Drug and Therapeutics Committee Training Course

Session 3.
Assessing Medicine Efficacy

Participants’ Guide
This document was made possible through support provided by the U.S. Agency for International Development, under the terms of cooperative agreement number HRN-A-00-00-00016-00. The opinions expressed herein are those of the author(s) and do not necessarily reflect the views of the U.S. Agency for International Development.

About RPM Plus

RPM Plus works in more than 20 developing and transitional countries to provide technical assistance to strengthen pharmaceutical and health commodity management systems. The program offers technical guidance and assists in strategy development and program implementation both in improving the availability of health commodities—pharmaceuticals, vaccines, supplies, and basic medical equipment—of assured quality for maternal and child health, HIV/AIDS, infectious diseases, and family planning, and in promoting the appropriate use of health commodities in the public and private sectors.

Recommended Citation

The materials may be freely abstracted, quoted and translated in part or in whole by non-profit or educational organizations (for reference or teaching only) provided the original source is acknowledged. They should not be sold nor used for any other commercial purpose.

ABBREVIATIONS AND ACRONYMS

ACS  acute coronary syndrome
CI   confidence interval
DTC  Drug and Therapeutics Committee
HR   hazard ratio
ITT  intention to treat
MI   myocardial infarction
n or N number
NNT  number needed to treat
RCT  randomized controlled trial
RRR  relative risk reduction
CONTENTS

Session 3. Assessing Medicine Efficacy................................................................. 7

Purpose and Content .......................................................................................... 7
Objectives ............................................................................................................. 7
Preparation and Materials .................................................................................. 7
Further Readings .................................................................................................. 8

Introduction ....................................................................................................... 8

Formulating the Question ................................................................................... 9

Finding the Evidence ......................................................................................... 9

What Sort of Evidence? ....................................................................................... 10

Assessing the Quality of the Evidence—What Makes a Good Clinical Trial? ........ 10
Why Was the Study Done? What Clinical Questions Were the Authors Addressing? ... 12
What Type of Study Was Done? ............................................................................ 12
Was the Study Design Appropriate to the Research Question? ......................... 14
Was the Design of the Study Sensible? .............................................................. 15
Who Is Studied? .................................................................................................. 16
How Well Was the Study Conducted? Was Systematic Bias Avoided or Minimized? ... 16

Quality of Randomized Controlled Trials: What to Look for ................................ 17
Randomization and Concealment of Allocation .................................................. 17
Double-Blind vs. Open Trials ............................................................................ 18
Inclusion of All Patients in the Statistical Analysis ............................................. 18

Quality of Non-Randomized Trials—What to Look for ....................................... 19

Understanding the Numbers ............................................................................. 20
What Did the Authors Think They Would Find? ................................................... 20
What Were the Results? ...................................................................................... 20
Different Types of Data Need Different Statistical Tests ..................................... 21
Understanding Metrics or the Bottom Line ......................................................... 23
What are the Standard Statistical Tests? .............................................................. 26

Summary Statement on Single Trials ................................................................. 29
Systematic Reviews and Meta-Analysis .............................................................. 29

Search Strategies and Inclusion Criteria ........................................................... 30

The Use of Systematic Reviews and Meta-Analysis ............................................ 31
The Use of Sensitivity Analysis in Interpreting the Results ................................... 32
Interpreting Inconsistent Results (Heterogeneity) .............................................. 32
Summary of Systematic Reviews ..................................................................... 33
Use of Evidence in Decision-Making................................................................. 33
Activity 1. Comparing Antimicrobial Medicines for Pneumonia......................... 34
Activity 2. Interpreting the Data: The Helsinki Heart Study.................................. 35
Activity 3. Critically Evaluating an Article.......................................................... 35
Activity 4. Interpreting the Data: A Medicine Trial to Compare Artesunate with Mefloquine to Treat Malaria................................................................. 35
Annex 1. Glossary............................................................................................... 38
Annex 2. Checklist to Detect Common Problems Encountered in Articles........... 41
SESSION 3. ASSESSING MEDICINE EFFICACY

Purpose and Content

Session 3 is designed to provide participants with a basic guide on how to determine medicine efficacy, primarily through review of the pharmaceutical literature with an emphasis on evaluating randomized controlled trials (RCTs), systematic reviews, and meta-analyses. Systematic and thorough evaluations of the pharmaceutical literature will provide the Drug and Therapeutics Committee (DTC) with the unbiased information necessary to select appropriate medicines for the formulary.

In most countries, evaluating the pharmaceutical literature is commonly done by physicians and pharmacists. Unfortunately, this review is often done incorrectly. With the tools presented in this session and practice at home, participants will be better equipped to evaluate the literature systematically and scientifically.

Objectives

After attending this session, participants will be able to—

- Understand the importance of determining efficacy and evaluating the clinical literature
- Discuss the major types of medicine study design
- Describe the key components of a journal article
- Understand how to evaluate and interpret results of a randomized controlled trial
- Discuss the use of systematic reviews and meta-analyses in evaluating medicines

Preparation and Materials

Read the following—

- This session—A thorough understanding of this session is necessary before the presentation.
- Review session 2, “Developing and Maintaining a Formulary” of the Participants’ Guide.
- Activity article—Randomized trial of artesunate and mefloquine alone and in sequence for acute uncomplicated falciparum malaria.
Further Readings


Introduction

There are thousands of drugs on the market in most countries. If an effective drug regulatory system is in place, there will have been some assessment of the quality, safety and efficacy of the drugs that are registered, but no single hospital or health care facility can afford to provide all the registered products. One of the key roles of a DTC, therefore, is to ensure that a selected list of medicines is provided. Medicines need to be selected so that the formulary—

- Meets the needs of the population being treated by the health facility
- Includes medicines that are effective and safe
- Includes medicines that are the most cost-effective of those that are available.

Every DTC needs to have a standard selection process for adding medicines to the formulary. The first consideration is whether the medicine being proposed is at least as effective as those currently being used for the same purpose. Making this judgment requires consideration of clinical trial evidence, which is complex and time consuming but critical for the effective functioning of a DTC.

Ideally, selection of a list of medicines should be closely aligned with standard treatment guidelines that are used locally for managing patients with the diseases that are important in the particular area, but often, DTCs are asked to consider new medicines that are not in the guidelines or medicines for conditions for which standard treatment guidelines do not exist.

This session covers the methods of assessing clinical evidence, including some information about different types of clinical studies, how to assess whether clinical trials are reliable and how to interpret the results of clinical studies including some important statistical concepts. These
skills require time and practice to acquire, but are essential for DTCs to adequately assess drugs for addition to the formulary.

Formulating the Question

The first step in assessing whether a new drug is effective is to formulate the clinical question that is important to your DTC. This should be constructed to specify—

- The patient population (P)
- The drug intervention you are interested in (I)
- The comparative treatment already available (C)
- The outcome that is important to clinicians and patients (O)

For example: In diabetic patients over age 60 (P), does metformin (I) compared to glibenclamide (C) reduce the risk of stroke (O)?

A well constructed clinical question means that—

- Your librarian, if you have one, or your pharmacist, or you can search the medical literature efficiently for relevant evidence.
- Your DTC can review only the evidence that is relevant, rather than being snowed under with many articles, reports, text book reviews, etc.!
- If the drug is added to your list, the drug’s use can be targeted to the indication and patients you have specified in your question, and vice versa; new products can be considered very specifically for precise purposes, rather than the broader approach of should we add metformin to our list.

Finding the Evidence

Once the DTC has considered the clinical question (which should be part of the proposal for any new product) there is a need to find the relevant clinical evidence. Depending on your resources, you may decide to ask for relevant clinical evidence to be submitted with the proposal to add a new drug, or you might have a member of the DTC undertake an independent search of the medical literature to find it. With an increasing availability of open access online databases of clinical studies, an independent review may offer the best approach, but it will take time!

There are many standardized ways now of searching databases such as PubMed and Medline to find clinical trials, systematic reviews, and summaries about new products. The important consideration is that, however the search is done, it attempts to find all relevant studies, not just a selection of information that presents the drug in the best light.
What Sort of Evidence?

