

Critical Care Clinics
An Evidence-Based Approach to Critical Care Medicine

Interpreting and Using Clinical Trials

Dr. Gordon S. Doig,
Division of Critical Care Medicine,
Department of Medicine,
University of Western Ontario

Mail to: London Health Sciences Centre,
375 South St.,
London, Ontario,
N6A 4G5

Phone: (519) 685-8500
Fax: (519) 432-7367
e-mail: gdoig@biostats.uwo.ca

Synopsis

This chapter provides a framework for quickly assessing the quality of a published clinical trial. It uses numerous previous publications of ‘users guides’ on critical appraisal as building blocks and combines them with new, important concepts published more recently in the primary clinical trials literature.

Introduction

The well-designed randomized controlled trial (RCT) is the most powerful tool available for the evaluation of the true benefits of care. The RCT has the potential to improve the quality of care and control costs through the careful comparison of alternative treatments^{2, 19}. Thus the RCT is the only tool that has the power to improve *both* the effectiveness and efficiency of care.

Given the importance of the RCT to modern medicine, there are many well-written texts dealing with the design, analysis and even the reporting of clinical trials aimed at the trialist^{2, 16, 20}. Despite the existence of these extensive methodological resources, there are numerous reports in the literature documenting design flaws and reporting deficiencies in published clinical trials^{9, 17}. Because of the RCT's importance to the dissemination and uptake of new therapies, some authors have written 'users guides' to assist the end user (the clinician) in reading and interpreting published trials^{3, 13, 14, 15}. Most recently, a version of 'users guides' has been published specifically for the Critical Care physician^{4,5}.

The most widely used users guides were published by the Evidence-Based Medicine Working Group^{13, 14} in 1993 and have not changed much in content since then. The purpose of this chapter is to provide a brief overview of the theoretical underpinnings of these users guides and to build on the original framework by incorporating some important recommendations published in the clinical trials literature since 1993. For a complete listing of the issues to be addressed when appraising a clinical trial, see Table 1.

Bias and the RCT

Bias is said to occur when the results or inferences of a research study deviate from *the truth*. In the context of clinical trials, *bias* can be defined as *anything besides our treatment or random chance that can modify the strength and/or direction of the association between*

treatment and outcome ⁶. Since the primary objective of a clinical trial is to evaluate and quantify the benefits of a novel experimental therapy, any true estimates of its effect can therefore only be achieved when all possible efforts have been made to minimize the impact of bias. There are 4 main design methodologies that can be used to minimize the effects of bias; 1) random allotment to intervention, 2) complete follow-up and reporting on all patients recruited into the trial 3) blinding and 4) the selection of unambiguous and appropriate end points.

If the trial under review does not adequately employ any one of the above four techniques, we should immediately be aware of the possibility that a bias may explain the findings the authors report. In other words, the results reported in the trial may be false. Therefore, when we are assessing the validity of a trial, the first set of issues we must address deal with potential sources of bias.

I. Is the trial's design valid?

1. Was the randomization technique adequately described?

The sole purpose of randomization is to remove bias from the allocation of patients to treatment groups. Since randomization is the single most effective way of reducing allocation bias, and there are many different approaches to randomization, the authors of the trial should stipulate explicitly which *specific method* is used (i.e. blocked, stratified, sealed envelope, central randomization), how randomization patterns were *concealed* from recruiting personnel and the exact *timing* of randomization within execution of the trial.

Details of the method are important so that the reader can ensure the randomization process was 'concealed' from the investigator who recruits the patients for the trial. If the randomization process is not concealed from the recruitment coordinator, he or she may be able to predict whether or not a potential subject will receive experimental or standard therapy *before*

the patient has given informed consent. In this situation, the trialist's preconceived bias about the therapies under investigation may influence his or her decision to approach the patient for participation.

One approach to determining if randomization 'worked' and actually reduced the potential for bias is to check if the experimental and control groups are similar with respect to important descriptive characteristics collected at the time of randomization. The trialist should make 'baseline comparisons' easy by providing a table of important prognostic factors so that the reader can tell if they were evenly distributed between the two groups. While statisticians will argue over the interpretation of p-values produced by comparison of randomized groups, most will agree that an imbalance in baseline characteristics should be explored for a potential effect on study outcome.

