the RD (p = 0.056 (Selective Decontamination of the Digestive Tract Trailists' Collaborative Group 1993)) and (ii) the adjusted rank correlation test in only one meta-analysis for the OR (p = 0.01 (Cook *et al.*

1996)). Using a threshold of > 1 filled study as evidence of publication bias (see Materials and methods; Publication bias, above), “Trim and fill” methodology identified publication bias in the OR metric (Table 9.4.4) in 3 meta-analyses (Cook *et al.* 1996; Wilkes *et al.* 2001c; Zeni *et al.* 1997). The number of “filled” studies varied from 1 to 8, but a change from non-significant to a significant treatment effect ($p < 0.05$) with “filling” occurred in one only (both fixed and random effects estimators (Cook *et al.* 1996)). In the RR metric, 4 meta-analyses (Cochrane Injuries Group Albumin Reviewers 1998; Heyland *et al.* 2001b; Wilkes *et al.* 2001c; Zaloga 1999) exhibited evidence of publication bias. The number of “filled” studies varied from 1 to 7, but a change from non-significant to a significant treatment effect ($p < 0.05$) with “filling” occurred in one only (fixed estimator (Cook *et al.* 1996)). In the RD metric, 4 meta-analyses exhibited publication bias (Cochrane Injuries Group Albumin Reviewers 1998; Heyland *et al.* 2001b; Wilkes *et al.* 2001c; Zaloga 1998), the number of “filled” studies varying from 1 to 5, with no change from non-significant to a significant treatment effect. Across the metrics the other (non-significant) change that occurred in treatment effects with “filling” were those of shrinkage to the null, although this was inconsistent.

Figure 9.4.3. Kernel density plots of bootstrap distribution of *Q* statistic for meta-analyses.



Composite kernel density plots of the bootstrap distribution of the Q statistic for the three treatment effect metrics; OR dashed line, RD solid line, RR dors. Vertical line shows the value of the Q statistic at p=0.05 for trial number for each meta-analysis.

Table 9.4.4. Tests of publication bias

Meta-analysis type	FP OR	FP RR	FP RD	$T\&F_{OR}$ Studies filled	Change	$T\&F_{RR}$ Studies filled	Change	$T\&F_{RD}$ Studies filled	Change
Parenteral Nutrition[1]	yes	yes	yes	0		0		0	
Selective gut decontamination[3]	no	yes	yes	0		0		0	
Albumin supplementation[6]	no	no	yes	0		2	1.46 to 1.41 *p: 0.008 to 0.01*	2	1.05 to 1.04** *p: 0.002 to 0.012*
Gastric bleeding prophylaxis[7]	yes	yes	no	3	0.77 to 0.70# *p: 0.08 to 0.01*	3	0.82 to 0.78## *p: 0.06 to 0.02*	0	
Antiinflammatory agents in sepsis[8]	no	no	no	1	0.90 to 0.89 *p: 0.05 to 0.04*	1	0.93 to 0.93 *p: 0.04 to 0.03*	0	
Colloid vs crystalloid[9]	yes	yes	yes	0		0		0	
Immunonutrients[10]	no	no	no	0		0		0	
Immunonutrition[11]	no	no	no	0		0		0	
Early Enteral nutrition[12]	no	no	yes	0		0		1	1.0 to 1.0 *p: 0.89 to 0.73*
Albumin supplementation_review[13]	yes	yes	yes	8	1.10 to 0.99 *p: 0.34 to 0.91*	7	1.06 to 1.01 *p: 0.44 to 0.88*	5	1.02 to 1.01 *p: 0.1 to 0.5*
Immunonutrition_review[14]	yes	no	no	0		0		3	0.99 to 1.0 *p: 0.9 to 0.59*

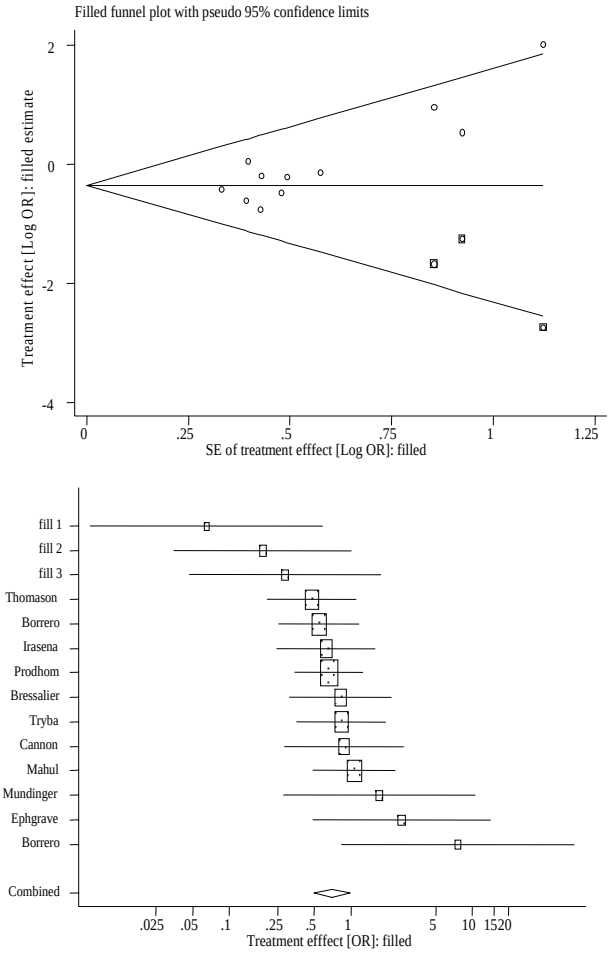
FP; funnel plot asymmetry assessed by visual inspection of funnel plot (treatment effect vs SE of effect) for the three metrics; odds ratio (OR), risk ratio (RR) and risk difference (RD). T&F; Trim and Fill methodology; evidence of publication bias for the three metrics; odds ratio (OR), risk ratio (RR) and risk difference (RD).
Studies; number of potential studies "filled" within each meta-analysis. Change; change in treatment effect, p value with "filling".
#; change for random effects estimator also (0.78 to 0.7; p: 0.08 to 0.05)
##; change for random effects estimator also (0.82 to 0.78; p: 0.06 to 0.05)
**; change for random effects estimator also (1.05 to 1.04; p 0.002 to 0.04)

9.4.4.1 Figure 9.4.4 shows the “filled” funnel and estimates (forest) plot for the Gastric bleeding prophylaxis meta-analysis (Cook *et al.* 1996). Three “studies” were found to be potentially missing and were “filled” into the funnel and forest plots; the treatment effect (OR metric) being modified from 0.77 (p=0.08) to 0.70 (p=0.01). These changes are consistent with those reported in the study of Sutton *et al* (Sutton *et al.* 2000).

9.4.5 Table 9.4.5 shows the estimates and SE of the β coefficients from the various meta-regressions for treatment effect (log OR and log RR and RD) versus control risk (log odds). Results are reported for weighted least squares regression and REML only where the slopes were “significant”; that is at a level of $p < 0.5$. The cross-over point(s) (intersection with log odds ratio = 0) should not be interpreted as a strict efficacy indicator, as analysis (not pursued here) has shown that such a point has wide 95% confidence intervals (Sharp *et al.* 2000). Differential sensitivity in the identification of significant relationships between the two conventional estimators and across metrics was evident; point estimates were similar, but SEs were larger, as expected, for REML regression. Bayesian analysis showed “significant” (95% credible intervals not spanning zero) regression slopes in various meta-analyses across the metrics with risk difference being the most sensitive; slope coefficients tended to be greater (more negative) compared with either weighted least squares or REML. Consistent risk related treatment effects across estimators and metrics was seen in only one meta-analysis (Selective Decontamination of the Digestive Tract Trailists' Collaborative Group 1993) and differential effects (risk related treatment effects by least squares and REML but not with Bayesian analysis) in a second (Cochrane Injuries Group Albumin Reviewers 1998). The oxygen supply in

sepsis meta-analysis (Heyland *et al.* 1996) was unique in that the slope coefficient was positive,

Figure 9.4.4. “Trim and fill” methodology for Gastric bleeding prophylaxis meta-analysis.



Top panel: Regression asymmetry test; theta (treatment effect as odds ratio) on the vertical axis against SE on the horizontal. Horizontal line shows pooled treatment effect with diagonal lines the 95% “pseudo” confidence levels. Small circles: trial data. Squared circles: filled data. Bottom panel: Forrest plot showing individual trials on the vertical axis and theta (treatment effect as odds ratio) on the horizontal. Point estimates are shaded boxes (size reflecting weight in the analysis) and horizontal lines as 95% CI. Filled studies are shown as fill1, fill2 and fill3. Overall estimate (including filled studies) with 95% CI is shown at the bottom as “Combined”

Table 9.4.5. Meta-regression estimates of significant ($p \leq 0.05$) β (slope) coefficients plus SE.

Estimator	WOLS log OR Coeff (SE)	WOLS log RR Coeff (SE)	WOLS RD Coeff (SE)	REML logOR Coeff) (SE)	REML logRR Coeff (SE)	REML RD Coeff (SE)	MCMC log OR Mcoeff (95%CrInt)	MCMC log RR Mcoeff (95%CrInt)	MCMC RD Mcoeff (95%CrInt)
Meta-analysis									
Selective gut decontamination[3]	-0.260 (0.09)*	-0.164 (0.061)*	-0.055 (0.015)**	-0.260 (0.112)*	-0.16 (0.078)*	-0.055 (0.023)*	-0.286 (-0.508 / -0.052)	-0.247 (-0.431 / -0.023)	-0.328 (-0.501 / -0.135)
Oxygen supply in sepsis[4]			0.181 (0.072)*	1.245 (0.589)*	0.738 (0.38)*	0.199 (0.09)*		0.978 (0.271 / 1.994)	0.661 (0.199 / 1.475)
Albumin supplementation[6]	-0.263 (0.123)*	-0.217 (0.06)*			-0.217 (0.1)*				
Gastric bleeding prophylaxis[7]	-0.392 (0.144)*		-0.029 (0.01)*	-0.392 (0.181)*		-0.029 (0.01)*			-0.372 (-0.642 / -0.091)
Anti-inflammatory agents in sepsis[8]									-0.305 (-0.690 / -0.015)
Albumin supplementation_review[13]	-0.334 (0.096)**	-0.197 (0.065)**		-0.332 (0.116)***	-0.201 (0.08)*				-0.202 (-0.438 / -0.010)

WOLS; weighted (inverse variance of dependent variable) least squares regression. REML; restricted maximum likelihood.
MCMC; Bayesian hierarchical regression using Gibbs sampling. Coeff(SE); β coefficient (slope) plus standard error.
LogOR; log odds ratio. logRR; log risk ratio. RD; risk difference
Mcoeff; median estimate of (slope) coefficient. 95%CrInt; 95% credible (Bayesian) interval
*; 0.01<p<0.05. **; 0.001<p<0.01

implying that treatment effects were progressively adverse (increase in log OR) for increase in base-line risk.

9.5 Discussion:

Our review of selected aspects of the statistical practice of meta-analysis has identified a number of areas of uncertainty and we will comment upon these in turn.

9.5.1 Heterogeneity: Tests for heterogeneity were performed in the original reports of all the published meta-analyses considered, the omissions being in two quantitative reviews (anti-inflammatory agents in sepsis (Zeni *et al.* 1997) and early enteral nutrition (Zaloga 1999)), which had the status of editorial commentary. Checking for evidence of heterogeneity is undoubtedly frequently performed in meta-analyses (Hahn *et al.* 2000a) and a general test of heterogeneity is the asymptotic Chi-squared based *Q* test (or its variants). The low power the *Q* test has often been often noted (Thompson *et al.* 1991) and it has been suggested that a p value of 0.1 is the more appropriate level (Fleiss 1986a), although in the studies under consideration, heterogeneity (assessed via the *Q* test) was observed at $p \leq 0.003$, except for the single case in the risk difference metric ((Cook *et al.* 1996), p=0.03). In the current study the sensitivity of the *Q* test was assessed as being less than the preferred (Emerson 1994) Zelen exact test, but formal demonstration of such would require a simulation study (Macaskill *et al.* 2001), which was not pursued at this stage. The recently described I^2 measure (Higgins *et al.* 2002), which is able to be calculated across metrics and is independent of the number of studies in a meta-analysis, unlike the *Q* statistic, allowed further insight into heterogeneity of the studies under consideration and may be a preferred measure.

9.5.1.1 An empirical study of 125 meta-analyses (Engels *et al.* 2000) found that the RD metric usually displayed more heterogeneity than the OR (and RR); Fleiss made a similar analytic observation on the basis that variation of RD across studies in a meta-analysis, being constrained by treatment and control risk, may generate an appearance of heterogeneity (Fleiss 1993). Formal statistical investigation of these issues has demonstrated that the *Q* statistic (a weighted least squares test procedure) has anti-conservative Type I error rates for the RD metric, especially with sparse data (Lipsitz *et al.* 1998) and is conservative with respect to the RR (Lui *et al.* 2000) when the number of individual studies within a meta-analysis is large and the group size is small. The findings above with respect to heterogeneity and RD metric were consistent with both the bootstrap distributions of the *Q* statistic (Figure 9.4.3) and the values of the I^2 measure (Table 9.4.2).

9.5.1.2 If the decision to employ the random effects estimator is made on solely on the basis of heterogeneity (however diagnosed), the low power of the asymptotic test may confound the choice, given that performance of the exact test requires specialised software. Thus it may be reasonable to present results from both estimators (Normand 1999) or simply defer to random effects (Takkouche *et al.* 1999), although the latter strategy has been criticised on the basis that heterogeneity may be overestimated (Sterne *et al.* 2002). That the random effects approach is not necessarily conservative, that is, associated with wider confidence intervals, has recently been demonstrated (Poole *et al.* 1999), although the commonly used DerSimonian and Laird method (method of moments estimator) may be associated with coverage below the nominal level of the overall measure of effect (Brockwell *et al.* 2001).

9.5.2 Metaregression:

9.5.2.1 The heterogeneity, diagnosed or suspected, of treatment effect between trials within a meta-analysis may be investigated by meta-regression, using either recorded trial covariates or surrogates for these. As noted above and pertinent to the paradigm of treatment efficacy in the critically ill, the control event rate (control mortality) would seem to reasonably approximate severity of illness and has been used to explore efficacy of therapy in two studies (Eichacker *et al.* 2002; Sun *et al.* 1996b) within the Critical Care literature, not withstanding the fact that weighted least squares estimates were used. A considerable, albeit controversial literature, has addressed the appropriate analysis and interpretation of the (meta)regression of treatment effect against control event rate, which analysis has also been addressed in this study. In its simplest form, such a regression amounts to fitting a least squares regression line to the L'Abbe plot, treatment versus control mortality, with appropriate weighting, either the square root of the study size (Sun *et al.* 1996b) or the inverse variance of the dependent variable (Hoes *et al.* 1995). Other variants have used treatment effect (log OR, log relative risk, RD) regressed against control event rate (Schmid *et al.* 1998) or control log odds (Sharp 2001).

9.5.2.2 Although such regressions have identified putative risk related treatment effects in meta-analyses, these approaches have been the subject of a sustained critique on the basis that they have failed to allow for both regression to the mean (here, the difference between outcome and baseline being correlated with baseline) and the stochastic nature of the control rate (regression dilution bias) (Arends *et al.* 2000). The attempt to avoid the consequences of regression to the mean by using the average of treatment and control log odds as the

independent variable (Brand 1994) affords no solution as it implicitly assumes that the (true) treatment effect does not vary between trials, which assumption, in the meta-analytic context, cannot be sustained (Sharp 2001). The circularity of assuming lack of variability in order to estimate how treatment effect varies with underlying risk has also been noted (Sharp *et al.* 1996). The stochastic nature of the control rate induces problems because the expected response in (ordinary) linear regression is conditional upon independent (fixed) variables and there is no inherent accounting for the random error in estimation of this control rate. To overcome this problem a "measurement error" linear regression model has been proposed (Walter 1997), but again, the methodology has been criticised (Bernsen *et al.* 1999; Sharp *et al.* 2000). This model also assumed that all the heterogeneity between the (true) treatment effects of trials was accounted for by the regression and no allowance was made for residual heterogeneity. The consequences of not accounting for residual heterogeneity (τ^2) are to under-estimate the SE's of regression coefficient(s); REML, as available with the Stata™ module "metareg", overcomes this problem (Thompson *et al.* 1999). For this estimator, a normal distribution for residual errors is assumed with both a within-trial and an additive between trial component of variance (τ^2). Within trial variances are obtained from the (trial) data and τ^2 is estimated iteratively. Although REML may have particular advantage over ordinary weighted linear regression, it does not address the fundamental concerns, above, of regression to the mean and measurement error. Such concerns may be surmounted by hierarchical modelling, using either Bayesian inference with Gibbs sampling (Sharp 2001;

Thompson *et al.* 1999) or the expectation-maximization algorithm (Schmid *et al.* 1998; Schmid 1999), to estimate unknown parameters.

9.5.2.3 Table 9.4.5 shows a number of features: marked sensitivity differences between the estimators and metrics in identifying significant risk related treatment effects; similar slope coefficients, but different standard errors, between least squares and REML, and regression dilution for least squares and REML estimates (slope coefficients biased towards zero compared with hierarchical regression). Neither significance of the pooled treatment effect nor the presence of heterogeneity (Table 2) were closely related to risk dependent treatment effects, as identified by hierarchical regression. These results are similar to those of Schmid *et al* (Schmid *et al.* 1998) who demonstrated (albeit using different estimation techniques), in a large study of 115 disparate meta-analyses, increased sensitivity of identifying significant treatment dependent relationships for both weighted least squares and the risk difference metric per se, and no correlation between the presence of risk related treatment effect and the actual significance of the pooled treatment effect. The current study, extending Bayesian techniques of "control rate" meta-regression to all metrics in a sizeable number of allied meta-analyses, demonstrates the potential "fragility" of such regressions to both estimator and metric. With respect to the particulars of the Bayesian analysis, both the influence of the nature of the prior(s) and the strategy of effective modelling of the trials as "fixed effect" have been the source of some, as yet, unresolved dispute (Arends *et al.* 2000; van Houwelingen *et al.* 1999).

9.5.3 Publication bias

9.5.3.1 A surprising finding in this study was that only 29% of the meta-analyses reported a test(s) for publication bias. Publication bias is both widespread and of various aetiology and numerous tests have been described (Song *et al.* 2000), but implementation in easily accessible software has limited routine use. The most commonly used assessment would appear to be the funnel plot, although, again, it is noted that there are other causes of funnel plot asymmetry (Egger *et al.* 1997) and we therefore define publication bias in an operational sense after Sutton *et al* (Sutton *et al.* 2000). Assessment of bias depends upon subjective interpretation of a graphic and the plot, treatment effect against a measure of study precision, shows construction dependence in terms of the choice of axes (Sterne *et al.* 2001b; Tang *et al.* 2000). The sensitivity of the funnel plot to detect bias is reported to be less than that of the quantitative adjusted rank correlation or regression asymmetry tests (Macaskill *et al.* 2001), which belies the more frequent assessment of asymmetry in the funnel plots for each metric in the current study (see Table 3). The assessments across metrics would also suggest that the regression asymmetry test is more sensitive than the adjusted rank correlation test, in agreement with the analysis of Sterne *et al* (Sterne *et al.* 2000b), who noted that the performance of the two quantitative tests depended upon factors such as the number and size of trials, the event rates in treatment and control groups and the presence of heterogeneity per se. They further recommended that clear trial size variation should exist, one or more trials be of medium or large size, which criterion would appear to be satisfied in the current study (see Table 1), and the methods not be used if fewer than 10 studies appeared in a meta-analysis, which recommendation was observed in this paper.