When we are talking about “clinical evidence” what exactly is meant? Articles published in reputable medical journals generally fall into one of four different types—

- Secondary research or reviews, which can be—
  - *Overviews*—usually non-systematic summaries of the literature on a particular topic, often heavily influenced by the opinion of the author
  - *Systematic reviews*—these are comprehensive and rigorous summaries of all the studies done on a particular topic, that may include statistical combination of the results of all the studies (meta-analysis) to come up with an overall estimate of the effect of one drug compared with something else
  - *Economic analyses*—a study that describes the costs and effects of a particular treatment or health care intervention

- Clinical trials, in which an intervention such as a drug treatment is given to a group of patients and compared to the effects of some other treatment

- Surveys, in which something is measured in a group of patients or some other sample of individuals

- Experiments, where some sort of maneuver (sometimes a drug treatment) is given to humans (or animals) in an artificial or controlled environment

For a DTC, the evidence that is most reliable to use to evaluate whether a new drug is effective is a systematic review that contains several randomized clinical trials and a meta-analysis. For evaluation of drug safety, additional studies may be needed. But the medical literature still contains many studies that have major flaws and are therefore unreliable, so it is important to be able to judge whether a study of any type is of good enough quality to be used as the basis of a drug choice. The systematic assessment of medical publications is known as critical appraisal and the next sections of this module will cover the basic approach to appraising an individual study as well as a systematic review.

Assessing the Quality of the Evidence—What Makes a Good Clinical Trial?

Most studies that report a clinical trial that are published in medical journals are in a standard format that has been adopted by the journal editors.

- *Abstract* (which also has a standard structure in many journals)
- *Introduction* (why the authors decided to do this research)
- *Methods* section (how they did it and how they analyzed their results)
- *Results* section (what they found)
- *Discussion* (what the results mean, in the opinion of the authors)
An example of an abstract in a high quality journal is shown in Box 1. There will also be an Acknowledgments section and a list of references. Judgments about the quality of a paper should be made on the basis of an assessment of the “Methods” section, not on the interest of the title or hypothesis or what the authors think the results mean.

**Box 1. Structure Abstract (Source: Journal of the American Medical Association, 2005)**

**Context** The SYNERGY trial comparing enoxaparin and unfractionated heparin in high-risk patients with acute coronary syndromes (ACS) showed that enoxaparin was not inferior to unfractionated heparin in reducing death or nonfatal myocardial infarction (MI) at 30 days.

**Objective** To evaluate continued risk in this patient cohort through 6-month and 1-year follow-up.

**Design, Setting, and Patients** Overall, 9978 patients were randomized from August 2001 through December 2003 in 487 hospitals in 12 countries. Patients were followed up for 6 months and for 1 year.

**Main Outcome Measures** Six-month outcomes were death, nonfatal MI, revascularization procedures, stroke, and site-investigator–reported need for re-hospitalization; 1-year outcome was all-cause death.

**Results** Six-month and 1-year follow-up data were available for 9,957 (99.8%) and 9,608 (96.3%) of 9,978 patients, respectively; 541 patients (5.4%) had died at 6 months and 739 (7.4%) at 1 year. Death or nonfatal MI at 6 months occurred in 872 patients receiving enoxaparin (17.6%) vs. 884 receiving unfractionated heparin (17.8%) (hazard ratio [HR], 0.98; 95% confidence interval [CI], 0.89-1.07; \( P = .65 \)). In the subgroup of patients receiving consistent therapy, that is, only enoxaparin or unfractionated heparin during the index hospitalization (n=6138), a reduction in death or nonfatal MI with enoxaparin was maintained at 180 days (HR, 0.85; 95% CI, 0.75-0.95; \( P=.006 \)). Re-hospitalization within 180 days occurred in 858 patients receiving enoxaparin (17.9%) and 911 receiving unfractionated heparin (19.0%) (HR, 0.94; 95% CI 0.85-1.03; \( P=.17 \)). One-year all-cause death rates were similar in the two treatment groups (380/4,974 [7.6%] for enoxaparin vs. 359/4,948 [7.3%] for unfractionated heparin; HR, 1.06; 95% CI 0.92-1.22; \( P=.44 \)). One-year death rates in patients receiving consistent therapy were also similar (251/3,386 [7.4%] for enoxaparin vs. 213/2,720 [7.8%] for unfractionated heparin; HR, 0.95; 95% CI, 0.79–1.14; \( P=.55 \)).

**Conclusions** In the SYNERGY trial, patients continued to experience adverse cardiac events through long-term follow-up. The effect of enoxaparin on death or MI compared with that of unfractionated heparin at 6 months was similar to that observed at 30 days in the overall trial and in the consistent-therapy group. One-year death rates were also similar in both groups. High-risk patients with ACS remain susceptible to continued cardiac events despite aggressive therapies.

**ClinicalTrials.gov Identifier:** NCT00043784.

*JAMA. 2005; 294:2594-600 www.jama.com*

There should be a statement about who funded the study. Sometimes this is in the Acknowledgments section only, but increasingly there will also be a description of the funding arrangements and the role of the funding source in the Methods section of the paper. This is important as it has been shown repeatedly that the funding arrangements for a study can have a big influence on the results.

There should also be a statement about the author’s affiliations and potential conflicts of interest as well as their individual contributions to the study and the report. If a study is funded by a pharmaceutical company and all the authors are employees and there was no independent investigator, analysis or review, it may be biased and probably not worth spending too much time on!
So you have decide to try to read the paper that you have found and critically appraise it to determine whether the results are reliable and valid and can be used by your DTC to make a decision about a drug. Where do you start? Ideally you should develop a systematic approach that you use for every study. There are many checklists of questions for evaluating scientific articles, but these ones below offer a simple and comprehensive approach, and have been widely used in learning and teaching evidence-based medicine (Greenhalgh 1997). Most of these questions should be able to be answered after a careful reading of the Methods section of a current paper.

**Why Was the Study Done? What Clinical Questions Were the Authors Addressing?**

In a research paper, the introductory sentence should give a summary statement of what the background to the study is. For example, “Use of antibiotics for the treatment of upper respiratory tract infections in children is very common but it has been suggested that mostly the antibiotics are not necessary.” You should then be able to find, either in the introduction or methods section, a statement of the hypothesis that the authors have decided to test. This statement might be, “Antibiotics do not improve the symptoms and recovery time of upper respiratory tract infections in children.” If it is presented in the negative like this example, it is known as a null hypothesis. Authors generally set out to disprove this hypothesis in their study.

**What Type of Study Was Done?**

There are many different types of studies and different ways of conducting them. As noted above, most published articles will be either reports of reviews, experiments, trials, or surveys. For each of these types of studies there are important features to look for in the design of the study. The first point to check is whether there is an intervention that is being tested, such as a drug, operation, or diagnostic procedure, or whether the authors are reporting the results of an observational study, that is, where they did not have control of what happened to the patients or subjects.

For studies testing whether new interventions are effective, the most reliable design is a randomized comparative trial, where the patients being studied are allocated by chance to either the new intervention that is being tested, or the alternative comparative treatment. Sometimes the comparative treatment is a placebo drug or procedure; sometimes it is the best treatment that is currently available. If the authors are testing a new intervention and have not used a randomized trial as their design, it is very important to consider why they have not done so, as it is likely that the result will be biased in favor of the new treatment. A typical diagram describing a randomized trial is shown in Box 2 below.

Ideally, any new drug that is being considered by a DTC should have been studied in several randomized controlled trials, which have then been summarized objectively in a systematic review.
For studies that are reporting the results of observations, there are generally three types—

- **Cohort studies.** These studies involve groups of patients or subjects are observed over a period of time. The period can range from an observation at a single time point\(^1\) to repeated observations over many years. Each cohort of patients or subjects should be defined in advance on the basis of common features. There are many ways of doing this, for example, geographical and date inclusion criteria (e.g., all patients living in a certain suburb in a certain year); criteria related to characteristics of a disease, or exposure to a common event (e.g., all employees of a factory closed on a certain date). The important feature to consider is whether the authors have specified these inclusion criteria in advance and have made reasonable efforts to ensure that all appropriate subjects have in fact been included. For a DTC, cohort studies can be useful in assessing safety of new drugs, and sometimes they can also provide helpful information about how drugs are used in practice. However, they are not a reliable source of information for assessing whether a new drug is effective.

- **Case-control studies.** These are studies where the authors have firstly defined a “case” that they are investigating, which is a particular outcome or event in a subject (such as admission to hospital because of a gastrointestinal hemorrhage). The authors then compare the frequency of exposure to a substance of interest (e.g., aspirin) in a group of “cases,” with the frequency of exposure to the same substance in a group of “controls,” that is subjects who do not have the outcome being studied. Such studies are complex and difficult to perform well without bias, but when done correctly, they can be extremely useful in assessing whether rare adverse events are caused by exposure to a particular drug or substance.