2. Were all patients who entered the trial properly accounted for and attributed at its conclusion?

Once informed consent is obtained, the patient is formally considered to have entered into the study. Often patients who give informed consent are not actually randomized and since this can reduce the generalizability of the results, the investigators should document all reasons for withdrawal. After consent has been obtained *and* the patient is randomized to receive an intervention, all attempts should be made to obtain complete follow-up and outcome assessment on *each and every patient*. Even if the patient did not actually receive the intervention to which they were randomized, their outcome and reasons for not receiving the intervention may shed important light on the actual effectiveness of the intervention.

The importance of *complete* documentation and reporting of patient loss and outcomes cannot be understated and was emphasized recently by the recommendations of the Consolidated

Standards of Reporting Trials (CONSORT) Group. The CONSORT members recommend that a flow chart, such as Figure 1, be used as a reporting device in all RCTs published in major peer reviewed journals ². If the trial authors do not provide an explicit, concise figure, then the information should be abstracted from the results section using Figure 1 as a template. If the information cannot be abstracted, the reader should be alerted to the possibility that the results reported are incomplete at best and should interpret the results presented with skepticism.

3. Were patients, health workers and study personnel ‘blind’ to treatment allocation?

Blinding reduces the chance that bias could enter the study through a systematic difference in supportive care, patient reported responses, outcome evaluations or other important, subtle or unrecognized ways *after randomization occurs*. Blinding is important enough that it should be used whenever it is not incompatible with optimal patient care. Reports of blinding techniques should be explicit and thorough enough to allow the reader to determine the different levels of blinding employed, such as: the patient, the healthcare team, outcome assessment panel other members of the research team and even the statistician performing the primary analysis.

By the nature of intensive care, most patients enrolled in clinical trials are blinded as to which treatment are they were allocated to. Blinding at the patient level is most important when outcomes are patient-reported, such as relief of symptoms, which is infrequent in ICU based studies. It should be noted that most quality of life (QoL) tools are ‘self-reported’ and somewhat subjective. When QoL tools are used in ICU trials, it becomes extremely important to blind the patient as to which treatment they were allocated to.

Blinding of the healthcare team is often under-reported in ICU trials, with the assumption being that the more objective the outcome being assessed, the less likely a lack of blinding will bias results. It should be noted, however, that even with the most objective outcomes, such as

mortality, the healthcare team plays a major role in determination since the major cause of mortality in the ICU is withdrawal of care. With more subjective outcomes, such as the diagnosis of ventilator-associated pneumonia, the healthcare team plays an even more important role. In these cases, a blinded outcome adjudication committee should be established to determine the diagnosis.

4. Apart from the experimental intervention, were the two groups treated equally?

A table should be provided to report the distribution of all co-interventions that occurred after randomization. The table should include important co-interventions such as antibiotic use, inotrope use, duration of ventilation, and the use of ancillary diagnostic testing. If there appears to be an important difference between the two groups with respect to these co-interventions, the reader must attempt to determine the mechanism that resulted in this difference.

First, the reader should closely examine the study protocol to determine if the recommended frequency of diagnostic interventions or co-interventions is different depending on the study arm the patient entered. For example, if the use of a new type of invasive monitoring device requires more frequent chest x-rays to determine exact placement, then this may help explain a finding such as an increase in the diagnosis of nosocomial pneumonia. If, however there are no differences in the study protocol with respect to diagnostic or supportive treatment recommendations, and the differences in co-intervention rates are statistically significant, we must start considering pathophysiological mechanisms.

In many studies investigating the use of steroids in sepsis as the primary intervention, it was found that antibiotic usage was more frequent in the steroid group. In a situation such as this, it is easy to rationalize that steroid use could increase the frequency of nosocomial infections, and it becomes reasonable to view such an increase as an important clinical outcome.

In situations where the pathophysiological link between the primary intervention and co-interventions is not so clear, then we should consider that this difference may be due to chance, and may actually be a source of bias.

