National Medicines Information Centre

VOLUME 9
NUMBER 5
2003

ST. JAMES'S HOSPITAL • DUBLIN 8
TEL 01-4730589 or 1850-727-727 • FAX 01-4730596 • e-mail: nmic@stjames.ie

INTRODUCTION TO CLINICAL RESEARCH AND CRITICAL APPRAISAL

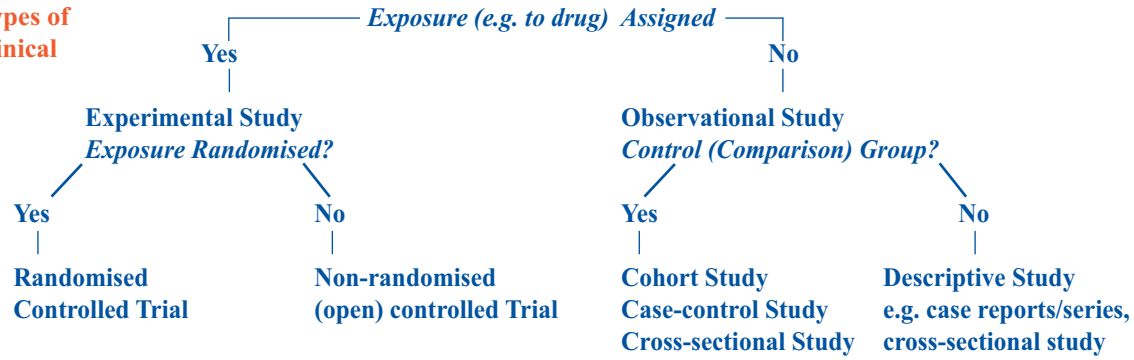
SUMMARY

- **Clinical studies are either experimental or observational**
- **Randomised controlled trials are the gold standard for evaluating drug treatments**
- **Study type and design is determined by the question under investigation**
- **Use of critical appraisal "checklists" can help in evaluating published papers**

INTRODUCTION

In current clinical practice, physicians come into regular contact with clinical research, particularly through their evaluation of new medicines. Moreover, the use of evidence-based medicine in clinical practice is seen as one of the key elements in future developments in the health service in Ireland[1]. Many physicians may have difficulty in understanding the design of clinical studies and in interpreting their results. This bulletin will give an overview of the most frequently used clinical study designs, the types of statistical methods commonly used and will provide guidance on how to read and critically appraise published papers.

Figure l. Types of Primary Clinical Research[2]



TYPES OF RESEARCH

Most published research is **quantitative** in nature (Figure 1). This type of research usually involves a narrow focus of research, e.g. comparing different drug treatment regimens[2]. It measures outcomes (endpoints) in a sample, with the aim of extrapolating the results to the larger population. **Qualitative** research is based on understanding human experience. Qualitative research is often used to ascertain patients' or clinicians' views on the availability or quality of various healthcare services[3,4]. Qualitative research uses small samples (e.g. focus groups) and is more flexible than quantitative research; however it may not be possible to extrapolate the results to the larger population.
Table 1 contains definitions which will be helpful in reading the following sections.

Clinical Trials:

Randomised controlled trials (RCTs) are regarded as the gold standard in clinical research[5]. They are principally used to evaluate drug treatment regimens. Randomisation removes the potential for selection bias by an investigator, therefore any difference in outcome between the treatment groups is likely to be valid. The randomisation process should be "concealed" so that the investigator cannot know or predict which treatment will be next allocated in the study. The recently published PROSPER study which compared the use of pravastatin with placebo in elderly patients with, or at high risk of developing, cardiovascular disease is an example of a RCT[6]. In a single-blind RCT the investigator is aware of which treatment has been allocated, but the patient is not; in a double-blind RCT neither

the patient nor the doctor is aware of the allocation[7]. Non-randomised clinical trials may be subject to selection bias and therefore will only be useful for providing supplementary efficacy or safety data.