9.5.3.2 The more recently described “trim and fill” method (a rank based data augmentation technique) has the potential advantage of both diagnosing (based upon the funnel plot) and correcting for publication bias, although it assumes that the suppression mechanism is outcome magnitude, not *P*-value, dependent. The asymmetric part of the funnel plot is first assessed for number of studies and then trimmed (using the ranks of the absolute values of the observed effect size and the sign of the effect sizes), the “true” centre of the remaining symmetrical part of the funnel is (re)estimated and then the trimmed studies and their missing counterparts are replaced around the (new) centre. The “filled” funnel plot is then used to generate a final estimate of the true effect size and its variance (Duval *et al.* 2000b). The method would appear to have a performance comparable with the other quantitative tests (Duval *et al.* 2000b; Duval *et al.* 2000a), although it has been suggested to possess a high false positive rate (Sterne *et al.* 2000a). Such a tendency may have been observed in the RD in the current study (see Table 3). However, only one (9% of studies) meta-analysis (Cook *et al.* 1996) had a subsequent change to a significant ($p<0.05$) treatment effect (with filling and re-estimation) for both fixed and random effect estimates. The number of meta-analyses identified with a significant change of treatment effect by trim and fill methods in the current study suggest a limited impact of this “missingness” and accords with the findings of the previously referenced empiric study of 48 meta-analyses, where, using a random effects model, 48% of the reviews were estimated to have missing studies (for fixed effect, 54%); but in only 4 of 48 did this result in changed estimate of statistical significance (a rate of 5-10%) (Sutton *et al.* 2000). Thus the central issue appears to be the robustness of the meta-analyses

to the estimated missing number of studies, not the mere elucidation of their frequency within or between meta-analyses (Sutton 2000); “trim and fill” methodology would appear to allow unique insight into this attribute.

9.5.4 Metrics:

9.5.4.1 The advantages of the various metrics (or scales) of treatment effect have been extensively discussed (Deeks *et al.* 2001; Engels *et al.* 2000; Fleiss 1993). Odds ratios have better statistical properties in terms of sampling distribution and are the key parameter in the linear logistic model (being a maximum likelihood estimator) (Walter 2000). The log odds scale is unbounded in both directions, but is numerically greater than the risk ratio when underlying event rates are frequent (Zhang *et al.* 1998). Risk ratio has been found empirically to be more intuitive than OR, but is bounded above in a manner dependent on the control group risk. Risk difference is immediately intuitive and expresses the consequences of no therapy (unlike both odds and risk ratios), but is constrained from –1 to 1 and suffers from potential bias with varying time to follow-up. One particular advantage of the risk difference is that it enables a number needed (NNT) and its confidence interval (Lesaffre *et al.* 2000) to be conveniently estimated. However, the NNT, as with the risk difference, is affected by the baseline risk and recent cautions have been expressed about the properties of this estimator, especially in the presence of measurement error (Hutton 2000; Smeeth *et al.* 1999).

9.5.4.2 The current study was unable to demonstrate a substantive advantage for a particular metric, similar to the findings of a recent survey of the Cochrane database (Deeks *et al.* 1997). The same authors, in a subsequent review of treatment metrics, concluded that the application of metric selection criteria,

either by a priori specification or post hoc selection, was “not straightforward”. That the metric should be a consistent estimator of treatment effect and be applicable to patients at different underlying baseline risks has been recently re-iterated (Deeks 2002; Walter 2000). This being said, it is apparent from this study that heterogeneity, risk related treatment effects and publication bias all demonstrate both estimator and metric dependence of varying degree; the RD metric would appear to be the most capricious in this regard. A strategy of choosing a metric with minimal heterogeneity would seem to be facilitated by relatively straightforward computation of the I^2 measure.

9.5.5 Consequences for meta-analytic practice: The above analyses would suggest a number of directions for meta-analyses, albeit the study sample was not inclusive of all possible meta-analyses of Critical Care practice (Gyldmark 1995; Steyerberg *et al.* 2000). The demonstration of heterogeneity and the consequent use (or not) of random effects estimates, would appear to be misplaced; such decisions are confounded by the treatment effect metric. Underlying risk related treatment effects are dependent upon neither the “significance” of the treatment effect nor the diagnosis of heterogeneity; the use of “naïve” standard regression analysis to demonstrate such a relationship cannot be recommended; appropriate estimation requires advanced modelling. The critical impact of publication bias (upon the pooled estimate) appears to be at the level of the “quantitative” effects of the (potentially) missing studies; identified and assessed by, for example, “trim and fill” methodology. Meta-analyses may, in fact, be robust to such missingness. The RD metric appears to be unduly sensitive to heterogeneity, publication bias and underlying risk dependence.

9.6 Conclusions

9.6.1 Only one Critical Care intervention considered here had a “significant” beneficial (mortality) effect (Keenan *et al.* 1997), whereas two (Cochrane Injuries Group Albumin Reviewers 1998; Heys *et al.* 1999) had (potentially) deleterious effects, although the “review” (Wilkes *et al.* 2001c) of the original albumin supplementation meta-analysis (Cochrane Injuries Group Albumin Reviewers 1998) did not demonstrate this and provoked editorial comment (Cook *et al.* 2001). That the subsequent SAFE study (The SAFE Study Investigators 2004) of albumin versus saline in resuscitation could not demonstrate a “deleterious” effect of albumin has been noted above. Flathers *et al* (Flather *et al.* 1997) suggested that large meta-analyses “are likely to be more reliable” and the “review” meta-analysis of albumin effect (Wilkes *et al.* 2001c) contained a considerably larger number of studies. The disparity between meta-analyses and subsequent large randomised controlled trials has been reported (LeLorier *et al.* 1997), although the conclusions of the latter study appeared to be dependent upon selection, end-points and agreement were defined (Ioannidis *et al.* 1998a). Similarly, Furukawa *et al* (Furukawa *et al.* 2000) in 2000, reviewing the Cochrane Library, noted that agreement among “mega-trials” (defined as those with > 1000 subjects) was as large as that reported between meta-analyses and mega-trials and concluded that “taking megatrials as the gold standard can be problematic and that there is no substitute for clear and hard thinking”.

9.6.2 Other meta-analyses in this sample have also been updated; non-invasive ventilation (Keenan *et al.* 1997) by Peter *et al* (Peter *et al.* 2002), and selective gut decontamination (Selective Decontamination of the Digestive Tract Trailists'

Collaborative Group 1993) by Liberati *et al* (Liberati *et al.* 2000); an alternate analysis of the colloid crystalloid controversy is also current (Choi *et al.* 1999). However, the purpose of the current study was not that of a compendium of meta-analyses, rather, an in-depth illustration of methodology.

9.6.3 The tests of publication bias, as conducted in the current study, may illuminate the question of updates. In the case of albumin supplementation, the adverse effect of albumin in the initial meta-analysis (Wilkes *et al.* 2001b) was moderated by filling with 2 studies, but a significant treatment effect was still present (p at least 0.012). Publication bias in the review meta-analysis (Wilkes *et al.* 2001a) served merely to move estimates to the null (consistent with the SAFE trial, above).

9.6.4 Risk related treatment effects would appear to exist for selective gut decontamination and oxygen supply in sepsis, with deleterious effects manifesting themselves at higher risks with increments in oxygen supply. Albumin supplementation and anti-inflammatory agents in sepsis may also exhibit this attribute.

10 OVERVIEW AND CONCLUSIONS

10.1 The first part of the thesis (Sections 1 to 5) focused on the interpretation of recent significant clinical trials in the Critical Care arena from a self-consciously critical-statistical perspective. Key aspects of the later paradigm (*P*-values, normality, design and conduct of clinical trials, equivalence, sample size) were initially delineated. To this extent we note a parallel endeavour by Villar *et al* (Villar *et al.* 2005), who attempted to "examine the issues that prevent critical care physicians from translating research results into their daily clinical practice". Their concerns were illustrated by a consideration of the same three trials (Bernard *et al.* 2001; Hebert *et al.* 1999; The ARDS Network Authors for the ARDS Network 2000) which have been subject to detailed scrutiny in Sections 4 and 5, above. This being said, their analysis, perhaps characteristically, displayed a lack of focus on salient statistical matters. In particular, the implications of:

10.1.1 sequential trial methodology and the reporting of results for the ARDSNet trial.

10.1.2 multi-centre trial methodology and site "outcome variability" for the PROWESS trial (DeLong *et al.* 2005)

10.1.3 the status of the Hebert *et al* trial (on red blood cell transfusion) as an equivalence trial

10.1.4 the status of the *P*-value and the meaning of the 95% CI (described as the "range of values of estimated effect for 95% of patients"). Similar interpretations of the 95% CI have been the object of critique in a recent publication (Moran *et al.* 2005a), which we include, bound with this thesis.

These problematic areas are precisely those that were addressed in the first part of the thesis and serve to remind us, again, to continue to "grapple" with these issues (see Section 1.6, above).

10.2 The controversy that accompanies the interpretation of clinical trials is also reproduced in the debate over institutional (for example, Intensive Care Unit) outcome measures. Although not a focus of our deliberations, we have addressed these matters elsewhere (Moran *et al.* 2003a) and include this paper, bound with the thesis.

10.3 The second part of the thesis presented concrete analyses of various data-sets with a view to demonstrating the operation of statistical techniques / estimators. In particular:

10.3.1 linear and generalized linear models; we have suggested that GLMs offer new opportunities for modelling outcomes and the utility of the bootstrap procedure has been explored.

10.3.2 model selection and missing data techniques; these are areas of intense ongoing research and will assume importance for the reporting of clinical research as specific routines are incorporated into standard statistical packages (Clark *et al.* 2003; Schafer 2003).

10.3.3 Cox regression; Cox regression has been the “traditional” medical survival model (Tibshirani 1982), but the proportional hazards assumption may be vitiated in the critically-ill, requiring either stratification or time-varying covariates. The latter still present challenges of implementation (multiple record-per-patient data) and understanding (Zhou 2001) and a paper illustrating this estimation technique s bound into the thesis (Moran *et al.* 2004d).

10.3.4 random effects, multilevel models and hierarchical regression are a very active area of statistical research (Hedeker 2005; Pinheiro *et al.* 2000). Such models have been incorporated into the social and economic sciences and are increasingly addressed in the medical literature. Section 8 has presented random

effects modelling in order to "identify patient heterogeneity with respect to both ventilation and cohort effect, where the some uncertainty existed about the adequacy of modelling of the effect" . We also canvas its use in the bound paper "Mortality and other event rates: what do they tell us about performance?" (Moran *et al.* 2003a).

10.3.5 meta-analysis is a dominant paradigm, with evidence-based-medicine, in the current medical literature. The core questions of heterogeneity, publication bias and metaregression have been surveyed in detail in Section 9 and the position of Bayesian estimation has been discussed with respect to inference from contro-arm risk. An empirical study on these matters (Moran *et al.* 2005a) is also bound into the thesis.

10.4 The central position of statistics in our interpretation of the medical literature has been reiterated in this thesis (Breslow 2003; Efron 2005).

10.5 Future directions:

10.5.1 with the rapid development of statistical software, the impact of random effects modelling both in linear and non-linear models (including survival analysis and Cox regression) will be of substantial interest. The utility of random effects estimation in outcome assessment, compared with standard modelling (logistic regression) techniques has recently been questioned (Hannan *et al.* 2005)

10.5.2 the advancement of the Bayesian paradigm as a self-conscoius challenge to the frequentist (Spiegelhalter *et al.* 2004)

10.5.3 as exemplified in this thesis, the conduct and interpretation of the results of clinical trials will be the subject of continued development and scrutiny. Recent cautions regarding the implications of early stopping (Montori *et al.* 2005) and reconsiderations of conditional and predictive power (stochastic curtailment) are

of paramount importance to the clinician (Bauer *et al.* 2006; Emerson *et al.* 2005b)

11. ACKNOWLEDGEMENTS

1. To my supervisors (Professor Ruffin and Associate Professor Solomon) for persistent encouragement in the generation and writing of this thesis.
2. Dr David Warn, MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK for the Bayesian estimates used in Section 9.3.4 (Meta-regession).

12. APPENDICES

The following papers are bound into this thesis, as appendices:

Moran JL, Bersten AD, Solomon PJ. Meta-analysis of controlled trials of ventilator therapy in acute lung injury and acute respiratory distress syndrome: an alternative perspective. Intensive Care Medicine. 2005;31:227-35

Moran JL, Peisach AR, Solomon PJ, Martin J. Cost calculation and prediction in adult intensive care: a ground-up utilisation study. Anaesthesia & Intensive Care. 2004;32:787-97

Moran JL, Solomon PJ, Fox V, Salagaras M, Williams PJ, Quinlan K *et al*. Modelling thirty-day mortality in the Acute Respiratory Distress Syndrome (ARDS) in an adult ICU. Anaesthesia & Intensive Care. 2004;32:317-29

Moran JL, Solomon PJ. Mortality and other event rates: what do they tell us about performance? Critical Care and Resuscitation. 2003;5:292-303

REFERENCE LIST

Abele-Horn M, Kopp A, Sternberg U, Ohly A, Dauber A, Russwurm W et al. A randomized study comparing fluconazole with amphotericin B/5-flucytosine for the treatment of systemic Candida infections in intensive care patients. Infection. 1996;24:426-32.

Agresti A, Hartzel J. Strategies for comparing treatments on a binary response with multi-centre data. Stat Med. 2000;19:1115-39.

Al Omran M, Groof A, Wilke D. Enteral versus parenteral nutrition for acute pancreatitis. Cochrane Database of Systematic Reviews. 2002.

Altman DG. Statistics in medical journals. Stat Med. 1982;1:59-71.

Altman DG. Comparability of randomised groups. The Statistician. 1985;34:125-36.

Altman DG. Categorising continuous variables. Br J Cancer. 1991a;64:975.

Altman DG. Statistics in medical journals: developments in the 1980s. Stat Med. 1991b;10:1897-913.

Altman DG. The scandal of poor medical research. BMJ. 1994;308:283-84.

Altman DG. Statistics in medical journals: some recent trends. Stat Med. 2000;19:3275-89.

Altman DG, De Stavola BL. Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. Stat Med. 1994a;13:301-41.

Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. Lancet. 1990;335:149-53.

Altman DG, Goodman SN. Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. JAMA. 1994b;272:129-32.

Altman DG, Goodman SN, Schroter S. How statistical expertise is used in medical research. JAMA. 2002;287:2817-20.

Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994c;86:829-35.

Altman DG, Bland JM. Statistics Notes: Absence of evidence is not evidence of absence. BMJ. 1995;311:485.

Amato MB, Barbas CS, Medeiros DM, Magaldi RB, Schettino GP, Lorenzi-Filho G et al. Effect of a protective-ventilation strategy on mortality in the acute respiratory distress syndrome. N Engl J Med. 1998;338:347-54.

American Journal of Respiratory and Critical Care Medicine. American Journal of Respiratory and Critical Care Medicine. American Journal of Respiratory and Critical Crae Medicine @ http://intl-ajrccm atsjournals org/. 2005.

Amos J. Anti-Infective Drugs Advisory Committee Meeting: Eli Lilly Slides. http://www fda gov/ohrms/dockets/ac/01/slides/3797s1_01_Lilly-CORE/index htm. 2001.

Anaesthesia and Intensive Care. Anaesthesia and Intensive Care. Anaesthesia and Intensive Care @ http://www aaic net au/. 2005.

Anaissie EJ, Darouiche RO, Abi-Said D, Uzun O, Mera J, Gentry LO et al. Management of invasive candidal infections: results of a prospective, randomized, multicenter study of fluconazole versus amphotericin B and review of the literature. Clin Infect Dis. 1996;23:964-72.

Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol. 1983;1:710-709.

Annane D, Sebille V, Charpentier C, Bollaert PE, Francois B, Korach JM et al. Effect of treatment with low doses of hydrocortisone and fludrocortisone on mortality in patients with septic shock. JAMA 288(7):862-71. 2002;288:862-71.

Annane D, Sebille V, Troche G, Raphael JC, Gajdos P, Bellissant E. A 3-level prognostic classification in septic shock based on cortisol levels and cortisol response to corticotropin. JAMA. 2000;283:1038-45.

ANZICS Clinical Trials Group and Institute for International Health SAFE Study Investigators. The Saline vs. Albumin Fluid Evaluation (SAFE) Study (ISRCTN76588266): Design and conduct of a multi-centre, blinded randomised

controlled trial of intravenous fluid resuscitation in critically ill patients. Br Med J. 2003;http://bmj.bmjjournals.com/cgi/content/full/326/7389/559/DC1.

Appleby JL. Why doctors must grapple with health economics. Br Med J. 1987;294:326.

Aras G. Superiority, noninferiority, equivalence, and bioequivalence - Revisited. Drug Inf J. 2001;35:1157-64.

ARDS Network. Protocol: Prospective, Randomized, Multi-Center Trial of Higher End-expiratory Lung Volume/Lower FiO2 versus Lower End-expiratory Lung Volume/Higher FiO2 Ventilation in Acute Lung Injury and Acute Respiratory Distress Syndrome. http://hedwig.mgh.harvard.edu/ardsnet/alveoli.pdf . 1999.
Ref Type: Electronic Citation

ARDS Network. Report: Prospective, Randomized, Multi-Center Trial of Higher End-expiratory Lung Volume/Lower FiO2 versus Lower End-expiratory Lung Volume/Higher FiO2 Ventilation in Acute Lung Injury and Acute Respiratory Distress Syndrome. http://hedwig.mgh.harvard.edu/ardsnet/ards04.html . 2002.
Ref Type: Electronic Citation

Arends LR, Hoes AW, Lubsden J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. Stat Med. 2000;19:3497-518.

Armitage P. Sequential Medical Trials. 2nd ed. New York: John Wiley & Sons; 1975.

Armitage P. Interim analysis in clinical trials. Stat Med. 1991;10:925-35.

Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. Journal of the Royal Statistical Society Series A. 1969;132:235-44.

Armitage P, Stratton IM, Worthington HV. Repeated significance tests for clinical trials with a fixed number of patients and variable follow-up. Biometrics. 1985;41:353-59.

Arntz HR, Willich SN, Schreiber C, Bruggemann T, Stern R, Schultheiss HP. Diurnal, weekly and seasonal variation of sudden death. Population-based analysis of 24,061 consecutive cases. Eur Heart J. 2000;21:315-20.

ASPEN Board of Directors and the Clinical Guidelines Task Force. Guidelines for the use of parenteral and enteral nutrition in adult and pediatric patients. Journal of Parenteral & Enteral Nutrition. 2002;26:Suppl-138SA.

