REVIEW ARTICLE

Evaluation of Clinical Trials

Alireza Khatami, MD, MSPH
Alireza Firooz, MD

Center for Research and Training in Skin Diseases and Leprosy, Tehran University of Medical Sciences, Tehran, Iran

Corresponding Author:
Alireza Khatami, MD, MSPH
Address: No. 79, Taleqani Avenue, Tehran 14166-13675 I.R. IRAN
Email: akhatami@tums.ac.ir

Received: November 2, 2007
Accepted: January 12, 2008

Abstract

In a number of important clinical issues such as evaluation of the efficacy or effectiveness of therapeutic or preventive interventions as well as for comparing the harms of interventions, randomized controlled trials (RCTs) provide the highest levels of evidence, either directly or indirectly. It is obvious that critical appraisal of these studies to assess their validity and precision is of paramount importance.

The aim of this review is to provide the readership an outline about different types of RCTs, the importance of proper appraisal of RCTs, an overview of the most important factors that have an influence on the validity of an RCT and a strategy for systematic evaluation of those factors, and to introduce some useful methods for improving design, implementation and reporting of RCTs as well as some tools that are used for the evaluation of these studies.

It is expected that after reading this review, the reader obtains some knowledge about different phases and types of RCTs, as well as being enabled to evaluate the four major factors: randomization sequence generation, randomization concealment, blinding and intention to treat analysis that affect the validity of an RCT. *(Iran J DErmatol 2008;11:76-85)*

Keywords: randomized controlled trials, dermatology, evidence-based dermatology

Introduction

According to Dr Sackett's definition of evidence-based dermatology (EBM) one of the three main components of EBM is to use the best currently available evidence from research[1]. In some important subjects of medicine such as efficacy and efficiency of therapeutic modalities, prevention of diseases and evaluation of harms of preventive or therapeutic interventions, randomized controlled trials (RCTs) provide the highest level of evidence, either directly or indirectly. It is obvious that proper use of the evidence from RCTs is only possible if the validity and the precision of these studies can be assessed correctly.

The objectives of this article are to familiarize readers with:

- Different types of RCTs
- The importance of the proper evaluation of the quality of an RCT
- The most important factors that can influence the validity and precision of an RCT and provide a strategy for systematic evaluation of those factors
- Enlisting other factors that can influence the validity of an RCT
- Some tools that have been developed for improving the quality of conducting and reporting of RCTs as well as some tools that have been developed for evaluation of the quality of an RCT

What is a Clinical Trial?

A clinical trial is an interventional epidemiological study; which in its simplest type, a researcher assesses the results of an intervention on an outcome.[2] A randomized controlled trial is a study in which a comparison between two or more interventions is possible through randomization and assigning at least one control group in the trial.[3,4] There is no need to mention that it is possible to compare an intervention with no treatment in an RCT. According to the provided definition, clinical trials encompass a wide spectrum of interventional studies. For sake of simplicity, different phases are

defined for trials which are conducted on humans to evaluate the efficacy or harm of an intervention.[4] Briefly, following phases have been defined:

Phase I

In which the safety and the mechanism of action of an intervention is studied on a limited number of volunteers. The number of participants at this phase is usually less than 100.

Phase II

This phase which is also known as the 'pilot efficacy study' is the first trial phase in which efficacy of an intervention is evaluated. While in vaccine trials immunogenicity is evaluated, the main outcomes in drug trials are efficacy and safety. In this phase, participants are usually allocated to experimental and control groups according to a generated random sequence and the number of participants is 100 to 500.

Phase III

This phase is known as 'extensive' clinical trial. In this phase, the efficacy and safety of an intervention are assessed on a larger number of participants. In phase III trials, random allocation of participants to experimental and control groups is accomplished. The total number of participants may be more than one thousand and the trial is commonly multi-centric.

Phase IV

This phase pertains to the studies which are performed after a vaccine or drug receives its marketing approval from authorized organizations. The aim of this phase is to look for particular adverse events or evaluation of the efficacy of an intervention in a long term. Since conduction of phase IV trials need ethical approval, they are different from those studies which are generally performed to assess the adverse effects of interventions and are known as 'post-marketing surveillances'[4].

The first published RCT was a clinical trial conducted by the British Medical Research Council, in which the efficacy of streptomycin in the treatment of pulmonary tuberculosis was assessed. In 1948, the results of this study were published in the British Journal of Medicine[5]. While in the late 1980s, about 5000 RCTs were published in medical journals each year. In 1998, 12000 RCTs were published. During the past two decades, small sample size RCTs, in which surrogate physiological outcomes were evaluated, were substituted by large sample size RCTs which are designed to measure clinically meaningful outcomes such as death.[5]

Clinical Trials: Different Design

Proper design of a clinical trial is of vital importance and is related to the aim of the study. In brief, the most commonly used designs for RCTs are as follow:[6]

RCTs with parallel arms

The simplest design for an RCT is a trial which has two parallel arms (figure 1). In these RCTs, eligible participants after giving their informed consent are randomly allocated to intervention (experimental) and control groups and predetermined interventions are taken on them. According to a predefined time schedule, the outcomes are measured. Principles of parallel design can be extended to studies with more than two arms.

