"LIES, DAMN LIES AND STATISTICS" (1): CRITICALLY APPRAISING CLINICAL TRIALS

Science by slight of hand...
Forging, fudging and finagling...
Sadistic statistics...
P-values with pee-wee value...
How to lie with statistics...
Dicing with death rates...
To p or not to p...

Helen Wood, M.D.

University of Texas Southwestern Medical Center at Dallas

Medical Grand Rounds

July 10, 1997

Biographical Information

Name: Helen Wood, M.D.
Rank: Assistant Professor of Internal Medicine
Division: General Internal Medicine
Training:

1) Residency and chief residency in Internal Medicine, Medical College of Wisconsin
2) Fellowship in General Internal Medicine, Medical College of Wisconsin

Interests:

1) Evidence-based medicine
2) Medical decision-making, including quality of life assessment with emphasis on utility assessment.
3) Women's Health
4) Medical education in primary care

INTRODUCTION

In 1993, in the September 2 issue of the New England Medical Journal, the results of the Global Utilization of Streptokinase and Tissue Plasminogen Activator in Occluded Arteries (GUSTO) trial were reported (2). This international collaborative trial of 41 021 patients, sought to definitively answer whether accelerated tissue plasminogen activator (t-PA) or steptokinase (SK) was the most efficacious thrombolytic therapy for treatment of patients presenting early with myocardial infarction associated with ST segment elevation. The authors concluded that t-PA was the preferred therapy due to its greater reduction in 30 day mortality compared with the combined SK groups. What are clinicians to do with this information and why is this important? Earlier reports by the Gruppo Italiano per lo Studio della Sopravvivenza nell" Infarcto Miocardico (GISSI - 2), the International Study Group, and the Third International Study of Infarct Survival (ISIS - 3) found no evidence of a survival advantage for any of the compared thrombolytic agents, which included SK and t-PA (3,4,5). Yet in the face of conflicting data, clinicians still need to make treatment decisions regarding which is the most efficacious thrombolytic treatment for the estimated 250 000 eligible patients who present with acute myocardial infarction in the United States per year (6). My purpose in this discussion is not to provide you with a current update on the most efficacious thrombolytic therapy for acute myocardial infarction, because there are much more capable people than myself for that task. Instead, my purpose is to use this trial, and the storm of controversy regarding its interpretation and application to clinical medicine that followed in its wake, to provide a meaningful context within which to discuss the appraisal of clinical trials.

Why do we need to talk about critically appraising clinical trials?

1) Clinical trials represent the pinnacle of the hierarchy of evidence when answering a clinical question (7, 8, 9).

Randomized trials have become the accepted standard for evaluating therapeutic efficacy. As a result, they are likely to have the greatest impact on our clinical decision-making, as they should. But they are not perfect, and given their potential ready application to medicine, some careful scrutiny should be brought to bear on them. The GUSTO trial has been considered one of the largest and best executed clinical trials ever performed and its results have had substantial impact on the choice of agent for thrombolytic therapy in the setting of acute myocardial infarction. Yet many questions have been raised regarding the proper interpretation of this trial and its subsequent integration into clinical practice.

2) Clinical trials have become increasingly more sophisticated in design and analysis.

This complexity is well illustrated by a review of 45 consecutive reports, 15 each from 3 medical journals (British Medical Journal, Lancet, New England Journal of Medicine) reported by Pocock and colleagues (10). They found multiplicity in many aspects of study design and analysis, including endpoints (a median of 6 assessed per study), measurements of outcome (repeated over time in 40%), treatment groups (Investigaton of 3 or more treatment groups in 20%), and statistical tests (the mean number of significance tests evaluating outcomes only was 4 per trial but increased to a median of 8 when those examing subgroups were included). One trial had 12 endpoints, 11 significance tests and only 12 patients. With this degree of complexity, it becomes easy for the results to obscured by the statistics. Where is the knowlegde that is lost in information? (T.S. Eliot)

3) There has been a tendency for new treatments and new technologies to be adopted before fully evaluated.

Laupacis and colleagues have developed a classification system that provides guidance on the use of clinical and economic evaluations in making decisions about the adoption and utilization of competing health care technologies (11). They applied this classification system to some published studies and found

that some approaches to health care that have been adopted prior to full evaluation appear, after their widespread dissemination, to be poorly supported by evidence. In addition, they demonstrated that an expensive, poorly effective therapy was harder to withdraw than introducing an equally expensive but more effective therapy. For example, the introduction of universal precautions against HIV transmission in health care workers costs about $565 000 per additional life-year saved (11).

4) Clinical trial results are often in conflict.

Conflicting results of multiple randomized clinical trials on the same topic is not an uncommon phenonmenom. Horowitz searched the literature in the disciplines of cardiology and gastroenterology, and found 36 topics, encompassing over 200 randomized trials, for which there were conflicting results in the reported relationship between a particular therapeutic agent and a clinical outcome (12). Some common and clinically important topics included the medical versus surgical treatment of stable coronary artery disease, beta-blockers and acute myocardial infarction, and steroids and alcoholic liver disease. What do we do when the conclusions from our highest forms of evidence don't agree with one another? Yet, such was the case with the GUSTO trial, which sparked the debate that subsequently ensued.

5) Finally, the quality of a clinical trial cannot be inferred from the reputation of the medical journal in which it is published.

There are a number of reports regarding the suboptimal methodological quality of clinical trials published in peer-reviewed journals (10). Gore reported the impact of implementing statistical review of submitted papers, which remain candidates for publication in Lancet after conventional review (13). Biostatisticians evaluated various methodological aspects of 191 reports in a standardized manner and judged the methods section to be inadequate in half. The most common major criticisms are presented in Table 1. After methodological review, the following recommendations were made: accept 8% (none had major adverse comments), accept after revision 46% (26% had major adverse comments), revise and re-review 32% (90% had major adverse comments) and reject 14% (100% had major adverse comments). Even with this process in place only 9 of 27 papers recommended for rejection were actually rejected.

Table 1: Summary of Methodological Review of Papers Passing Conventional Review

Methodological Area	% Inadequate	Specific Methodological Problems Cited
Abstract	24%	
Design	28%	- sampling scheme/eligibility of patients (10%) - power of study or sample size (9%) - comparability between groups (5%)
Analysis	38%	- need for better analysis (19%) - type of analysis not clear (13%) - inappropriate distribution shape (6%)
Inference	25%	- incorrect conclusions (13%) - multiple endpoints with selective emphasis (7%)
Presentation	8%	- failure to show raw data (3%) - failure to indicate scatter in tables or graphs (3%)

Just recently, the Consolidated Standards of Reporting Trials (CONSORT) guidelines were published (14). The goal of these guidelines is to increase the quality of carrying out and reporting clinical trials. Some journals have already adopted these guidelines and ask authors who submit papers to adhere to these recommendations. However, there is no mechanism for ensuring compliance with them in the review process and they do not constitute formal statistical review (15). Only a few journals, the Lancet, British

Medical Journal, the Journal of the American Medical Association and the Annals of Internal Medicine, are notable exceptions in employing methodological review as a component of the review process. Unfortunately, then methodologic issues that could impact the validity of the results and the inferences drawn may not be recognized by content experts, and thus adversely affect evolution of the knowledge base.

Having provided this rationale for why critical appraisal is necessary, I'd like to place my discussion in the appropriate perspective. Critically appraising a study puts one in the heady position of being an armchair quarterback. It is much easier after the fact to see what might have gone wrong with a trial or what might be less than ideal design. However, anyone who has ever been on the playing field, so to speak, in designing and carrying out a clinical trial appreciates the enormous complexity of the task. Issues such as feasibility with regard to study population, measurements, outcomes, cost, effort, time, etc. substantially influence the final study design that is implemented, which thus represents no small accomplishment. The goal, therefore, of critical appraisal is not statistical nihilism with regard to the quality of a trial. No trial is perfect, and no trial answers definitively all the clinical questions that might be asked of it. A less than perfect trial is not to be summarily abandoned. The goal of appraisal is to recognize the limitations of a study so that decision-making can be guided as objectively as possible, based on the study's validity and inferences that are supported by the data in the study. With careful analysis, we might better delineate where science ends and the art of medicine begins, a process which David Sackett refers to as "the science of the art of choosing better treatment." (16)

The paradigm that I'll be using to discuss the critical appraisal of clinical trials is that of evidence-based medicine (EBM). EBM represents a paradigm shift in the practice of medicine in its emphasis on using the literature more effectively in guiding medical practice (17). This paradigm does not replace the former paradigm which has emphasized clinical experience, the understanding of basic mechanisms of disease and pathphysioloigc principles, physical exam skills and content expertise. Instead, it expands and complements it with new skills such as precisely defining the patient's problem, efficient strategies for searching the literature, application of formal rules of evidence in the interpretation of the literature, appropriately applying the information to guide medical practice and increased emphasis on the psychosocial aspects of medicine. These formal rules of evidence for various types of studies have been developed and published by the Evidence-Based Medicine Working Group at McMaster University as a framework for critical appraisal of the literature. The full set of criteria have been published as a series of articles, collectively called "The Users' Guide to Interpretation of the Literature", over the last several years in JAMA (18-31). Today I'll be focusing on clinical trials, because they represent the highest form of evidence, and on those that investigate therapy and prevention since these questions are the most common ones asked in the clinical arena, although clinical trials can also be used to investigate diagnostic tests, prognosis, and other types of questions. Table 2 provides a summary of the criteria for evaluating studies assessing therapy and prevention which will provide a methodologic framework for our discussion. The analysis of clinical trials encompasses a broad scope of statistics, epidemiology and medical decision-making. I have therefore chosen to focus my discussion on some issues that are most relevant to each criteria or most frequently encountered by clinicians in their analysis of the literature. These issues are presented in italics in association with the appropriate criteria. Before moving on to discuss these criteria and apply them to the GUSTO trial, as a brief aside, more detailed and quantitative methods for assessing the quality of a randomized control trial have also been developed (32).

Table 2: Readers' Guides for an Article About Therapy or Prevention

Are the results of the study valid? *Primary guides:*	*Methodological issues*
Was the assignment of patients to treatments randomized?	• *Selection bias, what constitutes randomization, statistical advantages of randomization, assessing the "success" of randomization*
Were all patients who entered the trial properly accounted for and attributed at its conclusion? • *Was follow-up complete?* • *Were patients analyzed in the groups to which they were randomized?*	• *impact of loss to follow-up, cross-over in treatment, noncompliance,*
Secondary guides: Were patients, health workers and study personnel "blind" to treatment?	• *Diagnostic and outcome assessment bias*
Were the groups similar at the start of the trial?	• *Assessing the distribution of baseline covariates*
Aside from the experimental intervention, were the groups treated equally?	• the *impact of co-interventions*
What were the results? How large was the treatment effect?	• *Summary measures of benefit, significance/hypothesis testing, statistical parameters (α, β, p values), statistical significance vs clinical significance, modifying treatment effect for impact of covariatest, multiple comparisons*
How precise was the estimate of the treatment effect?	• *Confidence intervals, power and "negative studies"*
Will the results help me in caring for my patients? Can the results be applied to my patient care?	• *Subgroup analysis*
Were all the relevant outcomes studied?	• *clinical marker vs outcome of clinical interest, multidimensional outcomes, multiple outcomes, quality of life*
Are the likely treatment benefits worth the potential harms and costs?	• *Comorbidity of therapy, cost-effectivness/cost-utility analyses*

Overview of GUSTO Design

Before proceeding with an appraisal of the GUSTO trial, let me briefly review its design. 41 021 patient who presented to 1081 hospitals in 15 different countries were randomized to one of four treatment strategies: accelerated t-PA + IV heparin, SK + SC heparin, SK + IV heparin, t-PA (Table 3 presents the

details of dosing and administration of the main treatments). The inclusion criteria included chest pain due to suspected MI of ≤ 6 hours duration with ECG evidence of ST-segment elevation over the presumed area of infarct. Exclusion criteria included prior stroke or evidence of central nervous system damage, earlier entry into the trial, recent use of SK or anistreplase, active bleeding, and recent noncompressible puncture. All patients received at least 160 mg of aspirin on day 1, and 160 to 325 mg/d thereafter. Intravenous followed by oral beta-blockers using atenolol was also given in the absence of contraindications. The primary outcome for the trial was 30-day mortality. More details of the trial will be presented as they become relevant.

Table 3: Summary of the Treatment Groups in GUSTO

Treatment	Dose/administration of t-PA or SK	Heparin
t-PA	15 mg/kg bolus, the 0.75 mg/kg over 30 min and 0.5 mg/kg over next 60 min.	5000 U IV bolus, the IV infusion of 1000 U/hr (APTT 60-85)
SK	1 million U over 60 min	5000 U IV bolus, the IV infusion of 1000 U/hr (APTT 60-85)
SK	1.5 million U over 60 min	12 000 U SC BID
t-PA + SK	• t-PA: 1 mg/kg over 60 min (10% as bolus) • SK: I million U over 60 min	5000 U IV bolus, the IV infusion of 1000 U/hr (APTT 60-85)

I. ARE THE RESULTS OF THE STUDY VALID?

A. Primary Guides

1. Was the assignment of patients to treatments randomized?

I'm going to address two issues with regard to randomization: 1) what does randomization achieve and 2) how is the "success" of randomization assessed and what importance does this have? First of all, we need to understand what randomization is, and what it is not. Randomization is a procedure for assigning treatment to patients such that there is an equal probability of assignment to each treatment each time a patient is allocated (8). A good example of a randomization procedure is the use of a random table of numbers generated by a computer with odd numbers being assigned to one treatment and even numbers being assigned to the control group or standard treatment. Alternating assignment of patients to the treatment and the control groups is not randomization, because at the moment of allocation to treatment, if the last patient randomized was assigned to the control group, the patient now being allocated has a 100% probability, rather than an equal probability (50% in the case of two groups) of being assigned to the treatment group. Allocation of patients that is not truly random presents problems in the introduction of selection bias that results if the allocation scheme can be discerned by the patient or the clinician.

Randomization results in three major advantages (33). First, selection bias in the assignment of patients to treatments is eliminated. Randomization removes physician or patient preferences that might result in systematic assignment of the individual therapeutic agents to patients with predominantly poor (or good) prognoses for the outcome event (34). For example, whenever surgery is chosen for operable patients, and nonsurgical therapy reserved for those who are inoperable, the operable patients will usually have a better prognosis than the inoperable ones, even if surgery is not performed. In addition, whenever the use of a particular treatment requires that its recipients fulfill certain pretherapeutic criteria, the selection process may bias the assignment of patients with differing prognostic factors to one treatment group. For example, a nonrandomized study comparing thrombolytic therapy with angioplasty in the treatment of acute MI

may utilize selection criteria for the angioplasty group that selects patients with more favorable prognostic factors.