Austin PC, Ghali WA, Tu JV. A comparison of several regression models for analysing cost of CABG surgery. Stat Med. 2003;22:2799-815.

Avenell A, Handoll HH. Nutritional supplementation for hip fracture aftercare in the elderly. Cochrane Database Syst Rev. 2000;CD001880.

Azoulay E, Adrie C, De Lassence A, Pochard F, Moreau D, Thiery G et al. Determinants of postintensive care unit mortality: a prospective multicenter study. Crit Care Med. 2003;31:428-32.

Azoulay E, Alberti C, Bornstain C, Leleu G, Moreau D, Recher C et al. Improved survival in cancer patients requiring mechanical ventilatory support: impact of noninvasive mechanical ventilatory support. Crit Care Med. 2001;29:519-25.

Azoulay E, Moreau D, Alberti C, Leleu G, Adrie C, Barboteu M et al. Predictors of short-term mortality in critically ill patients with solid malignancies. Intensive Care Med. 2000;26:1817-23.

Bandt CL, Boen JR. A prevalent misconception about sample size, statistical significance, and clinical importance. J Periodontol. 1972;43:181-83.

Barber JA, Thompson JC. Multiple regression of cost data: use of generalised linear models. J Health Serv Res Policy. 2004;9:197-204.

Barber JA, Thompson SG. Analysis and interpretation of cost data in randomised controlled trials: review of published studies. BMJ. 1998;317:1195-200.

Barber JA, Thompson SG. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. Stat Med. 2000a;19:3219-36.

Barber JA, Thompson SG. Open access follow up for inflammatory bowel disease. Would have been better to use t test than Mann-Whitney U test. BMJ. 2000b;320:1730-1731.

Barnard GA. Discussion of 'Tests of significance in theory and practice' by D.J. Johnstone. The Statistician. 1986;35:499-502.

Barnard GA. Must clinical trials be large? The interpretation of P-values and the combination of test results. Stat Med. 1990;9:601-14.

Bartko JJ. Rationale for reporting standard deviations rather than standard errors of the mean. American Journal of Psychiatry. 1985;142:1060.

Bauer P, Koenig F. The reassessment of trial perspectives from interim data - a critical view. Stat Med. 2006;In press.

Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. Control Clin Trials. 1989;10:Suppl-175S.

Beal AL, Cerra FB. Multiple organ failure syndrome in the 1990s. Systemic inflammatory response and organ dysfunction. JAMA. 1994;271:226-33.

Beale RJ, Bryg DJ, Bihari DJ. Immunonutrition in the critically ill: a systematic review of clinical outcome. Crit Care Med. 1999;27:2799-805.

Becker RB, Zimmerman JE, Knaus WA, Wagner DP, Seneff MG, Draper EA et al. The use of APACHE III to evaluate ICU length of stay, resource use, and mortality after coronary artery by-pass surgery. J Cardiovasc Surg (Torino). 1995;36:1-11.

Begg CB. Significance tests of covariate imbalance in clinical trials. Control Clin Trials. 1990;11:223-25.

Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50:1088-101.

Begg CB, Leung DHY. On the use of surrogate end points in randomized trials. Journal of the Royal Statistical Society. 2000;163:15-28.

Benjamini Y, Krieger AM. Concepts and measures of skewness with data-analytic implications. The Canadian Journal of Statistics. 1996;24:131-40.

Berger JO. Could Fisher,Jeffreys, and Neyman have agreed on Testing. Statistical Science. 2003;18:1-32.

Berger JO, Delampady M. Testing precise hypotheses. Statistical Science. 1987a;2:317-52.

Berger JO, Sellke T. Testing a point null hypothesis: The irreconcilability of *P* values and evidence. Journal of the American Statistical Association. 1987b;82:112-22.

Berger VW. In defense of hypothesis testing. Br Med J. 2001;Rapid response 9 September @ http://bmj.bmjjournals.com/cgi/eletters/322/7295/1184/a#16449.

Bergmann R, Ludbrook J, Spooren WP. Different outcomes of the Wilcoxon-Mann-whitney test from different statistics packages. The American Statistician. 2000;54:72-77.

Berk RA. Regression analysis: A constructive critique. Thousand Oaks, CA: Sage Publications, Inc; 2004.

Berkson J. Tests of significance considered as evidence. Journal of the American Statistical Association. 1947;37:325-35.

Berlin JA, Antman EM. Advantages and limitations of metaanalytic regressions of clinical trials data. Online J Curr Clin Trials. 1994;Doc No 134:8425.

Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A et al. Efficacy and safety of recombinant human activated protein C for severe sepsis. N Engl J Med. 2001;344:699-709.

Bernard SA, Gray TW, Buist MD, Jones BM, Silvester W, Gutteridge G et al. Treatment of comatose survivors of out-of-hospital cardiac arrest with induced hypothermia. N Engl J Med. 2002;346:557-63.

Bernard SAM, MacC Jones BMB*, Horne MKB. Clinical Trial of Induced Hypothermia in Comatose Survivors of Out-of-Hospital Cardiac Arrest. Ann Emerg Med. 1997;30:146-53.

Bernsen RM, Tasche MJ, Nagelkerke NJ. Variation in baseline risk as an explanation of heterogeneity in meta-analysis by S. D. Walter, Statistics in Medicine, 16, 2883-2900 (1997). Stat Med. 1999;18:233-38.

Betensky R.A. Alternative derivations of a rule for early stopping in favor of H0. The American Statistician. 2000;54:35-39.

Birnbaum A. The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. Synthese. 1977;36:19-49.

Blackwelder WC. "Proving the null hypothesis" in clinical trials. Contol Clin Trials. 1982;3:345-53.

Blackwelder WC. Showing a treatment is good because it is not bad: when does "noninferiority" imply effectiveness? Control Clin Trials. 2002;23:52-54.

Blomqvist N, Svardsudd K. A new method for investigating the relation between change and initial value in longitudinal blood pressure data. II. Comparison with other methods. Scand J Soc Med. 1978;6:125-29.

Blough DK, Madden CW, Hornbrook MC. Modelling risk using generalized linear models. J Health Econ. 1999;18:153-71.

Blough DK, Ramsey SD. Using generalized linear models to assess medical care costs. Health Services & Outcomes Research Methodology. 2000a;1:185-202.

Blough DK, Ramsey SD. Using generalized linear models to assess medical care costs. Health Services & Outcomes Research Methodology. 2000b;1:185-202.

Bockenhoff A, Hartung J. Meta-analysis: different methods--different conclusions? Studies in Health Technology & Informatics. 2000;77:39-43.

Boneau CA. The effects of violations of assumptions underlying the t test. Psychol Bull. 1960;57:49-64.

Bonten MJ, Brun-Buisson C, Weinstein RA. Selective decontamination of the digestive tract: to stimulate or stifle? Intensive Care Med. 2003;29:672-76.

Borenstein M. The case for confidence intervals in controlled clinical trials. Control Clin Trials. 1994;15:411-28.

Bower RH, Cerra FB, Bershadsky B, Licari JJ, Hoyt DB, Jensen GL et al. Early enteral administration of a formula (Impact) supplemented with arginine, nucleotides, and fish oil in intensive care unit patients: results of a multicenter, prospective, randomized, clinical trial. Crit Care Med. 1995;23:436-49.

Box GEP. Science and statistics. Journal of the American Statistical Association. 1976;71:791-99.

Bradburn MJ, Deeks J, Altman DG. metan-sbe24 an alternative meta-analysis command. Stata Technical Bulletin Reprints. 1998;8:100.

Braga M, Vignali A, Gianotti L, Cestari A, Profili M, Di C, V. Benefits of early postoperative enteral feeding in cancer patients. Infusionsther Transfusionsmed. 1995;22:280-284.

Braitman LE. Confidence intervals assess both clinical significance and statistical significance. Ann Intern Med. 1991;114:515-17.

Brand R. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. Stat Med. 1994;13:295-96.

Braunschweig CL, Levy P, Sheean PM, Wang X. Enteral compared with parenteral nutrition: a meta-analysis. Am J Clin Nutr. 2001;74:534-42.

Breslow NE. Generalized linear models: Checking assumptions and strengthening conclusions. Statistica Applicata. 1996;8:23-41.

Breslow NE. Are statistical contributions to medicine undervalued? Biometrics. 2003;59:1-8.

Bridge PD, Sawilowsky SS. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and and Wilcoxon Rank-Sum test in small samples applied research. J Clin Epidemiol. 1999;52:229-35.

Briggs A, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. Journal of Health Services Research & Policy. 1998;3:233-45.

Britton A, McKee M, Black M, McPherson K, Sanderson C, Bain C. Choosing between randomised annd non-randomised studies: a systematic review. Health Technol Assess. 1998;2:1-123.

Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Threats to applicability of randomised trials: exclusions and selective participation. Journal of Health Services & Research Policy. 1999;4:112-21.

Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med. 2001;20:825-40.

Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey SG. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technology Assessment (Winchester, England). 2001a;5:1-56.

Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey SG. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technology Assessment (Winchester, England). 2001b;5:1-56.

BROSS ID. Statisitcal Criticism. Cancer. 1960;13:400.

Brown GW. Standard deviation, standard error. Which 'standard' should we use? American Journal of Diseases of Children. 1982;136:937-41.

Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA. 1987;257:2459-63.

Bruedel J, Diekmann A. The log-logisitc rate model. Two generalisations with an application to demgraphic data. Sociological Methods & Research. 1995;24:158-86.

Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group. JAMA. 1999;282:771-78.

Buchman AL. Glutamine: commercially essential or conditionally essential? A critical appraisal of the human data. Am J Clin Nutr. 2001;74:25-32.

Buchner DM, Findley TW. Research in physical medicine and rehabilitation. VIII. Preliminary data analysis. American Journal of Physical Medicine & Rehabilitation. 1990;69:154-69.

Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. Biometrics. 1997;53:603-18.

Buntin MB, Zaslavsky AM. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. J Health Econ. 2004;23:525-42.

Burns LR, Wholey DR. The effects of patient, hospital, and physician characteristics on length of stay and mortality. Med Care. 1991;29:251-71.

Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. [erratum appears in Biometrics 2000 Mar;56(1):324.]. Biometrics. 1998;54:1014-29.

Buyse ME. Analysis of clinical trial outcomes: some comments on subgroup analyses. Control Clin Trials. 1989;10:Suppl-194S.

Bycott PW, Taylor JM. An evaluation of a measure of the proportion of the treatment effect explained by a surrogate marker. Control Clin Trials. 1998;19:555-68.

Byington RP. Beta-blocker heart attack trial: design, methods, and baseline results. Beta-blocker heart attack trial research group. Control Clin Trials. 1984;5:382-437.

Cagnoni PJ, Walsh TJ, Prendergast MM, Bodensteiner D, Hiemenz S, Greenberg RN et al. Pharmacoeconomic analysis of liposomal amphotericin B versus conventional amphotericin B in the empirical treatment of persistently febrile neutropenic patients. J Clin Oncol. 2000;18:2476-83.

Califf RM, Ellenberg SS. Statistical approaches and policies for the operations of Data and Safety Monitoring Committees. American Heart Journal. 2001;141:301-5.

Capes SE, Hunt D, Malmberg K, Gerstein HC. Stress hyperglycaemia and increased risk of death after myocardial infarction in patients with and without diabetes: a systematic overview. Lancet. 2000;355:773-78.

Carlin JB, Wolfe R, Brown CH, Gelman A. A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. Biostat. 2001;2:397-416.

Carlton MA. Discussion to "Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing". The American Statistician. 2003;57:179-81.

Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med. 2000;19:1141-64.

Casella G, Berger RL. Comment on Testing precise hypotheses by J.O. Berger and M. Deampdy. Statistical Science. 1987a;2:344-47.

Casella G, Berger RL. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association. 1987b;82:106-11.

Cast Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. N Engl J Med. 1989;321:406-12.

Chalmers TC, Lau J. Changes in clinical trials mandated by the advent of meta-analysis. Stat Med. 1996;15:1263-68.

Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40:373-83.

Chastang C, Byar D, Piantadosi S. A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. Stat Med. 1988;7:1243-55.

Chatfield C. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society A. 1995;158:419-66.

Chatterjee S, Hadi AS, Price B. Regression analysis by example. 3rd ed. New York: John Wiley & Sons, Inc; 2000.

Chen CH, George SL. The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. Stat Med. 1985;4:39-46.

Chest. Chest. Chest @ http://www.chestjournal.org/ . 2005.
Ref Type: Generic

Chhikara RS, Folks JL. The Inverse Gaussian Distribution: Theory, Methodology, and Applications. New York: Marcel Dekker, Inc; 1989.

Chi GYH. Multiple testings: Multiple comparisons and multiple endpoints. Drug Inf J. 1998;32:1347S-62S.

Choi PT, Yip G, Quinonez LG, Cook DJ. Crystalloids vs. colloids in fluid resuscitation: a systematic review. Crit Care Med. 1999;27:200-210.

Choi SC, Smith PJ, Becker DP. Early decision in clinical trials when the treatment differences are small. Experience of a controlled trial in head trauma. Control Clin Trials. 1985;6:280-288.

Chow G. Testing equality between sets of coefficients in two linear regressions. Econometrica. 1960;28:591-605.

Chuang-Stein C. Clinical equivalence - A clarification. Drug Inf J. 1999;33:1189-94.

Chuang-Stein C. Testing for superiority or inferiority after concluding equivalence? Drug Inf J. 2001;35:141-43.

Clark DE, Ryan LM. Modeling injury outcomes using time-to-event methods. Journal of Trauma-Injury Infection & Critical Care. 1997;42:1129-34.

Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol. 2003;56:28-37.

Classen DC, Evans RS, Pestotnik SL, Horn SD, Menlove RL, Burke JP. The timing of prophylactic administration of antibiotics and the risk of surgical-wound infection. N Engl J Med. 1992;326:281-86.

Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association. 1979;74:829-36.

Clifton GL, Choi SC, Miller ER, Levin HS, Smith KR, Jr., Muizelaar JP et al. Intercenter variance in clinical trials of head trauma--experience of the National Acute Brain Injury Study: Hypothermia. J Neurosurg. 2001a;95:751-55.

Clifton GL, Choi SC, Miller ER, Levin HS, Smith KR, Jr., Muizelaar JP et al. Intercenter variance in clinical trials of head trauma--experience of the National Acute Brain Injury Study: Hypothermia. J Neurosurg. 2001b;95:751-55.

Clifton GL, Kreutzer JS, Choi SC, Devany CW, Eisenberg HM, Foulkes MA et al. Relationship between Glasgow Outcome Scale and neuropsychological measures after brain injury. Neurosurgery. 1993;33:34-38.

Clifton GL, Miller ER, Choi SC, Levin HS, McCauley S, Smith KR, Jr. et al. Lack of effect of induction of hypothermia after acute brain injury. N Engl J Med. 2001c;344:556-63.

Clifton GL. Lack of Effect of Induction of Hypothermia after Acute Brain Injury. N Engl J Med. 2001;345:66.

Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26:404-13.

COBALT Investigators. A Comparison of Continuous Infusion of Alteplase with Double-Bolus Administration for Acute Myocardial Infarction. N Engl J Med. 1997;337:1124-30.

Cochran WG. The combination of estimates from different experiments. Biometrics. 1954a;10:101-29.

Cochran WG. The combination of estimates from different experiments. Biometrics. 1954b;10:101-29.

Cochrane Injuries Group Albumin Reviewers. Human albumin administration in citically ill patients: systematic review of randomised controlled trials. BMJ. 1998;317:235-40.

Cohen J, Guyatt G, Bernard GR, Calandra T, Cook D, Elbourne D et al. New strategies for clinical trials in patients with sepsis and septic shock. Critical Care Medicine. 2001;29:880-886.

Cohen ME, Arthur JS. Randomization analysis of dental data characterized by skew and variance heterogeneity. Community Dentistry & Oral Epidemiology. 1991;19:185-89.

Colditz GA, Miller JN, Mosteller F. Measuring gain in the evaluation of medical technology. Int J Technol Assess Health Care. 1988;4:637-42.

Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. Stat Med. 1989;8:441-54.

Cole TJ. Sympercents: symmetric percentage differences on the 100 log(e) scale simplify the presentation of log transformed data. Stat Med. 2000;19:3109-25.

Concato J, Fenstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med. 1993;118:201-10.

Connors AF, Jr., Speroff T, Dawson NV, Thomas C, Harrell FE, Jr., Wagner D et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. JAMA. 1996;276:889-97.

Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician. 1981;35:124-29.

Cook D, Guyatt G. Colloid use for fluid resuscitation: evidence and spin. Ann Intern Med. 2001;135:205-8.

Cook DJ, Reeve BK, Guyatt GH, Heyland DK, Griffith LE, Buckingham L et al. Stress ulcer prophylaxis in critically ill patients. Resolving discordant meta-analyses. JAMA. 1996;275:308-14.

Cook D. Is Albumin Safe? N Engl J Med. 2004;350:2294-96.

Cooper J. Resuscitation fluid controversies - Australian trials offer new insights. Critical Care and Resuscitation. 2004;6:83-84.

Copas JB, Li HG. Inference in non-random samples. Journal of the Royal Statistical Society, Series B. 1997;59:55-95.

Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J Natl Cancer Inst. 1951;11:1269-75.

Cornfield J. The University Group Diabetes Program. A further statistical analysis of the mortality findings. JAMA. 1971;217:1676-87.

Cornfield J. Sequential trials, sequential analysis and the likelihood principle. The American Statistician. 2004;20:18-23.

Cornfield J, Haenszel W, Hammond C, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: Recent evidence and a discussion of some questions. J Natl Cancer Inst. 1959;22:173-203.

Corwin HL, Gettinger A, Pearl RG, Fink MP, Levy MM, Abraham E et al. The CRIT study: Anemia and blood transfusion in the critically ill-Current clinical practice in the United States. Crit Care Med. 2004;32:39-52.

Coste J, Fermanian J, Venot A. Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. Stat Med. 1995;14:331-45.

Cox DR. The role of significance tests. Scandanavian Journal of Statistics. 1977;4:49-70.

Cox DR. Foundations of statistical inference: The case for eclecticism. Australian Journal of Statistics. 1978;20:43-59.

Cox DR. Statistical significance tests. Br J Clin Pharmacol. 1982;14:325-31.

Cox DR. Another comment on the role of statistical methods. Br Med J. 2001;322:231.

Cox DR, Hinkley DV. Theoretical Statistics. London: Chapman & Hall; 1974.

Coyle D. Statistical analysis in pharmacoeconomic studies. A review of current issues and standards. Pharmacoeconomics. 1996;9:506-16.

Crawford SW, Petersen FB. Long-term survival from respiratory failure after marrow transplantation for malignancy. Am Rev Respir Dis. 1992;145:510-514.

Critical Care and Resuscitation. Critical Care and Resuscitation. Critical Care and Resuscitation @ http://som flinders edu au/FUSA/CriticalCare/AACCMHomePage html#d. 2005.

Critical Care Medicine. Critical Care Medicine. Critical Care Medicine @ http://www ccmjournal com/pt/re/ccm/home htm. 2005.