Observational Studies:

A cohort study involves the identification of a group with a common connection (e.g. born in the same week)[8]. The group (cohort) is followed over time to assess an "exposure" and "outcome" of interest (e.g. development of lung cancer in smokers as in the British Doctors Study[9]). Such studies enable calculation of true incidence and absolute risk of a disease. Since they identify exposure and then proceed to outcome (disease) they are less subject to bias than other observational studies. They are especially useful when it is not possible for ethical reasons to use a RCT design (e.g. randomising to cigarette/no cigarette use). Because cohort studies may run over many years and may need large numbers of participants, they are often very expensive.

Table 1: Useful Definitions

Bias: *Any factor which leads to a conclusion which differs from the truth. Major sources of bias in research are the use of non-random sampling, selection bias (using a sampling group not representative of the population), recall bias and non-response bias.*

Confidence Interval: *A range of values around a study result within which, at a given level of confidence, the true value of the population value is likely to be found (e.g. 95% CI means that the true population value has 95% chance of lying within the range of the confidence interval presented).*

Confounding: *Any factor which distorts the association between exposure and outcome.*

Critical Appraisal: *The process of assessing and interpreting evidence by systematically considering its validity, results and relevance.*

Evidence-based medicine: *The integration of clinical expertise with the best available evidence from systematic reviews and RCTs to reinforce clinical skills, judgement and management.*

Number needed to treat (NNT): *Number of people needed to be treated in order to achieve benefit in one person.*

Null hypothesis: *A test of significance, which is based on the premise that there is no association between 2 variables in a study (i.e. the reverse of what the study was designed to show). The aim is to reject this null hypotheses, thereby showing statistical significance.*

P value: *The probability of getting the observed result (or one more extreme) if the null hypothesis were true. $P<0.05 =<5\%$ probability, which is the conventional level of statistical significance.*

Power of a Test: *The probability that a test will produce a significant difference at a given significance level, e.g. many studies use a power of 80% and a significance level of 5%.*

Primary Research: *This relates to research that reports original results.*

Risk (Absolute Risk): *Chances of something happening in a specific population, i.e. number of events in a population in a time period, divided by the total population at the start of the time period.*

Secondary Research: *Studies such as overviews, systematic reviews and meta-analyses which summarise and draw conclusions from primary research.*

(Statistical) significance: *The observed result is unlikely to have occurred by chance.*

Validity: *Whether the study results are true or have been affected by bias.*

Case-control studies work backwards from the outcome (disease) in order to identify a possible association between the disease and a particular exposure[10]. Patients with the disease (e.g. vaginal carcinoma) are identified and matched (usually by age and gender) with a group without the disease to act as a control. Matching ensures confounding is taken into consideration. The level of exposure to the putative agent (e.g. in utero exposure to diethylstilboestrol) is ascertained through chart reviews, interviews or other means, in order to look for an association between drug exposure and disease. Case-control studies are valuable for evaluating possible causes of diseases that are rare (e.g. cancer) and/or take a long time to develop (e.g. cardiovascular disease). However, they are subject to several biases, including selection bias (selection of inappropriate control group) and recall bias (subjects' ability to accurately recall exposure)[11].

Cross-sectional studies are descriptive studies which provide a "snapshot" of the level of a disease or an exposure (e.g. level of community use of benzodiazepines) at a particular time point[12]. Data are normally collected via surveys or questionnaires. Cross-sectional studies are useful for defining the prevalence of a disease or lifestyle patterns (e.g. Survey of Lifestyles, Attitudes and Nutrition - SLAN) but they cannot provide information on possible cause(s) of a disease. Non-response bias may occur (potential differences between responders and non-responders will not be captured).