Secondly, randomization balances the treatment groups with respect to measurable patient characteristics, referred to as covariates, which may be prognostic factors for the outcomes of interest, whether or not these factors are known. Since most diseases are multifactorial in causation and prognosis, typically many factors, some known and some unknown, influence disease occurrence and progression. In other words, it is a very rare disease that has only one cause, and the outcome of which can be predicted with 100% certainty based on the knowledge of one clinical factor. The known factors can be measured. But randomization is particularly important for unknown prognostic factors, since it follows that they cannot be measured and thus cannot be adjusted for in the statistical analysis. If the reader is in doubt regarding the substantial impact of unknown prognostic factors on disease, i.e. you feel that most common diseases are well understood and explained, the reported inability to predict large proportions of the variance in outcomes given all the known prognostic factors is the best indicator of our generally incomplete understanding of most diseases. Therefore, the even distribution of covariates, especially unknown prognostic factors, among the treatment groups assures they will be truly comparable to one another.

One final advantage to randomization is that it provides the framework for comparing treatment effects among groups based on statistical inference (33). Although the groups compared are never perfectly balanced for important covariates, the distribution of prognostic factors by chance in the process of randomization allows a probability distribution to be ascribed to the difference in outcome between treatment groups. From a probabilistic perspective, averaging across many randomizations, covariates would be evenly distributed among the treatment groups. Although with a single randomization, it is unlikely that covariates will be distributed in an exactly equal proportion between the groups, their distribution by chance permits mathematical modeling to draw valid statistical inferences regarding the statistical significance of any differences.

Does randomization really make a difference? A number of investigators have found that nonrandomized studies yield larger estimates of treatment effects than studies using random allocation (35,36). Schultz extended this finding to clinical trials by evaluating a database of systematic reviews of 250 controlled trials published by the Pregnancy and Childbirth Group of the Cochrane Collaboration. Randomization schemes which were poorly executed, resulting in inadequate allocation concealment, also yielded exaggerated estimates of treatment effect (37). Petro has demonstrated that inadequate concealment of allocation can exaggerate odd ratios, on average, by about 40% (38). Thus, the process of randomization needs to be tamperproof , if a clinician or member of the research team is able to discern the randomization scheme, he can influence the channeling of patients with poor (or good) prognoses selectively into one treatment arm or another. This is easily accomplished by delaying the patient's entry into the trial until the next allocation or by causing the exclusion of eligible patients from the trial by encouraging them to refuse entry. Ideally, an outside randomization center should be established, but if not, then randomization treatment assignments should be placed in sealed envelopes in advance.

Once randomization is complete, how can we be certain that it was successful, i.e. resulted in truly comparable groups? There are methods to assess the 'efficacy' of randomization but I will defer their discussion for the moment except to point out the following. True comparability does not imply that the observed distribution of covariates is identical among all treatment groups, but rather that any observed imbalance is due to chance and not to some systematic bias. Therefore, a statistical test comparing treatment groups without regard for covariates is always "valid", if randomization has been done properly, and should always be reported (7) The central question is not whether this overall test should be replaced with others, but whether any useful insight can be gained by complementing this overall test by others that take covarites into account. More on this issue later. Keep in mind, then, that even randomization may result in an imbalanced distribution of covariates between the treatment groups simply due to chance (34). The greater the number of prognostic factors and the smaller the expected treatment effect, the larger the sample size needed to successfully evenly distribute the covarites by chance. The efficacy of randomization with regard to the latter can be evaluated by assessing the distribution of covariates between the treatment groups after the trial is completed. If known prognostic factors are not evenly distributed, it is also likely that any unknown factors are not. Therefore, it is customary in clinical trials to see a table reporting the distribution of important baseline demographic and

prognostic factors in each of the treatment groups. Table 4 is such a table reported by the GUSTO trial. Assessing the distribution of baseline factors is also one method of selecting variables to be entered into multivariate analysis so that statistical adjustment with regard to their impact on the treatment effect can be performed (8). This issue and the appropriate methods for assessing whether differences in the distribution of covariates have occurred will be discussed under "Were the groups similar at the start of the trial?"

If uneven distribution of covariates has occurred, what can be done about it? Before the study, if investigators are aware of prognostic factors that are strongly related to outcome, patients can be randomized as matched pairs or randomization can be stratified to guarantee that these factors will be evenly distributed without the potential interplay of chance. However, the complexity of these procedures are prohibitive for more than three to four factors (33). *After the study is completed,* statistical tests based on mathematical models are available to assist in adjusting for these differences which will be discussed later. The validity of conclusions based on these tests rests on the appropriateness of the assumed mathematical model. In general, only a small number of variables can practically be adjusted for and these must be identified and measured. Unknown prognostic factors cannot be adjusted for. What is the bottom line? Disparities in the treatment groups cannot always be removed by adjustment techniques and this conclusion underscores the importance of randomization.

How well did the GUSTO trial do in achieving randomization? The randomization process was exemplary. A centralized randomization center was utilized which was accessible by phone 24 hours per day, 7 days per week (2). Individual drug kits for each patient were forwarded to each study site for use according to the random assignment. These kits were sealed, coded with a numerical sequence, and active therapy was not identified until the seal was broken. As demonstrated in Table 4, there were no differences in baseline characteristics among the four treatment groups by the method the GUSTO investigators used to evaluate this, which I'll discuss in greater detail later. Keep in mind that there is still the potential for selection bias prior to randomization resulting from the often biased referral process of patients into clinical trials (33). GUSTO did not report what proportion of eligible patients were actually randomized. This selection bias has more impact on the generalizability of the study though since, due to randomization, the omission of certain types of patients would be expected to affect each treatment group equally. However, if a particular type of patient was never referred, and the response of this type of patient to treatment differed from other types of patients, the overall treatment effects may be biased.

Table 4 : Distribution of Base-Line Covariates in the GUSTO Trial

Characteristic	Streptokinase and Subcutaneous Heparin (N = 9841)	Streptokinase and Intravenous Heparin (N = 10,410)	Accelerated t-PA and Intravenous Heparin (N = 10,396)	Both Thrombolytic Agents and Intravenous Heparin (N = 10,374)
Age (yr)	62 (52, 70)	62 (52, 70)	62 (52, 70)	61 (52, 70)
Female sex (%)	25	25	25	25
Diabetes (%)	15	15	15	14
Cigarette smoker (%)	43	43	43	43
Hypertension (%)	39	38	38	38
Systolic blood pressure (mm Hg)	130 (111, 144)	129 (112, 144)	130 (113, 144)	130 (112, 143)
Heart rate (beats/min)	73 (62, 85)	74 (63, 86)	73 (62, 86)	74 (62, 86)
Previous infarction (%)	16	17	17	16
Previous CABG† (%)	4	4	5	4
Time to randomization (min)	120 (90, 180)	120 (90, 180)	120 (90, 180)	120 (90, 180)
Time to treatment (min)	164 (115, 232)	165 (120, 230)	165 (120, 230)	170 (121, 237)

*Values followed by numbers in parentheses are medians, with the 25th and 75th percentiles shown inside the parentheses. There were no differences in base-line characteristics among the four groups. Time to treatment, although not strictly a base-line characteristic, did differ among the groups (P<0.001).

†CABG denotes coronary-artery bypass surgery.

Page 10 was ommitted in pagination.

2. Were all Patients Who Entered the Trial Properly Accounted for and Attributed at its Conclusion?

Was follow-up complete?

Every patient who entered the trial should be properly accounted for at its conclusion (19). If this is not done, or if substantial numbers of patients are lost to follow-up, the validity of the study is open to question. The trial may be subjected to bias, since the patients who did not follow- up may have different prognoses than those who did. This difference may result from the reasons why patients did not follow-up, such as having suffered adverse outcomes or because they were doing well.

Readers must decide what degree of loss to follow-up is excessive and make a judgement as to whether the conclusions of the trial might have been substantially different if the outcomes of the patients who did not follow-up were known. The worst- and best-case case scenarios can be determined by assuming that all the patients lost to follow-up experienced the outcomes, or all of them did not experience the outcome, respectively. These provide the upper and lower limits for the impact of loss to follow-up on the results of the study. If the conclusions of the study would not change after considering the impact in this manner, then loss to follow-up is not excessive. The GUSTO trial reported that follow-up until death or 30 days after randomization was complete in 99.9% of patients, which was exemplary.

Were patients analyzed in the groups to which they were randomized?

This criteria is commonly referred to as the intention-to-treat-analysis. In this approach, all patients who were randomized to a specific treatment group are included in that group for the analysis of outcomes, regardless of whether the patients never received therapy, complied with therapy, experienced side effects and discontinued therapy, or crossed over to treatment in another group. This type of analysis protects the validity of the answer because it preserves the value of randomization (19). For example, a patient may crossover to another type of therapy due to lack of response to the therapy that was being received. If this cross-over occurred too late to allow adequate time to respond to the other therapy, and the outcome for this patient was analyzed in the treatment group to which he crossed over, the effectiveness of this treatment would erroneously appear diminished. However, the question addressed by the intention-to-treat analytic strategy is the relative effectiveness of *initiating a treatment policy*. The degree to which the analysis reflects *treatment actually taken* will vary.

Other strategies may be used to attempt to answer the question of treatment efficacy in subjects who have actually taken the therapy. For example, subjects in whom the intended treatment was modified may be excluded from the analysis as if they had never been randomized. However, differences in outcome susceptibility may invalidate the analysis as a result of using this strategy. In addition, the exclusion of major clinical events constrains the applicability of the study's findings. Another strategy is to exclude patients at a later point in time by means of a life-table method. However, bias can be introduced with strategies that exclude patients from analysis in the group to which they were randomized if the decision to modify treatment and the subject's prognosis for the outcome are related.

It is desirable to enable a fuller interpretation of the results of a study by incorporating information on the changes in treatment in the intention-to-treat analysis (i.e. control for these using multivariable analysis). More sophisticated methods of incorporating modification of treatment in the statistical analysis have been developed (39). The GUSTO trial appropriately utilized an intention-to-treat analysis. However, although this preserves the validity of the treatment assignment, substantial differences in intended therapy can influence outcomes. Most importantly in the GUSTO trial, 51% of the treatment group which received SK with subcutaneous (SC) heparin, 36% actually went on to receive heparin IV instead of subcutaneously (2). To the extent that IV heparin does not increase the efficacy of the treatment but increases the risk of hemorrhagic stroke, the net treatment effect in the SK + SC heparin group may have been adversely affected by this change in protocol. Nearly 60% of the time, this change was for a medical indication, for reinfarction etc, but 40% of the time it was not. Further discussion of the impact of this protocol change on the treatment effects observed will be addressed under "What were the results?"

B. *Secondary Guides*

Were Patients, Health Workers and Study Personnel "Blind" to Treatment?

Randomization only eliminates the influence of confounding variables that are present at the time of randomization. It does not protect the study from confounding variables that develop during the period of follow-up. The goals of blinding are to avoid intentional and unintentional bias in medical management and diagnostic interpretation, *occurring during follow-up*, that might affect the occurrence or assessment of outcomes across the different treatment arms (8). Blinding is important at four levels: 1) treatment allocation (was treatment assignment really randomized or was the allocation scheme discerned?), 2) patients (did knowledge of the treatment assignment influence compliance, report of symptoms or side effects?), 3) clinicians (did knowledge of the treatment assignment influence the overall medical management , the administration of co-interventions or the decision to assess outcomes?), and 4) investigators (did knowledge of the treatment assignment influence the assessment or interpretation of outcomes in the treatment groups?). However, many interventions can't be blinded (eg. surgery) or face logistical problems (7). For example, if blinding is used, a mechanism must be available to unblind the treatment if this should become necessary. Even when blinding is utilized, it may be unsuccessful, as when the treatment is associated with a discernible change in a physiologic response, such as the lowering of heartrate with beta-blocker therapy. After the study is over, it is a good idea to systematically assess whether the study patients or the investigators can "guess" the treatment assignment. If a higher than expected proportion guess correctly, then partial unblinding must be suspected.

The GUSTO trial was unblinded and so is subject to the potential introduction of the above biases. In an unblinded trial, a partial solution to the problem of unintended interventions is to specify and standardize the intervention. This was done for the study treatments in the GUSTO trial , but as already mentioned, half of the SK + SC heparin group received IV heparin. In addition, interventions other than the treatments under study, known as co-interventions, may be utilized differently between the treatment groups if they are not standardized, which was the case in the GUSTO trial. Only aspirin and beta-blocker use (if there was no contraindication) were standardized. The original GUSTO publication only reported the overall use of certain cardiac drug classes across all the treatment arms and stated that they were not significantly different between the treatment groups. The same was claimed for coronary revascularization procedures. The raw data for differences in the rates of these co-interventions among the treatment groups was not reported. This issue will be discussed further under "Aside from the experimental intervention, were the groups treated equally?"

The GUSTO investigators have offered a number of reasons as the rationale for the unblinded design of this trial, which include: 1) the need for a second infusion line in approximately 30 000 patients at increased expense and inconvenience to achieve blinding, 2) the use of a "hard"(objective) outcome such as death which is less susceptible to bias in its interpretation (i.e. a death is a death is a death...), 3) the confirmation of other unblinded thrombolytic trials by subsequent trials that were blinded, 4) the use of a randomized design to eliminate selection bias , 5) the use of an intention-to-treat analysis to eliminate bias due to changes in treatment, 6) the high rate of compliance with therapies (98%) which reduced bias due to side effects of treatment, 7) the blinded assessment of secondary endpoints which was based on objective information (eg. stroke definition was based on standardized criteria; diagnosis of stroke was made by MRI, CT, or autopsy; all data were examined by a blinded independent stoke review committee), 8) the independence of endpoint collection, processing, and interpretation from the trial sponsors, and 9) the on-site monitoring of a 10% random sample of all submitted data, which included comparing with and cross-checking source documentation in hospital records (40).

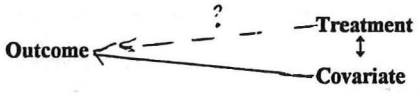
This rationale has been criticized by Ridker and colleagues (41). They point out that a second intravenous line does not seem too burdensome given the overall complexity and invasiveness of the trial and its use in the double-blinded ISIS-3 trial at a lower cost of intervention per patient than was accomplished in the GUSTO trial. Randomization *is not* sufficient assurance that bias is prevented. It only ensures, if done successfully, that all prognostic factors that may affect study endpoints are distributed among the various arms of the trial in an equal and unbiased manner at the start of the trial. It does nothing to ensure the comparability of treatment and diagnostic assessment of outcomes among the treatment groups once the trial is in progress.