Cronin L, Cook DJ, Carlet J, Heyland DK, King D, Lansang MA et al. Corticosteroid treatment for sepsis: a critical appraisal and meta-analysis of the literature. Crit Care Med. 1995;23:1430-1439.

Crowther MA, Marshall JC. Continuing challenges of sepsis research. JAMA. 2001a;286:1894-96.

Crowther MA, Marshall JC. Continuing challenges of sepsis research. JAMA. 2001b;286:1894-96.

Curfman GD. Hypothermia to Protect the Brain. N Engl J Med. 2002;346:546.

Cutler SJ, Greenhouse SW, Cornfield J, Schneiderman MA. The role of hypothesis testing in clinical trials. Biometrics seminar. J Chronic Dis. 1966;19:857-82.

Cytel Corporation. East V 3.0 Software. Cytel@http://www cytel com/new pages/EAST 2 html. 2003.

D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues- the encounter of academic consultants in statistics. Stat Med. 2003;22:169-86.

D'Amico R, Pifferi S, Leonetti C, Torri V, Tinazzi A, Liberati A. Effectiveness of antibiotic prophylaxis in critically ill adult patients: systematic review of randomised controlled trials. BMJ. 1998a;316:1275-85.

D'Amico R, Pifferi S, Leonetti C, Torri V, Tinazzi A, Liberati A. Effectiveness of antibiotic prophylaxis in critically ill adult patients: systematic review of randomised controlled trials. BMJ. 1998b;316:1275-85.

Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group. JAMA. 1998;279:545-49.

Davidson BL, Geerts WH, Lensing AW. Low-dose heparin for severe sepsis. N Engl J Med. 2002;347:1036-37.

Davis BR, Hardy RJ. Data monitoring in clinical trials: the case for stochastic curtailment. Journal of Clinical Epidemiology. 1994;47:1033-42.

De Groot MH. Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. Journal of the American Statistical Association. 1973;68:966-69.

De Gruttola V, Fleming T, Lin DY, Coombs R. Perspective: validating surrogate markers--are we being naive?. J Infect Dis. 1997;175:237-46.

de Jonge E, Schulz MJ, Spanjaard L, Bossuyt PM, Vroom MB, Dankert J. Effects of selective decontamination of digestive tract on mortality and acquisition of resistant bacteria in intensive care: a randomised controlled trial. Lancet. 2003a;362:1011-16.

de Jonge E, Schulz MJ, Spanjaard L, Bossuyt PM, Vroom MB, Dankert J. Effects of selective decontamination of digestive tract on mortality and acquisition of resistant bacteria in intensive care: a randomised controlled trial. Lancet. 2003b;362:1011-16.

DeCarlo LT. On the meaning and use of kurtosis. Psychological Methods. 1997;2:292-307.

Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med. 2002;21:1575-600.

Deeks JJ, Altman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care: Meta-analysis in context. 2nd ed. London: BMJ Publishing Group; 2001: 313-35.

Deeks JJ, Altman DG, Dooley G, Sackett DL. Choosing an appropriate dichotomous effect measure for meta-analysis: empirical evidence of the appropriateness of the odds ratio and relative risk. Control Clin Trials. 1997;18:84S-5S.

Delgado-Herrera L, Anbar D. A model for the interim analysis process: a case study. Control Clin Trials. 2003;24:51-65.

DeLong ER, Coombs LP, Ferguson TB, Peterson ED. The Evaluation of Treatment When Center-Specific Selection Criteria Vary with Respect to Patient Risk. Biometrics. 2005;61:942-49.

DeMets DL. Stopping guidelines vs stopping rules: A practitioner's point of view. Communications in Statistics - Theory and Methods. 1984;13:2395-417.

DeMets DL. Methods for combining randomized clinical trials: strengths and limitations. Stat Med. 1987;6:341-50.

DeMets DL. Sequential designs in clinical trials. Cardiac Electrophysiology Review. 1998;2:57-60.

DeMets DL, Gail MH. Use of logrank tests and group sequential methods at fixed calendar times. Biometrics. 1985;41:1039-44.

DeMets DL, Hardy R, Friedman LM, Lan KK. Statistical aspects of early termination in the beta-blocker heart attack trial. Control Clin Trials. 1984a;5:362-72.

DeMets DL, Lan G. The alpha spending function approach to interim data analyses. In: Thall PF, ed. Recent advances in clinical trial design and analysis. Boston: Kluwer Academic Publishers; 1995: 1-27.

DeMets DL, Lan KK. An overview of sequential methods and their application in clinical trials. Communications in Statistics - Theory and Methods. 1984b;13:2315-38.

DeMets DL, Pocock SJ, Julian DG. The agonising negative trend in monitoring of clinical trials. Lancet. 1999;354:1983-88.

Dempster AP. Comment on RA Fisher in the 21st century by B Efron. Statistical Science. 1998;13:120-121.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977;39:1-38.

Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br J Math Stat Psychol. 1992;45:265-82.

DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177-88.

Desgagne A, Castilloux AM, Angers JF, LeLorier J. The use of the bootstrap statistical method for the pharmacoeconomic cost analysis of skewed data. Pharmacoeconomics. 1998;13:487-97.

Dickersin K, Min YI. Publication bias: the problem that won't go away. Ann N Y Acad Sci. 1993;703:135-46.

Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. Methods for analyzing health care utilization and costs. Annu Rev Public Health. 1999;20:125-44.

Djulbegovic B, Clarke M. Scientific and ethical issues in equivalence trials. JAMA. 2001;285:1206-8.

Dominici F, Spiegelman D, Cole SR. Methodological Contributions to the American Journal of Epidemiology. Am J Epidemiol. 2004;160:197-98.

Dracup C. Hypothesis testing-What it really is. The Psychologist. 1995;8:359-62.

Duan N. Smearing estimate:A nonparametric retransformation method. Journal of the American Statistical Association. 1983;78:605-10.

Dudley RA, Harrell FE, Jr., Smith LR, Mark DB, Califf RM, Pryor DB et al. Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. J Clin Epidemiol. 1993;46:261-71.

Dudrick SJ, Wilmore DW, Vars HM, Rhoads JE. Long-term total parenteral nutrition with growth, development, and positive nitrogen balance. Surgery. 1968;64:134-42.

Dunnett CW, Gent M. An alternative to the use of two-sided tests in clinical trials. Stat Med. 1996;15:1729-38.

Duval S, Tweedie R. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. Journal of the American Statistical Society. 2000a;95:89-98.

Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000b;56:455-63.

Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Stat Med. 2002;21:1635-40.

Efron B. Why isn't everyone a Bayesian. The American Statistician. 1986;40:1-11.

Efron B. R.A. Fisher in the 21st Century. Statistical Science. 1998;13:95-122.

Efron B. Bayesians, Frequentists, and Scientists. Journal of the American Statistical Association. 2005;100:1-5.

Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman and Hall; 1993.

Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315:629-34.

Egger M, Smith GD, Altman DG. Systematic reviews in health care: Meta-analysis in context. 2nd ed. London: BMJ Publishing Group; 2001.

Eichacker PQ, Parent C, Kalil A, Esposito C, Cui X, Banks SM et al. Risk and the efficacy of antiinflammatory agents: retrospective and confirmatory studies of sepsis. Am J Respir Crit Care Med. 2002;166:1197-205.

Elashoff JD. Sample size Tables for Proportions. nQuery Advisor®Version 5.0 User's Guide. Cork, Ireland: Statistical Solutions Ltd; 2002: 15-1-15-46.

Elashoff JD. nQuery Advisor V 5.0. nQuery Advisor V5 0 @ http://www statsol ie/nquery/nquery htm. 2003.

Eli Lilly and Company. Briefing Document for XIGRIS for the Treatment of Severe Sepsis. http://www fda gov/ohrms/dockets/ac/01/briefing/3797b1_01_Sponsor pdf. 2001.

Ellenberg SS, Fleming TR, DeMets DL. Data Monitoring Committees in Clinical Trials : A Practical Perspective. Hoboken, NJ: John Wiley & Sons; 2002.

Ellenberg SSP, Temple RM. Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments: Part 2: Practical Issues and Specific Cases. Ann Intern Med. 2000;133:464-70.

Ely EW, Bernard GR, Vincent JL. Activated protein C for severe sepsis. N Engl J Med. 2002a;347:1035-36.

Ely EW, Bernard GR, Vincent JL. Activated protein C for severe sepsis. N Engl J Med. 2002b;347:1035-36.

Ely EW, Bernard GR, Vincent JL. Activated protein C for severe sepsis. N Engl J Med. 2002c;347:1035-36.

Ely EW, Bernard GR, Vincent JL. Activated protein C for severe sepsis. N Engl J Med. 2002d;347:1035-36.

Ely EW, Bernard GR, Vincent JL. Activated protein C for severe sepsis. N Engl J Med. 2002e;347:1035-36.

Ely EW, Bernard GR. Transfusions in Critically Ill Patients. N Engl J Med. 1999;340:467-68.

Emerson JD. Combining estimates of the odds ratio: the state of the art. Stat Methods Med Res. 1994;3:157-78.

Emerson, S. S+SEQTRIAL: Technical Overview. Research Report No 98. http://www.insightful.com/DocumentsLive/seqtech.pdf . 2000. Seattle, WA, Data Analysis Products Division, MarthSoft, Inc.
Ref Type: Electronic Citation

Emerson SS. Computation of the uniform minimum variance unbiased estimator of a normal mean following a group sequential trial. Computers & Biomedical Research. 1993;26:68-73.

Emerson SS. Stopping a clinical trial very early based on unplanned interim analyses: a group sequential approach. Biometrics. 1995;51:1152-62.

Emerson SS, Bruce A, Baldwin K. S+SeqTrial 2 User's Manual. Seattle, WA: Insightful Corporation; 2000.

Emerson SS, Fleming TR. Symmetric group sequential test designs. Biometrics. 1989;45:905-23.

Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. Biometrika. 1990;77:875-92.

Emerson SS, Kittelson JM, Gillen DL. Bayesian Evaluation of Group Sequential Clinical Trial Designs. UW Biostatistics Working Paper Series @ http://www bepress com/uwbiostat. 2005a;Working Paper 242.

Emerson SS, Kittelson JM, Gillen DL. On the Use of Stochastic Curtailment in Group Sequential Clinical Trials. UW Biostatistics Working Paper Series @ http://www bepress com/uwbiostat. 2005b;Working paper 243.

Enas GG, Enas HH, Spradlin CT, Wilson MG, Wiltse CG. Baseline comparability in clinical trials: prevention of "poststudy" anxiety. Drug Inf J. 1990;24:541.

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000;19:1707-28.

Esteban A, Anzueto A, Frutos F, Alia I, Brochard L, Stewart TE et al. Characteristics and outcomes in adult patients receiving mechanical ventilation: a 28-day international study. JAMA. 2002;287:345-55.

Evans C, Chalmers J, Capewell S, Redpath A, Finlayson A, Boyd J et al. "I don't like Mondays"-day of the week of coronary heart disease deaths in Scotland: study of routinely collected data. BMJ. 2000;320:218-19.

Evans SJ, Mills P, Dawson J. The end of the p value?. Br Heart J. 1988;60:177-80.

Evans TW. Hemodynamic and metabolic therapy in critically ill patients. N Engl J Med. 2001;345:1417-18.

Ewig S, Torres A, el Ebiary M, Fabregas N, Hernandez C, Gonzalez J et al. Bacterial colonization patterns in mechanically ventilated patients with traumatic and medical head injury. Incidence, risk factors, and association with ventilator-associated pneumonia. American Journal of Respiratory & Critical Care Medicine. 1999;159:188-98.

FDA/CBER. Biologics License Application: Recombinant Human Activated Protein C (rhAPC) [drotrecogin alfa (activated)] Xigris™ for Severe Sepsis. http://www fda gov/ohrms/dockets/ac/01/slides/3797s1_02_Forsyth/sld001 htm. 2001a.

FDA/CBER. Food and Drug Administration, ANTI-INFECTIVE DRUGS ADVISORY COMMITTEE, October 16, 2001; Xigris (Drotrecogin Alfa (Activated): Slides. http://www fda gov/ohrms/dockets/ac/01/slides/3797s1 htm. 2001b.

Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol. 1998;51:355-60.

Ferreira IM, Brooks D, Lacasse Y, Goldstein RS. Nutritional support for individuals with COPD: a meta-analysis. Chest. 2000;117:672-78.

Fetter RB, Shin Y, Freeman JL, Averill RF, Thompson JD. Case mix definition by diagnosis-related groups. Med Care. 1980;18:1-53.

Fewell Z, Hernan MA, Wolfe F, Tilling K, Choi HK, Sterne JA. Controlling for time-dependent confounding using marginal structural models. Stata Journal. 2004;4:402-20.

Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. [Article]. Lancet. 1994;343:311-22.

Finfer S, Bellomo R, Myburgh J, Norton R. Efficacy of albumin in critically ill patients. BMJ. 2003;326:559-60.

Firth D. Generalized linear models. In: Hinkley DV, Reid N, Snell EJ, eds. Statistical Theory and Modelling: In honour of Sir David Cox, FRS. London: Chapman and Hall; 1991: 55-82.

Fisher J, Magid N, Kallman C, Fanucchi M, Klein L, McCarthy D et al. Respiratory illness and hypophosphatemia. Chest. 1983;83:504-8.

Fisher LD. The use of one-sided tests in drug trials: an FDA advisory committee member's perspective. Journal of Biopharmaceutical Statistics. 1991;1:151-56.

Fisher RA. Statistical methods and scientific induction. Journal of the Royal Statistical Society, Series B. 1955;17:69-78.

Flather MD, Farkouh ME, Pogue JM, Yusuf S. Strengths and limitations of meta-analysis: larger studies may be more reliable. Control Clin Trials. 1997;18:568-79.

Fleiss JL. Analysis of data from multiclinic trials. Control Clin Trials. 1986a;7:267-75.

Fleiss JL. Confidence intervals vs significance tests: quantitative interpretation. Am J Public Health. 1986b;76:587-88.

Fleiss JL. Significance tests have a role in epidemiologic research: reactions to A. M. Walker. Am J Public Health. 1986c;76:559-60.

Fleiss JL. One-tailed versus two-tailed tests: Rebuttal. Control Clin Trials. 1988;10:227-28.

Fleiss JL. The statistical basis of meta-analysis. Stat Methods Med Res. 1993;2:121-45.

Fleming TR. Design and interpretation of equivalence trials. Am Heart J. 2000;139:S171-S176.

Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996;125:605-13.

Fleming TR, Green SJ, Harrington DP. Considerations for monitoring and evaluating treatment effects in clinical trials. Contol Clin Trials. 1984a;5:55-66.

Fleming TR, Harrington DP, O'Brien PC. Designs for group sequential tests. Control Clin Trials. 1984b;5:348-61.

Fligner MA, Policello II GE. Robust rank procedures for the Behrens-Fisher problem. Journal of the American Statistical Association. 1981;76:162-68.

Ford I, Norrie J, Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. Stat Med. 1995;14:735-46.

Fox J. Describing univariate distributions. In: Fox J, Long JS, eds. Modern methods of Data Analysis. Newbury Park, ca: Sage Publications; 1990: 58-125.

Foxman B, Frerichs RR. Epidemiology of urinary tract infection: I. Diaphragm use and sexual intercourse. Am J Public Health. 1985;75:1308-13.

Foxman B, Frerichs RR. Response from Drs Foxman and Frerichs. Am J Public Health. 1986;76:587.

Freedman DA, Petitti DB, Robins JM. On the efficacy of screening for breast cancer. Int J Epidemiol. 2004;33:43-55.

Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Stat Med. 1992;11:167-78.

Freeman PR. The role of p-values in analysing trial results. Stat Med. 1993;12:1443-52.

Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic?. BMJ. 2001;322:989-91.

Freidkin B, Gatswirth JL. Should the median test be retired from general use? The American Statistician. 2000;54:161-64.

Freidlin B, Korn EL. A comment on futility monitoring. Contol Clin Trials. 2002;23:355-66.

Freiman JA, Chalmers TC, Smith H, Jr., Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. New England Journal of Medicine. 1978;299:690-694.

Fried L, Bernardini J, Piraino B. Charlson comorbidity index as a predictor of outcomes in incident peritoneal dialysis patients. American Journal of Kidney Diseases [computer file]. 2001;37:337-42.

Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials. 3rd ed. New York: Springer-Verlag; 1998.

Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. Stat Med. 1992;11:1685-704.

Furst P. A thirty-year odyssey in nitrogen metabolism: from ammonium to dipeptides. Jpen: Journal of Parenteral & Enteral Nutrition. 2000;24:197-209.

Furukawa TA, Streiner DL, Hori S. Discrepancies among megatrials. J Clin Epidemiol. 2000;53:1193-99.

Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On the meta-analytic assessment of surrogate outcomes. Biostat. 2000;1:231-46.

Gallis HA, Drew RH, Pickard WW. Amphotericin B: 30 years of clinical experience. Rev Infect Dis. 1990;12:308-29.

Gallo PP. Center-weighting issues in multicenter clinical trials. J Biopharm Stat. 2000;10:145-63.

Gan FF, Koehler KJ. Goodness-of-Fit tests based on P-P probability plots. Technometrics. 1990;32:289-303.

Gan FF, Koehler KJ, Thompson JC. Probability plots and distribution curves for assessing the fit of probability models. The American Statistician. 1991;45:14-21.

Gans DJ. The search for significance: different tests on the same data. Journal of Statistical Computing and Simulation. 1984;19:1-21.

Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. British Medical Journal Clinical Research Ed. 1986;292:746-50.

Gastinne H, Wolff M, Delatour F, Faurisson F, Chevret S. A controlled trial in intensive care units of selective decontamination of the digestive tract with nonabsorbable antibiotics. The French Study Group on Selective Decontamination of the Digestive Tract. N Engl J Med. 1992;326:594-99.

Gattinoni L, Brazzi L, Pelosi P, Latini R, Tognoni G, Pesenti A et al. A trial of goal-oriented hemodynamic therapy in critically ill patients. SvO2 Collaborative Group. N Engl J Med. 1995;333:1025-32.

Gelman A, Rubin DB. Markov chain Monte Carlo methods in biostatistics. Stat Methods Med Res. 1996;5:339-55.

Gibbons JD, Pratt JW. P-values: Interpretation and Methodology. The American Statistician. 1975;29:20-25.

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L. The Empire of Chance : How Probability Changed Science and Everyday Life. New York: Cambridge University Press; 1989.

Gill JS, Zezulka AV, Beevers DG, Davies P. Relation between initial blood pressure and its fall with treatment. Lancet. 1985;1:567-69.

Gillen DL, Emerson SS. Information Growth in a Family of Weighted Logrank Statistics
Under Repeated Analyses. http://www insightful com/news_events/webcasts/pharm04/InformationGrowth pdf. 2003.

Ginsberg MDM. Therapeutic trials for traumatic brain injury-A journey in progress. Crit Care Med. 2002;30:935-36.

Glass GV. Primary, secondary, and meta-analysis of research. Educational Research. 1976;5:3-8.

Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. BMJ. 1995;311:1356-59.

