

Epidemiology and Critical Appraisal

Interpreting Information from Clinical Trials. What are the Results? Part I

Objectives:

1. Understand the problem of multiplicity in analysis.
2. Differentiate between primary and secondary outcomes, and apply this information to clinical trials.
3. Define composite outcome, valid use of composite outcomes, and the rationale for using composite outcomes.
4. Describe potential problems with subgroup analyses, and criteria for valid subgroup analyses.
5. Define interim analysis, and describe reasons for early termination of clinical trials.
6. Define surrogate outcome, and recognize use of surrogate outcomes in a clinical trial as well as potential drawbacks of the use of surrogates.
7. Know study phases in clinical trials, and problems of adverse event recognition including 'rule of three'
8. Apply the information to critical appraisal of clinical trials.

Lois Champion
lois.champion@lhsc.on.ca

Multiplicity

- the likelihood of finding at least one false-positive result (Type I error) increases with the number of comparisons made
- for example, if you do 10 statistical tests on data there is a 40% probability of finding at least one statistically significant (at p value of 0.05) result
- the problem of multiplicity can apply to trials in a variety of circumstances including:
 - multiple outcomes including surrogate endpoints
 - multiple treatments (multiarm trial)
 - subgroup analyses
 - repeated measures over time of the same outcome
 - interim analyses
 - exploratory analyses of data
 - adjusted analyses for prognostic factors
- the probability of at least one false positive result depends on the number of tests and the value which is selected as being statistically significant (the 'alpha value')
- for more information on multiplicity see Appendix I, however please understand you are not expected to know the information (equations etc.) provided in Appendix I for exam purposes, but it will help you to understand the concepts more clearly.

Types of outcomes (*synonym: endpoints*):

Primary outcomes

- clinical trials are done to answer a specific question, this is the primary endpoint
- primary endpoint must be clearly defined including the time frame, and established *a priori*; that is before the trial has started
- conclusions of the study should be based on the primary outcomes

Secondary outcomes

- many trials have additional endpoints that are investigated (secondary endpoints); these must also be established before the trial begins
- even when secondary outcomes are stated *a priori* they should only be considered meaningful when the primary outcome is positive
- secondary outcomes that are positive when the primary outcome shows no significant difference should be considered hypothesis-generating (that is a new study should be done to confirm the results)

Composite outcomes

- composite outcomes consist of multiple single endpoints that are combined, they avoid the need to select a single outcome when several related outcomes are expected to reflect the effects of the therapy
- the use of composite outcomes increases the event rate and therefore less patients are needed for a study
- composite outcomes should be:
 1. identified before the study begins
 2. clinically meaningful
 3. of similar importance to patients (*for example is shoulder dystocia during delivery as important as infant mortality?*)
 4. biologically plausible in terms of expected effects (*for example myocardial infarction and stroke are both cardiovascular endpoints*)
- when reviewing a study that has used a composite outcome the distribution of effect of the outcomes should be reviewed; ideally all components will be affected in the same way by the intervention (*for example, a composite outcome in which there is an increase in myocardial infarction but a decrease in stroke rate and no change in mortality is difficult to interpret clinically*)
- also compare the frequency of the component endpoints – are they similar, or did one endpoint happen much more frequently than the others skewing the results?
- generally composite outcomes are analyzed such that each endpoint is given equal weight in the analysis (although from the patient perspective some outcomes might be much worse than others)
- be aware of potential competing risks between endpoints; *for example studying nonfatal events without including death is not valid because patients who die may not have the nonfatal events (i.e. you have taken the worst outcomes out of the analysis)*

Multiple outcomes

- trials may use multiple outcomes, regardless these must be established prior to the trial beginning
- in addition statistical adjustment for multiple endpoints will be required

Surrogate endpoints

- surrogate (or intermediate) endpoints are laboratory measurements or physical signs used as a proxy for a clinically meaningful endpoint
- examples:
 - *cholesterol for myocardial infarction*
 - *bone mineral density for fractures due to osteoporosis*
- the use of surrogate endpoints allows trials to be smaller/shorter and less expensive
- a valid surrogate outcome must meet the following criteria:
 1. changes in the surrogate must be predictive of the relevant clinical outcome
 2. the surrogate must capture the effect of the intervention on the clinical outcome
- biomarkers do not always predict all possible toxicities of a systemically administered drug
- surrogate endpoints therefore may not accurately predict a clinically significant outcome, and must be interpreted with caution