The price of the unblinded design may have been differences in important co-interventions between the treatment arms that may at least partly account for the treatment benefit of accelerated t-PA. Even though the original GUSTO report stated that coronary revascularization procedures were similar in all treatment arms, the rate of coronary bypass in the t-PA arm was 9% compared to 8.3% in the SK arm (2). GUSTO investigators subsequently reported the specific rates of bypass in the treatment arms and claimed that the differences in bypass rates could not account for the 30-day mortality benefit observed. They analyzed the results of patients who did not receive bypass and compared them with those who did. The treatment benefit was preserved for patients who received t-PA vs SK but did not undergo bypass (30 day mortality 6.5% vs. 7.6% , respectively; $p < 0.001$). Among patients who underwent bypass, 30 day mortalities in the t-PA and SK arms were nearly identical (4% vs 4.1% , respectively). However, what then explains the reason for increased rates of bypass in the t-PA arm? If this procedure was not performed to reduce mortality, then the alternative explanation must be considered, i.e. that t-PA resulted in a higher rate of complications requiring emergent procedures for salvage. GUSTO investigators also subsequently reported that logistic regression analysis adjusting for differences in rates of bypass still demonstrated a significant difference in treatment benefit ($p = 0.001$) yet did not report the adjusted difference in mortality benefit. Given the size of the trial, calculations demonstrate that residual differences of 0.2-0.3% would still be statistically significant. Finally, specific rates for coronary arteriography and early percutaneous coronary angioplasty have not been reported. Raw data should be provided.

Were the Groups Similar at the Start of the Trial?

What is the importance of comparability among treatment groups and how is it assessed, particularly in the context of a randomized clinical trial? This criteria specifically addresses whether patient characteristics that might serve as important prognostic factors, which could influence the difference in treatment effect between groups, were evenly distributed, at the start of the study. Recall that one of the major objectives of randomization is to accomplish comparability of treatment groups with respect to these covaraites but that even chance may result in some imbalance in and thus confounding of any apparent treatment effect (42). Confounding is represented in Figure 1, and is seen when the treatment being studied and the covariate being considered are related to one another, and at least the covariate, either by itself or through some other factor, is related to the outcome being assessed. Although the treatment will appear to be related to the outcome, this may not actually be "true" but instead be the result of the treatment's association with the confounding factor. The randomization of treatment assignment in a clinical trial, particularly if it is large, should make confounding unlikely. But, in an occasional trial, the play of chance may create a gross maldistribution of prognostic factors that favors (or disfavors) one of the treatment groups.

Figure 1: Confounding



A number of methods are available to assess the difference in distribution of covariates among the treatment groups:1) simple comparison of the magnitude of the difference between the number of patients with the covariate in each treatment group, 2) statistical tests of significance in comparing the number of patients in each treatment group with regard to each covariate, 3) tests of interaction, and 4) others (43). The problem with the first method is that there is no consensus among clinicians as to what magnitude is sufficient to constitute an important difference. The second method applies a test for statistical significance to the distribution of each of these covariates. This was the method used to assess most of the covariates in the GUSTO trial (2). For example, the proportion of patients assigned to each treatment stratified by patient characteristics seen in Table 4 represents comparisons based on the chi square statistic for dichotomous (e.g. dead vs. alive) covariates. Although statistical tests for significance is the most common method, it is also the one that is most problematic from a statistical perspective. Its problems stem from three issues: 1) the impact of multiple comparisons, 2) the lack of power to detect

differences and 3) conceptual violation of the statistical foundations of significance or hypothesis testing. The latter, while important, is more theoretical and I will leave it to the interested reader to further explore the basis of this criticism.

The impact of multiple comparisons on the likelihood of finding a significant difference can be calculated as follows (44). In general, if α is the arbitrary cutoff value for significance and n covariates are examined for statistical significance, the probability that *at least one of them* will be found statistically significant is $1 - (1 - \alpha)^n$. With $\alpha = 0.05$ and $n = 20$, the probability of finding at least one statistically significant difference in the distribution of covariates *assuming that all the null hypotheses are true* (i.e. that there are no "true" differences between levels of a subgroup) is 0.64. Remember this assumption because it is the basis of criticism for those who oppose the use of adjustments for multiple comparisons. In practice, most large clinical trials examine the distribution of many covariates (often $n > 100$) among the treatment groups. The result is that chance alone can cause the difference. Adjustments for multiple comparisons can be made or tests for interaction can be performed but each approach leads to additional problems. These issues will be further addressed under "What Were the Results of the Study?: Subgroup Analysis".

In general, GUSTO patients, although similar with respect to important covariates among treatment arms, were a relatively good risk population. Only about 12% of patients were over the age of 75 years, all but 2% were Killip class I or II, only 16% had had a previous MI, and only 39% presented with an anterior infarction. Of striking importance, the time from onset of symptoms to onset of therapy was only 2.8 hours (2).

Aside from the Experimental Intervention, were the Groups Treated Equally?

After the onset of therapy, treatments under comparison may become unequal in the proficiency of therapy and in the detection of outcomes. The ways in which treatment is administered, maintained, or supplemented by concomitant therapy other than the treatment under study (co-interventions) may differ among treatment groups. Often, modification of intended treatment occurs in response to a change in the patient's clinical status in which the assigned treatment is no longer judged to be indicated by the physician and/or patient. Bias may be introduced if the decision to modify treatment and the subject's prognosis for the outcome are related. This bias in randomized clinical trials has generally been handled by the intention-to-treat approach in its analysis. (See "Were patients analyzed in the groups to which they were randomized?")(39).

The problems with detection bias arise from differences in monitoring the post-therapeutic course, in ordering subsequent tests and in the interpretation of the results of outcome assessments. To ascertain the occurrence of outcomes, subjects must: 1) remain at risk for the outcome until it can be ascertained; 2) be available for follow-up so that the outcome can be ascertained; and 3) if necessary, satisfactorily complete the procedures required for the determination of outcome status (39). For example, a competing event, such as death, prior to the time of ascertainment of outcome removes the patient from risk, i.e. as Sackett has pointed out, "dead patients cannot have strokes" (45). This problem is best dealt with by incorporating all competing events that could be related to the risk for the primary outcomes into the definition of the combined end point, since "...to confine analysis to fatal and nonfatal strokes could produce the situation in which the drug that decimates the ranks of the potential stroke victims by killing them from other causes will spuriously appear efficacious" (45).

When patients are not available for follow-up, the outcome status is known only up to the point in time at which the subject becomes unavailable. Removal of them from the analysis can produce a biased estimate of treatment effectiveness if the average prognosis of those unavailable for follow-up differs among treatments (39). A simple demonstration that the rates of unavailability are similar in the treatment groups, as is customarily done, may not protect the validity of this comparison. When outcomes require greater effort, such as an invasive procedure, some subjects may refuse to comply with the test for the outcome or be incapable of complying. Detection bias in randomized trials can be reduced with choosing outcomes that require less effort to ascertain, double-blinding and the use of "hard" endpoints (such as death), which are objective in nature and thus subject to less bias introduced by subjective interpretation of occurrence. But this strategy is unsuccessful for two reasons. First, problems can still occur in making determinations about the cause of death, and the use of "hard" endpoints limits the ability of studies to

more fully examine the impact of treatment, particularly with regard to quality of life. Feinstein recommends the improvement of the quality of "soft" data used to identify outcome events (34).

Were patients in the GUSTO trial treated similarly? I have already mentioned the disparity in the use of CABG, which was used more frequently in the t-PA arm, and the high utilization of IV heparin in the SK + SC heparin arm, particularly in the United States. The raw data for utilization of other co-interventions, such as cardiac classes of drugs, was not reported in the original publication. These were reported for the GUSTO angiographic subset study, but this study only encompassed about 6% of the original GUSTO enrollment (46).

II. WHAT WERE THE RESULTS?

A. How Large was the Treatment Effect?

Several issues are raised when considering the magnitude of the treatment effect.

1) How should clinical research results be summarized and what magnitude of clinical difference is important?

2) To what extent does statistical significance reflect clinical significance?

3) How should treatment effects in subgroups be assessed and what is the impact of multiple comparisons among subgroups?

4) What can be concluded from a negative study?

5) How should overall treatment effects be modified to accomodate the effects of prognostic factors?

1) How should research results be summarized?

How should the benefits and risks of therapeutic approaches be measured and compared? Several ways of summarizing the benefits of therapy are presented in Table 5 (47) . Aristotle captured the dilemma eloquently: "there are several possible ways of persuading people about any given subject." The absolute risk reduction (ARR) is the difference in event rates between the control and treatment groups. The relative risk reduction (RRR) is the reduction of adverse events achieved by a treatment, expressed as a proportion of the control rate. In other words, it is the difference in event rates between the control and treatment groups, divided by the event rate in the control group. The odds ratio (OR) is the traditional epidemiological expression of the relative likelihood of an outcome, and is expressed as the ratio of the odds of adverse outcomes in the two treatment groups being compared (where odds = 1 - probability). The number needed to treat (NNT) is the number of patients who must be treated in order to prevent one adverse event. Mathematically, it is the reciprocal of the absolute risk reduction. Approximate and exact confidence intervals can be calculated for all of the summary measures discussed (48).

Table 5 : Comparison of Summary Measures for Reporting Research Results

Summary Measure	Calculation	Advantages	Limitations
Absolute risk reduction (ARR)	Pc* - Pt*	• Captures magnitude of base-line risk without therapy	• Combines base-line risk and risk reduction into single number
Relative risk reduction (RRR)	Pc - Pt / Pc	• Conveys population impact	• Magnitude of base-line risk without therapy not conveyed
Odds ratio (OR)	Pt / 1 - Pt Pc / 1 - Pc	• Suitability for statistical modeling Conveys population impact	• Magnitude of risk without therapy not conveyed
Number needed to treat (NNT)	1 /ARR (decimal)	• Captures magnitude of base-line risk without therapy	• Combines base-line risk and risk reduction into single number • Need to specify duration of treatment

*where Pc and Pt are the event rates in the control group and treatment groups, respectively

Table 6 demonstrates how an odds ratio is derived (49). This summary measure is a ratio of the odds of an outcome in the treatment group to the odds of the outcome in the control group. The odds of an outcome is the ratio of the probability of the outcome over 1 - the probability of the outcome. For example, if a coin is tossed, the probability that it will come up tails is ½ or 50%. The odds of it coming up tails is 0.5 / 1 - 0.5 = 0.5 / 0.5 = 1:1.

Table 6: Illustration of the Derivation of the Odds Ratio for the GUSTO Trial

	Treatment (t-PA)	Control (SK)
Disease (30 day mortality)	a (6.3%)	b (7.35%)
No disease (30 day survival	c (93.7%)	d (92.65%)
Probability of outcome	a / a + c	b / b + d
**Odds of outcome*	a / a + c = a / c c / a + c	b / b + d = b / d d / b + d
Odds ratio	a / c = ad = 6.3 x 92.65 = b / d bc 7.35 x 93.7	
Relative risk	a / a + c = 6.3 / 6.3 + 93.7 = 0.85 b / b + d 7.35 / 6.3 + 92.65	

*Where: Odds = p / 1-p
Odds ratio = odds of outcome in the treatment group
Odds of outcome in the control group

A measure related to the odds ratio, the relative risk, is an epidemiological measure that is derived from a population (50). It is the ratio of the incidence of the outcome in the exposed members of a population to the incidence of the outcome in the unexposed members, where incidence is the number of individuals in the population that develop an outcome over the number of individuals at risk for the outcome in the population. Since clinical trials assign the exposure (the treatment interventions), the population gradient for exposure is no longer reflected but is instead fixed by the investigators. Therefore, true incidences , and thus relative risk, cannot be obtained from a clinical trial. However, the RR can be approximated with the odds ratio if the prevalence of disease is small. It is in the sense that relative risk reduction is used as a summary measure in clinical trials.

In order to illustrate the differences among the various measures of benefit, the Veterans Administration cooperative study on hypertension will be used as an example (51). This three-year study compared the efficacy of antihypertensive therapy (a combination of hydrochlorthiazide, reserpine, and hydralazine) with that of placebo. The rates of adverse events (sudden death, stoke, myocardial infarction, congestive heart failure, accelerated hypertension, and dissecting aneurysm) among patients with and without target-organ damage at the time of entry are as presented in Table 7.

Table 7: Measures of Efficacy in the Veterans Administration Trial

Patients' Condition at Entry	Rates of Adverse Events (%)		RRR (%)	OR	ARR (%)	NNT
	Placebo	*Treatment*				
Target-organ damage	22.2	8.5	62	0.325	13.7	7
No target-organ damage	9.8	4.0	59	0.384	5.8	17

Note that the treatment benefits expressed as relative risk reductions are about the same for both the target-organ damage group and the group without target-organ damage (~60%), despite the fact that the risk of adverse events with treatment was more than twice as high in the former as compared to the latter group. The disadvantage of RRR as a summary measure is that it doesn't accurately reflect the magnitude of risk without therapy, which in this study was much higher in the target-organ damage group than the no target-organ damage group. The RRR overestimates or underestimates the absolute impact of therapy when adverse events in untreated patients are very rare or very common, respectively. Overestimation of treatment effects thus becomes a problem in studies investigating conditions with low rates or differences in rates of adverse events, such as preventive therapy or clinical trials that compare treatment effects with some standard therapy in place of placebo.

The treatment effects expressed as odds ratios are also nearly equivalent, even though the two groups of patients have markedly different rates of adverse outcomes with treatment. Although the odds ratio has distinct statistical advantages over relative risk with regard to suitability for modeling both within a study and across studies when treatment effects are pooled, like RRR, it fails to capture the magnitude of risk without therapy.

In contrast, the absolute risk reduction for the group without target-organ damage is half that of the group with damage (5.8 vs 13.7%, respectively), thus expressing risk reduction with therapy and an additional measure of clinical effect, the consequences of giving no treatment. Similarly, the NNT indicates that more than twice as many patients in the no target-organ damage group would need to be treated to prevent one adverse outcome than in the target-organ damage group (17 vs 7, respectively).

Laupacis has proposed that the ideal summary measure would have the capacity to: 1) compare the consequences of doing nothing (the risk for the adverse event if no intervention is provided) with the potential benefits of doing something (extent to which this risk would be reduced by the use of a specific clinical treatment), 2) summarize the harm that would accompany the treatment in the form of side effects and toxicity to patients, 3) identify patients who are both at high risk for an event and responsive to therapy and 4) permit a comparison of the consequences of applying one approach to the prevention,

diagnosis, and treatment of one condition with the consequences of applying other approaches to other conditions (so can decide where best to focus efforts for individual clinicians and patients) (52). NNT is able to satisfy all these criteria, and is a useful measure of incorporating harm and in comparing treatment strategies within a disease and across diseases.