5. Were the study outcomes appropriate?

Primary outcomes for clinical trials should be *unambiguous, easily measured, and clinically important*¹¹. The most clinically meaningful outcome for the majority of ICU-based trials is mortality. In certain specific ICU patient populations however, mortality is very low and length of stay, quality of life and other clinically meaningful endpoints can become important. Since trials may consider outcomes other than mortality, it becomes important to understand the difference between a *clinically meaningful outcome* and a *surrogate outcome*.

A clinically meaningful outcome is defined as a direct measure of how a patient feels, functions or survives, whereas a surrogate outcome is a laboratory measurement or a physical sign used as a *substitute* for a clinically meaningful outcome. For example, in the field of HIV research, an improvement in CD4 cell counts has long been accepted as a surrogate measure for improvements in duration and quality of life.

In order for a surrogate to serve as a valid outcome, two criteria must be satisfied; 1) the surrogate must be a correlate of the true clinical outcome and 2) the surrogate *must fully capture the net effect of treatment on the clinical outcome*²¹. Most surrogate outcomes used in the critical care literature have been shown to correlate with their true clinical outcomes, such as the finding that patients with a shorter duration of ventilation develop pneumonia less frequently. However, few, if any, have been shown to satisfy the second more important criteria. The only way for a surrogate outcome to demonstrate that it satisfies the second validity criteria is to have

an adequately powered clinical trial prove that an intervention that improves the surrogate measure also results in an improvement of the actual clinical outcome.

When surrogate measures such as decreased time to extubation, PaO₂, O₂ delivery, decreased FiO₂ requirements, cardiac index, pHi and others are used in the critical care literature, the reader should be aware that unless the authors provide convincing evidence of the validity of these surrogate outcomes, that there is a real chance of being misled by the results ⁸.

II. Is the analysis valid?

1. Was the primary analysis appropriate?

Interpreting the statistical analysis of a trial can be a daunting task, but it should be remembered that a good trialist uses design features to make analysis as simple as possible. If all baseline characteristics are in balance at randomization, then the analysis of a single center trial resolves down to the application of a simple chi-square test, t-test or nonparametric test, depending on the study outcome. However, if baseline characteristics are not balanced, then their potential to bias or confound the trial findings should be assessed using a multivariate analysis ¹.

Excluding randomized patients with observable outcomes from the final analysis is the most common mistake made in the analysis of clinical trials. To avoid this important source of bias, the first analysis that should be presented is *the intention-to-treat analysis*. The intention-to-treat analysis includes all patients who were randomized into the trial and provides the most unbiased, objective evaluation of the true effectiveness of the new therapy ¹⁰.

After the results of the intention to treat analysis are presented, the authors may build an argument based on nonadherence to the protocol, the availability of poor quality outcome information, incomplete dosing or other reasons, to exclude certain patients from a secondary 'efficacy analysis'. The reasons for exclusion from this secondary analysis should be defined *a*

priori and the reader must be able to ascertain that equal numbers of patients were excluded from each arm of the trial (see figure 1). If the primary intention-to-treat analysis does not support the findings of the secondary efficacy analysis or the reader is not satisfied that equal numbers of patients were excluded from each arm, then the findings of the efficacy analysis should be viewed as *hypothesis generating* and should be re-confirmed in a subsequent clinical trial before being accepted as valid.

The final approach to analysis that the reader may encounter is a *subgroup analysis*. In a subgroup analysis, a patient group of special interest (ex: entry APACHE II > 24) is identified and analyzed separately from the rest of the trial. For a subgroup analysis to be considered a valid confirmatory analysis, two criteria must be fulfilled: 1) the special interest in the subgroup must be specified *a priori* and 2) the subgroup must be identifiable from baseline characteristics. If interest in the subgroup is developed after the results of the analysis are revealed or if the subgroup is only identifiable by characteristics recorded after randomization (ex: tolerance to feeds, gram-negative blood culture, etc.) then the results should be viewed only as *hypothesis generating* and must be re-confirmed in a subsequent clinical trial.