Subgroup analyses

- the best estimate of treatment effect to be expected from a patient treated outside a trial is the overall estimate from the primary outcome as opposed to results from subgroup analyses
- criteria for valid subgroup analyses:
 1. subgroups must be specified before the study begins (to avoid ‘mining’ for data)
 2. information about all the subgroups analyzed should be presented in the study
 3. there should be a biologic rationale for considering the subgroup separately from the rest of the patients in the study
 4. there is prior evidence or a belief that a different effect in the subgroup is plausible
 5. there should be statistical evidence of a difference in the effect of the treatment for the subgroups (*this should be done using tests for interaction looking to see if the treatment effecting the subgroup is significantly different from the overall population, rather than simple between group comparisons*)
- in general subgroup analyses should provide information for further study (i.e. be seen as hypothesis generating) rather than being seen as part of the primary study results

- **statistical issues with subgroup analysis:**
 1. problem of multiplicity – applying many statistical tests to the same data has the effect of increasing the chance that at least one of these comparisons will be called statistically significant even if there is no real difference (a false positive result); using tests for interaction avoids the problem of multiplicity (See Appendix I for more information about multiplicity)
 2. power – because subgroups represent a fraction of the study population subgroups may not have the statistical power to detect a genuine effect of the treatment (a false negative result, or Type II error)

Interim Analyses

- interim analysis is the repeated analyses of data as it accumulates
- large clinical trials often include interim analyses by an independent data monitoring committee with predefined early stopping rules for the study
- any interim analyses must be specified before the trial begins, and should be done by investigators blinded to the group allocation
- because interim analyses involve multiple testing of data, there must be statistical adjustments made for statistical significance (*i.e. if you were to ‘peek’ at the data 10 times during a trial there would be a 40% likelihood of finding a difference at the $p = .05$ level even if there was no real difference between the experimental and control groups; see Appendix for more information*).
- trials may be stopped early for reasons such as:
 1. futility: if it becomes apparent that even with continued accrual of patients that a significant effect of treatment will not be seen
 2. harm: if it is clear that the experimental group is experiencing an unanticipated adverse effect on the outcome; generally a lower level of evidence for harm is used relative to stopping for benefit
 3. benefit: if there is a clear benefit seen in the experimental group; be aware however that trials stopped early for benefit may represent a biased treatment effect (a random high)

Clinical vs. Statistical Significance

- statistical significance is not the same as clinical significance
- outcomes should be clinically relevant and important

Adverse Events

- an adverse event is an untoward medical occurrence in a patient or participant in a clinical investigation
- some adverse events are serious and/or life-threatening
- adverse events must be reported as part of the trial (sometimes less common adverse events are not recognized during a trial, and are recognized later with widespread use of the treatment)
- the **'Rule of Three'** says that the number of individuals who must be observed to be 95% confident of observing at least one case of an adverse effect is three times the denominator of the true probability of the occurrence of the adverse effect (for example if an adverse event occurs in 1 in 1 000 patients, then ~ 3 000 patients would be needed to be 95% confident of seeing at least one adverse event)
- in addition medications are often used (and sometimes marketed) for 'off-label' indications; that is indications that have not been studied/approved of clinically
- clinical trials are done in 4 phases, and medications will be approved based on relatively small clinical trials, so it is not surprising that adverse events may be recognized after widespread use and that sometimes medications which were approved will be withdrawn

Drug Development in Humans

*adapted from Arch Int Med 2006;166:1440

Phase	# of Patients	Length of Phase	Goal
1	20 – 100	several months	safety, dosage and efficacy
2	100 – 500	months to 2 years	effectiveness and short-term safety
3	500 – 3000	1 – 4 years	safety and effectiveness
4	> 3000	ongoing	long-term safety and rare adverse events (postmarketing surveillance, open label studies)

- a report from the Institute of Medicine published in September 2006 recommends changes to the system of drug approval and postmarketing surveillance in the United States
- see Appendix II to get an idea of why adverse events may be missed in a trial if the trial is small and/or the adverse event is uncommon