The major drawback of both ARR and NNT, however, is that they combine the baseline risk and risk reduction into a single number (52). Table 8 demonstrates the effect of the baseline risk of an adverse outcome in untreated patients and the RRR associated with treatment on the NNT. If the event rate is high in the control group, even a small RRR will produce a low NNT. Conversely if the baseline risk is low, the risk reduction must be large in order to produce a low NNT. In combining the baseline risk and risk reduction into a single number, nothing is revealed as to what happens to the other patients, i.e. NNT tells us that the patient *either* doesn't need therapy , will not respond to it or will not be able to continue therapy due to side effects. Table 9 demonstrates the limitations in interpreting the NNT. Additional disadvantages of NNT include its inability to capture or extrapolate the consequences of continuing therapy beyond trial completion and the necessity of expressing NNT per duration's of therapy, since a timeframe is not inherent to the calculation of this measure.

Table 8: NNT Resulting From Various Baseline Risks and RRR's

Base-Line Risk*	Relative Risk Reduction by a New Therapy (%)						
	50	40	30	25	20	15	10
	number needed to be treated						
0.9	2	3	4	4	6	7	11†
0.6	3	4	6	7	8	11	17
0.3	7	8	11†	13	17	22	33
0.2	10	13	17	20	25	33	50
0.1	20	25	33	40	50	67	100
0.05	40	50	67	80	100	133	200
0.01	200	250	333	400	500	667	1,000
0.005	400	500	667	800	1000	1333	2,000
0.001	2000	2500	3333	4000	5000	6667	10,000

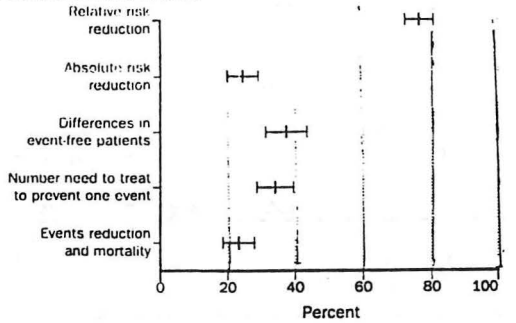
*Risk of an adverse event in control patients †Numbers used as examples in the text

Table 9: Illustration of the Limitations in Interpreting Number Needed to Treat

Example: NNT of 11		
How is this to be interpreted?	⇒	11 patients must be treated to prevent one adverse outcome.
What happens to other 10?	⇒	Do they develop outcome?
Compare the following:		
Base-line risk 0.9, RRR 10%	⇒	9/10 remaining patients have adverse event.
Base-line risk 0.3, RRR 30%	⇒	2/10 remaining patients have adverse event.
	∴ NNT is 11 in both!!	

Why is the format of numeric presentation in reporting results important? Studies in cognitive psychology have amply demonstrated that the manner in which clinical , particularly quantitative, information is presented can have profound effects on the likely interpretation of that information (53-55). In particular, physicians, and patients, may interpret quantitative information differently and inconsistently (56-59). Forrow investigated physicians' treatment decisions by presenting results of two clinical trials, one examing the effect of treatment for hypertension and another looking at treatment for hypercholesterolemia, in terms of absolute and relative changes in outcome rates (60). The physicians were asked to rate the likelihood that the information would change their treatment of patients. In response to the presentation of these trials, 49.2% of physicians presented with the hypercholesterolemia trial and 32.7% of physicians presented with the hypertension trial indicated a stronger likelihood of treating patients when presented with relative reduction in outcome rates than when presented with the absolute reduction . Other investigators have similarly found that relative reduction of risk influences readers to perceive treatment benefits as greater than they would if these same results were reported in terms of absolute differences (61). Bobbio has confirmed this finding when NNT was added to the formats of numerical presentation (see Figure 2) (62).

Figure 2: Relationship Between Physicians' Agreement to Prescribe Drug and Presentation Format



The summary measures of benefit for the GUSTO trial are presented in Table 10. Note that the ARR is considerably smaller than the RRR. Both were reported in the trial, but the RRR was more emphasized. The NNT was not presented in the strict definition, although a similar format, the number of lives saved per 1000 treated patients, was

Table 10: Comparison of Summary Measures of Benefit for t-PA Versus the Combined SK Groups in the GUSTO Trial.

Treatment Group	Accelerated t-PA	Combined SK groups	ARR (%)	RRR (%)	NNT
30-day Mortality Rates (%)	6.3	7.3	1	14	100
Calculation			7.3 - 6.5	7.3 - 6.5/ 7.3	1 / 0.01

How are clinicians to determine which summary measure should most influence our treatment decisions? This is a difficult question as there is no objective standard by which we can ultimately judge the "rightness" of our decisions. Even the experts haven't reached consensus on the best way to express the difference in treatment effect, nor on what constitutes a worthwhile difference. For example, Table 11 presents the NNT for various conditions, which can vary considerably (9). Consensus does not exist as to the threshold NNT above which treatment would not be considered beneficial enough to offer. This issue is further complicated by the mathematical characteristics of NNT in that conditions with low baseline risk, even in the face of large RRR in adverse outcomes, are associated with very large NNT. A good example of this is screening mammography in women ages 50-74 to prevent death due to breast cancer.

Harris and Leininger reviewed seven randomized controlled trials of breast cancer screening. The absolute risk reduction varied from 0.2% to 0.4%, corresponding to RRR from 15% to 30% and NNT of 1700 to 5000 (63). Despite the large NNT, breast cancer screening in this age group is widely recommended (64).

Table 11: Comparisons of Benefits of Various Conditions

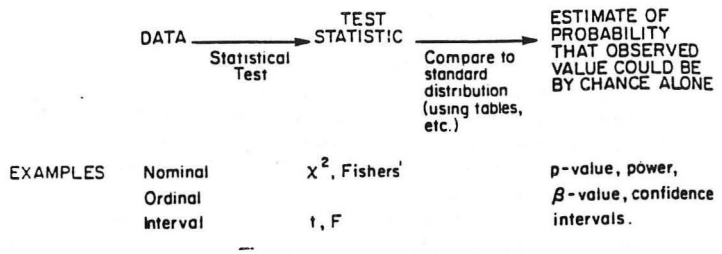
Therapy	Events	Base-line Risk	RRR (%)	NNT (per 5 yrs of F/U)
Stepped care: Dbp 115-129 mm Hg	Death, CVA, MI	0.13	89	3
Left main CABG	Death	0.32	56	6
Asa for TIA	Death, CVA	0.23	31	6
Cholesterol for dyslipidemia	Death, MI	0.12	14	89
INH for inactive Tb	Active Tb	0.01	75	96
Stepped care: dbp 90-109 mm Hg	Death, CVA, MI	0.05	14	141
Screening mammography	Death		15-30	1700-5000

There is also considerable controversy with regard to the GUSTO trial over whether an ARR of 1% is to be viewed as a clinically worthwhile benefit. The GUSTO investigators have responded to this criticism by pointing out that the absolute reduction in death associated with thrombolytic therapy when compared to placebo translated into saving 26 lives per 1000 treated patients. The GUSTO trial supported the further saving of nine to 11 additional lives, which represents approximately 40% further benefit than that originally ascribed to thrombolytic therapy, a benefit that was interpreted as important by the GUSTO investigators (40). This impact may be interpreted as important from a public health perspective, since small reductions in mortality due to common diseases summate over the many individuals who would experience the small benefit. However, in studies of individual patients, it is not clear that the small benefit that an individual would experience would be persuasive enough for them to choose thrombolytic therapy with t-PA over SK as a treatment option. Hux and colleagues found that when patients were asked if they would be willing to take a drug, provided at no cost and resulting in no side effects, that would increase heart disease-free survival an average of 15 weeks, 55% were unwilling to do so. When asked the same question but presenting the benefit in a stratified format that was arithmetically equivalent to the former format ("5% of patients would have an additional 2 to 6 years free of heart disease, 10% would have up to 2 additional disease-free years, and 85% would experience no change in the number of healthy years expected"), 44% of patients would be unwilling to take the drug (65). The controversy about what constitutes an important enough clinical difference to warrant treatment will not be easily settled. Many factors will play a role, such as the baseline risk of the disease in the population, the purpose of the treatment (preventive vs therapeutic), the effectiveness, toxicity, feasibility, and cost of the treatments being investigated, and the perspective (public health vs individual patient) being adopted. These factors will vary for each disease being studied. All of the summary measures provide potentially useful information. Relative risk reduction and odds ratio are helpful in conveying the population burden due to the illness that would be alleviated. Descriptions of expected benefits for individuals are best expressed as absolute risk reductions or number needed to treat. However, since many physicians do not seem to distinguish between the different presentations, and since different formats lead to different interpretations, the presentation of multiple formats is recommended, which would include at least the absolute risk reduction and some measure of event rate, or baseline risk, in the untreated group (57). Many experts strongly recommend that NNT be presented. Each summary measure reported should also be expressed with its associated confidence interval.

2) To what extent does statistical significance reflect clinical significance?

What does the concept of statistical significance refer to?(66) Statistical significance refers to what is essentially a decision rule regarding a probability threshold for deciding whether observed data could have been the result of random variation. The goal of the decision rule is to allow clinicians to make inferences about unknown parameters, which are generally population parameters, based on observed data . The foundation for the decision rule is a mathematical model of how observed data in the sample should be probabilistically distributed if only random variation is at play. The statistical test provides a means of comparing the observed data with the model, which represents what would occur on the basis of chance (see Figure 3). This test, in common usage, is performed by constructing two hypotheses: the null hypothesis (there is no "true" difference) and the alternative hypothesis (there is a "true" difference). For the statistical test, the null hypothesis is assumed, i.e. it is assumed that there is no difference. The probability of obtaining the particular value or greater in the observed data is predicted by comparing the observed data to that which would be expected based on the model. If this predicted probability, the P value, is less than an arbitrarily selected threshold, the null hypothesis is rejected and it is concluded that there is a significant difference, i.e. the likelihood that this difference could be due to chance is considered to be quite low, as long as the threshold for interpreting this as low is mutually accepted. If the P value is greater than the threshold probability then it is concluded that the null hypothesis cannot be rejected (note that this conclusion is not the same as accepting the null hypothesis, in effect proving that there is no difference). In the medical literature, this probability threshold is usually set at 0.05, or a 5 % chance that an observation this great or greater could be due to random variation and not the result of a "true" difference between the treatment groups.

Figure 3: Paradigm of Statistical Testing



The decision-making format of statistical testing lends itself to an analogy with diagnostic testing in its comparison between a test (the statistical test), and a gold standard or truth, which in this case is the population parameter if it could be known. The matrix for the possible decisions and their accuracy with regard to the "truth" is presented in Table 12.(67) Four combinations of agreement are possible. The statistical test may be positive when there is a "true" difference in the universal population (corresponding to a true positive test result), or the test may be positive when there is no "true" difference (a false positive test result or a Type I error). The statistical test may be negative when a "true" difference does exist (a false negative test result or a Type II Error), or the test may be negative when a 'true" difference does not exist (a true negative test result). Note that the probability threshold for statistical significance is α, which is the probability that one would incorrectly conclude that there is a "true" difference when there is not, i.e. a false positive test. α also then corresponds to the Type I error rate which is fixed in advance. The selection of the particular threshold α=0.05 was related to its convenience in that Fisher observed that two standard deviations encompassed 95% of the probability distribution of a Gaussian curve.

Table 12: The Decision Matrix for the Decision Regarding Statistical Significance

		Truth in the Universal Population	
		Difference Exists	No Difference Exits
Results of Statistical Test on Sample (statistical significance)	Positive	True Positive *Power* *Correctly conclude* that a difference exists when it does indeed exist	False Positive *Type I Error* (α) *Incorrectly conclude* that a difference exists when in "truth" it does not
	Negative	False Negative *Type II Error* (β) *Incorrectly conclude* that there is not a difference when in "truth" there is.	True Negative *Correctly conclude* that there is no difference when in "truth" there is not

As advanced as we believe we have become in our conceptualization of decision-making, Epictetus, as far back as the second century, had already captured this decision matrix (68):

"Appearances to the mind are of four kinds.

Things either are what they appear to be;
or they neither are, not appear to be;
or they are, and do not appear to be;
or they are not, yet appear to be.
Rightly to aim in all these cases
is the wise man's task."

How is this statistical theory to be applied to clinical research, and in particular, clinical trials? Figure 4 depicts the paradigm of the clinical research question, the focus of which is usually to provide an answer for all the patients with a particular disease of interest for whom the question has arisen, the universal population. However, this population is usually not accessible in its entirety for study. Instead, a sample of patients with the condition of interest is selected, from which inferences are to be drawn about the universal population. In a randomized clinical trial, the aim is to estimate the true response rate, P_t, for a treatment under study and compare it with the estimate of the true response rate, P_c, of a competing therapy, which may be a standard treatment or placebo (69). The treatment group of patients yields the observed response rate, P_t, which is an estimate of P_t, and the control group of patients produces the observed response rate, P_c, which is an estimate of P_c. The observed difference $\hat{P}_c$-$\hat{P}_t$ is then an estimate of the effectiveness of treatment (referred to as the point estimate). Even if P_t and P_c are truly equal, non-equal observed differences ($\hat{P}_c$ - $\hat{P}_t$) will occur due to random variation. The null hypotheses would be represented as $P_c - P_t = 0$, and the alternative hypothesis as $P_c - P_t \neq 0$. On the basis of a statistical test, a decision is made to reject or fail to reject the H_o. When we reject the H_o, we run the risk of making an erroneous decision. The true population rates, P_t and P_c, may indeed be equal, presenting a Type I error. However, if the test procedure leads us to fail to reject the H_o because $P > \alpha$ (fixed at 0.05), and we conclude the difference between $\hat{P}_t$ and $\hat{P}_c$ is not statistically significant, we run the risk of another error in our decision-making. This false negative or Type II error results when the true difference, $P_c - P_t$, is nonzero. Chance may have resulted in a difference not large enough to reject the H_o. Note that the statistical interpretation of failing to reject the H_o is not equivalent to accepting the H_o and to concluding that there is no difference in response between the two treatments. Unfortunately, this is often how it is interpreted. More about this later (70, 71).