Godard J, Guillaume C, Reverdy ME, Bachmann P, Bui-Xuan B, Nageotte A et al. Intestinal decontamination in a polyvalent ICU. A double-blind study. Intensive Care Med. 1990;16:307-11.

Goldstein R. Equivalence testing sg21. Stata Technical Bulletin Reprints. 1994;3:107-12.

Good.P. Permutation tests: a practical guide to resampling mehtods for testing hypotheses. Second ed. New York: Springer-Verlag; 2000.

Goodman SN. A comment on replication, p-values and evidence. Stat Med. 1992;11:875-79.

Goodman SN. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol. 1993;137:485-96.

Goodman SN. Confidence limits vs power calculations. Epidemiology. 1994;5:266-68.

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.

Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. 1994;121:200-206.

Goodman SN, Royall R. Evidence and scientific research. Am J Public Health. 1988;78:1568-74.

Gotzsche PC, Liberati A, Torri V, Rossetti L. Beware of surrogate outcome measures. Int J Technol Assess Health Care. 1996;12:238-46.

Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? Lancet 355(9198):129-34. 2000;355:129-34.

Gould AL. Multi-centre trial analysis revisited. Stat Med. 1998;17:1779-97.

Gould AL. Sample size re-estimation: recent developments and practical considerations. Stat Med. 2001;20:2625-43.

Green SB. How many subjects does it take to do a regression analysis. Multivariate Behavioural Research. 1991;26:499-510.

Greenberg Report. Organization, review, and administration of cooperative studies (Greenberg Report): a report from the Heart Special Project Committee to the National Advisory Heart Council, May 1967. Control Clin Trials. 1988;9:137-48.

Greene T, Beck GJ, Gassman JJ, Gotch FA, Kusek JW, Levey AS et al. Design and statistical issues of the hemodialysis (HEMO) study. Control Clin Trials. 2000a;21:502-25.

Greene WH. Heteroscedasticity. In: Greene WH, ed. Econometric Analysis. 4th ed. Upper Saddle River, NJ: Prentice-Hall, Inc; 2000: 499-524.

Greene WL, Concato J, Feinstein AR. Claims of equivalence in medical research: are they supported by the evidence? Annals of Internal Medicine. 2000b;132:715-22.

Greenhouse S. Jerome Cornfield's contributions to Epidemiology. Biometrics. 1982;Supplement 1982:33-43.

Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol. 1994;140:290-296.

Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierachical view of proposed solutions. Biostat. 2001;2:463-71.

Greenwald AG, Gonzalez R, Harris RJ, Guthrie D. Effect sizes and p values: what should be reported and what should be replicated? Psychophysiology. 1996;33:175-83.

Guiguet M, Blot F, Escudier B, Antoun S, Leclercq B, Nitenberg G. Severity-of-illness scores for neutropenic cancer patients in an intensive care unit: Which is the best predictor? Do multiple assessment times improve the predictive value? Crit Care Med. 1998;26:488-93.

GUSTO I Investigators. An International Randomized Trial Comparing Four Thrombolytic Strategies For Acute Myocardial Infarction. N Engl J Med. 1993;329:673-82.

GUSTO III Investigators. A Comparison of Reteplase with Alteplase for Acute Myocardial Infarction. N Engl J Med. 1997;337:1118-23.

Gyldmark M. A review of cost studies of intensive care units: problems with the cost concept. Crit Care Med. 1995;23:964-72.

Hacking I. The logic of statistical inference. Cambridge: Cambridge University Press; 1965.

Hahn S, Garner P, Williamson P. Are systematic reviews taking heterogeneity into account? An analysis from the Infectious Diseases Module of the Cochrane Library. J Eval Clin Pract. 2000a;6:231-33.

Hahn S, Williamson PR, Hutton JL, Garner P, Flynn EV. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. Stat Med. 2000b;19:3325-36.

Hampton JR. Mega-trials and equivalence trials: experience from the INJECT study. European Heart Journal. 1996;17:Suppl-34.

Hannan ELP, Wu CM, DeLong ERP, Raudenbush SWE. Predicting Risk-Adjusted Mortality for CABG Surgery: Logistic Versus Hierarchical Logistic Models. Med Care. 2005;43:726-35.

Hardin J, Hilbe J. Generalized linear models and extensions. College Station, TX: Stata Press; 2001.

Harlow LL, Mulaik SA, Steiger JH. What if there were no significance tests. Hillsdale, NJ: Lawrence Erlbaum Associates; 1997.

Harrell Jr FE. How should change be measured? http://hesweb1 med virginia edu/biostat/teaching/change pdf. 1997.

Harrell Jr FE. Design Library. Design Library @ http://biostat mc vanderbilt edu/twiki/bin/view/Main/Design. 2005.

Harrell FE, Jr. Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.

Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.

Hart A. Mann-Whitney test is not just a test of medians: differences in spread can be important. BMJ. 2001;323:391-93.

Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2001.

Hatala R, Holbrook A, Goldsmith CH. Therapeutic equivalence: all studies are not created equal. Canadian Journal of Clinical Pharmacology. 1999;6:9-11.

Hauck WW, Anderson S. A proposal for interpreting and reporting negative studies. Stat Med. 1986;5:203-9.

Hauck WW, Anderson S. Some issues in the design and analysis of equivalence trials. Drug Inf J. 1999;33:109-18.

Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials?. Control Clin Trials. 1998;19:249-56.

Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol. 1991;44:77-81.

Hauschke D. Choice of delta: A special case. Drug Inf J. 2001;35:875-79.

Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. British Journal of Radiology. 1971;44:793-97.

Hayes RJ. Methods for assessing whether change depends on initial value. Stat Med. 1988;7:915-27.

Hebert PC, Cook DJ, Wells G, Marshall J. The design of randomized clinical trials in critically ill patients. Chest. 2002;121:1290-1300.

Hebert PC, Wells G, Blajchman MA, Marshall J, Martin C, Pagliarello G et al. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical

care. Transfusion Requirements in Critical Care Investigators, Canadian Critical Care Trials Group. N Engl J Med. 1999;340:409-17.

Hedeker D. Generalized Linear Mixed models. Encyclopedia of Statistics in Behavioral Science. @ http://media.wiley.com/product_data/excerpt/04/04708608/0470860804-3.pdf: John Wiley & Sons, Ltd; 2005.

Henderson R, Oman P. Effect of frailty on marginal regression estimates in survival analysis. Journal of the Royal Statistical Society B. 1999;61:367-79.

Herndon JE, Harrell FE, Jr. The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. Stat Med. 1995;14:2119-29.

Heyland DK. Enteral and parenteral nutrition in the seriously ill, hospitalized patient: a critical review of the evidence. Journal of Nutrition, Health & Aging. 2000a;4:31-41.

Heyland DK. Parenteral nutrition in the critically-ill patient: more harm than good?. Proc Nutr Soc. 2000b;59:457-66.

Heyland DK, Cook DJ, Jaeschke R, Griffith L, Lee HN, Guyatt GH. Selective decontamination of the digestive tract. An overview. Chest. 1994;105:1221-29.

Heyland DK, Cook DJ, King D, Kernerman P, Brun-Buisson C. Maximizing oxygen delivery in critically ill patients: a methodologic appraisal of the evidence. Crit Care Med. 1996;24:517-24.

Heyland DK, MacDonald S, Keefe L, Drover JW. Total parenteral nutrition in the critically ill patient: a meta-analysis. JAMA. 1998;280:2013-19.

Heyland DK, Montalvo M, MacDonald S, Keefe L, Su XY, Drover JW. Total parenteral nutrition in the surgical patient: a meta-analysis. Can J Surg. 2001a;44:102-11.

Heyland DK, Novak F, Drover JW, Jain M, Su X, Suchner U. Should immunonutrition become routine in critically ill patients? A systematic review of the evidence. JAMA. 2001b;286:944-53.

Heys SD, Walker LG, Smith I, Eremin O. Enteral nutritional supplementation with key nutrients in patients with critical illness and cancer: a meta-analysis of randomized controlled clinical trials. Ann Surg. 1999;229:467-77.

Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21:1539-58.

Hilsenbeck SG, Clark GM. Practical p-value adjustment for optimally selected cutpoints. Stat Med. 1996;15:103-12.

Hilsenbeck SG, Clark GM, McGuire WL. Why do so many prognostic factors fail to pan out? Breast Cancer Research & Treatment. 1992;22:197-206.

Hinds CJ. Treatment of sepsis with activated protein C. Br Med J. 2001;323:881-82.

Hintze, J. L. PASS 2002. http://www.ncss.com/ . 2002. Kaysville, Utah, NCSS.

Ref Type: Electronic Citation

Hirsch RP. Validation samples. Biometrics. 1991;47:1193-94.

Hoenig JM, Heisey DM. The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician. 2001;55:19-24.

Hoes AW, Grobbee DE, Lubsen J. Does drug treatment improve survival? Reconciling the trials in mild-to-moderate hypertension. J Hypertens. 1995;13:805-11.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. Statistical Science. 1999;14:382-417.

Hogan JW, Laird NM. Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. Stat Methods Med Res. 1998;7:28-48.

Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6:65-70.

Holmgren EB. Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. J Biopharm Stat. 1999;9:651-59.

Holzer M, Behringer W, Schorkhuber W, Zeiner A, Sterz F, Laggner AN et al. Mild hypothermia and outcome after CPR. Hypothermia for Cardiac Arrest (HACA) Study Group. Acta Anaesthesiol Scand Suppl. 1997;111:55-58.

Horton NJ, Lipsitz SR. Multiple imputation in practice: Comparison of software packages for regression models with missing values. The American Statistician. 2001;55:244-54.

Horwitz RI, Singer BH, Makuch RW, Viscoli CM. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. J Clin Epidemiol. 1996;49:395-400.

Hosmer Jr DW, Lemeshow S. Applied Survival Analysis: regression modeling of time to event data. New York: JohnWiley & Sons, Inc; 1999.

Hoth T, Evans TW. Activated protein C: the cure for sepsis - again? Anaesthesia. 2001;56:1133-35.

Hougaard P. Frailty models for survival data. Lifetime Data Anal. 1995;1:255-73.

Huaringa AJ, Leyva FJ, Giralt SA, Blanco J, Signes-Costa J, Velarde H et al. Outcome of bone marrow transplantation patients requiring mechanical ventilation. Crit Care Med. 2000;28:1014-17.

Hubbard R, Bayarri MJ. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. The American Statistician. 2003;57:171-82.

Hughes MD. Reporting Bayesian analyses of clinical trials. Stat Med. 1993;12:1651-63.

Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. Stat Med. 1988;7:1231-42.

Hulme GJ, Columb MO. Mortality of medical oncological patients admitted to intensive care. Br J Anaesth. 1999;82:807-8.

Hung HM, O'Neill RT, Bauer P, Kohne K. The behavior of the P-value when the alternative hypothesis is true. Biometrics. 1997;53:11-22.

Hutton JL. Number needed to treat: properties and problems. Journal of the Royal Statistical Society Series A. 2000;163:403-19.

Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. Applied Statistics. 2000;49:359-70.

Hwang IK, Morikawa T. Design issues in noninferiority/equivalence trials. Drug Inf J. 1999;33:1205-18.

Insightful Corporation. S+SEQTRIAL 2 User's Manual. Seattle, Washington: Insightful Corporation; 2002.

Intensive Care Medicine. Intensive Care Medicine. Intensive Care Medicine @ http://www esicm org/PAGE_journalicm/?1h09. 2005.

Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. JAMA. 1998a;279:1089-93.

Ioannidis JP, Haidich AB, Lau J. Any casualties in the clash of randomised and observational evidence? BMJ. 2001;322:879-80.

Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. J Clin Epidemiol. 1997;50:1089-98.

Ioannidis JP, Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. Am J Epidemiol. 1998b;148:1117-26.

JAMA. Journal of the American Medical Association (JAMA). JAMA @ http://jama ama-assn org/. 2005.

Jennison C, Turnbull BW. Group Sequential Methods with applications to clinical trials. Boca Raton: Chapman & Hall/CRC; 2000a.

Jennison C, Turnbull BW. Group Sequential Methods with applications to clinical trials. Boca Raton: Chapman & Hall/CRC; 2000b.

Jiang J, Yu M, Zhu C. Effect of long-term mild hypothermia therapy in patients with severe traumatic brain injury: 1-year follow-up review of 87 cases. J Neurosurg. 2000;93:546-49.

Johnson DH. The insignificance of statistical significance testing. Journal of Wildlife Management. 1999;63:763-72.

Johnson NJ. Modified t tests and confidence intervals for asymmetrical populations. Journal of the American Statistical Association. 1978;73:536-44.

Johnstone DJ. Tests of significance in theory and practice. The Statistician. 1986;35:491-504.

Johnstone DJ. On the interpretation of hypothesis tests following Neyman and Pearson. In: Viertl R, ed. Probability and Bayesian Statistics. New York, NY: Plenum Press; 1987: 267-77.

Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ. 1996;313:36-39.

Julian DG, Camm AJ, Frangin G, Janse MJ, Munoz A, Schwartz PJ et al. Randomised trial of effect of amiodarone on mortality in patients with left-ventricular dysfunction after recent myocardial infarction: EMIAT. European Myocardial Infarct Amiodarone Trial Investigators. Lancet. 1997;349:667-74.

Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282:1054-60.

Kaiser L. Adjusting for baseline: change or percentage change? [see comments]. Stat Med. 1989;8:1183-90.

Kallen A. Treatment-by-center interaction: what is the isuue? Drug Inf J. 1997;31:927-36.

Kam LW, Lin JD. Management of systemic candidal infections in the intensive care unit. Am J Health Syst Pharm. 2002;59:33-41.

Karnofsky DA, Abelmann WH, Craver LF, Burchenal JH. The use of the nitrogen mustards in the palliative treatment of carcinoma. Cancer. 1948;1:634-56.

Keenan SP, Kernerman PD, Cook DJ, Martin CM, McCormack D, Sibbald WJ. Effect of noninvasive positive pressure ventilation on mortality in patients admitted with acute respiratory failure: a meta-analysis. Crit Care Med. 1997;25:1685-92.

Keene ON. The log transformation is special. Stat Med. 1995;14:811-19.

Kempthorne O. Some aspects of experimental inference. Journal of the American Statistical Association. 1966;61:11-34.

Kempthorne O. Theories of inference and data analysis. In: Bancroft TA, ed. Statistical Papers in honour of George W Snedecor. Ames, Iowa: The Iowa State Univerity Press; 1972: 167-91.

Kempthorne O. Of what use are tests of sinificnace and tests of hypotheses. Communications in Statistics-Theory and Methods A5. 1976;8:763-77.

Keuzenkamp HA, Magnus JR. On tests and significance in econometrics. Journal of Econometrics. 1995;67:5-24.

Kilian R, Matschinger H, Loeffler W, Roick C, Angermeyer MC. A comparison of methods to handle skew distributed cost variables in the analysis of the resource consumption in schizophrenia treatment. The Journal of Mental Health Policy & Economics. 2002;5:21-31.

Kim K. Point estimation following group sequential tests. Biometrics. 1989;45:613-17.

Kim MY, Buyon JP, Petri M, Skovron ML, Shore RE. Equivalence trials in SLE research: issues to consider. Lupus. 1999;8:620-626.

Kingman A, Zion G. Some power considerations when deciding to use transformations. Stat Med. 1994;13:769-83.

Kirshner B. Methodological standards for assessing therapeutic equivalence. Journal of Clinical Epidemiology. 1991;44:839-49.

Kittelson JM, Emerson SS. A unifying family of group sequential test designs. Biometrics. 1999;55:874-82.

Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer-Verlag; 1997.

Knapp G, Hartung J. Combined test procedures in the meta-analysis of controlled clinical trials. Studies in Health Technology & Informatics. 2000;77:34-38.

Knaus WA. Prognosis with mechanical ventilation: the influence of disease, severity of disease, age, and chronic health status on survival from an acute illness. Am Rev Respir Dis. 1989;140:S8-13.

Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985a;13:818-29.

Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985b;13:818-29.

Knaus WA, Draper EA, Wagner DP, Zimmerman JE. Prognosis in acute organ-system failure. Ann Surg. 1985d;202:685-93.

Knaus WA, Draper EA, Wagner DP, Zimmerman JE. Prognosis in acute organ-system failure. Ann Surg. 1985c;202:685-93.

Knaus WA, Harrell FE, Fisher CJ, Jr., Wagner DP, Opal SM, Sadoff JC et al. The clinical evaluation of new drugs for sepsis. A prospective study design based on survival analysis. JAMA. 1993;270:1233-41.

Knaus WA, Harrell FE, Jr., LaBrecque JF, Wagner DP, Pribble JP, Draper EA et al. Use of predicted risk of mortality to evaluate the efficacy of anticytokine therapy in sepsis. The rhIL-1ra Phase III Sepsis Syndrome Study Group. Crit Care Med. 1996;24:46-56.

Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991;100:1619-36.

Knottnerus JA, Bouter LM. The ethics of sample size: two-sided testing and one-sided thinking. Journal of Clinical Epidemiology. 2001;54:109-10.

Kolata EB. Controversy over study of diabetes drugs continues for nearly a decade. Science. 1979;203:986-90.

Kollef MH. Selective digestive decontamination should not be routinely employed. Chest. 2003;123:Suppl-8S.

Koretz RL. Is nutritional support worthwhile. In: Heatley RV, Green JH, Losowsky MS, eds. Consensus in Clinical Nutrition. Cambridge: Cambrideg University Press; 1994: 158-91.

Koretz RL. Nutritional supplementation in the ICU. How critical is nutrition for the critically ill? American Journal of Respiratory & Critical Care Medicine. 1995;151:t-3.

Korinek AM, Laisne MJ, Nicolas MH, Raskine L, Deroin V, Sanson-Lepors MJ. Selective decontamination of the digestive tract in neurosurgical intensive care unit patients: a double-blind, randomized, placebo-controlled study. Crit Care Med. 1993;21:1466-73.

Kress JP, Christenson J, Pohlman AS, Linkin DR, Hall JB. Outcomes of critically ill cancer patients in a university hospital setting. American Journal of Respiratory & Critical Care Medicine. 1999;160:1957-61.

Krueger WA, Lenhart FP, Neeser G, Ruckdeschel G, Schreckhase H, Eissner HJ et al. Influence of combined intravenous and topical antibiotic prophylaxis on the incidence of infections, organ dysfunctions, and mortality in critically ill surgical patients: a prospective, stratified, randomized, double-blind, placebo-controlled clinical trial. American Journal of Respiratory & Critical Care Medicine. 2002;166:1029-37.

Kruskal W. The significance of Fisher: a review of [Box JF] R.A. Fisher: the life of a scientist. Journal of the American Statistical Association. 1980;75:1019-30.

Kujath P, Lerch K, Kochendorfer P, Boos C. Comparative study of the efficacy of fluconazole versus amphotericin B/flucytosine in surgical patients with systemic mycoses. Infection. 1993;21:376-82.

Kupper LL. Estimation, Interval. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics. New York: John Wiley & Sons, Inc; 1998: 1391-94.

Kyburg Jr HE. The logical foundations of Statistical Inference. Boston, USA: D. Reidel publishing Company; 1974.