Cross-over studies

In these RCTs, a predetermined intervention is done on each group of participants after random allocation. After the end of the time of the interventions and a so-called 'wash-out period' which designates the needed time for elimination of the effects of the interventions, each group of the participants receives the intervention which was initially administered to the other group (figure 2). In chronic dermatoses such as psoriasis, cross-over RCT provides two advantages: (1) since each patient is his control as well, the total required number of samples is reduced and (2) some potential biases can be controlled. It should be kept in mind that cross-over design is not suitable for some diseases. For example this design is not a proper RCT design in acute cutaneous leishmaniasis which is a self limited disease in most cases.

RCTs with factorial designs

RCTs with a factorial design provide the possibility of comparing two interventions with each other as well as their combination to each intervention and to the control group. In its simplest design in which two interventions are under evaluation, the RCT contains 2 x 2 = 4 arms (figure 3).

The Importance of RCTs in Evidence-Based Medicine

As it was mentioned before, clinical trials are of crucial importance when the hierarchy of evidence

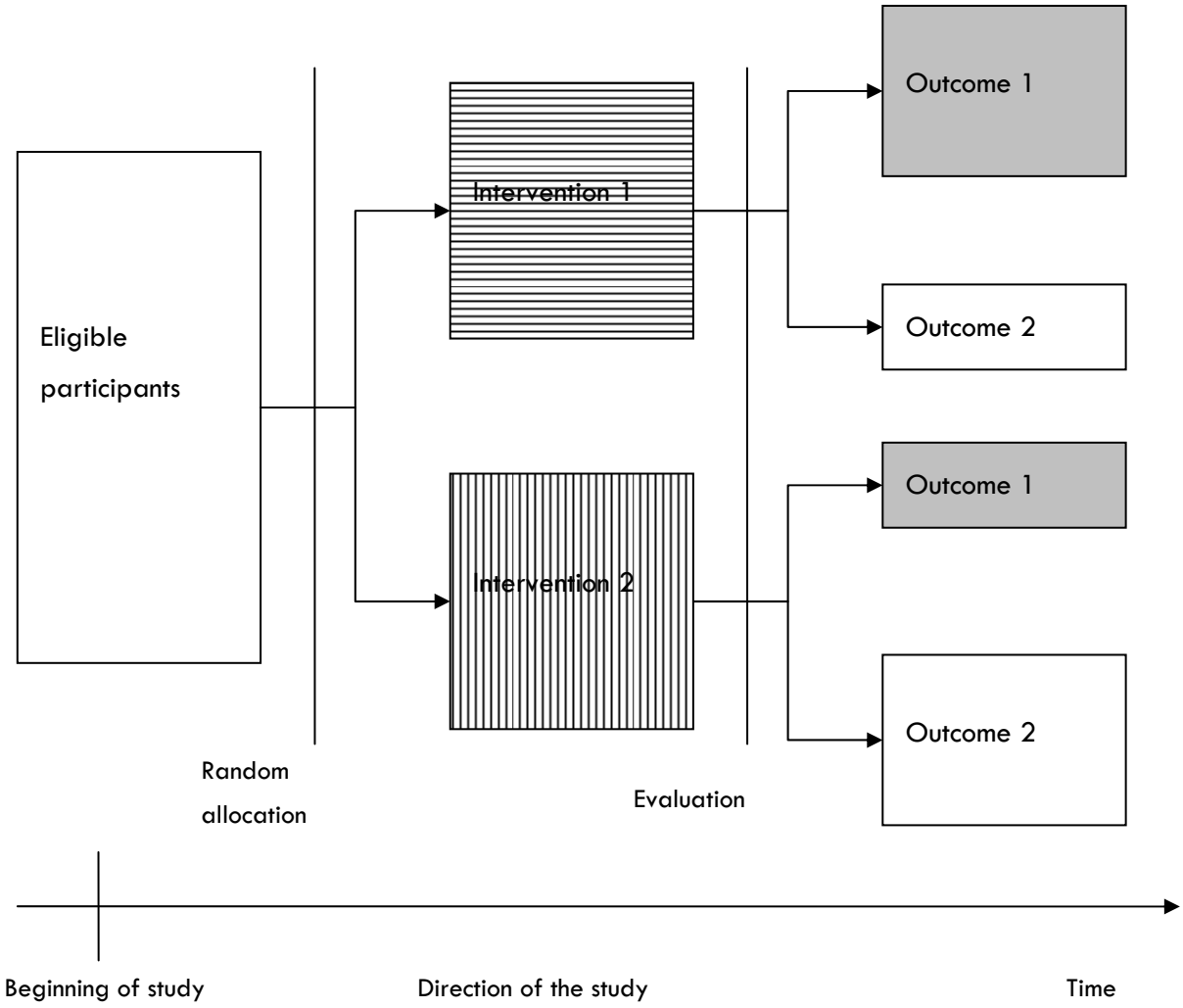

Figure 1: *An RCT with two parallel arms.*

with regard to treatment, prevention and harm are looked for.[7] It is because the highest level of evidence of aforementioned topics is systematic review of the homogenous RCTs followed by the individual RCTs with narrow confidence interval.