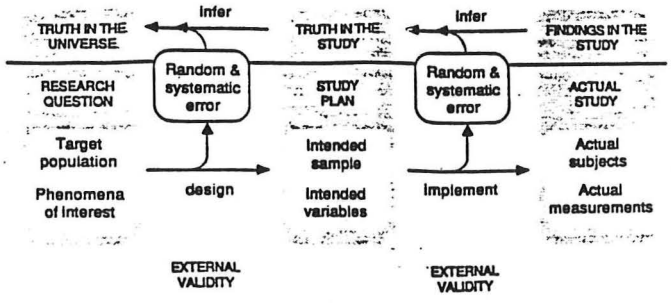


Figure 4: Paradigm Of Clinical Research Question

Significance testing , in its emphasis on an arbitrary cut-off to facilitate the decision of whether to reject or fail to reject the null hypothesis, poses distinct disadvantages. The decision rule fails to characterize the magnitude of differences between groups, and as clinicians, we are as interested in how much of a difference there is as we are in whether there is a difference. Let's assume that in comparing the treatment response between two drugs, we have made a correct decision to reject the H_o . This decision would be the same for a P value of 0.04 as it would be for one of 0.0001, and yet these P values likely represent differences in magnitude with regard to treatment effect. Now let's assume that we have failed to reject the Ho. This decision would be the result of a P value of 0.06 and one of 0.60, and thus fails to covey the differences in treatment effect that almost surely exist between these drugs. As can be gleaned from these examples, more information could be conveyed by reporting the actual P value which would convey how close to statistical significance the results were. However, the actual magnitudes of difference are still not apparent.

Another problem arises with significance testing and P values. When we fail to reject the H_o , and have made a Type II error, i.e. there is a true difference in treatment response but we failed to detect it, simply reporting the lack of statistical significance (or even the actual P value) tells us that a population difference is not likely at a particular high probability, but fails to provide a range of estimates for the population difference that might have lesser, but still clinically important, probability (70). In addition, statistical significance depends as much on sample size as it does on the observed difference (71). Table 13 demonstrates the relationship between sample size and the P value. Note that for a fixed difference in effect, the likelihood of it being declared statistically significant is strongly related to sample size. For small samples almost nothing is significant, and for very large samples quite small differences are significant. Therefore, statistical significance conveys little about the magnitude of the observed difference and consequently conveys little about clinical significance.

In the GUSTO trial, the P value for the *comparison of accelerated t-PA with the combined SK groups* with regard to 30-day mortality was 0.001, and yet the absolute difference in mortality was 1%. Although controversy as to whether this difference in treatment effect associated with this high level of statistical difference in GUSTO constitutes a clinically important difference has been heated, this example clearly illustrates that small magnitudes of difference can be associated with very small, and highly significant P values. *Clinical significance, therefore, is a judgement* (72). Clinicians weigh the relative the benefits of two (or more) treatments against the side effects, long-term complications, and other costs associated with these treatments.

Table 13: Relationship Between Sample Size and P Values With Fixed Treatment Effect

Trial Size	Relative Risk	95% Confidence Interval	P-Value
20	0.82	(0.12, 3.79)	0.33
200	0.82	(0.47, 1.43)	0.24
2000	0.82	(0.69, 0.98)	0.01
20 000	0.82	(0.77, 0.87)	< 0.00001

However, the choice to combine the two SK groups because there was no significant difference in the primary outcome between them has been sharply criticized (73,74,75). First of all, this comparison would have been underpowered to detect anything but very large differences between the SK groups. Secondly, the standard for thrombolytic therapy after ISIS-3 was reported was SK + SC heparin. There appeared to be little role for IV heparin when combined with SK, for several postulated reasons, and yet it posed a further increase in risk for bleeding. In fact, in the GUSTO trial, the SK + IV heparin group experienced a slightly greater mortality than the SK + SC heparin group, which was also combined with a higher rate of nonfatal strokes. Tables 14,15 demonstrate that combining the SK groups for comparison with t-PA results in an inflation of the number of lives saved per 1000 treated patients from 9 to 10.The P value for this comparison was 0.009. This comparison is further compounded when the complications of bleeding and strokes are considered, which will be addressed later.

Table 14: Comparison of Net Benefits of t-PA Versus SK + IV Heparin in GUSTO

Outcome	t-PA + IV heparin (Group 1)	SK + IV heparin (Group 2)	ARR (1 vs 2)	*No. of lives saved /1000 treated (1 vs 2)
30-day mortality	6.3	7.4	1.1	10

Table 15: Comparison of Net Benefits of t-PA Versus SK + SQ Heparin in GUSTO

Outcome	t-PA + IV heparin (Group 3)	SK + SQ heparin (Group 1)	ARR (3 vs 2)	*No. of lives saved /1000 treated (3 vs 2)
30-day mortality	6.3	7.2	0.9	9

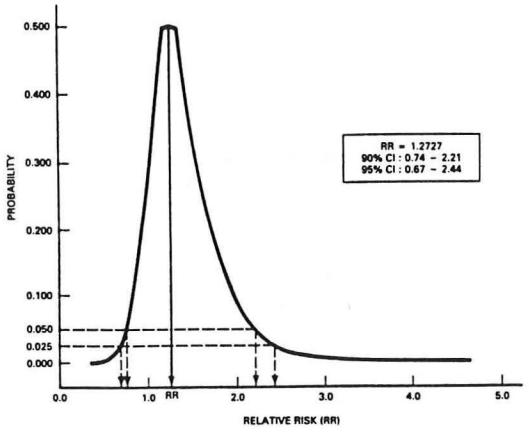
In summary, the problems with statistical significance include: 1) dichotomizing the "significance" of a study's results into a simple yes/no answer at some arbitrary cut-off level which cannot be equated with medical importance or biological relevance, 2) inability to encompass the magnitude of difference, 3) inability to reflect the precision in estimating the magnitude of difference and 4) inability to prove that there is no difference (the latter two I'll discuss shortly).

As an aside, I should mention, on behalf of the statistical purists among the readers, that the current use of significance testing is criticized by many experts for lacking a valid statistical framework (76). The procedures for significance testing as it is now used have foundations in the earlier formulations of significance testing, put forth by Fisher, and of hypothesis testing, put forth by Neyman and Pearson. These formulations have become somewhat merged, without the benefit of a mathematical or conceptual model, into a procedure that probably neither Fisher nor Neyman and Pearson would claim appropriately reflect the mathematical models they described (77). A detailed discussion of this discrepancy is beyond the scope of my purpose. Nonetheless, from a practical perspective, the current use of statistical testing is what the reader will be called upon to interpret.

If significance testing and P values are misleading, how should results be presented? A better way would be to choose a measure that quantifies the degree of effect and for which confidence intervals can be calculated (78). A confidence interval is the range of values, based on the data, that are plausible for the true treatment effect, at some specified level of confidence. A 95% confidence interval is customarily chosen, but others are possible. Put simply, this means that there is a 95% chance that the indicated range includes the population difference. Another way to interpret the confidence interval is from a frequentist perspective. If a series of identical studies (with 1000 patients each) were carried out on repeated samples (of 100) drawn from the same universal population (of 100 000 patients), and 95% confidence intervals for the difference between the means were calculated for each study, then 95% of the confidence intervals (or 95 out of the 100 confidence intervals), would contain the population difference in the means. Although the confidence interval defines the plausible range for the population difference, not all values

across the confidence interval have the same probability of being the population difference. Figure 5 depicts the probability density of a 95% confidence interval (78,79). The true value is more likely to be near the point estimate, and the probability falls off in a nonlinear manner in either direction away from the point estimate, although no values in a 95% confidence interval can be "rejected" as representing a true difference by a 5% significance level.

Figure 5: The Probability Density of a Confidence Interval

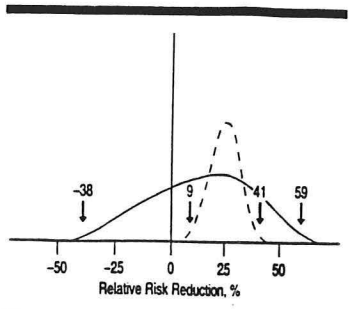


The width of the confidence interval depends on several factors that are inherent to its calculation (see Figure 6: sample size, the level of confidence desired, the variability in the sample and the magnitude of effect size observed (78,79). Figure 7 represents the effect of sample size on the confidence interval (78,79). As the sample size increases, in the face of a fixed effect size, the interval narrows, and corresponds to an increase in precision in the results. Precision can best be understood as the reliability, or reproducibility, of the results. How much variation is seen in the measurements? Figures 8 and 9 depict how precision is to be contrasted with accuracy ("true")(7). Precision relates to how close repeated measurements clutter, regardless of whether they are accurate. Accuracy refers to how close the measurements are to the true values, regardless of their variability. Rothman views the confidence interval as the approximate position of the true value, whereas the P value "is equivalent to funneling all interest into the precise location of one boundary of a confidence interval." (80) Analogously with sample size, as the effect size (eg. difference in treatment response rate) increases, the confidence interval narrows (79). In contrast, as the level of desired confidence increases, as would be expected, the confidence interval widens (see Figure 10).

Figure 6: Formula for Calculation of Confidence Interval

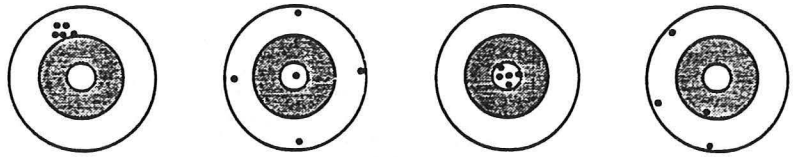
$$\frac{(p_c - p_e) +/- 1.96 \sqrt{\frac{p_c(1 - p_c)}{n_c} + \frac{p_e(1 - p_e)}{n_e}}}{p_c}$$

Figure 7: Impact of Sample Size on Confidence Interval



The solid line represents the confidence interval around the first example in which there were 100 patients per group and the number of events in the active and control groups were two and four, respectively. The broken line represents the confidence interval around the second example in which there were 1000 patients per group and the number of events in the active and control groups were 20 and 40, respectively.

Figure 8: Precision vs Accuracy



Good precision
Poor accuracy

Poor precision
Good accuracy

Good precision
Good accuracy

Poor precision
Poor accuracy

Figure 9: Eg. Plot of Blood Pressure Measurements; Precision vs Accuracy

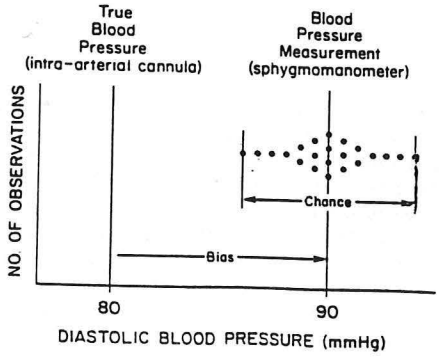


Figure 10: Relationship Between Confidence Interval Width and Level of Confidence Specified

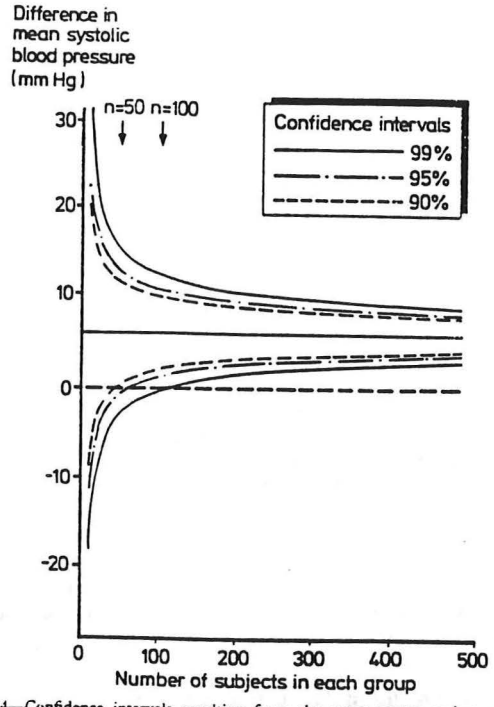


FIG 4—Confidence intervals resulting from the same means and standard deviations as in fig 1 and given in the worked example, but showing the effect on the confidence interval of sample sizes of up to 500 subjects in each group. The two horizontal lines show: --- zero difference between means, —— study difference between means of 6·0 mm Hg. The arrows indicate the confidence intervals shown in figs 1-3 for sample sizes of 100 and 50 in each group.

Confidence intervals thus provide information about the precision of the study results (the point estimate), but also, the magnitude of the treatment effect (as will be discussed shortly), and convey statistical significance. For an effect that is a difference, if either boundary of the confidence interval overlaps 0, then the possibility of zero, or no difference, is still plausible, and therefore the result is not statistically significant. Analogously, for a ratio, if one of the boundaries of the confidence interval overlaps one, the same is true. Keep in mind that confidence intervals convey information about the precision of the result due to sampling variation, but cannot control for issues related to non-sampling errors, such as biases in study design, conduct or analysis (79). Confidence intervals can be calculated for many statistics, including, means, proportions and their differences, regression slopes, and relative risks (70,79). Unfortunately in Pocock's study they were reported in only 6 of 48 trials reviewed (10).

In the GUSTO trial, the CI for the RRR in the comparison of t-PA and the combined SK groups was 5.9 to 21.3% (point estimate of 14%). This interval corresponds to a span of absolute risk reductions of 0.42% to 1.52% (73). Put another way, the results of the GUSTO trial are compatible with a true difference as low as 0.42% and as high as 1.52%. Although the true difference is likely to be closer to the point estimate of ARR of 1%, it may be more extreme in either direction at probabilities greater than 5%.

In summary then, how should the statistical analysis of a study be reported? 1) Show the raw data if the study small enough. 2) Provide the actual study point estimate (eg. mean, difference in means). 3) Provide the CI for the point estimate. 4) Provide the specific p value. Each of these pieces of information contributes unique and valuable insight into the study results.

3) How should treatment effects in subgroups be assessed and what is the impact of multiple comparisons among subgroups?

a) What is subgroup analysis?

Subgroup analysis has become a confusing term because it has been used to refer to a variety of procedures, the common focus of which is the impact of covariates on the treatment effect demonstrated. It essentially involves comparing the treatment effects among subgroups of patients. *Proper* subgroups consist of patients with a given set of *baseline* characteristics that include those not affected by treatment (such as age and sex) and disease characteristics defined prior to randomization, as well as, occasionally, outcomes after the trial is completed (81). For example, in the GUSTO Trial, among patients who received t-PA, the treatment effect was examined in a subgroup defined by age, and dichotomized into age < 75 and age > 75 (2). *Improper* subgroups are those defined by patient characteristics measured after randomization, *and* potentially affected by treatment. In the latter, a particular treatment effect, such as side effects, noncompliance or no response to therapy, may influence classification to the subgroup which may then influence treatment effect in a manner not specific to treatment but to the patient characteristic (81).

b) What are the intentions of subgroup analysis?