- **Case reports.** These are studies that usually report the experience of an individual doctor in relation to a patient or group of patients with the same disease, or treatment. The classic example of the value of case reports is the first case report of thalidomide causing limb deformities. This case report could not be used to prove that thalidomide was the cause, but it acted as a warning signal that was followed up in a number of case control studies that did establish the relationship beyond reasonable doubt. Case reports therefore provide valuable anecdotal information about individual events and can therefore be helpful in assessing the safety profile of a drug, but cannot be used to assess drug efficacy.

**Was the Study Design Appropriate to the Research Question?**

To answer this question, it is helpful to categorize the field of research in the paper that you are considering as one of the following types. You should then consider whether the study design used is the preferred design for the research question.

- **Therapy** (testing efficacy of drug treatments, surgical procedures, ways of delivering care)—preferred design is a randomized controlled trial

---

\(^1\) The difference between a cross-sectional study and a cohort study with observed measurement at a single point in time is that in the cohort study, the patients are defined in advance on the basis of common features and in the cross-sectional study the patients are not so defined in advance.
• Diagnosis (is a new diagnostic test valid and reliable)—preferred design is a combination of a randomized comparison of the new test with the old plus follow-up of a cohort or patients

• Screening (can a diagnostic test pick up a disease at a pre-symptomatic stage in a large population, and can that lead to better outcomes from treatment of the disease)—preferred design is a cross-sectional cohort survey\(^2\)

• Prognosis (what is the outcome of a disease over time)—preferred design is a longitudinal cohort study of patients with the disease

• Causation (is a particular agent—toxin, drug, environmental factor—related to the cause of a disease)—preferred design is a case control or cohort study

**Was the Design of the Study Sensible?**

A study can be the preferred design for the type of research and research question, but can still be designed in a way that reduces the usefulness and reliability of the results. There are two important aspects of study design to consider first.

• Exactly what specific intervention was being considered and what was it compared with? For studies of new drugs, do the authors describe exactly the dose, preparation, frequency and nature of administration, for example? Is the comparator something that is accepted as current best practice, or a reasonable alternative? For example, it would not be reasonable to compare a new product for hypertension with a half-dose of an ACE inhibitor, or drugs no longer used because of side effects, such as reserpine.

• What outcome was measured in the study, and how? Is it an outcome that is important to clinicians and patients, or is it a physiological measurement that has little connection with patient survival or quality of life? In drug trials particularly, there is a tendency for authors to measure either “surrogate outcomes,” that are supposed to be proxies for clinically relevant endpoints but may or may not be truly so. An alternative strategy that is used to boost a medicine’s chances of being “successful” is for authors to measure “composite outcomes,” such as the frequency of “death OR heart attack OR chest pain” in patients treated with the new and the old intervention. The problem with this approach is that often the effect of the drug is mainly on one, but not all, of these endpoints and they are all treated with equal importance by the investigators, which may not be clinically appropriate.

\(^2\) A cross-sectional cohort study is where a source population is sampled cross-sectionally and then the subjects' histories of exposures and outcomes are assessed retrospectively over a specified period of time. See Hudson, J. I., H. G. Pope Jr., and R. J. Glynn. 2005 The Cross-Sectional Cohort Study: An Underutilized Design *Epidemiology* 16(3):355–59.
**Who Is Studied?**

When making decisions about new drugs a DTC needs to consider how the patients in the trial relate to those in the setting where the drug will be used. Questions to consider include the following—

- How were the subjects for the study recruited—enthusiastic volunteers only or a more representative sample of a population?

- Who was included in the study—patients with only a single uncomplicated disease or patients who have multiple illnesses and are more representative of the population in which the drug will be used?

- Who was excluded from the study—and is it likely to lead to false conclusions about the effects of an intervention due to different responses in patients with different baseline severity of disease?

- What was the setting of the study and how does it relate to the local environment? The results of an intervention based on a study in tertiary care hospital may not translate well into a primary care setting without the same level of back-up care.

**How Well Was the Study Conducted? Was Systematic Bias Avoided or Minimized?**

There have been many “studies of studies” carried out over the past 20 years to try to define what features within a study will produce the most reliable results. In some respects, this question is the most important part of a critical appraisal of a study. To start with some definitions—

- Bias is defined as anything that leads to deviation of the results from the truth, or processes leading to such deviation.

- Randomization—the process of assigning patients to treatment groups by chance.

- Observer-blind—the person measuring the outcomes in a study is not told what treatment patients have received.

- Double-blind—neither the observer nor the patient in a trial knows what treatment the patients received.

- Allocation—the process of assigning patients to treatment groups.

- Intention-to-treat population—the total number of patients assigned to receive a particular treatment, irrespective of whether they actually received it or not.
• Confounding factor/variable—a variable that can cause or prevent the outcome of interest, is not an intermediate variable, and is associated with the factor under investigation.

Annex 1 contains a glossary of additional terms.

**Quality of Randomized Controlled Trials: What to Look for**

There are now empirical data that show that for a randomized controlled trial, the features that are most important in determining the reliability of the results (or the likelihood of bias) are the following—

- Exactly what method was used to randomize the patients to treatment groups
- Whether both patient and observer are kept “blind” to treatment
- Whether all patients are followed up and included in the analysis

We will now deal with each of these factors in turn in more detail.

**Randomization and Concealment of Allocation**

As noted above, randomization is the process used to assign subjects in study by chance to one treatment group or another. Assigning patients by chance removes the likelihood that the investigator will select patients, either consciously or unconsciously, for the experimental treatment who are more or less likely to respond to it. This is called selection bias and we know that studies where selection bias occurs are more likely to produce results that significantly favor the experimental treatment and also overestimate the size of benefit produced by the experimental treatment. Random allocation also should result in any possible confounding factors being equally distributed between the two groups, and therefore be less likely for them to have a differential effect on the results.

It is not just enough, however, to assign patients at random to treatment groups. There are many ways this can be done. For example, by tossing a coin and using the heads or tails to determine the treatment group; by putting numbers for each treatment group in sealed envelopes and pulling them out of a drawer at random, or by having a process (used in many trials now), where once the patient consents to enter the study, the treating physician phones a centralized service, that allocates the patient to treatment based on a computer generated code without telling either patient or doctor what it is. This last method is the most reliable because is least likely to be able to be influenced by the treating doctor and therefore is regarded as the “gold standard” for clinical trials. If the paper you are reviewing does not tell you how the patients were randomized and how the allocation process was concealed, it is more likely to have unreliable results that favor the experimental treatment.

---

A variable in a causal pathway that causes variation in the dependent variable and is itself caused to vary by the independent variable.
Double-Blind vs. Open Trials

There are three parties who can influence the trial results if they know what treatment a patient in a study is receiving: the patients themselves, the treating physician, and the person measuring the outcome. We know that if all three are aware of the treatment, then the results of the study are more likely to be biased in favor of the experimental treatment compared with the control. Various terms are used for this but the most common is observer bias (where the observer consciously or unconsciously influences the results). Ideally, therefore, the patient, treating doctor and observer should all be unaware of the treatment group so that any placebo should exactly match the active treatment. However, this may not always be feasible, in which case other steps should be taken to ensure that the results are as reliable as possible. These include the following—

- Making sure that the outcome is an objective, black and white event (such as death) rather than an event defined in a way that is open to subjective interpretation (such as “significant improvement in pain”).
- If possible, ensuring that the person measuring the outcome is unaware of the treatment, even if the patient and treating doctor know.
- Having dubious outcomes (for example, was the cause of death in a subject heart failure, which could have been prevented by the treatment, or was it due to a respiratory infection, unrelated to the treatment?) adjudicated by an independent committee (who are unaware of the treatment group for the patient concerned).

In summary, double blind placebo controlled trials with objective outcomes judged by an independent outcome committee are the gold standard. Less than this may practically be necessary, but will reduce your confidence in the results, and it will be a matter of judgment as to how much this matters for the assessment of an individual trial.

Inclusion of All Patients in the Statistical Analysis

A trial is less likely to have biased results if all patients recruited and allocated to treatment are accounted for. In the example shown in figure 1, the intention to treat (ITT) population is the 320 patients randomized to the two treatment groups, even though at the four month follow-up point there were only 230 patients left in the trial. By using the ITT population in the denominator for the analysis, the most conservative estimate of the effect of any treatment is obtained. Although in some instances this may be an underestimate, it is less likely to be subject to bias.