COMMONLY USED STATISTICAL TERMS

The aim of comparative clinical studies is usually to disprove the **null hypothesis** of no difference between the groups; the sample size is calculated on this basis and the statistical test used for analysis is chosen to suit the data[13]. **Sample size** refers to the desired number of subjects in a study and is calculated using the power of the study, the significance level chosen for the study and the level of departure from the null hypothesis (e.g. 10% difference in effect between drug treatments). Sample size also depends on the study design. There are standard formulae used for sample size estimation[14].

Results may be presented using *p* values and confidence intervals[15]. ***P* values** determine if the result is statistically significant or not while the **confidence interval** (CI) gives information about the magnitude of the result which may give a better indication of the clinical relevance of the findings. Statistical significance can also be inferred from the CI; for example, if a study compares the difference in blood pressure reduction between 2 drugs and the 95% CI for the difference in means does not cross zero (i.e. the null value meaning no difference between drugs), this result is regarded as being statistically significant at 5% level of significance. This would usually coincide with a significant *p* value ($p<0.05$). The method used to evaluate **risk** depends on the study type[16]. A cohort study can evaluate **absolute risk (AR)** of a disease as it follows a "population"; therefore **relative risk (RR)** can be calculated in these studies (see figure 2).

Figure 2: How to calculate relative risk

	Drug	No Drug
Adverse effect	20	5
No adverse effect	99,980	99,995
Total	100,000	100,000

AR_{drug} =20/100,000 =0.02%; $AR_{no\ drug}$=5/100,000 =0.005%

$RR=AR_{drug}/AR_{no\ drug}$ =0.02/0.005=4.0

However in a case-control study (which starts with the disease and works backwards to exposure), the absolute risk is not known. Therefore, the risk of developing the disease is calculated using the **odds ratio (OR)** which is the ratio of the odds of developing disease in the exposed group to the odds of developing disease in the non-exposed group. In the case of RR and OR calculations, the null value (i.e. no increased risk) is 1 and therefore if the 95% CI does not cross 1, the results are said to be statistically significant at the 5% level of significance.

Important Note: When evaluating RR it is useful to have an estimate of AR. If the AR is very small, then even a high RR (>2) may mean a small increased risk in absolute terms e.g. the increased risk of venous thromboembolism (VTE) with use of third generation combined oral contraceptives (COCs) was calculated as a RR of around 2 compared with second generation COCs; however the absolute risk of VTE in the group was low overall, therefore the absolute increased risk equated to 10 extra cases /100,000 women years (less than the estimated risk of VTE in pregnancy)[17].

HOW TO READ A PAPER – A STRUCTURED APPROACH

Most scientific papers are presented in a standard format: Introduction (why the research was done), Methods (how the study was done and what analysis was used), Results (what was found), and Discussion (what the results mean)[18].

General Considerations: When deciding if a paper is relevant to your practice, establish **what specific clinical question is being addressed and is it original work?** The Introduction should give the background to the research, and state the aims of the study[19]. Few clinical studies break entirely new ground; the majority seek to add new information to an existing hypothesis. Therefore, the practical question to ask about new research is "does this add to existing knowledge in any way?" For example: Is the study population different to that in previous studies? Is this study bigger/better than previous ones and if so how? Does the clinical issue warrant further investigation? The issue of author affiliation, sponsorship and potential conflict of interest should also be taken into account[20].

To evaluate the relevance of the study's benefit (=**efficacy**) to clinical usefulness (=**effectiveness**), examine carefully the **study population**. Look at how the subjects were recruited, what were the inclusion/exclusion criteria and were the subjects studied in real-life situations? For example hospital-based trials may recruit patients with more severe forms of disease and so results may not be transferable to primary care; the results of pharmacokinetic studies of new drugs in 30-year-old healthy male volunteers may not be applicable to an elderly population.

Next evaluate if an **appropriate study design** was used**.** The study design should be matched with the question being asked. RCT is the preferred study design to answer questions regarding drug treatment regimens, cohort and case-control studies look at disease aetiology and cross-sectional surveys are most appropriate to evaluate disease

screening, disease prevalence, drug usage or health attitudes.