2. Was a power (sample size) calculation performed?

Power is defined as the probability that the study will detect a true treatment effect of a pre-specified magnitude. In the face of positive results, the presentation of a power or sample size calculation provides evidence that the researchers were well organized and designed their investigation thoroughly. In the face of a negative clinical trial however, an *a priori* power calculation becomes much more important.

In a study of 71 published randomized control trials that failed to find a significant difference between treatment and control groups (negative findings), 67 of them were found to

be underpowered to detect the difference they set out to find⁹. The danger with studies that are underpowered is that potentially beneficial therapies may be discarded without adequate testing or therapies that truly are different may be accepted as being equivalent.

If a study reports that two therapies are ‘not significantly different’ or that ‘no beneficial effect was found’, an *a priori* sample size calculation must be documented before we are willing to consider these findings to support clinical decisions. If the original sample size calculation was performed at a power of 80%, we can be reasonably certain that we did not miss an important effect but until the power of clinical trials approaches 90% or higher, we should not be willing to rule out any possibility of missing an important difference.

Even when the power of a negative trial approaches 90% we should not be prepared to declare the therapies ‘equivalent’, we should merely conclude that we are 90% certain that a difference of a pre-specified magnitude does not exist. The magnitude of the difference we are prepared to rule out is defined by the ‘expected difference’ the trialists initially set out to investigate with their original sample size calculation. It is up to the reader to decide if this difference is clinically important or not.

III) What are the results?

1. How large was the treatment effect?

Besides reporting the finding of a statistically significant difference, the trialist must also provide an estimate of the magnitude of this treatment effect. There are many different ways that the researchers could report their estimated treatment effects, but they can basically be broken down into *absolute* or *relative* estimators.

Using an absolute estimator, a trial that finds a significant reduction in mortality rates from 50% in the control group to 35% in the treatment group would report a 15% *absolute*

risk reduction [50%-35%=15%]. If the same results were reported using *relative* estimators, the authors would report a 30% *relative* risk reduction [(50%-35%)/50%=30%]. Research has shown that when results are reported using measures of relative risk reduction, clinicians tend to overestimate the true benefit. Thus measures of absolute risk reduction (ARR) are preferred whenever possible¹⁸.

Some authors advocate the use of a measure called the *number needed to treat* (NNT) to interpret the clinical impact of new therapies. The NNT can be calculated directly from the absolute risk reduction, and using our example of a trial that reports a 15% absolute risk reduction, using the NNT approach, we would report that if we used the ‘new therapy’, 6.66 patients would require treatment in order to experience 1 additional clinical response [NNT=1/ARR=1/.15=6.667]. Similarly, if the example trial reported a 5% ARR, then we would calculate that 20 patients would require treatment in order to experience 1 additional response. The NNT initially appears appealing because it seems to give a measure of the ‘effort’ or amount of work required to reap a clinical benefit. Unfortunately the NNT is an extremely crude ‘effort-yield ratio’ and can be very misleading¹⁸.

We recommend that NNT’s be interpreted using extreme caution and only under the realization that they do not provide the full picture of resource consumption. If the reader is truly concerned about accurately measuring the true impact of a new therapy on resources, he or she should demand to see a fully-costed economic analysis⁷.

2. How precise was the treatment effect?

When a clinical trial reports the expected benefit of therapy, it is important for the reader to realize that the reported effect is simply *the best estimate of the average benefit*. In the example of a clinical trial reporting an absolute risk reduction of 15%, the accuracy of that

estimate depends on the inherent variability of the factor being studied *and* the sample size of the study itself. The authors can report how *accurate* their findings are by presenting a 95% confidence interval.

The concept of a 'confidence interval' is derived from statistical sampling theory, and is expressed as a numerical range between which we are 95% certain that the true results lie. If a study presents the finding of a 15% ARR, with a 95% confidence interval of 5% to 25%, we know with 95% certainty that the true results are not lower than a 5% benefit but also not higher than a 25% benefit. The 15% ARR is the *average* expected benefit. By knowing the lowest reasonable estimate of benefit and the highest reasonable estimate of benefit, we are able to better make a judgement of the potential impact of the therapy in our setting. For further reading on the utility of confidence intervals, the reader should refer to an excellent article published in *Statistics in Medicine* ¹².