RCTs are generally considered as the 'gold standard' studies for evaluation and/or comparison of the efficacy of different interventions.

Potential Errors in Clinical Trials

As it is pertinent to other epidemiological study designs, there is a probability of occurrence of different types of errors in an RCT.[8] These errors may occur during designing, conduction, analysis, and interpretation of an RCT. These potential errors are summarized in figure 4. Acknowledging these errors can enhance the understanding of the importance of the different strategies that are used to minimize the probability of the occurrence of such errors in RCTs and consequently can improve the ability to evaluate the validity and precision of RCTs.

Chance errors

They are categorized into:

- Type I error, which is also known as α error and is defined as the probability of rejection of the null hypothesis (H0) when it is correct.
- Type II error, which is also known as the β error and is defined as the probability of acceptance of H_0, when it is incorrect.

The probability of occurrence of chance errors is related to the sample size of the study and

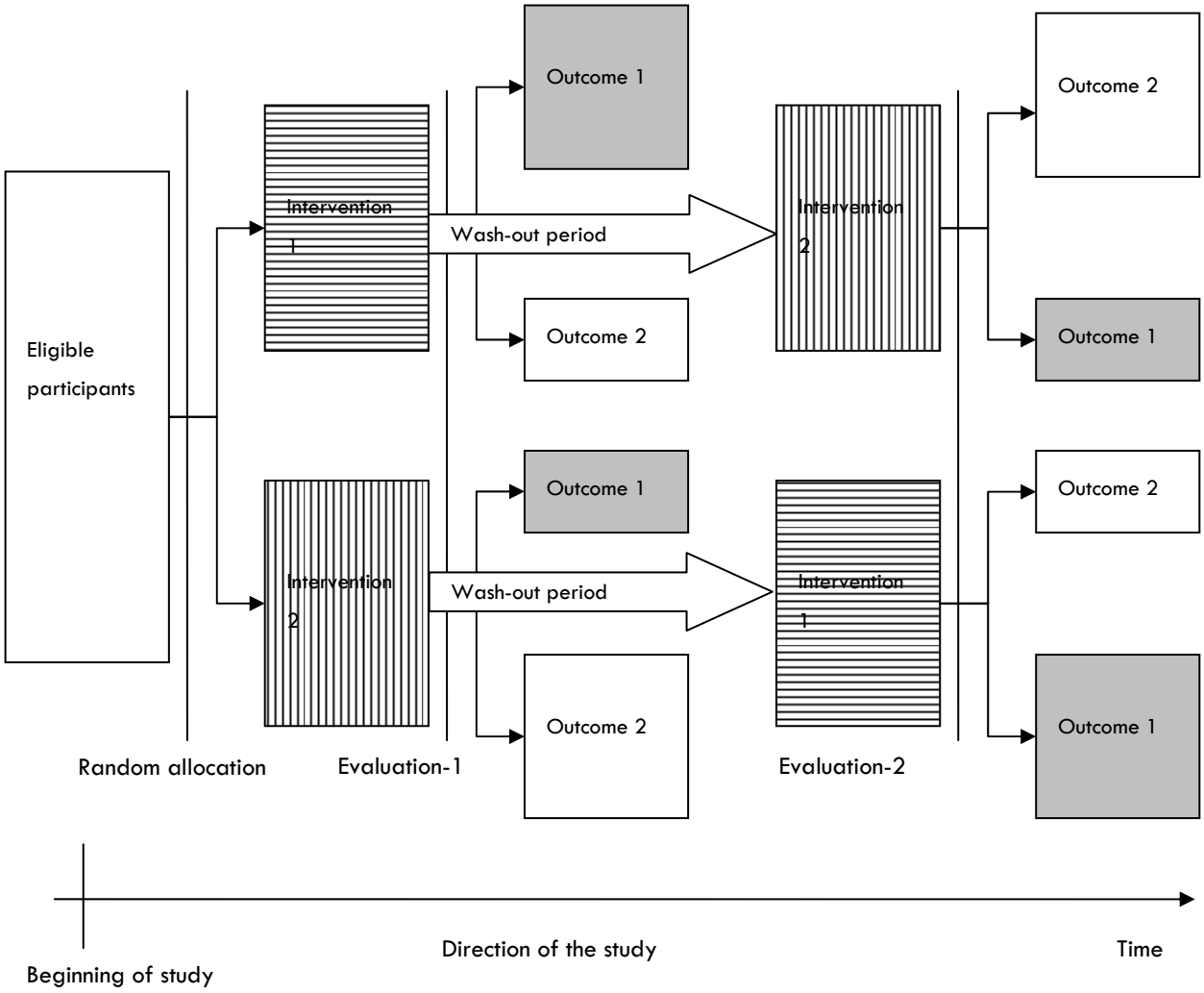

Figure 2: *An RCT with a cross-over design.*

increasing the sample size can reduce the probability of such errors.

Systematic errors

Systematic error which is also known as bias is an error which its occurrence is not related to chance so the probability of its occurrence is not affected by the sample size, which means that there is no decrease in the probability of such errors when the sample size is increased. Three main categories of biases can be defined:

- Selection bias: is a systematic error which is related to the factors affecting the selection of the participants for the study.
- Information bias: is a systematic error which results from the differences in data gathering from the participants.
- Confounding: This means mixing the effect of the understudy intervention with the effects of another variable (a confounder). There are several techniques for controlling the effects confounders. Some of these techniques should be considered during the design of a trial and some others can be used during data analysis.

The internal validity of an RCT is determined by strategies chosen to minimize the aforementioned errors.

Validity of an RCT: the most important determinants

There are four main factors which determine the validity of an RCT. They are:

- Generation of the randomization sequence: Random allocation provides the opportunity to allocate each participant in either the intervention or the control group according to a random sequence and independent of the researcher's preference for assigning a

