# Common statistical failures in analysis of clinical trials

March 2009

## Overview

Statistical deficiencies can occur in

- Study design
- Study conduct
- Data analysis
- Presentation of data analyses
- Interpretation of study findings



## Study design related deficiencies

- Inappropriate power
- Inappropriate choice of endpoints
- Inappropriate choice of statistical tests
- Non-representative sample (inappropriate inclusion/exclusion criteria)
- Ignoring multiple comparisons
- Unspecified or not detailed enough analyses
- Lack of “fallback positions” in protocol



## Inappropriate power

- Underpowered study
  - Higher probability of false negative result
  - Clinically meaningful effect may not be statistically significant
- Overpowered study
  - Statistically significant may not be clinically meaningful
  - Waste of resources



## Inappropriate choice of endpoints

- The most informative endpoint should be chosen
  - Prefer continuous/ quantitative endpoint over discrete/qualitative one
- Dichotomizing/categorizing a continuous variable reduces the power / increases the samples size
- “Change from baseline” endpoint is less efficient than baseline adjustment



## Ignoring multiple comparisons

- Multiple comparisons occur when
  - There are multiple endpoints (co-primary endpoint, several secondary endpoints)
  - The is more than one treatment group (e.g. in a dose-finding study)
  - Interim looks are planned
  - Adaptations are planned
- There are many acceptable methodologies to handle the multiplicity problem
- Ignoring the multiple comparisons results in Type I error inflation – increasing the probability of false positive result

## Unspecified or not detailed enough analyses

- All planned analyses should be pre-specified in detail within the study protocol
- Failure to do so can result in rejection of the results

7

## Lack of "fallback positions" in protocol

- Statistical tests implicitly make assumptions on the nature of the data
- In case the assumptions are not met, alternative methods of analysis should be pre-specified in the protocol
- Failure to do so can result in declaring the analysis as "invalid" and the alternative analyses as "data driven"

8

## Study conduct related deficiencies

- Recruitment is too fast
- Centers that recruit too many subjects
- Unblinding / leak of information
- Failure in randomization

9

## Recruitment is too fast

- Fast recruitment is good from operational point of view
- However, too fast recruitment will result in insufficient information for testing whether the study assumptions are met
- This can jeopardize the study goals. For example: if the population is not as active as assumed, the power of the study may be smaller than expected

10

## Centers that recruit too many subjects

- If something "goes wrong" in a center that recruited many subjects, it ay affect the results of the whole study
- Additionally, when there are large centers, there are also small centers that recruit just a few subjects. This has implications on the balance of the study – actual treatment allocation may differ from the planned one
- In small centers there is also higher potential of unblinding by guessing the actual treatment
- It is recommended not a allow a single center to recruit more than 2.5% of the sample size

11

## Unblinding / leak of information

Unblinding can occur as result of

- Exposure to detailed adverse event information
- Inadequate randomization method
- Leak of information from company personnel who are unblinded as part of their job (clinical supplies, independent statistician, pharmacovigilance)

12

## Data analysis related deficiencies

- Failure to validate statistical assumptions
- Lack of baseline comparisons and adjustments
- Inappropriate handling of multiplicity issues*
- Inappropriate handling of missing values / dropouts
- Post-hoc/data driven analyses

All of these issues should be addressed in advance in the study protocol and or in the Statistical Analysis Plan before revealing the blind of the study

* Already discussed



## Data presentation related deficiencies

- Failure to present measures of uncertainty/variability
- Use of inappropriate descriptive statistics
- Presentation of confidence interval for groups but not for group differences
- Non-meaningful precision
- Misleading graphs



## Failure to present measures of uncertainty/variability

- An appropriate measure of uncertainty or variability should be presented along with every reported statistic, e.g.
  - Mean – Standard Deviation
  - Median – Range or Inter-Quartile Range
  - Point estimate of effect – Confidence interval and p-value



## Use of inappropriate descriptive statistics

- Mean and standard deviation should be calculated only for quantitative variables
- For ordinal variables – only the median and other percentiles are appropriate
- For categorical variable – only frequency tables are appropriate



## Presentation of confidence interval for groups but not for group differences

- To establish difference between treatment groups, one must test this difference and present an appropriate confidence intervals
- Calculating the confidence intervals for each the group means and showing that they do not intersect is only a rule of thumb
- It is possible that the confidence intervals for the group means do not intersect, yet the difference between the groups is not statistically significant



## Non-meaningful precision

- Only a meaningful number of decimal points should be presented
- For example, when presenting the mean body temperature, one or two decimal points should be sufficient
- 36.8º or 36.75º is fine, but 36.748445 is inappropriate

## Misleading graphs

- A whole presentation can be dedicated to this subject
- Most misleading graphs are due to:
  - Inappropriate scale
  - Disproportional bars
  - Use of three-dimensional presentation
  - Cutting the origin off the presentation
  - Excluding outliers/extreme values
  - Presenting a sub-group of the data

19

## Result interpretation related deficiencies

- Statistical significance vs. clinically meaningful results
- Misinterpretation of association as causality
- Misinterpretation of p-values
- Misinterpretation of confidence intervals

20

## Misinterpretation of association as causality

- Statistical tests (such as the correlation coefficient, chi-square test, regression analysis, etc.) can establish association between two or more variables
- However, they do not establish causality. When association between two phenomena is observed there are many possibilities, such as:
  - The first phenomenon is causing the second one
  - The second phenomenon is causing the first one
  - There may be a third phenomenon that is affecting the ones we observed
  - The observed association is an artifact
- Statistical analysis alone can not establish causation
- A pre-specified theory/hypothesis and a well controlled experiment is needed

21

## Misinterpretation of p-values

- A p-value is a convenient way to determine whether a result is statistically significant
- However, the numerical value of the p-value itself is not measuring “how significant” were the results
- Terms such as “highly significant” or “borderline significant” are not well defined statistically speaking

22

## Misinterpretation of confidence intervals

- A confidence interval for a parameter does not provide a probability for the real value of the parameter
- This is because a parameter is constant by definition, and a probability is associated with random variables

23