IV) Will the results help me in caring for my patients?

When the reader is satisfied that the essential validity issues have been addressed, they are ready to consider if the trial can or should be incorporated into their daily practice.

1. Can the results be applied to my patient care?

It is important to consider if the patients that you see on a daily basis and the ICU where you work are similar to those reported in the study. Reading the basic description of the hospital that conducted the trial (community, tertiary care, teaching etc.) and the ICU that cared for the patients (surgical, neurological, long term ventilation, mixed etc.), it is easy to determine if the services and level of care offered in the study can be reproduced in your own situation.

It is also important to get a good description of the patients who entered the study to be able to decide if they are similar to your own. The study should describe its patients with respect to

age, demographics, major diagnosis, disease stage, disease severity and outcomes. If you feel that the patients you care for on a daily basis would qualify for entry into the study you just read, then in all likelihood, your own patients will benefit from the intervention in the same magnitude as those in the study. If however you are concerned that your patients are different with respect to an important consideration (ex: the study did not include patients receiving dialysis), then caution may be warranted when applying the results to these groups of patients.

2. Are the likely benefits worth the potential harms and costs?

Once you have decided that your hospital can offer the level of care delivered in the study and you are satisfied that the patient population in the study is representative of the type of patient seen in your institution, you now face a final decision: How much will the new therapy impact on scarce resources in your hospital and are the potential benefits worth these costs?

If the study was fully-costed, your decision is relatively easy since the trialists will present you with their estimates of resource consumption (costs), but in the face of a trial with no costs the decision becomes more difficult. The best way to approach this situation is to map out the proposed improvement in outcomes, other clinical benefits, any documented harm, changes in diagnostic test use, and impact on person time. Once this process map is developed, try to best assess the impact of these changes in your specific situation.

Summary

In 1754, aboard HMS Salisbury, James Lind conducted a simple, controlled clinical trial. He took 12 patients with “pale and bloated skin, listlessness, an aversion to exercise, swollen gums, halitosis, ecchymotic mucous membranes, and limb edema” and allocated them to receive treatment with one of six different therapies. Since the patients receiving two of his six chosen

interventions had such a dramatic recovery he felt ethically obligated to end his trial and administer these treatments to all the remaining sailors.

Today we fully recognize the impact that the controlled clinical trial can have on the development of new interventions. Unfortunately for us though, very few of these interventions are likely to have as dramatic an impact on outcomes as lemons and oranges did on scurvy. Because the interventions we study tend to have relatively small treatment effects and because the design and reporting of published RCTs has consistently been documented to be less than perfect, there is a real need for us to develop critical appraisal skills. This chapter is by no means the only approach to critical appraisal, but hopefully it serves as an adequate starting point for your journey.

Table 1: Appraising a Randomized Controlled Trial

I) Is the trial's design valid?

1. Was the randomization technique adequately described?
 - a) what was the method of randomization?
 - b) how was concealment achieved?
 - c) when did randomization occur?
 - d) was baseline balance achieved?
2. Were all patients who entered the trial properly accounted for and attributed at its conclusion?
 - a) was follow-up and reporting of losses complete?
 - b) were patients analyzed in the groups to which they were randomized?
3. Were patients, health workers, and study personnel "blind" to treatment?
4. Aside from the experimental intervention, were the groups treated equally?
 - a) were there any *planned* diagnostic work-ups, monitoring activities or supportive interventions that differed between the two groups?
 - b) were there any *unplanned* diagnostic work-ups, monitoring activities or supportive interventions that differed between the two groups?
5. Were the study outcomes appropriate?
 - a) were all clinically important outcomes considered?
 - b) were surrogate outcome measures validated in previous RCTs?

II) Is the analysis valid?

2. Was the primary analysis appropriate?
3. Was a power (sample size) calculation performed?

III) What are the results?