References

1. Identifying outcome reporting bias in randomized trials on PubMed: review of publications and survey of authors. *BMJ* online first. Jan. 28, 2005.
2. Empirical evidence for selective reporting of outcomes in randomized trials. Comparison of protocols to published articles. *JAMA* 2004;291:2457.
3. Validity of composite endpoints in clinical trials. *BMJ* 2005;330:594.
4. Composite outcomes in randomized trials. Greater precision but with greater uncertainty? *JAMA* 2003;289:2554.
5. How should clinicians interpret results reflecting the effect of an intervention on composite endpoints: should I dump this lump? *EBM* vol 10; 2005.
6. Interpreting results from secondary endpoints and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? *BMJ* 2001;322:989.
7. Reported outcomes in major cardiovascular clinical trials funded by for-profit and not-for-profit organizations: 2000 – 2005. *JAMA* 2006;295:2270.
8. Lessons learned from recent cardiovascular trials. Part I and Part II. *Circulation* 2002;106:746, *Circulation* 2002;106:880.
9. Randomized trials stopped early for benefit. A systematic review. *JAMA* 2005;294:2203.
10. Interpretation of subgroup results in clinical trial publications: Insights from a survey of medical specialists in Ontario, Canada. *Am Heart J* 2006;151:580.
11. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? *Am Heart J* 2006;151:257.
12. Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. *Lancet* 2005;365:176.
13. Multiplicity in randomized trials I: endpoints and treatments. *Lancet* 2005;365:1348.
14. Multiplicity in randomized trials II: subgroup and interim analyses. *Lancet* 2005; 365:1657.
15. The challenge of subgroup analyses – reporting without distorting. *NEJM* 2006;354:1667.

Appendix I Multiple Comparisons

- the likelihood of finding at least one false-positive result (Type I error) increases with the number of comparisons made
- the problem of multiplicity can apply to trials in a variety of circumstances including:
 - ❑ multiple outcomes including surrogate endpoints
 - ❑ multiple treatments (multiarm trial)
 - ❑ subgroup analyses
 - ❑ repeated measures over time of the same outcome
 - ❑ interim analyses
 - ❑ exploratory analyses of data
 - ❑ adjusted analyses for prognostic factors
- the probability of at least one false positive result depends on the number of tests and the value which is selected as being statistically significant (the so-called alpha value)

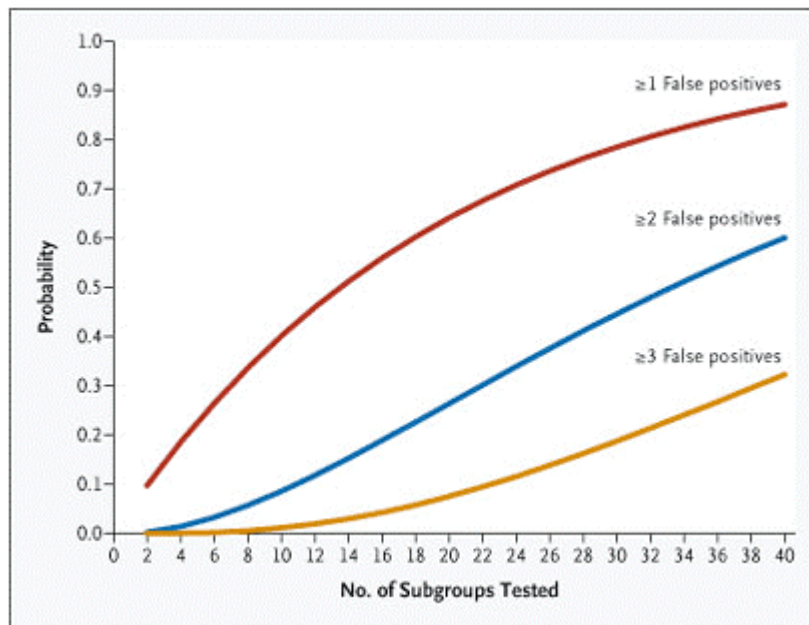
$$P (FP) = 1 - (1 - \text{alpha value})^n$$

- where:
 - P is probability of at least one false positive result
 - alpha value is level selected for statistical significance
 - n is number of independent comparisons
- since alpha value is typically 0.05 the equation becomes:

$$P (FP) = 1 - (1 - 0.05)^n$$

**Probability of at least one significant result at the 5% significance level
given no true difference**

Number of Tests	Probability
1	.05
2	.10
3	.14
5	.23
10	.40
20	.64



Probability that multiple subgroup analysis will yield at least one, two or three false positive results.
NEJM 2006;354:16

- this graph shows the effect of testing multiple subgroups; for example if you were to test 40 subgroups you would have close to a 90% probability of at least one false positive result and a > 25% chance of having at least three false positive results (at $p < .05$)

Statistical adjustments for multiple comparisons:

- an example of a method for correcting for multiplicity is the Bonferroni correction
- this is a simple method which divides the significance level by the number of tests

$$\text{alpha level corrected} = \text{alpha level} / n$$

- for example if we are presuming that statistical significance is .05, and 10 tests have been done then the corrected significance level is $0.05 / 10$ which is 0.005
- one disadvantage of this technique is that it is quite conservative, and therefore real differences might not be detected (Type II error)

Adjustments for multiple interim analyses:

- interim analyses involves analyzing data before the planned completion of the trial, and should be prespecified, done by independent committee and have rules for early stopping of the trial (for futility, harm or benefit)
- statistical adjustments must be made when interim analyses are part of a trial because they involve multiplicity
- examples of interim stopping rules include Pocock, Peto and O'Brien-Fleming; examples of these are shown in table on following page

from: Multiplicity in randomised trials II: Subgroup and Interim Analyses
Lancet 2005;365:1657.