Laaban JP, Waked M, Laromiguiere M, Vuong TK, Rochemaure J. Hypophosphatemia complicating management of acute severe asthma. Ann Intern Med. 1990;112:68-69.

Lan KK, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. Communications in Statistics - Sequential Anlaysis. 1982;1:207-19.

Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70:659-63.

Lane PW. Generalized linear models in soil science. European Journal of Soil Science. 2002;53:241-51.

Lang JM, Rothman KJ, Cann CI. That confounded P-value. Epidemiology. 1998;9:7-8.

Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med. 1992;327:248-54.

Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med. 1988;318:1728-33.

Lausen B, Schumacher M. Evaluating the effect of otimized cutoff values in the assessment of prognostic factors. Computational Statistics & Data Analysis. 1996;21:307-26.

Lawlor DA, Davey SG, Ebrahim S. Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? Int J Epidemiol. 2004;33:464-67.

Lee AH, Xiao J, Vemuri SR, Zhao Y. A discordancy test approach to identify outliers of length of hospital stay. Stat Med. 1998;17:2199-206.

Lehman EL. Testing statistical hypotheses. New York: John Wiley and Sons; 1959.

Lehman EL. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two. Journal of the American Statistical Association. 1993;88:1242-49.

LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med. 1997;337:536-42.

Lesaffre E, Boon P, Pledger GW. The Value of the Number-Needed-to-Treat Method in Antiepileptic Drug Trials. Epilepsia. 2000;41:440-446.

Leung DHY. Statistical methods for clinical studies in the presence of surrogate end points. Journal of the Royal Statistical Society A. 2001;164:485-503.

Lewis JA, Machin D. Intention to treat--who should use ITT? Br J Cancer. 1993;68:647-50.

Lewis SJ, Egger M, Sylvester PA, Thomas S. Early enteral feeding versus "nil by mouth" after gastrointestinal surgery: systematic review and meta-analysis of controlled trials. BMJ. 2001;323:773-76.

Li Y, Shi L, Roth HD. The bias of the commonly-used estimates of variance in meta-analysis. Communications in Statistics-Theory & Methods. 1994;23:1063-85.

Liberati A, D'Amico R, Pifferi S, Leonetti C, Torri V, Brazzi L et al. Antibiotics for preventing respiratory tract infections in adults receiving intensive care. [update of Cochrane Database Syst Rev. 2000;(2):CD000022]. Cochrane Database of Systematic Reviews. 2000;CD000022.

Liberati A, D'Amico R, Pifferi S, Leonetti C, Torri V, Brazzi L et al. Antibiotics for preventing respiratory tract infections in adults receiving intensive care. Cochrane Database of Systematic Reviews. 2003.

Light R, Pillemer DB. Summing up: the science of reviewing research. Cambridge: Harvard University Press; 1984.

Lindley DV. Discussion of 'Tests of significance in theory and practice' by D.J. Johnstone. The Statistician. 1986;35:502-4.

Lindsey JK, Jones B. Choosing among generalized linear models applied to medical data. Stat Med. 1998;17:59-68.

Lipman TO. Grains or veins: is enteral nutrition really better than parenteral nutrition? A look at the evidence. Journal of Parenteral & Enteral Nutrition. 1998;22:167-82.

Lipsitz SR, Dear KB, Laird NM, Molenberghs G. Tests for homogeneity of the risk difference when data are sparse. Biometrics. 1998;54:148-60.

Little R. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association. 1988a;83:1198-202.

Little R. Regression with missing X's: A review. Journal of the American Statistical Association. 1992;87:1227-37.

Little R. Robust estimation of the mean and covariance matrix from data with missing values. Applied Statistics. 1988b;37:23-28.

Liu A, Tan M, Boyett JM, Xiong X. Testing secondary hypotheses following sequential clinical trials. Biometrics. 2000;56:640-644.

Lizan-Garcia M, Garcia-Caballero J, Asensio-Vegas A. Risk factors for surgical-wound infection in general surgery: a prospective study. Infection Control & Hospital Epidemiology. 1997;18:310-315.

Long, J. S. and Freese, J. Scalar measures of fit. Stata Technical Bulletin 56(July), 34-40. 2000. TX, Stata Corporation.
Ref Type: Serial (Book,Monograph)

Louis TA. Assessing, accommodating, and interpreting the influences of heterogeneity. Environ Health Perspect. 1991;90:215-22.

Ludbrook J. Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. Clinical & Experimental Pharmacology & Physiology. 1994;21:673-86.

Ludbrook J. The Wilcoxon-Mann-Whitney test condemned. Br J Surg. 1996;83:136-37.

Ludbrook J. Statistics in physiology and pharmacology: a slow and erratic learning curve. Clinical & Experimental Pharmacology & Physiology. 2001;28:488-92.

Ludbrook J, Dudley H. Issues in biomedical statistics: statistical inference. ANZ Journal of Surgery. 1994;64:630-636.

Ludbrook J, Dudley H. Why permutation tests are superior to t and F tests in biomedical reearch. The American Statistician. 1998;52:127-32.

Lui KJ, Kelly C. Tests for homogeneity of the risk ratio in a series of 2x2 tables. Stat Med. 2000;19:2919-32.

Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health. 2002;23:151-69.

Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. Stat Med. 2001;20:641-54.

Macdonald RR. The Incompleteness of Probability Models and the Resultant Implications for Theories of Statistical Inference. Understanding Statistics. 2002;1:167-89.

MacFie J. Enteral versus parenteral nutrition: the significance of bacterial translocation and gut-barrier function. Nutrition. 2000;16:606-11.

Makuch R, Johnson M. Issues in planning and interpreting active control equivalence studies. Journal of Clinical Epidemiology. 1989;42:503-11.

Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. Cancer Treatment Reports. 1978;62:1037-40.

Makuch RW, Johnson MF. Some issues in the design and interpretation of 'negative' clinical studies. Archives of Internal Medicine. 1986;146:986-89.

Makuch RW, Simon RM. Sample size considerations for non-randomized comparative studies. Journal of Chronic Diseases. 1980;33:175-81.

Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. J Health Econ. 1998;17:283-95.

Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ. 2001;20:461-94.

Manns BJ, Lee H, Doig CJ, Johnson D, Donaldson C. An economic evaluation of activated protein C treatment for severe sepsis. N Engl J Med. 2002b;347:993-1000.

Manns BJ, Lee H, Doig CJ, Johnson D, Donaldson C. An economic evaluation of activated protein C treatment for severe sepsis. N Engl J Med. 2002a;347:993-1000.

Mantel N. Evaluation of survival data and two new rank-order statistics arising in its consideration. Cancer Chemotherapy Reports. 1966;50:163-70.

Mantel N. Why stepdown procedures in variable selection. Technometrics. 1970;12:621-25.

Mantel N. Active and passive smoking and pathological indicators of lung cancer--a report of limited value? JAMA. 1993;270:1689.

Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22:719-48.

Marik PE, Zaloga GP. Early enteral nutrition in acutely ill patients: a systematic review. Crit Care Med. 2001;29:2264-70.

Marion DW, Penrod LE, Kelsey SF, Obrist WD, Kochanek PM, Palmer AM et al. Treatment of traumatic brain injury with moderate hypothermia. N Engl J Med. 1997;336:540-546.

Marks HM. Rigorous uncertainty: Why RA Fisher is important. International Journal of Epidemiology Vol 32(6)()(pp 932-937), 2003. 2003;932-37.

Marshall LF. Intercenter variance. J Neurosurg. 2001;95:733.

Martin R, Quinton P, Hinds CJ. Prognosis of patients admitted to the intensive care with life threatening medical complications of haematological malignancy: has it changed? Br J Anaesth. 1998;81:813P-4P.

May K. A Note on the Use of Confidence Intervals. Understanding Statistics. 2003;2:133-35.

Maynard ND, Bihari DJ. Postoperative feeding. BMJ. 1991;303:1007-8.

McAlister FA, Sackett DL. Active-control equivalence trials and antihypertensive agents. American Journal of Medicine. 2001;111:553-58.

McCloskey RV, Straube RC, Sanders C, Smith SM, Smith CR. Treatment of septic shock with human monoclonal antibody HA-1A. A randomized, double-blind, placebo-controlled trial. CHESS Trial Study Group. Ann Intern Med. 1994;121:1-5.

McNutt A, Wu C, Xue X, Hafner JP. Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. Am J Epidemiol. 2003;157:940-943.

McPherson K. Statistics: the problem of examining accumulating data more than once. N Engl J Med. 1974;290:501-2.

McPherson K. On choosing the number of interim analyses in clinical trials. Stat Med. 1982;1:25-36.

McPherson K. Sequential stopping rules in clinical trials. Stat Med. 1990;9:595-600.

Meduri GU, Headley AS, Golden E, Carson SJ, Umberger RA, Kelso T et al. Effect of prolonged methylprednisolone therapy in unresolving acute respiratory distress syndrome: a randomized controlled trial. JAMA. 1998;280:159-65.

Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting & Clinical Psychology. 1978;46:806-34.

Meldrum ML. A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. Hematology - Oncology Clinics of North America. 2000;14:745-60.

Mick R, Ratain MJ. Bootstrap validation of pharmacodynamic models defined via stepwise linear regression. Clinical Pharmacology & Therapeutics. 1994;56:217-22.

Milberg JA, Davis DR, Steinberg KP, Hudson LD. Improved survival of patients with acute respiratory distress syndrome (ARDS): 1983-1993. JAMA. 1995;273:306-9.

Millns H, Woodward M, Bolton-Smith C. Is it necessary to transform nutrient variables prior to statistical analyses? American Journal of Epidemiology. 1995;141:251-62.

Moeschberger ML, Klein JP. Statistical methods for dependent competing risks. Lifetime Data Anal. 1995;1:195-204.

Montori VM, Devereaux PJ, Neill KJA, Karen EAB, Christoph HE, Matthias B et al. Randomized Trials Stopped Early for Benefit: A Systematic Review. JAMA: Journal of the American Medical Association. 2005;294:2203-9.

Moore RA, Gavaghan D, Tramer MR, Collins SL, McQuay HJ. Size is everything--large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. Pain. 1998;78:209-16.

Mora-Duarte J, Betts R, Rotstein C, Colombo AL, Thompson-Moya L, Smietana J et al. Comparison of caspofungin and amphotericin B for invasive candidiasis. New England Journal of Medicine. 2002;347:2020-2029.

Moran JL, Bersten AD, Solomon PJ. Meta-analysis of controlled trials of ventilator therapy in acute lung injury and acute respiratory distress syndrome: an alternative perspective. Intensive Care Med. 2005a;31:227-35.

Moran JL, Peake SL. Further reflections on clinical trials in critical care. Critical Care and Resuscitation. 2001a;3:226-29.

Moran JL, Peake SL, Solomon P. Hypothermia as therapy in cerebral injury. Critical Care and Resuscitation. 2002a;4:93-102.

Moran JL, Peake SL, Solomon P. Learning new lessons or repeating old mistakes? Critical Care and Resuscitation. 2002b;4:257-60.

Moran JL, Peake SL, Solomon PJ. Reporting of clinical trials using group sequential methods. Critical Care and Resuscitation. 2001b;3:146-47.

Moran JL, Peake SL, Warn D. Heterogeneity of treatment effects in meta-analysis of intensive care practice. Intensive Care Med. 2001c;27:170.

Moran JL, Peisach AR, Solomon P. Modelling total costs in adult intensive care units: new models for old questions. Intensive Care Med. 2001d;27:S142.

Moran JL, Peisach AR, Solomon PJ, Martin J. Cost calculation and prediction in adult intensive care: a ground-up utilisation study. Anaesthesia & Intensive Care. 2004a;32:787-97.

Moran JL, Peter JV, Solomon P. Nutrition as Therapy: let's look at the evidence. Critical Care and Resuscitation. 2002c;**4**:164-69.

Moran JL, Solomon P. Worrying about normality. Critical Care and Resuscitation. 2002d;4:316-19.

Moran JL, Solomon P, Ay Yeung KW, Pannall PR, John G, Eliseo A. Phosphate metabolism in intensive care patients with acute respiratory failure. Critical Care and Resuscitation. 2002e;4:93-103.

Moran JL, Solomon PJ. Mortality and other event rates: what do they tell us about performance? Critical Care and Resuscitation. 2003a;5:292-303.

Moran JL, Solomon PJ. Selective digestive decontamination: once again. Critical Care and Resuscitation. 2003b;5:241-46.

Moran JL, Solomon PJ. Some aspects of the design and monitoring of clinical trials. Critical Care and Resuscitation. 2003c;5:137-46.

Moran JL, Solomon PJ. The interpretation of lack of evidence of a difference in efficacy: equivalence trials and the treatment of fungal infections. Critical Care and Resuscitation. 2003d;5:216-23.

Moran JL, Solomon PJ. A farewell to P-values? Critical Care and Resuscitation. 2004b;6:130-138.

Moran JL, Solomon PJ. Critical care trials: sample size, power and interpretation. Critical Care and Resuscitation. 2004c;6:239-42.

Moran JL, Solomon PJ, Fox V, Salagaras M, Williams PJ, Quinlan K et al. Modelling thirty-day mortality in the Acute Respiratory Distress Syndrome (ARDS) in an adult ICU. Anaesthesia & Intensive Care. 2004d;32:317-29.

Moran JL, Solomon PJ, Williams PJ. Assessment of Outcome Over a 10-year Period of Patients Admitted to a Multidisciplinary Adult Intensive Care Unit with Haematological and Solid Tumours. Anaesthesia & Intensive Care. 2005b;33:26-35.

Morgan TJ. Life without the PA catheter. Critical Care and Resuscitation. 2004;6:9-12.

Morgan TM. Analysis of duration of response: a problem of oncology trials. Control Clin Trials. 1988;9:11-18.

Morgan TM, Elashoff RM. Effect of categorizing a continuous covariate on the comparison of survival time. Journal of the American Statistical Society. 1986;81:917-21.

Morris C. Activated protein C in the treatment of sepsis. Anaesthesia. 2002a;57:502-4.

Morris RW. Does EBM offer the best opportunity yet for teaching medical statistics? Stat Med. 2002b;21:969-77.

Morrison DE, Henkel RE. The significance test controversy-a reader. Chicago,Ill: Aldine Publishing; 1969.

Moss AJ, Hall WJ, Cannom DS, Daubert JP, Higgins SL, Klein H et al. Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia. Multicenter Automatic Defibrillator Implantation Trial Investigators. New England Journal of Medicine. 1996;335:1933-40.

Moye LA, Tita AT. Defending the rationale for the two-tailed test in clinical research. Circulation. 2002;105:3062-65.

MPS Research Unit. PEST 4. reading, UK: MPS Research Unit; 2002.

Murray GD. Reply from BJS Statistical Adviser. Br J Surg. 1996;83:137.

Myers RH, Montgomery DC. A tutorial on generalized linear models. Journal of Quality Technology. 1997;29:274-91.

Narayan RK. Hypothermia for traumatic brain injury--a good idea proved ineffective. N Engl J Med. 2001;344:602-3.

Nathens AB, Marshall JC. Selective decontamination of the digestive tract in surgical patients: a systematic review of the evidence. Arch Surg. 1999b;134:170-176.

Nathens AB, Marshall JC. Selective decontamination of the digestive tract in surgical patients: a systematic review of the evidence. Arch Surg. 1999c;134:170-176.

Nathens AB, Marshall JC. Selective decontamination of the digestive tract in surgical patients: a systematic review of the evidence. Arch Surg. 1999a;134:170-176.

NCSS. NCSS 2004. NCSS 2004 @ http://ncss com. 2004.

Nester MR. An applied Statistician's creed. Applied Statistics. 1996;45:401-10.

Neyman J. Outline of a theory of statistical estimation based on the classical theory of probabilty. Philosophical Transactions of the Royal Society of London Series A: Mathematical and Physical Sciences. 1937;236:333-80.

Neyman J. Fiducial argument and the theory of confidence intervals. Biometrika. 1941;32:128-50.

Neyman J. Note on an article by Sir Ronald Fisher. Journal of the Royal Statistical Society, Series B. 1956;18:288-94.

Neyman J. Silver Jubilee of my dispute with Fisher. Journal of the Operational Research Society of Japan. 1961;3:145-54.

Neyman J. Frequentist probability and frequentist statistics. Synthese. 1977;36:97-131.

Neyman J, Pearson ES. On the problem of the most efficient test of statistical hypotheses. Philosophical Transactions of the Royal Society of London Series A: Mathematical and Physical Sciences. 1933;231:289-337.

Neyman J, Pearson ES. Joint Statistical Papers. London, UK: Cambridge University Press; 1967.

Ng T-H. A specification of treatment difference in the design of clinical trials with active controls. Drug Inf J. 1993;27:705-19.

Ng T-H. Conventional null hypothesis testing in active control equivalence studies. Control Clin Trials. 1995;16:356-58.

Ng T-H. Choice of delta in equivalence testing. Drug Inf J. 2001;35:1517-27.

Nguyen MHM, Peacock JEJM, Tanner DCM, Morris AJM, Nguyen MLM, Snydman DRM et al. Therapeutic Approaches in Patients With Candidemia: Evaluation in a Multicenter, Prospective, Observational Study. Arch Intern Med. 1995;155:2429-35.

Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods. 2000;5:241-301.

Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med. 1999;18:321-59.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979;35:549-56.

O'Connell B, Craig JI, Marcus RE, Ludlam H. Cost-effective use of liposomal amphotericin B. Clinical & Laboratory Haematology. 2002;24:317-19.

O'Hagan A, Stevens JW. Assessing and comparing cost: How robust are the bootstrap and methods based on asymptotic normality? Health Econ. 2003;12:33-49.

Oldham PD. A note on the analysis of repeated measuements of the same subjects. Journal of Chronic Disease. 1962;15:969-77.

Olkin I. Meta-analysis: reconciling the results of independent studies. Stat Med. 1995;14:457-72.

Opal SM. Clinical impact of novel anticoagulation strategies in sepsis. Current Opinion in Critical Care. 2001b;7:347-53.

Opal SM. Clinical impact of novel anticoagulation strategies in sepsis. Current Opinion in Critical Care. 2001a;7:347-53.

Ostrosky-Zeichner L, Marr KA, Rex JH, Cohen SH. Amphotericin B: time for a new 'gold standard'. Clin Infect Dis. 2003.

Overall JE. Tests of one-sided versus two-sided hypotheses in placebo-controlled clinical trials. Neuropsychopharmacology. 1990;3:233-35.

Overall JE, Atlas RS. Selecting an interim analysis procedure. Psychopharmacology Bulletin. 1993;29:141-47.

Pampallona S, Tsiatis AA. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis. Journal of Statistical Planning and Inference. 1994;42:19-35.

Pampallona S, Tsiatis AA, Kim KM. Interim monitoring of group sequential trials using spending functions for the Type I and Type II error probabilities. Drug Inf J. 2001;35:1113-21.

Parkhurst DF. Arithmetic versus geometric means for environmental concentration data. Environmental Science & Technology. 1998;92-98.

Pearson ES. Statistical concepts and their relation to reality. Journal of the Royal Statistical Society, Series B. 1955;17:204-7.