Trials of new interventions that do not report what happened to all patients and do not report an ITT analysis should be treated with more uncertainty than those that do.

Figure 1 summarizes the key sources of bias that can occur in randomized trials.
Quality of Non-Randomized Trials—What to Look for

There are fewer empirical studies that can be used to help us assess what is most important in deciding whether a non-randomized study has been done well or done poorly, but if you think about it, it should be apparent that the main difference between randomized and non-randomized studies is the likelihood that in non-randomized studies, the risk of selection bias is higher. This is such a critical difference that it tends to overwhelm all other considerations, but in assessing the quality of a non-randomized study, it is important to try to make a judgment about how much selection bias has been introduced in the process used to include subjects in the study. If, in your judgment, it is significant, then the results are more likely than not to be due to selection and confounding. Although there are statistical techniques that can be used to attempt to compensate for this, no amount of statistical adjustment can overcome the fundamental bias that is introduced by absence of randomization.
Understanding the Numbers

So you have got this far through the methods section of the paper and have decided that the study is of sufficient quality for you to persist with assessing the results. In the results there will be some statistical analyses. Before getting overwhelmed by statistical jargon, it is worth stating some basic principles. The purpose of these analyses is to determine whether any differences in outcomes that have been found in one group compared with another are real and due to the difference in treatment or whether they are simply the result of the play of chance. We cannot avoid talking about some basic statistical concepts here, but for practical purposes, there are a few key numbers to master that will greatly improve the reliability of decisions of a DTC.

We will divide these considerations into two groups—what did that authors think they were going to find, and how do we interpret what they did find?

What Did the Authors Think They Would Find?

A trial should be big enough and long enough to have a high chance of detecting, if it exists, an effect of treatment that is both clinically and statistically significant (the real effect rather than one due to chance). So somewhere in the methods section there should be a statement about the difference between the two treatments that the authors want to find and therefore an estimated sample size.

For example, if you are comparing two drugs for the treatment of high blood pressure, we know from long-term observational studies that a difference of 5 mm Hg in systolic pressure that is sustained over several months is enough of a difference to result in a change in the likelihood of stroke or cardiovascular disease. Trials comparing treatments in hypertension therefore need to include enough subjects to show a 5 mm Hg difference if it exists. It would be more reliable if strokes and cardiovascular disease were actually measured, but such a trial would need to take longer and probably be larger as well. What you as the reader are looking for is, what did the authors decide was an important difference before they did the trial and therefore (using a standardized statistical algorithm available in most statistical software packages), how many patients did they calculate they would need. Did they get that many subjects into the study or not, and if not, why not?

What Were the Results?

The key concepts for a DTC to consider in relation to understanding results of studies are as follows—

- Different types of data need different statistical tests.
- But most types of studies can be analyzed using one of a relatively few standard approaches.
- It is necessary to compare the effect of treatment in one group relative to the effect in the other.
• It is also critical to compare the absolute value of the results in one group with those in the other.

• The difference in the effects of the treatments (if any) can be described as the estimate of the effect size, and the range of plausible results is expressed in the confidence interval; the probability that the difference is real and not due to chance is expressed in the p value.

Again, some detailed discussion of each of these points follow. We cannot provide a complete statistical textbook; further reading for those of you who are interested is listed in the reference material.

**Different Types of Data Need Different Statistical Tests**

For trials of new drugs, there are four considerations. First, are there descriptive statistics that you can use to compare the different patient groups at baseline? This should include, for example, the mean age of patients in each group, perhaps the weight, the proportion of patients from different ethnic backgrounds, the proportion of patients with other illnesses, and so on. An example of a standard type of table is shown in figure 2.
In the table in figure 2, you can see that there are \( p \) values given for the comparison of each variable. For \textit{continuous variables}, (where each value has a meaning, such as centimeters or kilograms) such as weight, the \( p \) value has been calculated from a \textit{t test}. For \textit{categorical variables} (the variable exists or not, e.g., current smoker), a chi-squared test has been used. All the \( p \) values are larger than 0.05, which is conventionally accepted as the cutoff value for statistical significance in most trials. A \( p \) value of 0.05 means that there is a 1 in 20 probability that the differences between the groups are due to chance alone. In this context the purpose of such comparisons is to show that as far as possible, there are no real differences between the groups at baseline that could influence the results.

The second consideration is whether the main outcome of the trial (technically this should be specified in the methods as the primary outcome) is a continuous variable or a categorical one, and whether the data are \textit{normally distributed} or not. These factors then should determine what type of statistical test should be used, which will be discussed further below.
Third, the authors should have specified in the methods what their planned analysis was, both in terms of the tests to be used based on the considerations noted above, but also in terms of what groups and comparisons are to be made. For example, were they only going to compare the results of all patients in the intervention group with the control, or were they planning in advance to compare the results in men and women as well? If you do enough comparisons, you will end up with one of them having a statically significant $p$ value by chance, so it is important that the authors plan the analyses before they get the data and do not go “data dredging” afterward.

And fourth, what metric (numerical expression) have the authors used to describe their results for dichotomous data (outcomes with a yes/no response)?

**Understanding Metrics or the Bottom Line**

Generally, the simplest way of describing the results of a trial is to use proportions to describe the number of patients in a treatment group who have the outcome (as the numerator) compared with the total number in the group. In a theoretical trial that compares the effect of a new drug with an old one in terms of the number of people who die from heart failure after five years of treatment, the results might look like what is shown in Table 1.

**Table 1. Results**

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Patients Who Died</th>
<th>Number of Patients Who Did Not Die</th>
<th>Total in Each Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>New medicine group</td>
<td>5 (a)</td>
<td>95 (b)</td>
<td>100 (a+b)</td>
</tr>
<tr>
<td>Old medicine group</td>
<td>10 (c)</td>
<td>90 (d)</td>
<td>100 (c+d)</td>
</tr>
</tbody>
</table>

The risk of dying if you are treated with the old drug is 10/100 or 0.10 or 10 percent. This is also called the control event rate or control event risk. (Strictly speaking, if you use the term rate you should also specify the time period, e.g., per five years, but often the time frame is omitted when specifying the figures in articles or reports!)

The risk of dying within 5 years if you are treated with the new drug is 5/100 or 0.05 or 5 percent; the experimental event rate.

As a DTC, you are interested in what the effects are in one group compared to the other. The relative risk of death in the new drug group compared with the old is the ratio of the experimental event rate to the control event rate—$0.05/0.10 = 0.5$.

That is, you are half as likely to die if you get the new drug versus the old (or you are twice as likely to die if you get the old drug rather than the new). Other terms used for this metric are rate ratio or risk ratio; these can be helpful reminders of how it is calculated.

As a DTC, however, you are not just interested in the relative performance of a drug: you need to consider the absolute benefit, which takes into account how common the events are. This is expressed as the absolute risk difference, which as the term implies is calculated as the difference in risks—$0.05 - 0.10 = -0.05$. 

---

Session 3. Assessing Medicine Efficacy
That is, there is a 5 per 100 (5 percent) reduction in the risk of dying if you are treated with the new drug compared with the old. You will see this called absolute risk reduction, or risk difference as well.

\[
\text{Absolute Risk Difference (reduction)} = \text{risk reduction with Drug A} - \text{risk reduction with Drug B}
\]

But not all patients who were treated with either the new drug or the old for the five-year period died (or lived)—if the benefit from these treatments is preventing death, we actually had to treat quite a lot of patients to prevent five deaths. This way of thinking about the treatment benefit leads to another metric: the number of patients needed to treat (to prevent one death)—

\[
\frac{1}{\text{absolute risk difference, in this case}} = \frac{1}{0.05} = 20
\]

So the number needed to treat (NNT) here is 20 patients need to be treated for 5 years to prevent 1 death.

\[
\text{Number of Patients NNT (to prevent 1 death)} = \frac{1}{\text{absolute risk difference}}
\]

In the 2007 SPARCL trial, atorvastatin was studied to determine the effects on the incidence of stroke and cardiovascular events. This study showed that the drug was associated with a significant reduction in the number of cardiovascular events. The NNT was determined to be 16 with a range of 12-24.