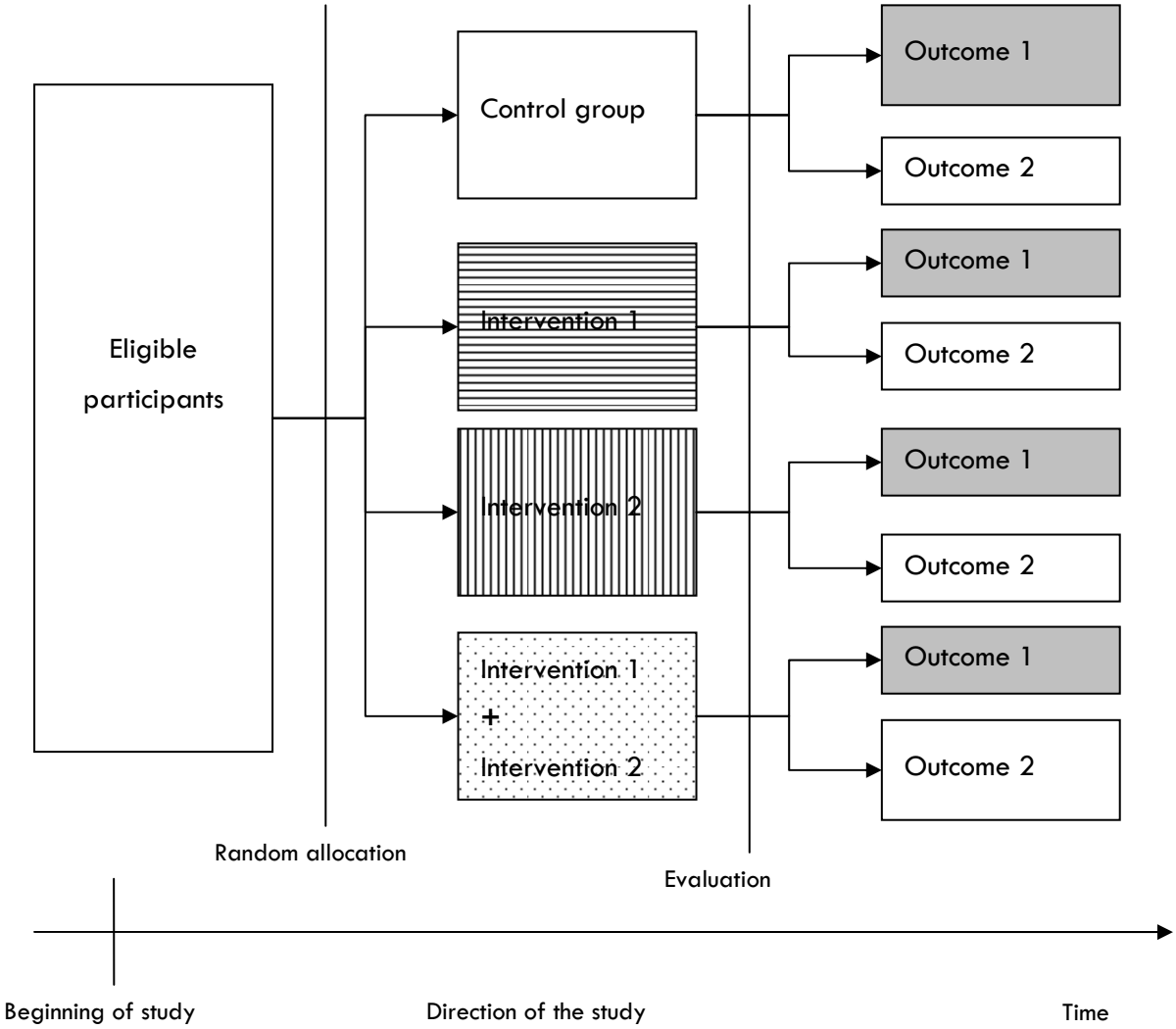

Figure 3: *An RCT with a factorial design.*

particular intervention to a certain individual. The main objective of random allocation is to eliminate the selection bias. It may also decrease the effect of some confounders. When the quality of an RCT is appraised, it is important that the appraiser look for detailed description of random sequence generation. If such a detailed description is not provided in an RCT, the validity of that RCT is questionable[9,10]. Some acceptable methods for the generation of a random sequence include:

- Computer generation of a random sequence
- Using tables of random numbers
- Random selection of numbers or envelops which contain numbers
- Flipping a coin
- Using a dice or shuffling cards with printed numbers

It is important to remember that methods such as recruiting participants according to their file number, date of birth or date of admission may not prevent selection bias. Such methods are known as pseudorandomization methods.[10,11] Adetugbo and Williams[11] studied the quality of the published RCTs in one of the leading dermatology journals in the world and reported that the method of the generation of the random sequence was only reported in 1 out of 73 published RCTs.