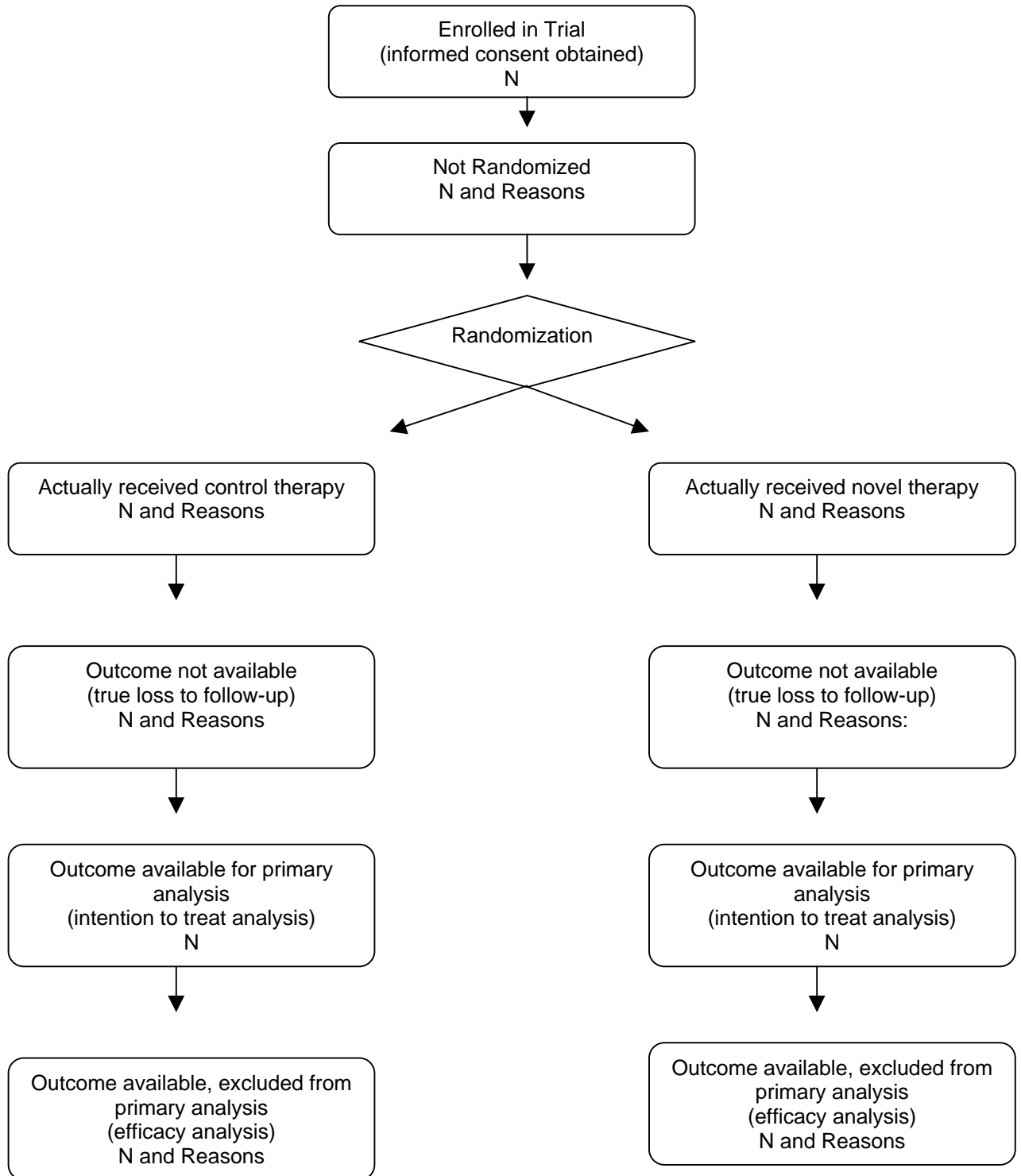
1. How large was the treatment effect?
2. How precise was the treatment effect?

IV) Will the results help me in caring for my patients?

1. Can the results be applied to my patient care?
2. Are the likely benefits worth the potential harms and costs?

In: Critical Care Clinics Volume 14, Number 3: An Evidence-based approach to Critical Care Medicine, Cook DJ and Levy MM eds., W. B. Saunders Company, Philadelphia, 1998:513-524.