You may see these results presented other ways. For example—the relative risk reduction would be 1- relative risk, i.e., 1— 0.5, or 50 percent.

\[
\text{Relative Risk Reduction (RRR)} = 1- \text{relative risk}
\]

A 50 percent reduction in risk sounds far more attractive the 5 percent absolute risk difference! And needless to say, relative risk reductions are fare more frequently quoted in pharmaceutical advertising material than are NNTs or absolute risk differences. In the SPARCL study described above, the RRR was 23 percent for any cardiovascular event occurring in patients taking atorvastatin.

Yet another way is to see results expressed as odds ratios. If you understand gambling, odds ratios are also easy to follow: the odds of an event is the number who have the event divided by the number who do not.
In our example, in the new treatment group this would be 5/95, or 0.0526. For the odds ratio of the new treatment compared to the old, therefore—

\[
\frac{5}{95} = 0.0526 = 0.474 \\
\frac{10}{90} = 0.111
\]

The odds ratio is not quite the same as the relative risk, but close. However, interpreting odds ratios is not straightforward. For most purposes think of them like a relative risk, but the results will not always be identical, and the more common an event is, the less similar the two metrics will be.

These calculations are simple and important to do yourself. If authors do not give you the NNT, you should work it out as it gives you an idea of costs and benefits—do you, for example have to pay for treatment for 20 patients for 5 years for 1 to have a benefit? If results are only expressed in relative terms in a study, is that because the events are so uncommon that any absolute benefit is tiny? Even in complicated studies, it is a helpful exercise to do the calculations to get an approximate estimate of the effect of the new treatment compared to the old and see how your approximations compare with the quoted results. Depending on the precise statistical analysis used, they won’t necessarily be exactly the same—but they should at least be in the same order of magnitude.

Table 2 below provides the general formula for the calculations assuming you set up the results in the same way as in the example. If the outcome is a benefit (i.e., deaths prevented), the current convention is to express relative risks as values < 1.0; if the outcome is a harm or adverse effect (i.e., deaths caused), the convention is to express the value as > 1.0. As it is easy to confuse the direction, it is important that you always check what the treatment group is and what the control group is for each calculation.
### Table 2. Formulas Commonly Used to Assess Treatment Effect Size Compared to Control

<table>
<thead>
<tr>
<th>Group or Parameter</th>
<th>Number with the Event</th>
<th>Number without the Event</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental group</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Control group</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Experimental event rate</td>
<td>a/[a + b]</td>
<td>c/[c + d]</td>
<td></td>
</tr>
<tr>
<td>Control event rate</td>
<td>a/(a+b) / c /[c+d]</td>
<td>1-[(a/(a+b) / c /[c+d)]</td>
<td></td>
</tr>
<tr>
<td>Relative risk (experiment/control)</td>
<td>(a/(a+b) - (c/(c+d)))</td>
<td>1/[(a/(a+b) - (c/(c+d))]</td>
<td></td>
</tr>
<tr>
<td>Absolute risk difference (experiment—control)</td>
<td>(a/b)/(c/d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number needed to treat</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odds ratio(treatment/control)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each of these metrics is a way of expressing the **point estimate** for the size of the effect of the treatment compared to the control group. Equally important is to have an estimate of the **95 percent confidence interval**, which can be considered to be the plausible possible range of the results. That is, if the same study were repeated 100 times in different samples of patients, the 95 percent confidence interval would include the true mean point estimate 95 times out of 100.

For those of you who are enthusiasts, the 95 percent confidence interval for the difference between two proportions is as follows—

\[
95\%\ CI = +/- (1.96 \times \text{Standard Error of the Difference})
\]

Where the SED is:

\[
SE_{diff} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}
\]

And \( P \) is probability in group, \( n \) is number in group, subscript indicates group 1 or 2.

The point of considering a confidence interval is that if the metric is the relative risk and the 95 percent CI includes 1.0, then any difference between treatments is probably due to chance (i.e., the \( p \) value would be > 0.05). Similarly, if the metric is the absolute difference and the 95 percent include 0.0, there is no real difference between treatments. However, the upper and lower limits of the confidence interval will give an indication of what the maximum and minimum differences might be and therefore a clinical judgment can be made about the importance of the difference if it exists. See figure 4 for an example—

**What are the Standard Statistical Tests?**

For the clinician and DTC, there are about half a dozen statistical methods that are worth understanding. For purposes of this training, we will list and define (Table 3) the important statistical tests used in drug studies. Participants are referred to the further readings section to obtain more information.
Table 3. Statistical Tests Commonly Used

<table>
<thead>
<tr>
<th>Statistical Test</th>
<th>Use</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Two sample t test</td>
<td>Compares the results for a single measurement of a variable in 2 groups of patients</td>
<td>Girls’ height at age 10 with boys’ height at age 10</td>
</tr>
<tr>
<td>(2) <strong>Mann-Whitney U test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Paired t test</td>
<td>Compares the results of two measurements in the same group</td>
<td>Weight before and after feeding in 10-day-old infants</td>
</tr>
<tr>
<td>(2) <strong>Wilcoxon matched paired test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1- or 2-way analysis of variance:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) <strong>Fisher ANOVA</strong></td>
<td>Compares the results of multiple measurements in a group</td>
<td>Repeated measurements of BP over time</td>
</tr>
<tr>
<td>(2) <strong>Kruskal-Wallis ANOVA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) <strong>Pearsons correlation coefficient</strong></td>
<td>Assesses the strength of association between one or more factors and another</td>
<td></td>
</tr>
<tr>
<td>(2) <strong>Spearmans correlation coefficient</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression modeling (in a variety of forms)</td>
<td>Describes the numerical relationship between one or more factors and another</td>
<td></td>
</tr>
</tbody>
</table>

Of those tests mentioned, the first (1) in each pair is a parametric test and the second (2) is its *nonparametric* equivalent. Nonparametric tests (or distribution-free) should be used when one cannot assume that the data for the variable is normally distributed. Nonparametric tests have less power than the appropriate parametric test, but are stronger when the assumptions used for parametric tests are not satisfied.

All of the tests can be used to provide a *p value*, which, as noted above, is an indicator of whether a result is likely to be due to the play of chance.

In papers published in current journals, the additional method that is used in survival analysis, usually presented graphically as shown in figure 3. This type of analysis will be presented with a hazard ratio (usually derived from Cox’s proportional hazard model) which can be interpreted in a similar way to relative risks.
Figure 3. Example of survival analysis.

A useful rule of thumb is that if a paper uses a method that doesn’t seem to be one of the standard ones, have the authors explained why?

Finally, a very useful list of “errors” to remember when assessing the statistics and results in any study, thanks to Trisha Greenhalgh, is shown in box 3.
Box 3. Ten Ways to Cheat on Statistical Tests When Writing Up Results

1. Throw all your data into a computer and report as significant any relation where $P < 0.05$.
2. If baseline differences between the groups favor the intervention group, remember not to adjust for them.
3. Do not test your data to see if they are normally distributed. If you do, you might get stuck with non-parametric tests, which aren't as much fun.
4. Ignore all withdrawals (drop outs) and non-responders, so the analysis only concerns subjects who fully complied with treatment.
5. Always assume that you can plot one set of data against another and calculate an "r value" (Pearson correlation coefficient), and assume that a "significant" r value proves causation.
6. If outliers (points which lay a long way from the others on your graph) are messing up your calculations, just rub them out. But if outliers are helping your case, even if they seem to be spurious results, leave them in.
7. If the confidence intervals of your result overlap zero difference between the groups, leave them out of your report. Better still, mention them briefly in the text but don't draw them in on the graph—and ignore them when drawing your conclusions.
8. If the difference between two groups becomes significant four and a half months into a six month trial, stop the trial and start writing up. Alternatively, if at six months the results are "nearly significant," extend the trial for another three weeks.
9. If your results prove uninteresting, ask the computer to go back and see if any particular subgroups behaved differently. You might find that your intervention worked after all in Chinese women aged 52-61.
10. If analyzing your data the way you plan to does not give the result you wanted, run the figures through a selection of other tests until you get the result you want.


Summary Statement on Single Trials

Critical review of a single trial takes skill, time, and resources, although the skills can be acquired with practice. Even if you only have a few trials for each drug that you are considering, you can expect your DTC to need technical support in critical appraisal. It is not enough just to be able to "trash" a trial—in the end, a committee still has to make a judgment based on the best available information, even if this is a poor quality study.

So ideally you will be able to access more than one trial in relation to each drug, and it is better still there if there is a systematic review.