- Randomization concealment: This means unawareness of the recruiters of the participants about the intervention which is assigned to be administered to them. Without implementation of concealment of randomization sequence, it may

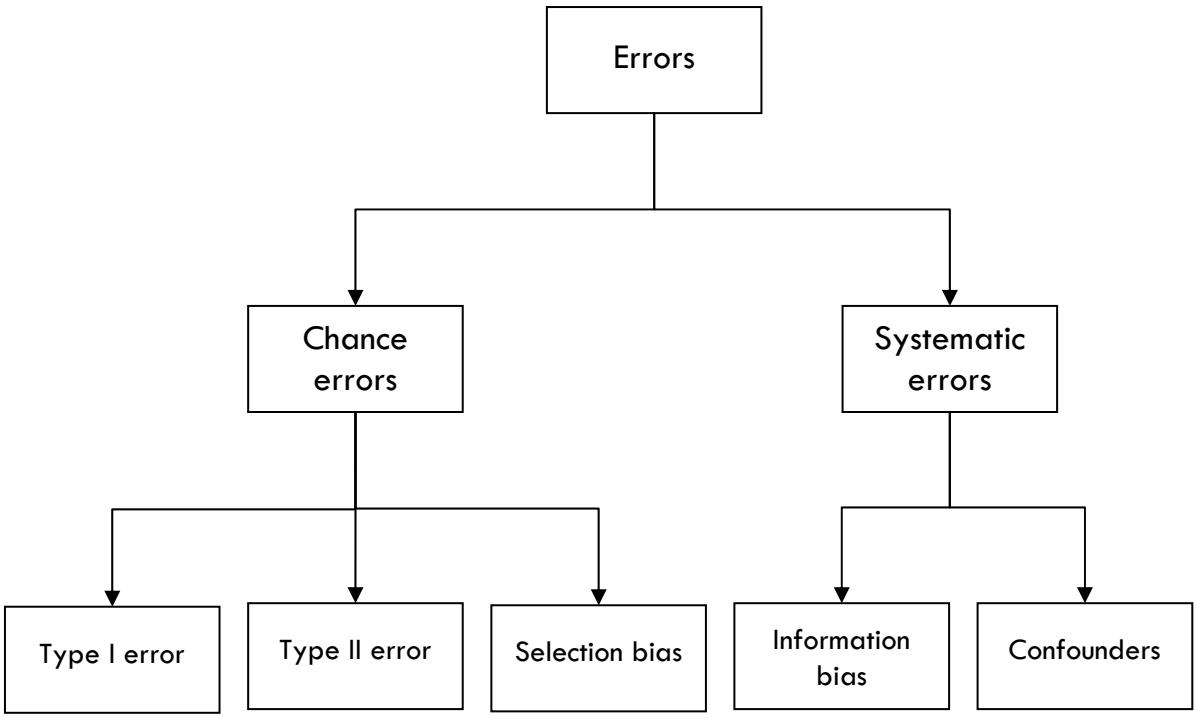

Figure 4: *Potential errors in an RCT.*

be possible for those who are responsible for recruitment of the participants to decide to administer certain interventions for certain participants. In some pseudorandomization methods such as using date of admission, it may not be possible to implement concealment of randomization sequence. It was reported that those RTCs that had not reported or incompletely reported the concealment of randomization sequence showed up to 40% overestimation of the efficacy of the understudy intervention in comparison with those RCTs that had described the concealment of randomization sequence in details.[10] In addition, those RCTs that had poorly reported the concealment method showed the most heterogeneity in their findings[10].