Number of planned interim analyses	Interim analysis	Pocock	Peto	O'Brien-Fleming
2	1	0.029	0.001	0.005
	2 (final)	0.029	0.05	0.048
3	1	0.022	0.001	0.0005
	2	0.022	0.001	0.014
	3 (final)	0.022	0.05	0.045
4	1	0.018	0.001	0.0001
	2	0.018	0.001	0.004
	3	0.018	0.001	0.019
	4 (final)	0.018	0.05	0.043
5	1	0.016	0.001	0.00001
	2	0.016	0.001	0.0013
	3	0.016	0.001	0.008
	4	0.016	0.001	0.023
	5 (final)	0.016	0.05	0.041

Appendix II Adverse Events

Numbers of patients that need to be exposed to a medication to ensure that an adverse drug reaction has a 95% probability of being observed at least once. *

Frequency of Adverse Drug Reaction	Minimum # of Patients Required
Very common (10%)	29
Common (1%)	299
Uncommon (0.1%)	2994
Rare (0.01%)	29956

- you can also use a quick rule known as the ‘**Rule of Three**’ which says that the number of individuals who must be observed to be 95% confident of observing at least one case of an adverse effect is three times the denominator of the true probability of the occurrence of the adverse effect
- for example if adverse event is 1 / 100 then you need ~ 300 patients to be 95% sure that you will observe at least one case of the adverse event
- or similarly if the event is rare (1 / 10 000) you would need ~ 30 000 patients to be 95% confident of observing at least one case of the adverse event
- some adverse events are not unusual (for example a heart attack) and may not be considered an adverse effect of a medication and therefore not reported
- this explains why rare adverse events may not be seen in clinical trials, but recognized later with post-marketing surveillance
- this also explains why sometimes medications that have been approved are withdrawn when adverse events are recognized as the medication is used in the general population
- in addition, medications are often used (and sometimes marketed) for ‘off-label’ indications; that is indications that were not studied/approved
 - for example Pfizer/Parke-Davis company was taken to court because of off-label promotion of the medication gabapentin (Neurontin®)
 - this promotion was multifaceted and included targeting specific physician groups, training physicians for ‘peer-to-peer’ selling programs, resident programs, arranging teleconferences with surreptitious monitoring by pharmaceutical employees who were in ‘listen only’ mode and not named as participants, paying for physicians to attend conferences, undertaking exploratory trials in neuropathic pain ‘which if positive will be publicized and published’ (company memo), and not publishing negative trials, using ‘seed’ trials the purpose of which was to have physicians prescribe the study medication (Narrative Review: The Promotion of Gabapentin: An Analysis of Internal Industry Documents. Ann Int Med 2006;145:284–293)

- examples of medications that have been withdrawn from the market include:
 - cerivastatin (cholesterol lowering medication) approved in 1997, withdrawn in 2001 because of increased risk of rhabdomyolysis (2 – 6 cases/ 100 000 patients) relative to other statins (medications of the same class)
 - cisapride (gastric motility agent) approved 1993, withdrawn 2000 because of risk of fatal heart arrhythmias
 - thalidomide found to cause congenital malformations

- postmarketing surveillance refers to the system intended to detect adverse drug events once new medications are in widespread use
- unfortunately this system is not ideal, and there may be a significant time lag between when adverse events are recognized and the withdrawal of medications from the market
- concerns also include the potential for conflict of interest with pharmaceutical companies
- examples in literature in which conflict of interest was apparent include:
 - withdrawal of cerivastatin
 - references:
 - Potential for conflict of interest in the evaluation of suspected adverse drug reactions. Use of cerivastatin and risk of rhabdomyolysis. JAMA 2004;292:2622.
 - Bayer's Response JAMA 2004;292:2655.
 - withholding of clinical trials suggesting harm with SSRIs and teenage risk of suicide (GlaxoSmithKline published the clinical trials after a lawsuit)
 - references:
 - Drug Company experts advised staff to withhold data about SSRI use in children. CMAJ 2004;170:783
 - Did regulators fail over SSRIs? BMJ 2006;333:92

- due to these concerns the Institute of Medicine published a report in September 2006 that recommends widespread reform of the system