Systematic Reviews and Meta-Analysis

A systematic review is an overview of individual or primary studies that contains an explicit statement of objectives, materials, and methods, and has been conducted according to explicit and reproducible methodology. They should be distinguished from traditional, or narrative reviews, which are usually written by an author to prove or disprove a particular point and are
therefore likely to be more selective in the studies for citing or inclusion. The advantages of systematic reviews are—

- The use of explicit methods limits bias in identifying and rejecting studies.
- The conclusions are generally more reliable and accurate because of the methods that are used.
- Large amounts of information can be assimilated quickly by health care providers, researchers, and policymakers.
- Results of different studies can be formally compared to establish the generalizability of findings and consistency of results.
- Reasons for heterogeneity (inconsistency in results across studies) can be identified and new hypotheses generated about particular subgroups.
- Where appropriate, results of individual studies can be statistically combined using meta-analysis to provide a single summary estimate of the effect of an intervention.

A major source of a large number of systematic reviews is the Cochrane Library, which contains over 2,500 complete reviews on a wide variety of topics (see: http://www.mrw.interscience.wiley.com/Cochrane) and these are added to every three months. These reviews are also regularly updated to incorporate new evidence as it becomes available.

To fully understand and interpret a systematic review, there are some technical points to consider. These include—

- How the trials included in the systematic review were found, and the potential for publication bias: search strategies and inclusion criteria
- The use of meta-analysis
- The use of sensitivity analysis in interpreting the results
- Interpreting inconsistent results (heterogeneity)

**Search Strategies and Inclusion Criteria**

Any systematic review should include a description of the criteria used by the authors to select studies to include in the review and a description of how they went about finding the studies. This should include not only the definition of the interventions and population in the studies, but whether studies in all languages were included, and also whether unpublished studies were sought for and added to the list. It is well established that negative studies do not get published as often as positive ones, and thus the assessment of the efficacy of an intervention based only on published trials can sometimes be unduly positive.
How the authors went about finding the studies should also be described. There are many electronic databases that index medical studies but searching only the electronic sources such as PubMed can still result in important studies being missed. It is important that authors of a systematic review use multiple strategies to find all the relevant trials, including contacting authors, and if relevant, pharmaceutical manufacturers.

Although systematic reviews have generally focused on pulling together only RCTs, there is increasing interest in systematic reviews of other study types. This is particularly important when assessing the safety of a product, as RCTs may not be sufficient to evaluate adverse effects. However, the same principles should apply to systematic reviews of non-randomized studies: authors should specify in advance what types of studies will be included and how they will find them.

The Use of Systematic Reviews and Meta-Analysis

The term meta-analysis is often used interchangeably with systematic review, but the two should not be confused. Meta-analysis refers to the statistical techniques used to combine the results of clinical trials into a single estimate of effect. It is best thought of as a weighted average effect, with larger trials counting more than smaller ones. Meta-analysis can be used to calculate pooled or summary estimates for all the metrics described above (relative risk, risk difference, and so forth). There are also ways of combining the results of trials that report only continuous outcomes (means, medians, and so forth) but these are more complex and difficult to interpret. Not all systematic reviews can (or should) include a meta-analysis. If the trials appear to be in significantly different populations, of different design, or have different interventions, it may not be clinically sensible to try to combine them.

Meta-analysis results should generally be presented graphically as well as in tables. The standard presentation is the forest plot (figure 4). Each dot represents the results for a single trial with the 95 percent confidence interval and the size of the dot is proportional to the number of subjects in the trial. The summary estimate is the diamond at the bottom of the graph, with the middle of the diamond as the point estimate and the rest of it representing the 95 percent CI. In this example, the summary estimate is the odds ratio for the risk of cancer in patients treated with statins compared with those not treated with statins, and as it is 1.02, with a confidence interval of 0.97 to 1.07, we can be reasonably confident that there is no relationship.
The Use of Sensitivity Analysis in Interpreting the Results

The results of a meta-analysis can be very sensitive to the inclusion and exclusion of trials. For example, if the analysis shown in figure 4 for some reason exclude the HPS and LIPID trials, the pooled estimate would be likely to be quite different. Sometimes it is important for authors to carry out the analyses including and excluding trials—these are called sensitivity analyses. They can be useful tools to understand inconsistent trial results, but should again be specified beforehand.

Interpreting Inconsistent Results (Heterogeneity)

In the language of meta-analysis, homogeneity means that the results of each trial are mathematically compatible with each other. It can be assessed from the forest plot of a meta-analysis: if the confidence limits of all the trials overlap, the trials are probably homogeneous, but if they do not, statistical heterogeneity exists. This is tested for by the q statistic. The challenge for the reviewer is then to determine why there are differences in the results: is it due to chance, or is it a signal that there are real differences between the trials that lead to the different results, such as duration, design, or population sampled?
Summary of Systematic Reviews

Systematic reviews, like all other scientific studies can be good, bad, or indifferent. The key points summarized above will help you assess them. A good systematic review that pulls together the results of all relevant clinical trials is the most robust and reliable evidence we currently have and should be a key part of any DTCs decision and assessment process.

Use of Evidence in Decision-Making

It is worthwhile to spend a few minutes considering how best to use evidence in decision making. As noted above, it is not sufficient to be able to “trash a trial”—a DTC, like any other decision making group still has to decide what to do. The following approach to using evidence in making treatment recommendations tries to summarize all of the aspects of assessing studies as well as taking into account uncertainty, values, and costs. It has been developed over the past five years by an international group and is gaining increasing acceptance from guidelines groups and other who make treatment recommendations. See also Annex 2 for a checklist to detect common problems encountered in articles.

- First of all, what is the clinical question in terms of patient population, intervention, comparator, and outcome?
- What are the critical outcomes for the intervention? These may not necessarily be measured in the trials.
- For these critical outcomes, what type of evidence do you have? Ideally a systematic review, but this will not always be the case.
- What is the quality of the evidence that you have, including the assessment of study design, concealment of allocation, randomization method, blinding, and completeness of follow-up?
- Is the evidence consistent? Do all the studies generally give the same results? If not, is there a reason for the differences?
- Is the evidence directly related to your setting? Is the comparator relevant, is the intervention the same as the one you are interested in, is the setting of the studies similar? If not, what impact is it likely to have on the overall estimate of effect?
- What is the size of the effect? What is the 95 percent CI? Does it include or exclude clinically important differences?
- What are the harms?
- What are the likely trade-offs between benefit and harm?
• What are the costs?

• Are there local values and preferences that need to be considered?

Then, taking all of this into account, would you recommend that the intervention should be adopted or not?

Good luck and happy appraising.

We will now work through some practical examples to put all of the theory into practice.

**Activity 1. Comparing Antimicrobial Medicines for Pneumonia**

For this activity, assume your DTC is considering the formulary addition of a new antimicrobial drug for treating lower respiratory tract infections in children. The drug study abstract you have just read concludes that this drug’s efficacy is equal to a combination of antibiotics in treating pneumonia in hospitalized children.

This study looked at 35 children in the treatment group and 43 in the control group. The setting was a large university hospital. This was an open label study, and children receiving a new antimicrobial were compared with other children in the hospital who were receiving different antibiotic combination regimens to treat pneumonia. Patients were chosen to receive this antibiotic by the physician depending on the severity of the pneumonia. The drug requested for the formulary was typically given to children with less severe pneumonia (based on the judgment of the physician) while the combination drug therapy was reserved for children who appeared to be sicker and at higher risk.

Results showed that the study drug was equally effective as a combination of antibiotics and was less costly. There was no difference in the incidence of adverse drug reactions. The manufacturer of the drug sponsored the study.

You are especially interested in such a drug since it is less costly and the study shows that it is effective. Safety information is limited at the early stages of its marketing.

• How would you describe the study design? Is it valid?
• What are the controls in the study?
• How are patients randomized?
• What kinds of bias can be introduced in this type of study?
• Are the results of this study usable in your country?
Activity 2. Interpreting the Data: The Helsinki Heart Study

The Helsinki Heart Study identified treatment modalities for elevated lipoproteins through the use of gemfibrozil. Subjects: 4,081 asymptomatic men aged 40–55 with dyslipidemia (total cholesterol minus high-density lipoprotein > 5.2 mmole/liter)

Treatment: gemfibrozil 600 mg twice daily (2,051 men) or matched placebo (2,030 men) in a five-year randomized double-blind study

Results: number of events (fatal, nonfatal myocardial infarction, or cardiac death)

- Gemfibrozil—56 events
- Placebo—84 events

Please calculate the following—

- Event rate for placebo group (%): 
- Event rate for active drug group (%): 
- Relative risk: 
- Relative risk reduction (%): 
- Absolute risk reduction (%): 
- NNT for five years to prevent one event:

Activity 3. Critically Evaluating an Article

One or more articles will be distributed in class. You should work in groups to evaluate the one assigned to you and present your findings in class.