It is worth mentioning that in 93%, 89%, 48% and 45% of RCTs which were published in dermatology, rheumatology, obstetrics and gynecology and general medical journal, respectively, there were no description of concealment of randomization sequence.[10] Acceptable methods for concealment of randomization sequence include:

- Sequentially numbered, opaque, sealed envelops (SNOSE)
- Serially numbered containers
- Pharmacist control
- Central randomization in which who recruits the participants contacts the center where the randomization sequence is available and asks for the assigned intervention after recruiting each participant.

As it is evident, proper random allocation depends on appropriate implementation of concealment of the randomization sequence, so it is of paramount importance that researchers who design and conduct RCTs pay meticulous attention to the implementation and reporting of the used methods in these regards. Adetugbo and Williams[11] reported that only in 5 out of 73 evaluated RCTs for their quality, proper description of methods for concealment of randomization sequence were available.

- Blinding (masking): describes a situation in which participants, investigators, or outcome assessors are unaware about the interventions that are administered to the participants. The history of blinding backs to 200 years ago and use of blinding is not limited to RCTs. The main objective of blinding is to eliminate information bias. Blinding may be applied to participant, health care providers, data collectors, outcome assessors, data analyzers and the authors of the

manuscript.[12] Devereaux et al.[13] in their study on different interpretations of blinding reported 10, 17 and 15 different interpretations for single-blind, double-blind, and triple-blind, respectively; in physicians who had participated in their study. They also found 5, 9 and 7 different definitions for single-blind, double-blind and triple-blind, respectively; in the textbooks they included.[13] In addition, they assessed 200 RCTs for the proper description of blinding and found that out of 5 and 83 RCTs which had been described as single-blind and double-blind, only in 2 and 11 RCTs, one and two groups, respectively were really blinded.[13] According to the same study, blinded groups were not described in 41 out of 83 studies which had been reported as double-blind. According to the International Conference on Harmonization (ICH) guideline, blinding is the preferred terminology in comparison with masking.[12]

In brief, studies can be categorized into following groups according to their blinding status:

- *Non-blind or open-label*: in which all groups that are involved in a study are aware of the assigned interventions.
- Single-blind: in which participants, investigators or outcome assessors are unaware of the assigned interventions. Although the term 'single-blind' may suggest that participants are blinded in a study, it must be emphasized that this term may also indicate that either the investigators who recruit and administer the interventions or the outcome assessors are blinded.
- *Double-blind*: is usually used in those trials in which participants and investigators or outcome assessors are unaware of the assigned interventions.
- *Triple-blind*: is usually used to describe a double-blind trial in which the data-analyzers are also unaware of the administered interventions.
- *Quadruple-blind*: is a rarely used term to describe a trial in which participants, investigators, outcome assessors and data-analyzers are unaware of the administered interventions.

As it has been previously mentioned, the main objective of blinding is to decrease information bias. Contrary to randomization and concealment of the randomization sequence which should be accomplished before allocating participants in the experimental and control groups, blinding is implemented after participants are allocated into their assigned groups. In addition, unlike randomization and concealment, which can always be accomplished, it may not be possible to perform blinding to some groups involved in an RCT. For example when a surgical method is compared to a medical therapeutic regimen, it is not possible to keep all involved groups blinded. It should be emphasized that a double-blind RCT is not always synonymous with a high quality RCT.[13,14] Even more, it is generally accepted that proper generation of randomization sequence and appropriate concealment sequence are more important determinants of the quality of an RCT. It is obvious that terms such as single-, double- or triple blind might be interpreted differently, so it is important that while an RCT is being read, particular attention is given to the groups that are kept blinded in that trial.