These analyses are used to 1) examine the baseline distribution of covariates among the treatment groups, which has already been discussed in the context of randomization, 2) explore the influence of the subgroups on treatment effects (i.e. are there any interactions between the covariate and the effect of treatment which would be a prerequisite to 3), 3) adjust the overall treatment effect for differences among the subgroups, and 4) determine the effect of treatment in subgroups corresponding to the different values of the covariates, or patient characteristics (34). This latter purpose has been increasingly emphasized and is often the driving force for subgroup analysis. It represents the attempt to extend the generalizability of the trial's results to individual decision-making in "my patient" by looking at treatment effects in patients with specific characteristics that more closely resemble clinicians' individual patients. In other words, the response of the "average" patient to therapy is not necessarily the response of the patient being treated since patients with a specific disease often differ greatly from one another.

c) How pervasive is subgroup analysis?

In a survey of 45 clinical trials reported in three leading medical journals, Pocock and colleagues found at least one subgroup analysis that compared the response to treatment in different categories of patients in 51% of reports (10). What has the impact of subgroup analysis been on clinical practice? This question has been addressed by comparing treatment recommendations generated from early trials of new treatments based on subgroup analysis with treatment recommendations that would have been made had subgroup analysis been ignored, and then determining whether later trials confirmed

the earlier report of subgroup analysis. Yusuf reviewed 65 randomized trials examing the impact of beta-blockers in acute myocardial infarction. Many of these trials claimed benefits in particular subgroups. For example, an earlier trial claimed that treatment was beneficial in patients <65 years of age, but actually harmful in those > age 65 years. Most subsequent trials showed similar benefits of treatment among both age groups (81).

Results of subgroup analysis, then, can have a major impact on clinical care, especially when a particular category of patients is denied treatment or when ineffective or harmful therapy is given to a subgroup. In fact, at times the results of subgroup analysis have been more emphasized than the results of the overall treatment effect and more influential in their impact on clinical care.

d) Examining the impact of treatment effects in subgroups

Subgroup analysis using statistical tests of difference between response to treatment with a subgroup

Subgroup analysis, then, appears to be both informative, but potentially misleading. Even given a rigorous study design, the extent to which subgroup analyses should be done, or believed, is highly controversial. What are some of the concerns from a statistical perspective? Much of the problem stems from the methods of comparing the treatment effect in the subgroups with the overall treatment effect. Comparisons among subgroups most commonly take the form of examining the differences in the treatment outcome among the various levels of a subgroup, without relating it to the overall treatment effect. For example, in the GUSTO trial, the 30 day mortality among the patients age ≤ 75 years was compared among patients treated with each of the treatments. This analysis was then repeated for patients > age 75 years. In a typical trial, this analysis will be repeated multiple times for many other subgroups. The two other prespecified subgroup analyses in GUSTO were infarct location(anterior versus inferior) and time to randomization (2).

I've already addressed the statistical problem that arises when multiple comparisons are made. Table 16 (82) demonstrates that the probability of observing a significant (but not necessarily "true") treatment effect in at least one subgroup increases as the overall treatment effect and the number of covariated being tested increases. For 10 covariates, the probability of obtaining at least one significant difference in treatment effect within a subgroup varies from 0.61 to >0.99, depending on the size of the overall treatment effect. Table 17 demonstrates the impact of subgroup analysis based on astrological sign in the International Study (4). For those readers with strong beliefs in the influence of your astrological sign on your health, here's a study to support what you've already known! Patients born under Libra or Gemini were harmed by the treatment with aspirin. There are several methods available to adjust statistical significance for the number of subgroup comparisons made. A simple conservative approach, the Bonferroni correction, divides the overall significance level by the number of comparisons actually made. For example, if 10 comparisons were made, then the newly specified α to achieve statistical significance would be 0.005 (83).

Table 16: Probability of Obtaining a Significant Result in At Least One Subgroup, as a Function of the Overall Treatment Effect (Z) and the Number of Covariates Considered (N)

N	"Trend" in Overall Effect ($Z = 1.0$, $p < 0.32$)	"Significant" Overall Effect ($Z = 2.0$, $p < 0.05$)
1	0.05	0.32
2	0.09	0.53
5	0.21	0.85
10	0.37	0.98
20	0.61	> 0.99

Table 17: Subgroup Analysis of Astrological Sign vs Aspirin Treatment in "International Study"

	Vascular Mortality at Week 5		
	Asa (%)	Placebo (%)	Odds Decrease (% +/- SD)
Patients born under Libra and Gemini	11.1	10.2	8 (adverse) (NS)
Patients born under other "birth" signs	9.0	12.1	26 (+/- 5) ($p < 0.00016$
Overall results	9.4	11.8	23 (+/- 4) ($p < 0.00001$)

Simply correcting for multiple comparisons among subgroups in a trial that shows an overall treatment effect leads to comparisons concerning individual subgroups that are lacking in power for the demonstration of any "true" differences that may exist, due to the more stringent α adopted as the threshold for statistical significance. The sample size of a well-designed clinical trial is optimally designed large enough to ensure a high probability, or power, of detecting a clinically important overall difference between the treatment groups. Generally, this sample size is insufficient to detect effects within even relatively large subgroups and are very unlikely to detect interactions. In fact, as we'll see shortly, most trials have insufficient size to detect overall effects.

I should mention that epidemiologists have expressed little enthusiasm for such formal correction methods for multiple comparisons due to their reduction in power (44,84). Briefly, the rationale for this position contends that the statistical formulation of the multiple comparisons problem is predicated on the universal null hypothesis. This means that in order to reject the alternative hypothesis', that some associations are present in the data, one first of all assumes that if only purely random processes govern the variability of all the observations in hand, then a certain number of significant associations would be found by chance. However, for a large number of associations, it does not make logical sense to assume that none of them could be "true". The validity of the data produced by a single study is logically independent of the motivations for having asked a question. In addition, the timing of articulating the question (a priori or a posteriori) is, of itself, incapable of influencing the results. Distinctions in timing are considered to merely reflect the investigators' knowledge at a given point in time that relates to the cumulative evidence in support of the hypothesis. Rothman recommends that each specific hypothesis be evaluated individually, even if there are many, according to the quality of the results of the study and their compatibility with other evidence only with respect to that specific hypothesis. Limited enthusiasm for pursuing certain associations would be recommended based on the lack of cumulative evidence in support of the hypothesis. I will defer further discussion of this issue and leave it to the interested reader to pursue a more comprehensive treatment of this issue.

Subgroup analysis using statistical tests for interaction

A better method of assessing differences in outcomes between levels of a subgroup are tests for interaction (44,81,85). An interaction exists between treatment and risk group (subgroup) if the true treatment effect is different among the various levels of the risk subgroup. This approach generally takes the form of statistically evaluating the outcome (represented as the dependent variable) in the face of treatment *and* the subgroup, represented with its different levels, as independent variables in the equation. This is often referred to as bivariate analysis, since two variables (treatment and a subgroup) are examined with regard to their relation to outcome. This analysis is repeated for each subgroup of interest. The advantage of this

comparison is that it relates the difference between the subgroup levels to the overall treatment effect. In the study by Pocock examing the methodological quality of reported clinical trials, subgroup analysis was performed in about 50% but only 3 of these studies used tests of statistical interaction as the method of testing for differences (10). *In addition, the reports tended to make overall treatment comparisons insufficiently prominent.*

Not all interactions are created equal though. Peto (86) has stressed the important distinction between interactions that are quantitative (the direction of treatment effect is the same but differs in degree) and those which are qualitative (the direction of treatment effect is not the same among different subgroups). These latter interactions are much more serious because they imply that the treatment effect is beneficial in some subgroups but harmful in others. Although important to discover, "true" qualitative interactions are usually very unlikely to occur in most circumstances. Most formal statistical testing designed to detect them lack power except for the most extreme interactions. The lack of power stems from the greater variance in the treatment difference (δ) than for α or for β. Similarly, and most unfortunately, the same is the case with tests for quantitative interaction. Therefore, even if the observed interaction is statistically significant, it is more likely to be due to chance since *some* variations in the true treatment effect are naturally to be expected between different risk groups (82). In other words, tests of interaction also face problems due to multiple comparisons, and they also lack power to detect "true" differences.

Multivariate techniques can also be used to assess the impact of important cofounders, i.e. subgroups that influence treatment effects in addition to the treatments. These methods have the advantage of assessing the impact of each subgroup in the context of simultaneous consideration of all other subgroups of interest and need not be dependent on positive tests for interaction for their entry into the model, since the lack of power in tests for interaction may erroneously lead to excluding confounders from consideration. For example, Lee and coworkers performed a simulated randomized trial in coronary artery disease to illustrate the need for clinical judgement in using modern statistical methods in assessing the therapeutic claims of studies investigating complex diseases (87). 1073 consecutive, medically treated coronary artery disease patients from the Duke University data bank were randomized into two groups, designated "treatment group 1" and "treatment group 2". The groups were 'comparable' at baseline as assessed by statistical tests for significant differences. As expected, no overall difference between the two groups, who were essentially 'treated' with the same therapy but were merely arbitrarily designated into separate groups, were found. Interestingly, subgroup analysis revealed that there was a difference in survival for those 397 patients who had three-vessel coronary artery disease and abnormal left ventricular contraction when group 1 patients with this characteristic were compared with group 2 patients. However, multivariable adjustment procedures that incorporated a number of known confounders (the baseline distribution of which had not been statistically significantly different) revealed that this difference resulted from the combined effect of small but cumulative imbalances in a number of important confounders between those patients in group 1 versus group 2. (Multivariable methods will be discussed further under "Adjustment of the overall treatment effect for confounders"). Other methods for subgroup analysis have been developed but are beyond the scope of this discussion (88).

What is the bottom line with regard to subgroup analysis? Since any differences in treatment effect among levels of a subgroups are most likely due to chance, and since the power of subgroup analysis for the detection of " true" differences is quite limited, many experts recommend that subgroups of interest should be defined a priori. These groups should be selected based on known pathophysiology of the disease being studied or prior reported subgroup analyses that suggested differential treatment effects and they should be limited in number (81,85). *The assessment of any difference in treatment effect among subgroups should be evaluated by tests of interaction, preferably carried out with multivariate methods that evaluate the combined impact of all important prognostic factors on treatment effect simultaneously.* Ultimately, this approach underscores the realization that most clinical trials can at best answer a very small number of questions, perhaps even only one, with confidence.

Subgroups defined a posteriori should generally be considered to be hypothesis generating and require confirmation in another clinical trial these hypotheses concerning subgroup effects should have strong biological rationale. In addition, hypotheses in ordered subgroups should generally respect the natural ordering (81). An example of this is presented in Table 18 from the GUSTO trial, where the timing of thrombolytic therapy is examined with regard to treatment effect (2). The reduction in mortality was no longer statistically significant for treatment > 2 hours after on set of chest pain. One way of interpreting

this data, based on the statistical significance of difference in effect among the subgroup categories, is that treatment with t-PA within 4 hours of the onset of chest pain saves lives, but that treatment beyond 4 hours does not, since the odds ratios associated with longer duration's of treatment were not statistically significant. A more cohesive way to interpret this data would be to preserve the graded relationship between time of administration and impact on mortality that would be suggested by the proposed underlying mechanism which related time to patency and mortality due to MI (81).

All associations examined with respect to subgroup analysis should be reported, both negative and positive, in conjunction with the odds ratios and associated confidence intervals. Either adjustment for multiple comparisons should be performed or a comparison of the expected number of positive associations with the number actually observed should be reported. Any credence given to these subgroup differences should remain conservative and based on prior cumulative evidence. This approach is not inconsistent, then, with Rothman's recommendations. Eliminating the role of adjustments for multiple comparisons generally leads to conservative conclusions based on most subgroup analyses due to the need to weigh the validity of subgroup differences based on prior evidence. Yusuf has demonstrated that had these more stringent rules been followed, the numerous reports of a mortality benefit in patients treated with a beta-blocker would not have been supported (81).

Table 18: Relationship Between Time to Treatment and Mortality in the GUSTO Trial

Time to Treatment (hours)	% of Patients	30 day Mortality (%) t-PA	SK
< 2	27	4.3	*5.4
2 - 4	51	5.5	6.7
4 - 6	19	8.9	9.3
> 6	4	10.4	8.3

*Indicates statistically significant.

However, it is often not apparent how many hypotheses were formulated and tested (since some may not have been reported) or even of those reported, which hypotheses were specified in advance. The prudent stance is to rely more on the overall results to indicate the likely "true" effect in a particular subgroup, rather than on actual observations, unless the trial was specifically designed to have sufficient power within subgroups of interest. Despite all this debate, clinicians are faced with the practical task of having to make the best use of the data before them. Oxman and Guyatt (85) have proposed criteria for deciding whether apparent differences in subgroup response are real many of which we have already discussed. Keep in mind that these criteria assume that the underlying methodology of the study is sound.

1. Is the magnitude of the difference clinically important?
   Given the extent of biologic variability, it would be surprising not to find interactions between treatment effects and various other factors. The litmus test in answering this question is whether the size of the difference is sufficient to lead to different clinical decisions for the different subgroups compared. In general, the larger the difference between the effect in a particular subgroup and the overall effect, the more plausible it is that the difference is real. A problem arises when a large number of comparisons are made and only the most extreme differences are reported. This problem is compounded when treatment effect is only modest. Authors should report how many comparisons were made and how they decided which ones to report.
2. Was the difference statistically significant?
3. Did the hypothesis precede rather than follow the analysis?
4. Was the subgroup analysis one of a small number of hypotheses tested?
   For example, one study investigating the role of digoxin in congestive heart failure looked at 16 variables, but we can't always know how many were looked at. BHAT: 146 comparisons. Overall pattern of effects approximates a normal distribution.

5. Was the difference suggested by the comparisons within rather than between studies?
   Since patient and measurement characteristics in other studies would be expected to be more variable than in just one study, and thus less comparable, comparisons of subgroups between studies would be expected to be more likely to lead to chance differences than comparisons among subgroups within a study.

6. Was the difference consistent across studies?
   Replication of an interaction in another independent, unbiased study is compelling evidence for a "real" difference but the power of the studies being analyzed must be taken into consideration as well as the differences between the studies (in setting, population, definitions of outcomes, etc.).

7. Is there indirect evidence that supports the hypothesized difference?
   This criteria addresses the biologic plausibility of the difference. However, researchers can be very imaginative in postulating potential mechanisms to explain incidentally discovered relationships. Other indirect evidence can also help to support the hypothesized relationship, such as results of other related (intermediary) outcomes, results in different populations (eg animal), and results of similar interventions.