Validity: Having established that the paper is attempting to answer a relevant question and is using a correct study design, the validity of the study should be evaluated. Firstly was the **randomisation** process or the **recruitment** correct? Was **bias** avoided or minimised? Possible sources of bias in RCTs include selection bias at the time of randomisation, exclusion bias (at time of withdrawal) or detection bias (in terms of "outcome" assessment)[19]. Selecting a comparable control group may be difficult in cohort and case-control studies. Secondly, was the study **blinded** and if so, was it double-blinded? Double blinding removes **investigator bias** i.e. the possibility that the investigator, knowing that the patient has been randomised to active drug (e.g. anti-hypertensive agent) rather than placebo, might be more likely to recheck an unexpected result. Thirdly, in the case of systematic reviews and meta-analyses, was the **retrieval process** comprehensive enough to ensure that good quality studies were included? Finally, the overall **study design and conduct** should be evaluated to ensure the study was valid and reproducible - this includes sample size, type of outcome(s) evaluated, duration and completeness of follow-up. The chosen outcome(s) will determine the **duration of the follow-up**, e.g. post-operative analgesia studies require a short follow-up of 24-48 hours, whereas a study of screening for lung cancer will need many years of follow-up. Study participants **drop out** for a variety of reasons such as inappropriate randomisation, adverse reactions, loss of motivation or just lost to follow-up. These people should be included in the analysis (this is called **intention to treat** analysis) otherwise the results may be biased (usually in favour of the test drug/other intervention under evaluation)[21].

Results: The results section of most papers usually contain a variety of tables, graphs and figures to aid understanding and highlight the significance of the results. The **presentation** and **interpretation** of the results section vary depending on the study design. In RCTs, look for the table showing the **baseline characteristics** of the groups. Note if there are any differences between the groups, as this may alter the interpretation of the final results. While there are many statistical tests available to analyse data, the majority of scientific papers use a small core of tests[13]. Therefore a study that uses an obscure test should be questioned – is it that the usual analyses did not produce the anticipated result and the authors tried other analyses to secure the desired results? The *p* value and CI are the two most frequently used statistics. It is important to remember that **statistical significance** does not always equate to **clinical relevance.** In addition, the larger the study size, the narrower the CI and the more likely the result is to be definitive[15]. The authors may also use RR, AR or OR to express the effects of an intervention in terms of the likely benefit or harm. The **number needed to treat (NNT)** calculation is important as it puts the results into context in terms of overall benefit to the particular target population[22].
All studies have **flaws**. The main areas where flaws may occur are in the content, the detail and the conclusion(s) not being in keeping with the results. Papers should deal with these issues in the discussion section, which should help in the evaluation of the impact of the results on clinical practice.

Applicability: The applicability of any study should be assessed before adopting a policy change[19]. Are the results relevant to your patient population? Are the proposed benefits worth the potential harm and/or cost? Are the results in keeping with other available evidence? A single scientific paper rarely leads to change in clinical practice; rather it is a physician's training, experience and common sense, together with a growing body of scientific data that leads to the **implementation of change** in clinical practice.

Table 2: CHECKLIST FOR READING A PAPER - SUMMARY

General	**Validity**	**Results**	**Applicability**
Clear question?	Randomisation }	Presentation } correct?	Relevant to your
Study design OK?	Sample size } correct?	Statistics }	practice?
	Study conduct }	Size of result relevant?	

CONCLUSION

Medical journals publish just a fraction of the literature sent for publication. Although peer-reviewed journals greatly reduce the likelihood of poor quality research being published, some published papers may be incorrectly designed and/or draw incorrect conclusions from the results. Therefore it is vital to use a structured approach when reading a paper in order to critically appraise and learn from the study findings.

References available on request. Date prepared: Nov. 2003