Peck CC, Weschler J. Report of a workshop on confirmatory evidence to support a single clinical trial as a basis for new drug approval. Drug Inf J. 2002;36:517-34.

Peduzzi P. Termination of the Department of Veterans Affairs Cooperative Study of steroid therapy for systemic sepsis. Control Clin Trials. 1991;12:395-407.

Pena EA, Rohatgi VK. Most powerful test. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics. New York: John Wiley & Sons, Inc; 1998: 2703-6.

Perneger TV. Sifting the evidence. Likelihood ratios are alternatives to P values. Br Med J. 2001;322:1184-85.

Peter JV, Moran JL, Phillips-Hughes J. A metaanalysis of treatment outcomes of early enteral versus early parenteral nutrition in hospitalized patients *. [Review]. Crit Care Med. 2005;33:213-20.

Peter JV, Moran JL, Phillips-Hughes J, Warn D. Noninvasive ventilation in acute respiratory failure--a meta-analysis update. Crit Care Med. 2002;30:555-62.

Peto R. Statisitcal aspects of cancer trials. In: Halnan KE, ed. Treatment of Cancer. London: Chapman & Hall; 1982: 867-71.

Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol. 1995;48:23-40.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. British Journal of Cancer. 1976;34:585-612.

Phillips P, Shafran S, Garber G, Rotstein C, Smaill F, Fong I et al. Multicenter randomized trial of fluconazole versus amphotericin B for treatment of candidemia in non-neutropenic patients. Canadian Candidemia Study Group. European Journal of Clinical Microbiology & Infectious Diseases. 1997;16:337-45.

Piantadosi S. Hazards of small clinical trials. J Clin Oncol. 1990;8:1-3.

Pickles A, Crouchley R. Generalizations and applications of frailty models for survival and event data. Stat Methods Med Res. 1994;3:263-78.

Piggott TD. Methods for handling missing data in research synthesis. In: Cooper H, Hedges LV, eds. The handbook of research synthesis. New york: Russell Sage Foundation; 1994: 163-75.

Pinheiro JC, Bates DM. Mixed-effects models in S and S-Plus. Rensselaer, NY: Springer-Verlag New York, Inc; 2000.

Pinheiro JC, DeMets DL. Estimating and reducing bias in group sequential designs with Gaussian independent incremental structure. Biometrika. 1997;84:831-45.

Pittet D, Thievent B, Wenzel RP, Li N, Gurman G, Suter PM. Importance of pre-existing co-morbidities for prognosis of septicemia in critically ill patients. Intensive Care Med. 1993;19:265-72.

Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977;64:191-99.

Pocock SJ. Allocation of patients to treatment in clinical trials. Biometrics. 1979;35:183-97.

Pocock SJ. When to stop a clinical trial. Br Med J. 1992;305:235-40.

Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. Control Clin Trials. 1989;10:209S-21S.

Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. Stat Med. 1990;9:657-71.

Polis MA, Blackwelder WC. Trimethoprim-sulphamethoxazole or Pentamidine for Pneumocystis carinii Pneumonia. Ann Intern Med. 1987;106:475.

Poole C. Beyond the confidence interval. Am J Public Health. 1987;77:195-99.

Poole C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology. 2001;12:291-94.

Poole C, Greenland S. Random-effects meta-analyses are not always conservative. Am J Epidemiol. 1999;150:469-75.

Posavac EJ. Using p Values to Estimate the Probability of a Statistically Significant Replication. Understanding Statistics. 2002;1:101-12.

Pratt JW. Review of "Testing statistical hypotheses" 1959. E.L. Lehman. Journal of the American Statistical Association. 1961;56:163-67.

Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med. 1989;8:431-40.

Price KJ, Thall PF, Kish SK, Shannon VR, Andersson BS. Prognostic indicators for blood and marrow transplant patients admitted to an intensive care unit. American Journal of Respiratory & Critical Care Medicine. 1998;158:876-84.

Proschan MA. Statistical methods for monitoring clinical trials. J Biopharm Stat. 1999;9:599-615.

Putten WV. srd: Stata module for Survival Regression Diagnostics. http://home planet nl/~wimvanputten/stata/. 2002.

Quezado ZM, Natanson C, Alling DW, Banks SM, Koev CA, Elin RJ et al. A controlled trial of HA-1A in a canine model of gram-negative septic shock. JAMA. 1993;269:2221-27.

Quezado ZM, Natanson C, Hoffman WD. Looking back on HA-1A. Arch Intern Med. 1994;154:2393.

Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Control Clin Trials. 2000;21:330-342.

Rabe-Hesketh S, Pickles A, Taylor C. sg129: Generalized linear latent and mixed models. The Stata Technical Bulletin Reprints. 2000;9:293-306.

Raftery AE, Volinsky CT. bicreg: S-PLUS function to apply Bayesian Model Averaging to variable selection for linear regression models. @ http://www research att com/~volinsky/software/bicreg. 1996.

Rao CR. R.A. Fisher: The founder of modern statistics. Statistical Science. 1992;7:34-48.

Rascati KL, Smith MJ, Neilands T. Dealing with skewed data: an example using asthma-related costs of medicaid clients. Clinical Therapeutics. 2001;23:481-98.

Reboussin DM, DeMets DL, Kim KM, Lan KK. Computations for group sequential boundaries using the Lan-DeMets spending function method. Control Clin Trials. 2000;21:190-207.

Reinhart K, Menges T, Gardlund B, Harm ZJ, Smithes M, Vincent JL et al. Randomized, placebo-controlled trial of the anti-tumor necrosis factor antibody fragment afelimomab in hyperinflammatory response during severe sepsis: The RAMSES Study. Crit Care Med. 2001;29:765-69.

Reinhart K, Wiegand-Lohnert C, Grimminger F, Kaul M, Withington S, Treacher D et al. Assessment of the safety and efficacy of the monoclonal anti-tumor necrosis factor antibody-fragment, MAK 195F, in patients with sepsis and septic shock: a multicenter, randomized, placebo-controlled, dose-ranging study. Crit Care Med. 1996;24:733-42.

Rex JH, Bennett JE, Sugar AM, Pappas PG, van der Horst CM, Edwards JE et al. A randomized trial comparing fluconazole with amphotericin B for the treatment of candidemia in patients without neutropenia. Candidemia Study Group and the National Institute. N Engl J Med. 1994;331:1325-30.

Rex JH, Pappas PG, Karchmer AW, Sobel J, Edwards JE, Hadley S et al. A randomized and blinded multicenter trial of high-dose fluconazole plus placebo versus fluconazole plus amphotericin B as therapy for candidemia and its consequences in nonneutropenic subjects. Clin Infect Dis. 2003;36:1221-28.

Rex JH, Walsh TJ. Estimating the true cost of amphotericin B. Clin Infect Dis. 1999;29:1408-10.

Rex JH, Walsh TJ, Nettleman M, Anaissie EJ, Bennett JE, Bow EJ et al. Need for alternative trial designs and evaluation strategies for therapeutic studies of invasive mycoses. Clin Infect Dis. 2001;33:95-106.

Richard C, Warszawski J, Anguel N, Deye N, Combes A, Barnoud D et al. Early use of the pulmonary artery catheter and outcomes in patients with shock and acute respiratory distress syndrome: a randomized controlled trial. JAMA. 2003;290:2713-20.

Rindskopf DM. Testing "small", not null, hypotheses: Classical and Bayesian approaches. In: Harlow LL, Muliak SA, Steiger JH, eds. What if there were no significance tests? Hillsdale, NJ: Lawrence Erlbaum Associates; 1997: 319-22.

Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. N Engl J Med. 2001;345:1368-77.

RM87179. Supplemental Slides, Eli Lilly and Company. http://www fda gov/ohrms/dockets/ac/01/slides/3797s1_01_Lilly-backup/index htm. 2001.

Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology. 1992;3:319-36.

Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11:550-560.

Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. International Statistical Review. 1991;58:227-40.

Rocha LA, Martin MJ, Pita S, Paz J, Seco C, Margusino L et al. Prevention of nosocomial infection in critically ill patients by selective decontamination of the digestive tract. A randomized, double blind, placebo-controlled study. Intensive Care Med. 1992;18:398-404.

Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimantal groups. Psychol Bull. 1993;113:553-65.

Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by treatment. Journal of the Royal Statistical Society Series A. 1984;147:656-66.

Rothman KJ. Writing for epidemiology. Epidemiology. 1998b;9:333-37.

Rothman KJ. Writing for epidemiology. Epidemiology. 1998a;9:333-37.

Rothman KJ, Michels KB. The continuing unethical use of placebo controls. New England Journal of Medicine. 1994;331:394-98.

Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet. 1995;345:1616-19.

Royall RM. The effect of sample size on the meaning of statistical tests. The American Statistician. 1986;40:313-15.

Rubenfeld GD, Crawford SW. Withdrawing life support from mechanically ventilated recipients of bone marrow transplants: a case for evidence-based guidelines. Ann Intern Med. 1996;125:625-33.

Rubin DB. Multiple imputation after 18+ years. Journal of the American Statistical Association. 1996;91:473-89.

Rutten-van Molken MP, van Doorslaer EK, van Vliet RC. Statistical analysis of cost outcomes in a randomized controlled clinical trial. Health Econ. 1994;3:333-45.

Safar P, Kochanek PM. Lack of effect of induction of hypothermia after acute brain injury. N Engl J Med. 2001;345:66.

Safar PJ, Kochanek PM. Therapeutic hypothermia after cardiac arrest. N Engl J Med. 2002;346:612-13.

Salsburg D. Hypothesis testing. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics. New York: John Wiley & Sons, Inc; 1998: 1969-76.

Salsburg D. The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. New York: W.H. Freeman and Company; 2001.

Sanchez GM, Cambronero Galache JA, Lopez DJ, Cerda CE, Rubio BJ, Gomez Aguinaga MA et al. Effectiveness and cost of selective decontamination of the digestive tract in critically ill intubated patients. A randomized, double-blind, placebo-controlled, multicenter trial. American Journal of Respiratory & Critical Care Medicine. 1998;158:908-16.

Sankoh AJ, D'Agostino RB, Huque MF. Effficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. Stat Med. 2003;22:3133-50.

Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. Applied Statistics. 1999;48:313-29.

Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society A. 1999;162:71-94.

Savage LJ. On rereading RA Fisher. The Annals of Statistics. 1976;4:441-500.

Sawilowsky SS, Blair RC. A more realistic look at the robustness and type II error properties of the t test to departure from population normality. Psychol Bull. 1992a;111:352-60.

Sawilowsky SS, Hillman SB. Power of the independent samples t test under a prevalent psychometric measure distribution. Journal of Consulting & Clinical Psychology. 1992b;60:240-243.

Schafer JL. Multiple imputation of incomplete multivariate data under a normal model, version 2.03. Software for Windows 95/98/NT. NORM @ http://www stat psu edu/~jls/misoftwa html. 2001b.

Schafer JL. Multiple imputation of incomplete multivariate data under a normal model, version 2.03. Software for Windows 95/98/NT. NORM @ http://www stat psu edu/~jls/misoftwa html. 2001a.

Schafer JL. Multiple imputation in multivariate problems when the inputation and analysis models differ. Statistica Neerlandica. 2003;57:19-35.

Schafer JL, Olsen MK. Multiple imputation for multivariate mising-data problems: a data analyst's perspective. http://www stat psu edu/~jls/. 1998.

Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. The American Statistician. 2001;55:182-86.

Schervish MJ. P values: what they are and what they are not. The American Statistician. 1996;50:203-6.

Schierhout G, Roberts I. Fluid resuscitation with colloid or crystalloid solutions in critically ill patients: a systematic review of randomised trials. BMJ. 1998;316:961-64.

Schmid CH. Exploring heterogeneity in randomized trials via meta-analysis. Drug Inf J. 1999;33:211-24.

Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med. 1998;17:1923-42.

Schoenfeld DA. A simple algorithm for designing group sequential clinical trials. Biometrics. 2001;57:972-74.

Schulman KA, Glick HA, Rubin H, Eisenberg JM. Cost-effectiveness of HA-1A monoclonal antibody for gram-negative sepsis. Economic assessment of a new therapeutic agent. JAMA. 1991;266:3466-71.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical Evidence of Bias: Dimensions of Methodological Quality Associated With Estimates of Treatment Effects in Controlled Trials. JAMA Feb 1, 1995;273(5):408-412. 1995;273:408-12.

Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis. 1967;20:637-48.

Schwemer G. General linear models for multicenter clinical trials. Control Clin Trials. 2000;21:21-29.

Sculier JP, Paesmans M, Markiewicz E, Berghmans T. Scoring systems in cancer patients admitted for an acute complication in a medical intensive care unit. Crit Care Med. 2000;28:2786-92.

Seidenfeld T. Fiducial probability. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics. New York: John Wiley & Sons, Inc; 1998: 1510-1515.

Selective Decontamination of the Digestive Tract Trailists' Collaborative Group. Meta-analysis of randomised controlled trials of selective decontamination of the digestive tract. BMJ. 1993;307:525-32.

Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. The American Statistician. 2004;55:62-71.

Senn S. Statistical Issues in Drug Development. Statistical Issues in Drug Development. Chichester, West Sussex: John Wiley & Sons Ltd; 1997: 187-206.

Senn S. Covariate imbalance and random allocation in clinical trials. Stat Med. 1989a;8:467-75.

Senn S. Discussion of the role of P-values in analysing trial results by PR Freeman. Stat Med. 1993;12:1453-57.

Senn S. Testing for baseline balance in clinical trials. Stat Med. 1994a;13:1715-26.

Senn S. Base logic: tests of baseline balance in randomised clinical trials. Clinical Research and Regulatory Affairs. 1995;12:171-82.

Senn S. Some controversies in planning and analysing multi-centre trials. Stat Med. 1998;17:1753-65.

Senn S. Consensus and controversy in pharmaceutical statistics. The Statistician. 2000a;49:135-76.

Senn S. The many modes of meta. Drug Inf J. 2000b;34:535-49.

Senn S. Two cheers for P-values? Journal of Epidemiology & Biostatistics. 2001;6:193-204.

Senn S. A comment on replication, p-values and evidence, S.N.Goodman, Statistics in Medicine 1992; 11:875-879. Stat Med. 2002;21:2437-44.

Senn S, Harrell F. On wisdom after the event. J Clin Epidemiol. 1997;50:749-51.

Senn S, Harrell FE, Jr. On subgroups and groping for significance. Journal of Clinical Epidemiology. 1998a;51:1367-68.

Senn S, Walter S, Olkin I. Odds ratios revisited. Evidence Based Medicine. 1998b;3:71-72.

Senn SJ. The use of baselines in clinical trials of bronchodilators. Stat Med. 1989b;8:1339-50.

Senn SJ. Methods for assessing difference between groups in change when initial measurements is subject to intra-individual variation. Stat Med. 1994b;13:2280-2283.

Shao J. Bootstrap model selection. Journal of the American Statistical Society. 1996;91:655-65.

Sharp SJ. metareg, sbe23: Meta-analysis regression. Stata Technical Bulletin Reprints. 1998;7:148-55.

Sharp SJ. Analysing the relationship between treatment benefit and underlying risk: precautions and recommendations. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care: Meta-analysis in context. 2nd ed. London: BMJ Publishing Group; 2001: 176-88.

Sharp SJ, Thompson SG. Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. Stat Med. 2000;19:3251-74.

Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. BMJ. 1996;313:735-38.

Sheiner LB, Beal Sl. Some suggestions for measuring predictive performance. Journal of Pharmacokinetics and Biopharmaceutics. 1981;8:503-12.

Shih JH. Sample size calculation for complex clinical trials with survival endpoints. Control Clin Trials. 1995;16:395-407.

Shiozaki T, Hayakata T, Taneda M, Nakajima Y, Hashiguchi N, Fujimi S et al. A multicenter prospective randomized controlled trial of the efficacy of mild hypothermia for severely head injured patients with low intracranial pressure. Mild Hypothermia Study Group in Japan. J Neurosurg. 2001;94:50-54.

Shlaes DM, Moellering RC, Jr. The United States Food and Drug Administration and the end of antibiotics. Clinical Infectious Diseases. 2002;34:420-422.

Shun Z, Yuan W, Brady WE, Hsu H. Type I error in sample size re-estimations based on observed treatment difference. Stat Med. 2001;20:497-513.

Siddiqui O, Ali MW. A comparison of the random-effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. J Biopharm Stat. 1998;8:545-63.

Siegel JP. Assessing the use of activated protein C in the treatment of severe sepsis. N Engl J Med. 2002a;347:1030-1034.

Siegel JP. Assessing the use of activated protein C in the treatment of severe sepsis. N Engl J Med. 2002b;347:1030-1034.

Siegel JP. Assessing the use of activated protein C in the treatment of severe sepsis. N Engl J Med. 2002c;347:1030-1034.

Silvestri L, Mannucci F, van Saene HK. Selective decontamination of the digestive tract: a life saver. J Hosp Infect. 2000;45:185-90.

Simon R. Patient heterogeneity in clinical trials. Cancer Treat Rep. 1980;64:405-10.

Simon R. Why confidence intervals are useful tools in clinical therapeutics. J Biopharm Stat. 1993;3:243-48.

Simpsom F, Doig GS. Parenteral vs. enteral nutrition in the critically ill patient: a meta-analysis of trials using the intention to treat principle. Intensive Care Med. 2005;31:12-23.

Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. J Clin Epidemiol. 1994;47:881-89.

Skovlund E, Fenstad GU. Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? J Clin Epidemiol. 2001;54:86-92.

Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses--sometimes informative, usually misleading. BMJ. 1999;318:1548-51.

Smith SM. Clarifying the evidence. Br Med J. 2001;Rapid response: 28 January.

Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med. 1995;14:2685-99.

Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technology Assessment (Rockville, Md). 2000;4:1-115.

Speigelhalter D, Thomas A, Best N. WINBUGS. WINBUGS @ http://www mrc-bsu cam ac uk/bugs. 2000.

Speigelhalter DJ, Abrams K, Myles JP. Multiplicity, exchangeability and hierarchical models. In: Speigelhalter DJ, Abrams K, Myles JP, eds. Bayesian approaches to clinical trials and health care evaluation. Chichester, West Sussex: John Wiley &Sons Ltd; 2004: 98.

Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester: John Wiley & Sons, Ltd; 2004.

Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? Control Clin Trials. 1986;7:8-17.

Spielman S. A refutation of the Neyman-Pearson theory of testing. British Journal of Philospohy of Science. 1973;24:201-22.

Spielman S. The logic of tests of significance. Philosophy of Science. 1974;41:211-26.

Spiessens B, Lesaffre E, Verbeke G, Kim K, DeMets DL. An overview of group sequential methods in longitudinal clinical trials. Stat Methods Med Res. 2000;9:497-515.

Spriet A, Beiler D. When can 'non significantly different' treatments be considered as 'equivalent'? British Journal of Clinical Pharmacology. 1979;7:623-24.

Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. Prev Med. 1991;20:47-63.

Stat Xact 4. StatXact 4 for Windows: Statistical software for exact nonparametric inference. StatXact 4 @ http//www cytel com. 1999.

Stata Statistical Software V8. Stata Statistical Software, Version 8. Stata Statistical Software @ http://stata com. 2003.