   How well did the GUSTO trial conform to these criteria? I've already mentioned several subgroup analyses reported by the GUSTO investigators. They prespecified a priori three subgroups of interest age, location of infarct and time to randomization, based on prior evidence of their potential impact on the outcome. The analysis examining age dichotomized into two categories, patients age $\leq$ 75 years and age > 75 years, revealed a significant P value for the difference in the former group ($p < 0.05$) but not in the latter ($p > 0.05$). Subsequent interpretations of the GUSTO trial have focused on this lack of statistical significance for treatment effect in the older age group (41). The GUSTO investigators have responded to the claim that t-PA is not effective in reducing mortality in patients age > 75 years quite appropriately by pointing out that a test for *interaction* between age and treatment was performed and not found to be significant (40). In addition, they point out that the overall treatment effect is *smaller* in patients $\leq$ 75 years (absolute difference of 1.1%) than for those >75 years (absolute difference of 1.3%) and found that the treatment effect in each of these subgroups is not significantly different from the overall treatment effect of 1% when tested formally for statistical significance.

   The conflicting results of these comparisons makes it imperative that the reader understand what is being compared and what methods are being used to make the comparison. Although a limited number of subgroup analyses were prespecified, which reduces the impact of multiple comparisons, the test for interaction and the specific comparison of the subgroup treatment effect with the overall effect warrant a conservative interpretation as to whether a "true" difference exists in the efficacy of treatment effect due to age. Although as discussed previously, the lack of statistical significance for treatment differences in the older age group may be due to underpowering of the study for examining this association since only 12% of enrolled patients were older than 75, not being able to reject the null hypothesis is not the same as proving there is no difference.

4) *What, then, can be concluded from a "negative" study?*

When a study is reported as "negative", it really means that the P value for the difference between the treatment and control groups did not reach statistical significance. *When we fail to reject the null hypothesis, what can be concluded about a possible difference between the treatment and control groups?* Does it mean that we "accept" the H_o and imply that there is a 5% chance of being wrong in concluding that there is no difference? When the H_o cannot be rejected, there is the potential to make a Type II error which is closely related to the sample size, as well as to other factors. Table 19 demonstrates the relationship between Type II error (β) and sample size. As sample size increases, β decreases. Like α, β should be prespecified in planning a study, so that an adequate sample size can be determined that will be needed to detect a difference of an expected magnitude. Unfortunately, too often sample size has not been planned for in this manner. In the study by Pocock, sample size was

planned for in only 11% of the studies (10). The GUSTO trial exhibited exemplary planning with regard to sample size. The methods section indicated that the suede was designed to detect a 1% absolute reduction in mortality in the t-PA group as compared with the SK group, that the expected baseline mortality in the "control" group (SK + SQ heparin) was 8%, and specified an α of 0.05 and a β of 0.1 (power = 90%). On the basis of these parameters, 41 000 patients were needed.

Tables, which differ according to the type of statistical test, are readily available to assist in estimating the minimum sample to satisfy a variety of predetermined values for α, β and Δ (the estimated effect size). Unfortunately, planning for sample size is often either not done, or the appropriate sample size not achieved. This leads to an increase in the probability of the Type II error, or β. Frieman et al. reviewed the findings of 71 clinical trials that reported no significant difference ($P > 0.05$) between the compared treatments (69). The investigators for these trials incorrectly interpreted their data as indicative of no effect, when, in a great majority of the trials reviewed, the data were consistent with a reasonable strong effect of the new treatment. The latter was determined by calculating the 90% confidence intervals for each of the trials, which are plotted in Figure 11. 65% of the trials reported a reduced mortality in the treatment group and the plot of the confidence intervals show that they overlap 0 but are skewed to favoring treatment.

Figure 11: 90% CI Limits for the True % Difference for the 71 Trials

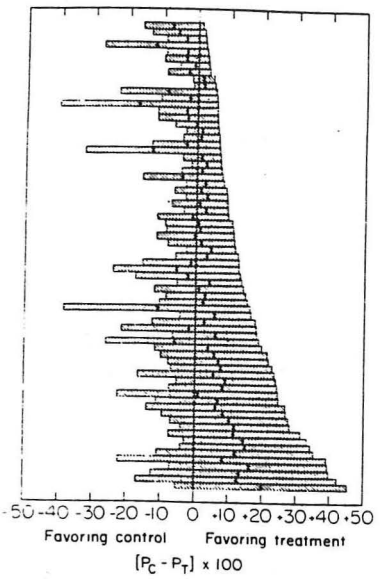


Figure 2. Ninety per Cent Confidence Limits for the True Percentage Difference for the 71 Trials.
The vertical bar at the center of each interval indicates the observed value, $\hat{P}_c - \hat{P}_t$, for each trial.

Freiman analyzed this data in another way to better illustrate the large Type II errors associated with these studies (69). She calculated the 25 % and 50% relative reductions in adverse outcome, which are commonly used standards for evaluating the efficacy of therapy, as well as the associated β, for each of these trials. In 80% of these trials, the confidence interval included the 25% relative reduction in endpoint (i.e. 20% were not designed to detect even a 25% relative difference in outcome). In 49% of

these trials, the confidence intervals included the 50% relative reduction in endpoint (i.e. 51% were not designed to detect a relative reduction in outcome of 50%).

The rationale for Freiman's approach is as follows. The probability of Type II error is not analogous to the probability of a Type I error. A high risk of a Type I error (false positive conclusion) does not guarantee a low risk of a Type II error (false negative conclusion). The probability of Type I error is calculated on the basis that the Ho is correct ($P_c - P_t = 0$). The probability of Type II error is based on the H_o being false ($P_c - P_t \neq 0$). But if $P_c - P_t$ is non zeo, then there is an infinity of possible values for this difference. For each value of the difference, there is probability of Type II error. Figure 12 depicts the relationship between Type II error (β) and the difference between treatments ($P_c - P_t - \Delta$) in a clinical trial. This curve of β as a function of Δ is known as the operating characteristic of the test. It highlights the probabilities, the β's, of missing a 25% or a 50% reduction in an adverse outcome due to treatment, in the study depicted, were associated with $\beta = 0.77$ $\beta = 0.42$, respectively. Tables 20 and 21 represent comparisons between the event rates in the treatment and control groups to estimate whether the study was powered to detect 25% and 50% reductions in outcome (67).

Figure 12: Operating Characteristic Curve of a Representative Clinical Trial

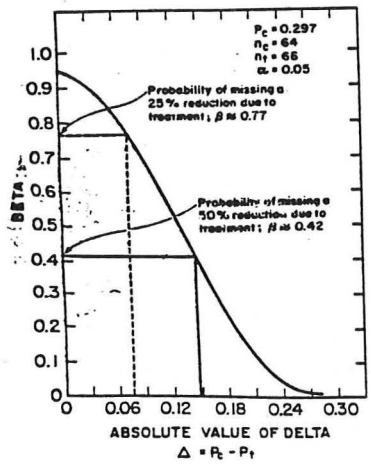


Table 19: Was the Trial Big Enough to Show a RRR ≥ 25% if it Had Occurred?

Observed rate of events in the control group	Observed rate of events in the experimental group																		
	.95	.90	.85	.80	.75	.70	.65	.60	.55	.50	.45	.40	.35	.30	.25	.20	.15	.10	.05
.95		14	27	68	391														
.90		11	18	38	110	1057													
.85			14	25	54	185	4889												
.80			11	18	33	78	326												
.75				13	22	44	112	635											
.70				11	16	28	57	165	1524										
.65					13	20	35	75	250	6349									
.60					10	15	24	43	99	402									
.55						12	17	28	53	132	722								
.50							13	20	33	65	180	1607							
.45							10	15	22	38	79	254							
.40								11	16	25	44	98	381						
.35									12	18	28	50	121	634					
.30										13	19	30	57	1296					
.25										10	13	20	33	64	196	4537			
.20											10	14	20	34	71	261			
.15												10	14	20	35	78	371		
.10													10	13	20	34	80	589	
.05															12	17	30	74	1245

Trials up here have risk reductions of 25% or more

Trials down here needed fewer than 10 patients per group

*This table displays how many patients would be needed per group, to be confident (α = .05, 1 tailed) that you have not missed a 25% relative risk reduction in the experimental group. (A reduction from .40–.30 is a 25% relative risk reduction, as is a reduction from .20 to 0.15.)

Table 20: Was the Trial Big Enough to Show a RRR ≥ 50% if it Had Occurred?

Observed rate of events in the control group	Observed rate of events in the experimental group																
	.70	.65	.60	.55	.50	.45	.40	.35	.30	.25	.20	.15	.10	.08	.06	.04	.02
.98	14	24	50	165	5803												
.95	12	19	37	102	921												
.90		14	26	58	236												
.85		12	19	38	108	995											
.80		10	15	27	63	256						Trials up here have risk reductions of 50% or more					
.75			12	21	41	116	1059										
.70				16	29	66	268										
.65				13	22	43	120	1082									
.60				11	17	30	68	270									
.55					13	22	42	119	1059								
.50					11	17	30	66	260								
.45						13	22	42	113	987							
.40						11	16	28	62	239							
.35							13	20	38	102	867						
.30							10	15	26	55	205						
.25								12	18	33	86	699					
.20									13	22	45	160					
.15		Trials down here needed fewer							10	15	26	64	482				
.10		than 10 patients per group								11	17	32	102	254	2017		
.08											14	25	66	131	453		
.06											12	20	44	76	179	1313	
.04											10	16	31	47	87	274	
.02												12	22	30	47	97	561

*This table displays how many patients would be needed per g... ...05, 1-tailed) that you have not missed a 50% relative risk reduction in the experimental group (A reduction from .40 to .20 is a 50% relative reduction, as is a reduction from .20 to .10.)

The hallmark of these studies was inadequate sample size and there is no surer way to demonstrate no difference between treatments than to underpower a study. This is particularly a problem when a new treatment is being evaluated as a replacement for standard therapy, not because it is presumed to be more efficacious, but because it is less costly, has fewer side effects or is more convenient than the standard therapy (70). An underpowered study for this purpose can all too easily lead to the substitution of an inferior therapy for the standard therapy. Therefore, failure to reject the H_o is not equivalent to its acceptance and proving that a true difference does not exist.

5) How should overall treatment effects be modified to accomodate the effects of prognostic factors?

What is the goal of adjusting the treatment effect? Adjustment of the overall treatment effect is intended to remove the effects of confounding factors since, as we have seen, the distribution of prognostic factors may be unbalanced between treatment groups due to chance alone . When adjustment is performed after a study is completed, it generally involves the use of multivariable methods. The four main multivariable methods used in the medical literature are presented in Table 21, which highlights the differences in the methods with respect to the type of outcome variable appropriate to each. The technical details of these methods are described elsewhere. The use of these methods continues to grow and even several years ago, 20% of all studies used one of these methods to adjust the overall treatment effect (89).

Table 22 : Comparison of Commonly Used Multivariable Methods

Multivariable methods	Outcome variable	Example
Multiple linear regression	• Continuous	• Blood pressure
Multiple logistic regression	• Binary event at a fixed point in time	• Alive vs dead
Discriminate function analysis	• Category or group to which a patient belongs	• Race
Proportional hazards analysis (Cox regression or survival curve)	• duration of time to occurrence of a binary event	• Death

What do multivariable methods represent when applied to medical data? The basis of "medical data" rests on the description of biological systems. Most mathematical models that we use were not developed to specifically describe these systems, and in fact, most systems do not conform to the mathematical models we use to describe them. There is no "inherent" link between these models and any "reality" they represent in the biological systems. Nonetheless, the use of these mathematical models abounds, and some understanding is necessary. However, the reader is cautioned to have some measure of skepticism when examing the output of these models, rather than viewing them as a "magical" black box yielding the "truth". Statistics should serve reason, rather than replace it.

In general, multivariable analysis mathematically relates independent variables X_1, X_2, X_3,...X_n to an outcome variable via a model expressed as a combination: $G + b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + \varepsilon$, where G is a function arranged in various mathematical forms; b_j is a regression coefficient indicating the impact of each X_j variable on the outcomes; b_0 is the intercept term, which is usually included in the model (33). If a particular $b_j = 0$, then the variable X_j has no impact on the outcome; a positive value of b_j indicates that higher values of X_j are associated with the outcome expressed as G; and negative values have the reverse effect. A random variable ε is an "error" term representing the increment by which any individual G value deviates from the calculated value of G. The validity of an adjustment technique depends on the correctness of the assumed mathematical model. Another serious problem arises if the covariates are measured after the treatment has been administered, since the possibility exits that the covariates have been influenced by the therapy. An example of this is when the rate of adherence to therapy differs.

The most commonly used multivariable method is multiple logistic regression (89). What is the appeal of multiple logistic regression? From a mathematical standpoint, it is extremely flexible and easily used. It also lends itself especially well to a biologically meaningful interpretation. Since many outcomes of interest in clinical research have only two outcomes. Each regression coefficient turns out to be the odds ratio for the variable with which it is associated. Despite the restriction to only two possible outcomes, the logarithmic function distributes the odds ratio over a large range of values, due to the asymptotic nature of the function (90).

Multivariable methods can be used for at least five purposes (89). The overwhelming majority of the time, they are used to quantify the risk of individual variables for their specific effects among other independent variables, as has been discussed. In a study assessing the frequency of and reasons for the use of multivariable methods, 75% of the publications used multivariable analysis to quantify risk estimates

reported as regression coefficients, odds ratios, or relative risks for individual variables (89).Another use of multivariable methods involves the confirmation of the risk factors identified in the bivariate analyses as independent risk factors i.e. do they retain their importance in the simultaneous context of the other variables? An example of this use is presented in Table 22. In this study assessing risk factors for death due to coronary artery disease, when risk factors were assessed individually for their impact on the outcome, many of them were found to be independent (chi square value > 3.84)(87). However, when these factors were combined in a multivariable model, only two remained independent. Multivariable methods are also used to confirm the results of non-regression analyses, such as standardization, as well as screening for significant risk factors when the total number of variables may make bivariate analysis too time-consuming and difficult to assimilate. Finally, multivariable methods can be used to create risk scores that combine the impact of multiple variables into a single risk score that is used for predicting outcomes of individual patients.

Table 22: Comparison of Individual and Joint Prognostic Significance of Baseline Variables

	*Individually	*Jointly
Treatment	5.4	2.4
History of CHF	36.1	3.8
Cardiomegaly on CXR	15.7	0.3
Resting ST-T-wave abnormalities	8.4	2.2
Mitral insufficiency	25.1	2.7
AV O_2 difference	52.0	11.1
LV diffusely abnormal contraction	17.3	0.3
Left main stenosis	6.9	5.7

A number of issues surround the use of multivariable methods (91). Controversy exists regarding how the variables (risk or prognostic factors) should be selected for entry into the model (92-95). Variables can be selected by choosing those in which there is a large disparity in their distribution between groups (based on a statistical test of their difference), tests of interaction, or based on evidence-based judgement as to which factors are known to strongly influence the outcome, and others. Each strategy has its merits and disadvantages. In general, either tests for interaction, or simultaneous inclusion of all factors considered to be important in affecting the outcome of interest are preferred. In addition, the type of model used and how the model is built can affect both the direction and the magnitude of the results. This complexity in the methods and the variability that can result in what is reported has lead some to refer to multivariable methods as "science by sleight-of-hand" (91).