Activity 4. Interpreting the Data: A Medicine Trial to Compare Artesunate with Mefloquine to Treat Malaria


Box 4. Exercise: Managing Malaria

You are working for the Department of Communicable Diseases. In your country, as in many others, one of the major health problems is malaria. As is the case in other settings, resistance to standard treatment is becoming a significant problem and you are therefore trying to decide what alternative treatments should be included in your standard treatment guidelines. You are aware that artesunate is a relatively new drug that may offer some advantages in the treatment of Falciparum malaria. You have been asked to present a summary of the clinical trial evidence and an assessment of the potential cost-effectiveness of using artesunate compared to mefloquine. Your librarian has identified a possibly useful study from Thailand for you to use as the basis for your assessment.
Complete the following questions.

- What type of study is this? What are the key study design characteristics? Does the study design allow you to address clinical efficacy and potential cost-effectiveness? Is the study relevant to the question that you have been asked to address?

- Were patients, those administering treatment, and those assessing outcomes blinded to the patients’ treatment allocation? Which patients were included in the analyses—all of those randomized? Only those who completed the trials? Are there any baseline differences between the treatment groups?

- How reliable or valid do you think these studies are? What are the reasons for your answer?

Fill in the grid below. Are the study populations similar to those in your country?

**Dosage Regimen and Baseline Characteristics in Trials**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Looareesuwan et al. 1992</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Artesunate N=</td>
</tr>
<tr>
<td></td>
<td>Mefloquine N=</td>
</tr>
<tr>
<td>Dosage Regimen</td>
<td></td>
</tr>
<tr>
<td>Proportion male [n (%)]</td>
<td></td>
</tr>
<tr>
<td>Age [mean (standard deviation)]</td>
<td></td>
</tr>
<tr>
<td>Parasite count [Mean(range)]</td>
<td></td>
</tr>
</tbody>
</table>

- What are the outcomes that are presented in the trial?
- Are they relevant to the question that you are trying to answer?
- Are the outcome measures surrogate measures, and if so, has the surrogate been validated?

To help in your interpretation of these studies, we have summarized some of the results from each of them. However, you will still have to review the “Methods” section in each study to confirm exactly how each outcome was measured.

The following grid is the comparison of artesunate and mefloquine (from Looareesuwan et al., 1992). For the moment, we will ignore the combination treatment. Total numbers are the number of patients randomized to each treatment group, and where possible, the results for the “intention to treat” analyses have been presented.
### Efficacy results from Looareesuwan et al. (1992)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Artesunate N = 42</th>
<th>Mefloquine N = 43</th>
<th>Difference in means (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fever clearance time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Mean (standard deviation)]</td>
<td>35.1 (23.4)</td>
<td>69.7 (37.5)</td>
<td></td>
</tr>
<tr>
<td>Parasite clearance time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Mean (standard deviation)]</td>
<td>35.9 (10.1)</td>
<td>63.5 (25.5)</td>
<td></td>
</tr>
<tr>
<td>Outcome</td>
<td>Artesunate N = 42</td>
<td>Mefloquine N = 43</td>
<td>Relative risk (95% CI)</td>
</tr>
<tr>
<td>Patients cured at 28 days [n (%)]</td>
<td>35</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Patients completing 28 days follow-up [n (%)]</td>
<td>40</td>
<td>37</td>
<td></td>
</tr>
</tbody>
</table>

- How have the outcomes been expressed? Based on the data in the grid, how would you determine whether there is a difference in the effects of the treatments, and what the size of the difference is?

- Fill in the gaps in the table (the columns for difference in means, relative risk, risk difference, and 95 percent CI).

- What is the size of the effect of artesunate compared with mefloquine treatment? What is the plausible range of results?

- For each of the outcomes: is the size of the effect clinically important? Are all of the values included in the 95 percent CI also clinically important?

- In terms of effectiveness, do you think this trial provides evidence of superiority of artesunate compared with mefloquine?
Annex 1. Glossary

Allocation—The process of assigning patients to treatment groups.

Bias—Anything that leads to deviation of the results from the truth or processes leading to such deviation. This deviation can be either conscious or subconscious, but it allows for systematic error to enter a clinical trial and leads to an incorrect estimate of the outcome of interest. A bias can be a prejudice or a specific opinion favoring an issue before there is adequate information to support the position. Different types of bias occur in all studies and careful design, blinding, and randomization will effectively limit bias.

Blinding—An experiment is “double-blinded” if neither the investigator nor the patient is aware of the treatment a patient is receiving. The goal of blinding is to try to exclude investigator or participant biases that may influence outcomes or the reporting of outcome variables. To achieve blinding, a common practice is to administer identical appearing (and tasting) medication at identical times to all subjects.

Observer-Blind—The person measuring the outcomes in a study is not told what treatment patients have been received.

Double-Blind—Neither the observer nor the patient in a trial knows what treatment the patients received.

Cohort study—Patients are grouped according to their exposure or lack of exposure to a particular factor (e.g., a therapy). Outcomes over time then are compared in the different groups. Most cohort studies are prospective, but cohort analyses also can be performed retrospectively.

Confidence Intervals—The CI indicates the range within which the true study results lie. By convention 95 percent CIs are used and indicate that there is a 95 percent chance that the true results lies within the observed or estimated range. The larger the sample size, the narrower the CI of an observed value (e.g., mean reduction in blood pressure or percent of patients with pain relief). A smaller CI means that we can be more confident that the results of a study are indeed reliable and not due to chance alone. Confidence intervals are useful in determining what values could be expected for the population in general as opposed to those found with the smaller study sample.

Confounding factor or variable—A variable that can cause or prevent the outcome of interest, is not an intermediate variable and is associated with the factor under investigation. Confounders can also be described as an alternative explanation for a result in the study.

Control groups—Comparative trials utilize control groups as a reference standard against which to compare outcomes observed in the treatment (intervention) under investigation. Examples of control groups include placebo, alternative treatment, and historical. Control groups should closely resemble the group receiving experimental therapy in baseline demographic and clinical characteristics.
**Descriptive studies**—Include case reports and clinical series that present outcomes in treated patients. Benefits of treatment or adverse drug reactions may be reported.

**Intention-to-treat population**—The total number of patients assigned to receive a particular treatment, irrespective of whether they actually received it or not.

**Null Hypothesis**—All drug trials test a hypothesis that says Drug A is superior or safer or different due to some other parameter as compared to Drug B or no drug (placebo). An opposing hypothesis, the null hypothesis, states that there is no difference between Drug A and Drug B or placebo and the drug trial must seek to substantiate or refute it. Formulation of the null hypothesis is a vital step in testing statistical significance. Having formulated such a hypothesis, one can establish the probability of observing outcomes more different from the prediction of the null hypothesis, if the null hypothesis is true. That probability is what is commonly called the "significance level" of the results. That is, in our drug trial, we may predict that Drug A will produce a greater effect on our outcome of interest as compared to Drug B — this is our alternative hypothesis. We then consider how often we would expect to observe our experimental results (outcomes), or results even more extreme, if we were to take many samples from a population where there was no effect (i.e., we test against our null hypothesis). If we find that this happens rarely (up to, say, 5 percent of the time), we can conclude that our results support our experimental prediction — we accept our alternative hypothesis that drug A is different from Drug B. Even so any significant difference observed between drugs may be a result of error inherent in the study and not due to any differences of outcome due to the drugs themselves.

**Observational studies**—Studies where the investigators do not have control over what happens to patients or subjects. They are used to detect the cause of health problems or to describe consequences of treatment. Examples include case control, cohort, and cross-sectional studies. These studies may use comparison groups but do not control for the effects of variables as well as randomized controlled trials. This type of study can provide useful information about drug safety, but not usually on efficacy.

**The power of a study**—Indicates the likelihood of a hypothesized result being observed and is dependent on sample size. A value of 80 percent is taken by convention to be the minimum and indicates that there is an 80 percent chance of observing a real difference, for example, between the drug of interest and the comparator drug, meaning that there is a 20% chance of not observing a difference that really exists in the population.