Steichen T. Tests for publication bias in meta-analysis. Stata Technical Bulletin Reprints. 1998;7:125-33.

Steichen T. Nonparametric "trim and fill" analysis of publication bias in meta-analysis. Stata Technical Bulletin. 2000;61:8-14.

Stern JM, Egger M, Smith GD. Investigating and dealing with publication and other biases. In: Egger M, Smith GD, Altman DG, eds. Systematic Reviews in Health Care: Meta-analysis in context. 2nd ed. London: BMJ Publishing Group; 2001: 189-208.

Sterne J. sbe22. Cumulative meta-analysis. Stata Technical Bulletin Reprints. 1998;143-47.

Sterne JA, Davey SG. Sifting the evidence-what's wrong with significance tests? BMJ. 2001a;322:226-31.

Sterne JA, Egger M. High false positive rate for trim and fill method. BMJ. 2000a;http://bmj.com/cgi/eletters/320/7249/1574.

Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol. 2001b;54:1046-55.

Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000b;53:1119-29.

Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med. 2002;21:1513-24.

Stewart LA, Parmar MKB. Bias in the analysis and reporting of randomized controlled trials. Int J Technol Assess Health Care. 1996;12:264-75.

Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19:1059-79.

Stonehouse JM, Forrester GJ. Robustness of the t and U test under combined assumption violations. Applied Statistics. 1998;25:63-73.

Stoutenbeek CP, van Saene HK, Miranda DR, Zandstra DF. The effect of selective decontamination of the digestive tract on colonisation and infection rate in multiple trauma patients. Intensive Care Med. 1984;10:185-92.

Stoutenbeek CP, van Saene HK, Zandstra DF. Prevention of multiple organ system failure by selective decontamination of the digestive tract in multiple trauma patients. In: Faist E, Baue AE, Schildberg FW, eds. The immune consequences of trauma, shock and sepsis - mechanisms and therapeutic approaches. Lengerich: Pabst Science Publishers; 1996: 1055-66.

Streiner DL. Do you see what I mean? Indices of central tendency. Canadian Journal of Psychiatry - Revue Canadienne de Psychiatrie. 2000;45:833-36.

Sullivan LM, D'Agostino RB. Robustness of the t test applied to data distorted from normality by floor effects. Journal of Dental Research. 1992;71:1938-43.

Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996a;49:907-16.

Sun X, Wagner DP, Knaus WA. Does selective decontamination of the digestive tract reduce mortality for severely ill patients? Crit Care Med. 1996b;24:753-55.

Sutton AJ. Re: High false positive rate for trim and fill method. BMJ. 2000;http://bmj.com/cgi/eletters/320/7249/1574.

Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. BMJ. 2000;320:1574-77.

Sutton CD. Computer-intensive methods for tests about the mean of an asymmetrical distribution. Journal of the American Statistical Association. 1993;88:802-10.

Sznajder M, Leleu G, Buonamico G, Auvert B, Aegerter P, Merliere Y et al. Estimation of direct cost and resource allocation in intensive care: correlation with Omega system. Intensive Care Med. 1998;24:582-89.

Takkouche B, Cadarso-Suarez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. Am J Epidemiol. 1999;150:206-15.

Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. J Clin Epidemiol. 2000;53:477-84.

Temple R. Problems in interpreting active control equivalence trials. Accountability in Research. 1996;4:267-75.

Temple RM, Ellenberg SSP. Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments: Part 1: Ethical and Scientific Issues. Ann Intern Med. 2000;133:455-63.

The ARDS Network Authors for the ARDS Network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute

respiratory distress syndrome. The Acute Respiratory Distress Syndrome Network. N Engl J Med. 2000;342:1301-18.

The Coronary Drug Project Research Group. Practical aspects of decision making in clinical trials: the coronary drug project as a case study. The Coronary Drug Project Research Group. Contol Clin Trials. 1981;1:363-76.

The Hypothermia after Cardiac Arrest Study Group. Mild therapeutic hypothermia to improve the neurologic outcome after cardiac arrest. N Engl J Med. 2002;346:549-56.

The NHLBI ARDS Clinical Trials Network. Higher versus Lower Positive End-Expiratory Pressures in Patients with the Acute Respiratory Distress Syndrome. N Engl J Med. 2004;351:327-36.

The SAFE Study Investigators. A Comparison of Albumin and Saline for Fluid Resuscitation in the Intensive Care Unit. N Engl J Med. 2004;350:2247-56.

The Veterans Affairs Total Parenteral Nutrition Cooperative Study Group. Perioperative total parenteral nutrition in surgical patients. N Engl J Med. 1991;325:525-32.

The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. Control Clin Trials. 1998;19:61-109.

Thompson SG, Pocock SJ. Can meta-analyses be trusted? Lancet. 1991;338:1127-30.

Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med. 1999;18:2693-708.

Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med. 1997;16:2741-58.

Thompson WD. Statistical criteria in the interpretation of epidemiologic data. Am J Public Health. 1987;77:191-94.

Tibshirani R. A plain man's guide to the proportional hazards model. Clinical & Investigative Medicine. 1982;5:63-68.

Tomz M, King G, Zeng L. **ReLogit: Rare Events Logistic Regression**. http://gking harvard edu/stats shtml. 1999.

Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Impact of missing data due to selective dropouts in cohort studies and clinical trials. Epidemiology. 2002;13:347-55.

Tramer MR, Reynolds DJ, Moore RA, McQuay HJ. When placebo controlled trials are essential and equivalence trials are inadequate. BMJ. 1998;317:875-80.

Tremblay LN, Hyland RH, Schouten BD, Hanly PJ. Survival of acute myelogenous leukemia patients requiring intubation/ventilatory support. Clinical Investigative Medicine. 1995;18:19-24.

Tukey JW. Use of many covariates in clinical trials. International Statistical Review. 1991;59:123-37.

Unertl K, Ruckdeschel G, Selbmann HK, Jensen U, Forst H, Lenhart FP et al. Prevention of colonization and respiratory infections in long-term ventilated patients by local antimicrobial prophylaxis. Intensive Care Med. 1987;13:106-13.

United States Federal Drug and Food Administration. FDA briefing document: Anti-infective Advisory Meeting. http://www fda gov/ohrms/dockets/ac/01/briefing/3797b1_02_FDAbriefing pdf. 2001a.

United States Federal Drug and Food Administration. Transcript of Anti-Infective Drugs Advisory Committee Meeting; October 16 2001. http://www fda gov/ohrms/dockets/ac/01/transcripts/3797t1_01 pdf. 2001b.

van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M et al. Intensive insulin therapy in the critically ill patients. N Engl J Med. 2001;345:1359-67.

van Houwelingen H, Senn S. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med. 1999;18:110-115.

van Nieuwenhoven CA, Buskens E, van Tiel FH, Bonten MJ. Relationship between methodological trial quality and the effects of selective digestive decontamination on pneumonia and mortality in critically ill patients. JAMA. 2001;286:335-40.

van Saene HK, Petros AJ, Ramsay G, Baxby D. All great truths are iconoclastic: selective decontamination of the digestive tract moves from heresy to level 1 truth. Intensive Care Med. 2003;29:677-90.

Vaupel JW, Yashin AI. Heterogeneity's Ruses: Some surprising effects of selection on population dynamics. The American Statistician. 1985;39:176-85.

Verwaest C, Verhaegen J, Ferdinande P, Schetz M, van den BG, Verbist L et al. Randomized, controlled trial of selective digestive decontamination in 600 mechanically ventilated patients in a multidisciplinary intensive care unit. Crit Care Med. 1997;25:63-71.

Villar J, Perez-Mendez L, Aguirre-Jaime A, Kacmarek RM. Whay are physicians so skeptical about positive randomized controlled clinical trials in ctitical care medicine. Intensive Care Med. 2005;31:196-204.

Vincent JL. Selective digestive decontamination: for everyone, everywhere? Lancet. 2003;362:1006-7.

Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol. 1986;123:174-84.

Wagenmakers AJ. Muscle function in critically ill patients. Clin Nutr. 2001;20:451-54.

Wald A. Sequential analysis. New York: John Wiley & Sons; 1947.

Walsh TJ, Finberg RW, Arndt C, Hiemenz J, Schwartz C, Bodensteiner D et al. Liposomal amphotericin B for empirical therapy in patients with persistent fever and neutropenia. National Institute of Allergy and Infectious Diseases Mycoses Study Group. N Engl J Med. 1999;340:764-71.

Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. Stat Med. 1997;16:2883-900.

Walter SD. Choice of effect measure for epidemiological data. J Clin Epidemiol. 2000;53:931-39.

Walter SD. Number needed to treat (NNT): estimation of a measure of clinical benefit. Stat Med. 2001;20:3947-62.

Walton RJ, Bijvoet OL. Nomogram for derivation of renal threshold phosphate concentration. Lancet. 1975;2:309-10.

Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics. 1987;43:193-99.

Ward JD, Becker DP, Miller JD, Choi SC, Marmarou A, Wood C et al. Failure of prophylactic barbiturate coma in the treatment of severe head injury. Journal of Neurosurgery. 1985;62:383-88.

Ware JH, Antman EM. Equivalence trials. N Engl J Med. 1997;337:1159-61.

Ware JH, Muller JE, Braunwald E. The futility index. An approach to the cost-effective termination of randomized clinical trials. Am J Med. 1985;78:635-43.

Warn DE, Thompson SG, Spiegelhalter DJ. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. Stat Med. 2002;21:1601-23.

Warren HS, Danner RL, Munford RS. Anti-endotoxin monoclonal antibodies. N Engl J Med. 1992;326:1153-57.

Warren HS, Suffredini AF, Eichacker PQ, Munford RS. Risks and benefits of activated protein C treatment for severe sepsis. N Engl J Med. 2002b;347:1027-30.

Warren HS, Suffredini AF, Eichacker PQ, Munford RS. Risks and benefits of activated protein C treatment for severe sepsis. N Engl J Med. 2002e;347:1027-30.

Warren HS, Suffredini AF, Eichacker PQ, Munford RS. Risks and benefits of activated protein C treatment for severe sepsis. N Engl J Med. 2002c;347:1027-30.

Warren HS, Suffredini AF, Eichacker PQ, Munford RS. Risks and benefits of activated protein C treatment for severe sepsis. N Engl J Med. 2002d;347:1027-30.

Warren HS, Suffredini AF, Eichacker PQ, Munford RS. Risks and benefits of activated protein C treatment for severe sepsis. N Engl J Med. 2002a;347:1027-30.

Webb CH. Selective decontamination of the digestive tract, SDD: a commentary. J Hosp Infect. 2000;46:106-9.

Weesie J. Score test for omitted variables: Stata module. http://www fss uu nl/soc/iscore/stata/. 1999.

Weiss GB, Bunce H, III, Hokanson JA. Comparing survival of responders and nonresponders after treatment: a potential source of confusion in interpreting cancer clinical trials. Control Clin Trials. 1983;4:43-52.

Weissman C. Analyzing intensive care unit length of stay data: problems and possible solutions. Crit Care Med. 1997;25:1594-600.

Welch BL. The significance of the differnece between two means when the population variances are unequal. Biometrika. 1937;29:350-362.

Wenzel RP. Anti-endotoxin monoclonal antibodies--a second look. N Engl J Med. 1992;326:1151-53.

Wenzel RP. Treating sepsis. N Engl J Med. 2002;347:966-67.

Westlake WJ. Symmetrical confidence intervals for bioequivalence trials. Biometrics. 1976;32:741-44.

Wharton JM, Coleman DL, Wofsy CB, Luce JM, Blumenfeld W, Hadley WK et al. Trimethoprim-sulfamethoxazole or pentamidine for Pneumocystis carinii pneumonia in

the acquired immunodeficiency syndrome. A prospective randomized trial. Annals of Internal Medicine. 1986;105:37-44.

Wheatley K, Clayton D. Be skeptical about large apparent treatment effects: the case of an MRC AML12 randomization. Control Clin Trials. 2003;24:66-70.

White HD. Thrombolytic therapy and equivalence trials. Journal of the American College of Cardiology. 1998;31:494-96.

White MH, Bowden RA, Sandler ES, Graham ML, Noskin GA, Wingard JR et al. Randomized, double-blind clinical trial of amphotericin B colloidal dispersion vs. amphotericin B in the empirical treatment of fever and neutropenia. Clin Infect Dis. 1998;27:296-302.

Whitehead J. The design and analysis of sequential clinical trials. 2nd ed. Chichester: Ellis Horwood; 1992.

Whitmore GA. The inverse Gaussian distribution as a model of hospital stay. Health Serv Res. 1975;10:297-302.

Wiens BL. Something for nothing in noninferiority/superiority testing: A caution. Drug Inf J. 2001;35:241-45.

Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. Control Clin Trials. 2002;23:2-14.

Wiens BL, Iglewicz B. Testing noninferiority of response rates for regulatory filings using transformations. Drug Inf J. 2001;35:1165-71.

Wilcox RR, Keselman HJ, Kowalchuk RR. Can treatment group equality be improved?: The bootstrap and trimmed means conjecture. Br J Math Stat Psychol. 1998;51:123-34.

Wilkes MM, Navickis RJ. Patient survival after human albumin administration. A meta-analysis of randomized, controlled trials. Ann Intern Med. 2001c;135:149-64.

Wilkes MM, Navickis RJ. Patient survival after human albumin administration. A meta-analysis of randomized, controlled trials. Ann Intern Med. 2001b;135:149-64.

Wilkes MM, Navickis RJ. Patient survival after human albumin administration. A meta-analysis of randomized, controlled trials. Ann Intern Med. 2001a;135:149-64.

Williams RL, Chen M-L, Hauck WW. Equivalence approaches. Clinical Pharmacology & Therapeutics. 2002;72:229-37.

Windeler J, Trampisch H-J. Recommendations concerning studies on therapeutic equivalence. Drug Inf J. 1996;30:195-200.

Wingard JR, Kubilis P, Lee L, Yee G, White M, Walshe L et al. Clinical significance of nephrotoxicity in patients treated with amphotericin B for suspected or proven aspergillosis. Clin Infect Dis. 1999;29:1402-7.

Winston DJ, Schiller GJ, Territo MC. Liposomal amphotericin B for fever and neutropenia. N Engl J Med. 1999;341:1154-55.

Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. Stat Med. 1990;9:65-71.

Yanagawa T, Bunn F, Roberts I, Wentz R, Pierro A. Nutritional support for head-injured patients. Cochrane Database Syst Rev. 2000;CD001530.

Yashin AI, Iachine I. How long can humans live? Lower bound for biological limit of human longevity calculated from Danish twin data using correlated frailty model. Mechanisms of Ageing & Development. 1995;80:147-69.

Yates F. The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. Journal of the American Statistical Association. 1951;46:19-34.

Young KD, Lewis RJ. What is confidence? Part 2: Detailed definition and determination of confidence intervals. Ann Emerg Med. 1997;30:311-18.

Yu M, Burchell S, Hasaniya NW, Takanishi DM, Myers SA, Takiguchi SA. Relationship of mortality to increasing oxygen delivery in patients > or = 50 years of age: a prospective, randomized trial. Crit Care Med. 1998;26:1011-19.

Yusuf S. Meta-analysis of randomized trials: looking back and looking ahead. Control Clin Trials. 1997;18:594-601.

Zachos M, Tondeur M, Griffiths AM. Enteral nutritional therapy for inducing remission of Crohn's disease. Cochrane Database Syst Rev. 2001;CD000542.

Zaloga GP. Immune-enhancing enteral diets: where's the beef? Crit Care Med. 1998;26:1143-46.

Zaloga GP. Early enteral nutritional support improves outcome: hypothesis or fact? Crit Care Med. 1999;27:259-61.

Zarich S, Fang LS, Diamond JR. Fractional excretion of sodium. Exceptions to its diagnostic value. Arch Intern Med. 1985;145:108-12.

Zeni F, Freeman B, Natanson C. Anti-inflammatory therapies to treat sepsis and septic shock: a reassessment. Crit Care Med. 1997;25:1095-100.

Zhang J, Yu K. What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. Journal of the American Medical Association. 1998;290:1690-1691.

Zheng B. Summarizing the goodness of fit of generalized linear models for longitudinal data. Stat Med. 2000;19:1265-75.

Zheng B, Agresti A. Summarizing the predictive power of a generalized linear model. Stat Med. 2000;19:1771-81.

Zhou M. Understanding the Cox regression model with time-change covariates. The American Statistician. 2001;55:153-55.

Zhou XH, Brizendine EJ, Pritz MB. Methods for combining rates from several studies. Stat Med. 1999;18:557-66.

Zhou XH, Gao S, Hui SL. Methods for comparing the means of two independent log-normal samples. Biometrics. 1997;53:1129-35.

Zhou XH, Li C, Gao S, Tierney WM. Methods for testing equality of means of health care costs in a paired design study. Stat Med. 2001a;20:1703-20.

Zhou XH, Stroupe KT, Tierney WM. Regression analysis of health care charges with heteroscedasticity. Applied Statistics. 2001b;50:303-12.

Ziegler EJ, Fisher CJ, Jr., Sprung CL, Straube RC, Sadoff JC, Foulke GE et al. Treatment of gram-negative bacteremia and septic shock with HA-1A human monoclonal antibody against endotoxin. A randomized, double-blind, placebo-controlled trial. The HA-1A Sepsis Study Group. N Engl J Med. 1991;324:429-36.

Ziegler EJ, Smith CR. Anti-endotoxin monoclonal antibodies. N Engl J Med. 1992a;326:1165.

Ziegler EJ, Smith CR. Anti-endotoxin monoclonal antibodies. N Engl J Med. 1992b;326:1165.

Zimmerman DW. Invalidation of parametric and nonparametric statistical test by concurrent violation of two assumptions. The Journal of Experimental Education. 1998;67:55-68.

Zimmerman DW, Zumbo BD. Parametric alternatives to the Student t-test under violation of normality and homogeneity of variance. Perceptual & Motor Skills. 1992;74:835-44.

Zumbo BD, Hubley AN. A note on misconceptions concerning prospective and retrospective power. The Statistician. 1998;47:385-88.

Moran, John L., Bersten, Andrew D. & Solomon, Patricia, J. (2005) Meta-analysis of controlled trials of ventilator therapy in acute lung injury and acute respiratory distress syndrome: an alternative perspective.
Intensive Care Medicine v.31 (2) pp. 227-235

Moran, John L., Peisach, A.R., Solomon, Patricia, J. & Martin, J. (2004) Cost calculation and prediction in adult intensive care: a ground-up utilisation study. *Anaesthesia & Intensive Care v.32 (6) pp. 787-797*

Moran, John L., Solomon, Patricia, J., Fox, V., Slagaras, M. Williams, P.J. & Quinlan, K., Bersten, A.D. (2004) Modelling thirty-day mortality in the Acute Respiratory Distress Syndrome (ARDS) in an adult ICU.
Anaesthesia & Intensive Care v.32 (3) pp. 317-329

Moran, John L.& Solomon, Patricia, J. (2003) Mortality and other event rates: what to they tell us about performance?
Critical Care & Resuscitation v. 5 (4) pp. 292-303