Given the increasing application of multivariable methods in the medical sciences and its associated complexity, minimum criteria have been established to evaluate the appropriateness of the multivariable methods used in a study. Table 23 presents criteria that have been proposed by Cancato, et al. as a minimum set of guidelines by which studies using multivariable methods are to evaluated with regard to their execution, interpretation and reporting of multivariable methods (89). I highly recommend this paper to readers for a more thorough discussion of these criteria. Cancato and colleagues assessed the quality of multivariable methods use by examining studies using one of the four methods previously described that were published in the Lancet or New England Journal of Medicine sometime between 1985 and 1989. The rotation of these criteria, as presented in parentheses in the first column of Table 23, is exceedingly common.

Table 23. Problems and Issues in the Application and Reporting of Logistic Regression and Proportional Hazards Analysis

Problem or issue (% invalidated)	Description
Problem	
Overfitting of data (42%)	Fewer than 10 outcome events per independent variable in the model
Nonconformity to a linear gradient (29%)	Nonconstant impact of variables in different zones of ranked data
Nonproportional risk (81%)	Violation of assumption of proportional hazard function over time (in the proportional hazards method)
No report of tests for interaction (73%)	Check not mentioned for interactions between independent variables
Unspecified coding of variables (84%)	Unknown classification or codings for independent variables
Unspecified selection of variables (86%)	Unknown method of selecting among candidate independent variables
Issue	
Collinear variables	Independent variables with high correlation to one another
Influential observations	"Outlier" observations that have substantial impact on results
Validation of the model	Separate method of confirming analytic results

A. How Did the GUSTO Perform with Regard to These Criteria?

The original GUSTO report presented a survival curves for each of the treatment groups, and compared them with proportional hazards regression. The investigators indicated that logistic regression was used to assess the consistency of treatment effects for age, location of infarct and time to treatment. In applying the above criteria, there were sufficient numbers of outcomes events to analyze these three subgroups, even when subcategorized, as they were. Interactions for these three variables were reported and they were prespecified. The remainder of the criteria were not addressed, nor was the process for building the regression model described. Purportedly, the impact of the three prespecified prognostic factors were adjusted for in the reported overall treatment effect, although the "unadjusted" treatment effect was not reported.

B. How Precise was the Estimate of the Treatment Effect?

The discussion regarding precision and its application to the GUSTO trial has been presented in conjunction with "What Were the Results of the Study?"

III. WILL THE RESULTS HELP ME IN CARING FOR MY PATIENTS?

A. Can These Results be Applied to My Patient Care?

1. Are the patients similar to my own? (20).

The test for this criteria is whether your patient would have been enrolled in the trial if she had been there, i.e. does she meet the inclusion criteria without violating any of the exclusion criteria? This criterion may be too restrictive as clinical trials generally must reduce the complexity that is usually characteristic of clinical practice to provide for better control (12). Alternatively, you could ask, is there a compelling reason why the results should not be applied to my patient? If the answer is no, then generalization of the study's results to your patient seems reasonable. This criteria is also closely related to the report of outcomes for subgroups of patients, since subgroups with particular characteristics may be more representative your patient. The appropriate integration of subgroup analysis into clinical practice has already been discussed.

I have already mentioned the relatively low risk nature of the patients studied in the GUSTO trial. Patients at higher risk may not experience the same net benefit from treatment. I've also allude addressed whether the treatment effect is to be interpreted differently in patients age $\leq$ 75 years as opposed to age > 75 years. A particularly important characteristic of the GUSTO trial that limits its generalizability greatly is the graded relationship between time to treatment and reduction in mortality. This relationship is strongest within the first two hours and falls off thereafter. Most of the patients in the GUSTO trial presented within 2-3 hours of chest pain and received thrombolytic therapy in the third and fourth hours. In usual clinical practice, patients present later, about 4 hours after the onset of chest pain (96).

B. Were all Clinically Important Outcomes Considered?

Were outcomes evaluated that are important to patients? Studies often evaluate surrogate outcomes, or physiological markers, in lieu of important outcomes because they are more feasible (less cost, effort, etc.). Although it is tempting to assume that a marker for an event would be a reliable predictor for the event, this is not always the case. For example, since the use of antiarrhythmic drugs following myocardial infarction reduced ventricular depolarizations in the short-term, it was assumed that they would reduce the occurrence of life-threatening arrhythmias in the long-term. However, the trial investigating the long-term efficacy of these drugs found an increase in mortality associated with their use (97).

It is also important to interpret the report of a favorable benefit with regard to one outcome cautiously unless it is clear that there were no deleterious effects on other outcomes. For example, a reduction in cardiovascular deaths might be offset by an increase in noncardiovascular deaths. The overall mortality, death due to all causes, would most appropriately capture the impact of therapy. Increasingly, quality of life, as defined from the patient's perspective, is being emphasized as perhaps the most important outcome. Chemotherapy for cancer may prolong a patient's life, but if it does so at the expense of the patient's quality of life, the treatment may not be perceived as beneficial by the patient.

Since thrombolytic therapy is associated with an increased risk of bleeding, including hemorrhagic stroke accompanying an MI, the GUSTO trial also appropriately evaluated these endpoints. It reported an absolute excess of hemorrhagic strokes of 0.2% (P = 0.03) for t-PA as compared with the combined SK groups (2 strokes per 1000 patients treated). However, comparison of t-PA with SK + SC heparin reveals an absolute difference of 0.33 % in all strokes and 0.18% for hemorrhagic strokes, which translates to 3.3 and 1.8 excess strokes in the t-PA group, respectively. Rather than combining the 30-day mortality with one of these endpoints, the GUSTO investigators chose to use a combined event rate that encompassed 30-day mortality with disabling strokes (which were defined as substantial limitation of activity and capabilities or inability to live independently or work) instead. This endpoint is rather subjective in nature, and trivializes the impact of lesser degrees of deficit on the quality of life of patients. In choosing this combined endpoint, the number of excess strokes was less impressive for t-PA compared to SIC and SC heparin. In general, the incidences of other bleeding events tended to favor t-PA.

Subsequent to the preliminary report of GUSTO, the results of the angiographic substudy were published (46). As already mentioned, this substudy examined angiographic patency after thrombolysis in about 6% of the original GUSTO patients. The importance of this study is that it addresses the validity of the biologic mechanism proposed to explain the survival benefit associated with t-PA. These investigators confirmed that t-PA opened arteries more rapidly than SK, but that the patency rates were equivalent at 3 hours, the so-called "catch up" phenomenon. Integration of the results of GUSTO with other thrombolytic trials have lead some experts to conclude that t-PA opens up arteries about 45 minutes earlier than SK (73). The reduction in mortality per hour associated with lysis begun at different times, based on the fibrinolytic therapy trialists (FTT) meta-analysis of the results from the large randomized trials of thrombolytic therapy versus no treatment, were calculated (98). Lysis in the first hour saves 39/1000 lives and in the second hour saves 30/1000, resulting in a net benefit of only 9/1000 lives if treatment is begun in the first hour (one hour earlier). From 2.5 hours to 20 hours, the gain of lysis one hour earlier is only 1.6/1000 lives. The majority of GUSTO patients received thrombolytic therapy in the first 2 to 3 hours, and this may positively account for 10/1000 lives saved that was reported. However, according to the FTT

analysis, even if therapy at 2.5 hours is compared with therapy at 1.5 hours (one hour earlier), only 5/1000 lives are saved, which is substantially less than that reported by GUSTO. Additional skepticism must be brought to bear on the role of patency in mediating the survival benefit observed. Despite earlier patency in the t-PA group, left-ventricular ejection fraction did not differ among the four treatment groups (74).

C. Are the Likely Treatment Benefits Worth the Potential Harms and Costs?

The beneficial impact of treatment on patients has already been discussed as well as some of the major complications of therapy, bleeding and stroke. It is equally important to combine the benefit and harm in a quantitative manner to assess the net benefit of therapy. NNT can be used in such a manner, by directly integrating the NNT quantifying the "harm" of therapy with the NNT quantifying the benefit of therapy. A quantitative estimate of the combination of benefit and harm, the net benefit, can thus be produced. For example, therapy that is modestly effective, but without side effects may be more appealing than more efficacious therapy which carries significant risks of serious side effects. Tables 24 and 25 presents the net benefit resulting from integrating survival benefit and the harm of nonfatal strokes for the comparisons of t-PA with SK + IV heparin and t-PA with SK + SC heparin, respectively reported in GUSTO. Note that, as already discussed, the decision to combine the SK groups in the comparison with t-PA favors treatment with t-PA (10 additional lives saved per 1000 treated), whereas comparison with SK + SC heparin alone results in only 7 additional lives saved per 1000 treated patients, if *all strokes* are considered as opposed to just severely disabling ones. In fact, the P values for this comparison was 0.04, which means that the 95% confidence interval ranges from a near zero difference, to perhaps a 30% or greater relative advantage (73). Keep in mind the previous discussion regarding a likely exaggeration in stroke complications in the SK + SC heparin group due to the large proportion that actually received IV heparin, and the FTT estimation of the expected benefit of patency for the time after the onset of chest pain that most GUSTO patients experienced.

Table 24 : Comparison of Net Benefits of t-PA Versus SK + IV Heparin in GUSTO

Outcome	t-PA + IV heparin (Group 1)	SK + IV heparin (Group 2)	ARR (%) (1 vs 2)	*No. of lives saved /1000 treated (1 vs 2)
30-day mortality	6.3	7.4	1.1	10
†Or CVA	7.2	8.2	1.0	10
†Or hemorrhagic CVA	6.6	7.6	1.0	10
†Or disabling CVA	6.9	7.9	1.0	10

*if stroke counted as equivalent to death
† nonfatal strokes

Table 25 : Comparison of Net Benefits of t-PA Versus SK + SC Heparin in GUSTO

Outcome	t-PA + IV heparin (Group 3)	SK + SQ heparin (Group 1)	ARR (3 vs 2)	*No. of lives saved /1000 treated (3 vs 2)
30-day mortality	6.3	7.2	0.9	9
†Or CVA	7.2	7.9	0.7	7
†Or hemorrhagic CVA	6.6	7.4	0.8	8
†Or disabling CVA	6.9	7.7	0.8	8

*if stroke counted as equivalent to death
† nonfatal strokes

Finally, once the net clinical benefit has been determined, the cost of utilizing accelerated t-PA in place of SK must be considered. There have been a number of cost-effectiveness analyses , including one by the GUSTO investigators, that have addressed this issue (6,99,100) The conclusion of these analyses have varied, depending on the assumptions modeled. The cost-effectiveness of thrombolytic therapy is sensitive to the comparative net clinical benefit assumed, which is quite controversial, as we've discussed. The most useful way, then, of presenting the cost-effectiveness of t-PA versus SK + SC heparin is to present the cost associated with a continuum of ARR 's that might be entertained. Table 26 presents data in this manner from a cost-effectiveness analysis by Naylor and colleagues. If t-PA results in ARR of 2.5% or less, then the cost associated with its use would not be considered compatible with the threshold for use of other widely accepted health interventions, which is generally $100 000 per quality life year. However, the decision to accept the cost of a new treatment is both an individual one, and one that will be made from the perspective of public health policy. As yet, there has not been resolution between the tension experienced by these two different perspectives. As a basis of comparison, Table 27 presents the cost effectiveness ratios associated with other cardiac interventions. Of interest, the grading system of new treatments and technologies proposed by Laupacis (11) graded the use of t-PA versus SK to treat acute MI as E, which includes new therapies or technologies which are less effective than or as effective as the existing one but more costly.

Table 26: One-way Sensitivity Analysis for t-PA versus SK

ARR	Marginal Cost Per Short-Term Survivor
3.00%	$92 620
2.50%	$111 144
2.00%	$138 930
1.50%	$185 241
1.00%	$277861
0.50%	$555722

Naylor CD, et al. Can J Cardiol 1993;9:553-8.

*Table 27: Comparative Cost-Effectiveness Ratios Interventions to Reduce *CHD Mortality in Middle-aged Males*

Intervention	~ Marginal cost/life-year gained
Beta-blockers post-MI (55 yo male @ medium risk)	$5000
CABG for left main disease.	$9000
CABG for 3-vessel disease	$20 000
Treatment of severe HTN	$20 000
Treatment of moderate HTN	$45 000
CABG for 2-vessel disease	$120 000
Life-long cholestyramine (serum cholesterol = 6.85 mmol/L; average risk profile)	>$250 000

*CHD = coronary heart disease

How should the results of the GUSTO trial be placed into context?

A number of experts (73,74,75) have reviewed the GUSTO trial and summarized its advantages, its problems, and how it might be placed into appropriate context with the following:

1) The choice of thrombolytic therapy is much less important to ultimate survival than the delay in time to onset of treatment.

2) There is probably an estimated net benefit of 3 lives saved per 1000 patients treated (perhaps less) associated with the use of accelerated t-PA in the setting of acute MI over standard therapy. This adjusted estimate is based on the more appropriate comparison with the SK + SC heparin arm, the heavy contamination of this arm with IV heparin which would be expected to increase the hemorrhagic stroke rate, the inclusion of all strokes, and consideration of the likely benefit of achieving patency one hour earlier based on the FTT analysis (the apparent difference between t-PA and SK).

3) The small absolute differences in efficacy among thrombolytic therapies are not likely to be applicable to most patients with MI who usually present more than 4 hours after the onset of chest pain.

4) The more important issues to address now are how to achieve thrombolytic therapy faster and how to increase the utilization of thrombolytic therapy in eligible patients.

Osler said: "Let us agree that good clinical medicine will always blend the art of uncertainty with the science of probability. This has been expanded upon by Naylor: "But let us also hope that the blend can be more heavily towards science, whenever and wherever sound evidence is brought to light" (101). The goal of critical appraisal is not to dampen enthusiasm, but to further our understand of the limitations of our evidence. Having said this, let me finish with the following:

"Clinical medicine seems to consist of a few things we know, a few things we think we know (but probably don't) and lots of things we don't know at all. When evidence alone cannot guide clinical actions, some will espouse minimalism whereas others will favor intervention based on inference and experience." So goes the science of the art of choosing better treatment...