Figure 1: Patient Flow: documenting reasons for incomplete follow-up.



References

1. Beam TR, Gilbert DN and Kunin CM: General guidelines for the clinical evaluation of anti-infective drug products. *Clin Infect Dis* 15(Suppl 1):S5-S32, 1992
2. Begg C, Cho M, Eastwood S, et al: Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *JAMA* 276(8):637-639, 1996
3. Chalmers TC, Smith H, Blackburn B, et al: A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials* 2:31-49, 1981
4. Cook DJ, Guyatt GH, Heyland DK, et al: How to use an article on therapy or prevention: Pneumonia prevention using subglottic secretion drainage Part II: What are the results and will they help me in patient care? *Crit Care Med* (in press)
5. Cook DJ, Hebert PC, Heyland DK, et al: How to use and article on therapy or prevention: Pneumonia prevention using subglottic secretion drainage Part I: Are the results of the study valid. *Crit Care Med* (in press)
6. Doig GS and Rochon J: Statistical Considerations for the Design of the Optimal Clinical Trial. *In* Vincent JL and Sibbald WJ (eds): *Update in Intensive Care and Emergency Medicine Volume 19: Clinical Trials for the Treatment of Sepsis*. Berlin, Springer-Verlag, 1995, p345
7. Drummond MF, Stoddart GL and Torrance GW: Why is economic evaluation important? *In* *Methods for the Economic Evaluation of Health Care Programmes*. Oxford, Oxford University Press, 1987, p 6
8. Fleming TR and DeMets DL: Surrogate end points in clinical trials: Are we being misled? *Ann Intern Med* 125:605-613, 1996

9. Freiman JA, Chalmers TC, Smith H, et al: The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: survey of 71 'negative' trials. *N Engl J Med* 299:690-694, 1978
10. Friedman LM, Furburg CD and DeMets DL: Issues in data analysis. *In* *Fundamentals of Clinical Trials*, ed 3. New York, Mosby-Year Book Inc, 1996, p 184-317
11. Friedman LM and Schron EB: Statistical problems in the design of antiarrhythmic drug trials. *J Cardiovas Pharm* 20(Suppl 2):S114-S118, 1992
12. Gardner MJ and Altman DG: Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 292:746-750, 1986
13. Guyatt GH, Sackett DL, Cook DJ for the Evidence-Based Medicine Working Group: Users' Guides to the Medical Literature II. How to use and article about therapy or prevention A. Are the results of the study valid? *JAMA* 270(21):2598-2601, 1993
14. Guyatt GH, Sackett DL, Cook DJ for the Evidence-Based Medicine Working Group: Users' Guides to the Medical Literature II. How to use and article about therapy or prevention B. What were the results and will they help me in caring for my patients? *JAMA* 271(1):59-63, 1993
15. Ingelfinger JA, Mosteller F, Thibodeau LA and Ware AH: Reading a report of a clinical trial. *In* *Biostatistics in Clinical Medicine*, ed 3. New York, McGraw-Hill, Inc., 1994, p 259-279
16. Meinart CL: Single-center versus multicenter trials. *In* *Clinical trials: Design, conduct and analysis*. New York, Oxford University Press, 1986, p 23-30
17. Moher D, Dulberg CS, Wells GA: Statistical power, sample size and their reporting in randomized controlled trials. *JAMA* 272:122-124, 1994
18. Naylor CD, Chen E and Stauss B: Measured enthusiasm: Does the method of reporting trial results alter the perceptions of therapeutic effectiveness. *Ann Int Med* 117:916-921, 1992

19. NIH inventory of clinical trials: fiscal year 1979, Volume I. National Institutes of Health, Division of Research Grants, Research Analysis and Evaluation Branch, Bethesda, MD.
20. Pocock SJ: The size of a clinical trial: multicentre trials. *In* Clinical trials: a practical approach. Chichester, John Wiley & Sons, 1983, p 123-142
21. Prentice RL: Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 8:431-40, 1989