- Intention to treat analysis: commonly abbreviated as ITT is used when the primary outcome of a study is measured in all participants according to their original random allocation at predetermined times during the follow-up period. Two main causes of exclusion of the participants from their assigned groups after the beginning of the study are protocol deviation and loss to follow-up. Some reasons for protocol deviation are changes in the eligibility criteria of a participant which results in a withdrawal criterion and not to follow the assigned intervention. Withdrawals of consent, refusal to complete the study, unavailability of the participants for example as a result of migration are among common causes of loss to follow-up.[15] Proponents of ITT mention following reasons for their support for ITT[16, 17]:
  - ITT assists to maintain prognostic balance between study arms
  - ITT limits possible explanations for observed difference among different subgroups of participants in a study
  - ITT minimizes the effects caused by withdrawals
  - ITT minimizes type I errors
  - ITT maximizes the generalizability of a study

Opponents of ITT mention following reasons in support of their idea[16, 17]:

- ITT is a too conservative approach and may increase the probability of committing a type II error

- ITT decreases the probability of showing the positive effect(s) of an intervention
- Approach to efficacy is more important than approach to effectiveness

However, the importance of ITT can be explained through considering the fact that those participants that are excluded from a study after they have randomly allocated into their intervention groups are probably different from those remained in the study and that those participants that do not follow the assigned interventions usually are different from other participants in terms of prognosis.[15] For further discussion on ITT reading Hollis and Campbell's article in the British Medical Journal is recommended[17]. It is important to know that ITT can only be performed properly when all data on the outcome of interest is available in all randomly allocated participants.[16,17]

Other Factors That May Affect the Quality of an RCT

Disease definition

It is important that the diagnostic criteria for confirming the clinical diagnosis of a certain disease (e.g. performing immunofluorescence studies to diagnose an immunobullouse disease) and diversity of the clinical manifestations in diseases such as atopic dermatitis which presents with a wide spectrum of clinical features among different age groups be acknowledged.

Clinically sensible an meaningful outcomes

During assessment of the quality of an RCT, it is important that the clinically meaningful and measurable outcomes both for the physician who might use that trial and for the patients to whom the results of the trial might be applied be looked for. For example in an RCT which is designed for the evaluation of the efficacy of a systemic intervention in the treatment of genital warts, complete clearance of the warts is a meaningful outcome while reduction in the total number of the lesions is not. Until better scales are developed, the best outcomes in RCTs are those which are the most objective and simplest to measure. Outcomes with categorical scales such as death vs. alive or cured vs. not cured are the most understandable outcomes. In those RCTs that such a simple and objective outcome definition is not possible, an ordinal scale may be the best option.

Comparable baseline characteristics among the study groups

It is possible that some unknown confounders be present in every research study. In addition, random allocation does not warrant the similar distribution of all baseline groups among the study groups; in other words, even the proper implementation of random allocation may not eliminate the different distributions of some characteristics among the study groups which may be caused by chance alone. If the study groups are small, the arbitrary 0.05 level of significance may not be able to detect the differences between one or more characteristics among the study groups. Using techniques such as block randomization or stratification can minimize such problems. It is also important to notice that it is not necessary that all study groups be the same regarding all baseline characteristics. It is important that those characteristics that may have a considerable effect on the response to a certain intervention be comparable at the baseline. It is worth mentioning that differences in the baseline characteristics among study groups may also be addressed at data analysis using techniques such as multi-varaite analysis.

Data dredging

'Data dredging' is applied when a large number of the statistical tests are performed on several outcomes in various subgroups of participant and there is an overemphasis on one found significant difference (i.e. $p < 0.05$). A considerable number of RCTs in dermatology report up to 10 different outcomes that are measured at different times. It should be kept in mind that when the level of significance is considered at 0.05, there is the possibility to get one 'significant' result in 20 performed statistical tests by chance. So, when someone reads an RCT, looking for possible existence of data dredging is quite important. Looking for a clearly predefined meaningful outcome measure as the criterion for the success of an intervention may assist to avoid misleading caused by data dredging.

Inappropriate selection and use of statistical tests

It is not infrequent that parametrical statistical tests are used to test data that do not follow Normal distribution. In data with skewed distribution (also known as non-parametrically distributed) such as number of the acne lesions, non-parametrical statistical tests should be used. A quick method for determining whether a data set follows Normal distribution or not is to subtract the standard

deviation of the distribution multiplied by 2 from the arithmetic mean of the given data. If the result is less than zero, it is probable that the data be skewed and use of the proper non-parametrical tests is necessary.

Selection of an inappropriate variable for statistical analysis

Sometimes statistical analysis is performed on a wrong variable. If someone is not aware of this possibility, this mistake may be overlooked easily. For example, in an RCT which is designed and conducted to compare the efficacy of two therapeutic interventions in a certain disease, it is the difference between the two groups that should be statistically tested, while in some studies the research team compares the difference in each group with the baseline of the same group separately and if there is a significant difference in one of these comparisons, they conclude that one intervention is significantly better than the other one which is obviously wrong.

Interpretation of negative RCTs

It is important to acknowledge that finding no significant difference between administered interventions does not mean that there is no difference in the efficacy of those interventions. To obtain a better understanding about the non significant differences among different intervention groups, using 95 % confidence intervals are really helpful for making the right decision.

Conditions under which an RCT has been conducted

The most important factors include: similar follow-up of the study groups in terms of interval and duration and that the groups, except for the understudy interventions, receive the same care.