**P Value**—Expresses the probability that any results observed in a study are real and not due to chance. A p-value of 0.05 indicates that there is a 1 in 20 probability that any study result is due to chance meaning that there is a 5 percent chance of observing a result which does not exist in the population. This means that there is a 95 percent chance that any difference observed, for example, between the drug of interest and the comparator drug, is a true difference in the population.

**Randomization**—The process of assigning patients to treatment groups by chance. True random selection occurs when patients have an equal chance of receiving standard or experimental therapy, assuming equal numbers of patients are to be assigned to each group.
Randomized comparative trial—Patients being studied are allocated by chance to either the new intervention that is being tested, or the alternative comparative treatment. This is the “gold standard” as it minimizes bias and control for confounding variables. However, randomized controlled trials are expensive and time-consuming and can raise ethical concerns about the treatment strategies. Therefore, some investigators may elect not to use this type of study design.

Sampling—The term used to describe the selection of patients for a clinical study. Many clinical studies aim to obtain a cross-section of the population, one that represents different ages, ethnicity, regions of the country, life styles, and health conditions. In other studies, the sampling may be more targeted and look at a particular subpopulation or group.

Statistical Power—The probability that the test will reject a false null hypothesis. In other words, statistical power is the likelihood that the experiment will detect a treatment effect that truly exists in the population. A power of 80 to 90 percent is considered acceptable in most clinical studies and this means that 5 out of 5 to 9 out of 10 trials will detect a difference that truly exists. A type I error is where a study detects a difference that does not truly exist in the population, that is, a false positive result (or a wrong rejection of the null hypothesis. A type II error is where a study does not detect a difference that truly exists, that is, a false negative result (or a wrong acceptance of the null hypothesis. As power increases, the chances of a Type II error decrease, and vice versa. A power of 80 to 90 percent means that 1 in 5 to 1 in 10 trials will fail to detect a difference that truly exists in the population.
## Annex 2. Checklist to Detect Common Problems Encountered in Articles


<table>
<thead>
<tr>
<th>Checklist</th>
<th>Potential Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objectives</strong></td>
<td></td>
</tr>
<tr>
<td>Are the objectives stated in the abstract, introduction or methods?</td>
<td>A drug may be tested only against a placebo, or against a drug with poor past performance and not against the standard or most effective drug in its class.</td>
</tr>
<tr>
<td>Is sufficient information given about the disease outcomes and the effects of the drug studied, so that you may judge how clinically important they are?</td>
<td>Clinically unimportant outcomes may be used.</td>
</tr>
<tr>
<td><strong>Methods</strong></td>
<td></td>
</tr>
<tr>
<td>Was a randomized control trial (RCT) done? • Best design for efficacy</td>
<td>The study design may be insufficient to be able to ascribe observed differences to the new drug being tested.</td>
</tr>
<tr>
<td>Was a case control study done? • Commonest design for safety</td>
<td></td>
</tr>
<tr>
<td>Was the study blinded? If not, is this explicitly discussed? And are the confounders accounted for?</td>
<td>Study participants or investigators are not blinded leading to possible bias in interpreting the results.</td>
</tr>
<tr>
<td>Is sufficient information given on the drugs used and the disease states treated in order to judge whether the study is relevant to your patient population?</td>
<td>Study patients may not be representative of the population that will take the drug. Often the patients in studies are fitter and have a more certain diagnosis and fewer concurrent diseases than the population who would take the drug.</td>
</tr>
<tr>
<td>Was the sample size of patients sufficient to detect significant differences in outcomes between intervention and control groups?</td>
<td>The number of patients may be too small to ensure any differences are not due to chance.</td>
</tr>
<tr>
<td>• Were the inclusion and exclusion criteria of patients specified? • Were the assignment of patients randomized? • Were the control subjects appropriate?</td>
<td>Patients may not have been randomized to study and treatment groups so that patients treated with the new drug may not be similar to those treated with the comparator drug.</td>
</tr>
<tr>
<td>• Is the drop-out rate of patients in the intervention and control groups reported? • Were the rates the same? If not, is any explanation given for the different rates?</td>
<td>Patients randomized to take the new drug may not have completed the study so that side-effects or less effect of the drug may not be reported. Patients with more side-effects or less effect may be more likely to drop out.</td>
</tr>
<tr>
<td>How many dose regimes were compared for each drug? Were they equivalent?</td>
<td>Different drugs may be compared using fixed non-equivalent doses; the comparator drug may be under-dosed.</td>
</tr>
<tr>
<td><strong>Review articles and meta-analysis</strong> (analysis across different RCT studies) • What criteria were used to find the articles? • How was the search done? • Which databases were used and were</td>
<td><strong>Review articles and meta-analysis</strong> may be biased by which studies are included and which not, and how each study was appraised. Studies with negative results may have been excluded.</td>
</tr>
<tr>
<td><strong>Checklist</strong></td>
<td><strong>Potential Problems</strong></td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>unpublished articles included?</td>
<td>Economic articles may be biased due to incomplete reporting of all the costs associated with a drug treatment, that is, non-drug costs (e.g., equipment) and outcomes, including negative ones (e.g., side-effects). Different discounting rates for drug costs and future benefits may be used to emphasize a drug’s cost-effectiveness ratio.</td>
</tr>
<tr>
<td>• Is there a description of how individual studies were appraised, and, if relevant, meta-analysis done?</td>
<td></td>
</tr>
<tr>
<td>Economic articles</td>
<td>Economic articles may be biased due to incomplete reporting of all the costs associated with a drug treatment, that is, non-drug costs (e.g., equipment) and outcomes, including negative ones (e.g., side-effects). Different discounting rates for drug costs and future benefits may be used to emphasize a drug’s cost-effectiveness ratio.</td>
</tr>
<tr>
<td>• Are all the costs associated with drug treatment, including good and bad outcomes, described? (not just prices)</td>
<td></td>
</tr>
<tr>
<td>• Has discounting been used to reflect the costs of any future benefits or consequences in present day values?</td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>The presentation and analysis of data may be misleading.</td>
</tr>
<tr>
<td>• What measures of outcome were used?</td>
<td>Differing efficacy can only be assessed by using established measures, for example, relative or absolute risk reduction or number of patients needed to treat. Economic evaluation requires using standard analyses.</td>
</tr>
<tr>
<td>• Were any differences shown due to real differences between intervention and control groups or just due to chance from small sample size or by selecting a small subset of patients?</td>
<td></td>
</tr>
<tr>
<td>• For economic studies:</td>
<td></td>
</tr>
<tr>
<td>What type of analysis was done? cost minimization? Cost-effectiveness analysis?</td>
<td>The statistical significance of a trial may be valid but the clinical significance may be weak.</td>
</tr>
<tr>
<td>Has a sensitivity analysis been done?</td>
<td></td>
</tr>
<tr>
<td>Were the differences in clinical outcome between groups large, important and relevant as well statistically significant?</td>
<td>Confounding variables may not have been adequately controlled so that any differences seen are due to the confounders not the new drug.</td>
</tr>
<tr>
<td>Were all recruited patients taken into account in the analysis? If patients who died or dropped out of the study are excluded from the analysis, there may be a bias towards greater efficacy.</td>
<td></td>
</tr>
<tr>
<td>Conclusions</td>
<td>The conclusions may not agree with the results or may be extrapolated too widely.</td>
</tr>
<tr>
<td>Were the populations for which conclusions were drawn represented by the subjects in the study?</td>
<td>There may be little discussion of safety in relation to efficacy.</td>
</tr>
<tr>
<td>Was there any discussion of whether the potential benefits were worth the potential harm? If not, maybe the likely benefits are not worth the risk.</td>
<td></td>
</tr>
<tr>
<td>Funding</td>
<td>The study may have been funded by a drug company for their own product; often drug companies do not publish negative studies.</td>
</tr>
<tr>
<td>• Is there a description of how the study was funded?</td>
<td>Study may not be peer reviewed but published either in a “throw-away” journal or in symposia proceedings; alternatively it may be published in a journal with less rigorous peer review.</td>
</tr>
<tr>
<td>• What is the reputation of the authors and are their affiliations described?</td>
<td></td>
</tr>
<tr>
<td>• Is the study published in a peer-reviewed journal that is listed in Index Medicus, which covers all major reputable journals?</td>
<td></td>
</tr>
<tr>
<td>• Are references cited and are they reputable?</td>
<td></td>
</tr>
</tbody>
</table>