Role of sponsors

It must be kept in mind that billions of dollars of benefit from an intervention may depend on convincing the health-care decision makers. So, particular attention should be paid to the financial resources of an RCT.

Efforts to Improve the Conduct and Reporting Of RCTs

Several strategies have been developed for improving the quality of RCTs through optimization of conduct, report and publication of these studies. Herein, two of these strategies are briefly discussed.

- Good clinical practice (GCP), which is a standard for design, conduct, auditing and monitoring, recording, analysis and reporting clinical trials, has been developed to warrant the accuracy and validity of the data as well as assuring that the rights of the participants are protected and their data is kept confidential.
- There are some tools that are developed for the standardization reporting RCTs. One of them is known as Consolidated Standards On Reporting Trials (CONSORT), which is an important tool in clinical researches. CONSORT assists to improve the quality of reporting RCTs through an evidence-based viewpoint. CONSORT is available at the following website: www.consort-statement.org/Statement/revisedstatement.htm. Nowadays, a large number of well known medical journals ask for the completed CONSORT checklist and flow diagram with a manuscript of an RCT[18]. This step should be accomplished before the manuscript is sent for being peer-reviewed.

Tools for the assessment of the quality of clinical trials

Some tools have been developed for assessing the quality of RCTs. Two of these tools are briefly described:

- Jadad scale: is a tool which gives one score to each of the following items:
  - Randomization
  - Blinding
  - Description of Randomization methods
  - Description of blinding methods
  - Data on follow-up (withdrawal)

  According to Jadad Scale, RCTs which have a total score of 4 or 5 are considered as good-quality RCTs. If the total score of an RCT is 2 or 3, it is of fair-quality and if an RCT gets a total score of 0 to 1, it is considered as an RCT of poor-quality[19].
- There are several tools that have been developed for critical appraisal of the quality of different types of studies and are available at the Critical Appraisal Skills Programme (CASP). One of these tools has been developed for critical appraisal of RCTs and is available at the CASP website[20].

Conclusion

Clinical trials are among the most important clinical studies particularly in an evidence-based approach to studies concerning treatment,

prevention or harm. Anyone who intends to use the results of clinical trials should be able to critically appraise such studies.

References

1. Sackett DL, Rosenberg WM, Richardson WS, et al. Evidence-Based Medicine: What it is and what it isn't. BMJ 1996; 312:71-2.
2. Cummings SR, Grady D, Hulley SB. Designing an experiment: clinical trials In Hulley SB, Cummings SR, Browner WS, et al(eds). Designing clinical research: an epidemiologic approach. Philadelphia: Lippincott Williams & Wilkins; 2001: 143-55.
3. Clinical trials. In: Greenberg RS, Daniels SR, Flanders WD, et al (eds). Medical epidemiology. New York: Mc Graw-Hill; 2005: 107-31.
4. Last JM, Abramson JH, Friedman GD, et al (eds). A dictionary of epidemiology. New York: Oxford University Press; 1995.
5. Devereaux PJ, Yusuf S. The evolution of the randomized controlled trial and its role in evidence-based decision making. J Intern Med 2003; 254:105-13.
6. Machin D. General issues. In: Machin D S, Green S (eds). Textbook of clinical trials. Chechister: John Wiley & Sons, Ltd; 2004 :11-44.
7. Sackett DL(ed). Evidence-based medicine. How to practice and teach EBM Edinburgh: Churchill Livingstone; 2000: 173-7.
8. Rothman KJ. Epidemiology: an introduction. New York: Oxford University Press; 2002.
9. Williams HC. How to critically appraise a study reporting effectiveness of an intervention. Evidence-based dermatology. In: Williams HC, Bigby M, Diepgen T, et al (eds). Evidence-based dermatology. Oxford: BMJ Publishing Group; 2003: 56-63.
10. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet 2002; 359: 614-8.
11. Ádetugbo K, Williams HC. How well are randomized controlled trials reported in the dermatology literature? Arch Dermatol 2000; 136:381-5.
12. Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. Ann Intern Med 2002; 136:254-9.
13. Devereaux PJ, Manns BJ, Ghali WA, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 2001; 285:2000-3.
14. Mills EJ, Wu P, Gagnier J, et al. The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement. Contemp Clin Trials 2005; 26:480-7.
15. Fergusson D, Aaron SD, Guyatt G, et al. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ 2002; 325:652-4.
16. Schoenfeld PS. Evidence-based medicine in practice: applying intention-to-treat analysis and perprotocol analysis. Am J Gastrenterol 2005; 100:3-4.
17. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999; 319; 670-4.
18. CONSORT. Available at: http://www.consort-statement.org. Last accessed: July 7, 2007.
19. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996; 17:1-12.
20. Critical appraisal tools. Available at: http://www.phru.nhs.uk/casp/critical_appraisal_tools . htm. Last accessed: July 7, 200