Understanding Statistics for Clinicians

ESMO Clinical Trials Course
Vienna, Austria
27-28 March 2015

Urania Dafni
Laboratory of Biostatistics, Dept of Nursing,
School of Health Sciences,
University of Athens, Greece

Frontier Science Foundation-Hellas (FSF-H)
Athens, Greece
Definition of a clinical trial

• A *clinical trial* is an experiment testing medical treatments on human subjects

• An *experiment* is a series of observations made under conditions *controlled* by the investigator
Classification by Objective

Phase III: Compare new treatment to standard therapy or placebo.

Scientific question

• Experiments answer a scientific question by isolating the intervention and the outcome from extraneous influences

• **Goals:**
  – Eliminate systematic error (Bias)
    • any effect rendering the observed results not representative of the treatment effect.
  – Minimize random error (increase precision, decrease variability)
    • inaccuracy of results due to sampling
  – Ensure the generalizability of study results

Study Design is the methodology for achieving these goals
Bias - Variability

No bias, small variability

Bias, small variability

No bias, larger variability

Bias, larger variability
Utilize statistical design to minimize bias & variability

- **Bias**: Randomization, stratification, blinding, choice of design

  [Image 138x429 to 246x537]

  [Image 135x234 to 243x396]

  http://www.dcscience.net/?p=239

- **Variability**: Establish a *sample size* sufficient to achieve study goals:
  - The number of patients that provide *sufficient power* to detect a difference between treatment arms if such a difference exists.
  - A negative study may reflect a lack of benefit, or simply a lack of sufficient number of patients to detect a difference.
Gold Standard: RCTs

Randomized clinical trials are the gold standard for evaluating therapeutic interventions.

Randomization:

• Provides a treatment assignment that is independent of outcome and patient/disease features, thus **balancing treatment groups of known and unknown factors associated with outcome.**

• The intention-to-treat (ITT) analysis approach is the gold standard for phase III randomized, controlled clinical trials: analyzes all patients in the treatment groups as randomized without regard to treatment actually received.

*JCO* 2012- Riwani and Hilsenbeck; 30(4): 453-458
Question 1.

• In a Phase III Clinical trial, to **address bias** of the estimate, the design needs to include at least:

  A. Randomization
  B. Stratification & Blinding
  C. Adequate Sample Size & Stratification & Blinding
  D. Randomization & Stratification & Blinding
Answer to Question 1.

• In a Phase III Clinical trial, to **address bias** of the estimate, the design needs to include at least:

  A. Randomization
  B. Stratification & Blinding
  C. Adequate Sample Size & Stratification & Blinding
  D. Randomization & Stratification & Blinding
Phase III - Parallel design

- Superiority Trials
- Non-Inferiority Trials

- **Null hypothesis or** $H_0$ is a statement we would like to reject and generally we do not want to be true.
  - **Superiority Trial** $H_0$: “no effect” or “no difference”
  - **Non-inferiority Trial** $H_0$: “different effect” or “difference”

- **Alternative hypothesis or** $H_1$ is the statement we would like to prove if true

Our final conclusion will always be one of these:
- Reject the null hypothesis or
- Fail to reject the null hypothesis
Question 2

We can conclude that two treatments do not differ if:

A. In a well powered large superiority trial the conclusion reached was a statistically significant result
B. In a non–inferiority trial the conclusion reached was that there is no difference
C. In a well powered large superiority trial the conclusion reached was of a non-statistically significant result
D. In B and C
Question 2

We can conclude that two treatments do not differ if:

A. In a well powered large superiority trial the conclusion reached was a statistically significant result

B. In a non–inferiority trial the conclusion reached was that there is no difference

C. In a well powered large superiority trial the conclusion reached was of a non-statistically significant result

D. In B and C
Superiority Trials

• **Aim:** To demonstrate the superiority of a new therapy compared to an established therapy or placebo
  
  – By how much should the new therapy be better than the established?
  
  – This extra effect of the new compared to the reference therapy is called the *least relevant difference* or the *clinical significance* ($\Delta$)

**ASCO:** Recommended Targets for Meaningful Trial Goals

*Ellis et al. 2013. JCO. American Society of Clinical Oncology Perspective: Raising the Bar for Clinical Trials by Defining Clinically Meaningful Outcomes*

**ESMO:** Development of the Magnitude of Clinical Evaluation Scale

*ASCO 2014, ESMO 2014; Annals of Oncology 2015*
Recommended Targets for Meaningful Trial Goals

Table 1. Summary of Recommended Targets for Meaningful Clinical Trial Goals

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Patient Population</th>
<th>Current Baseline Median OS (months)</th>
<th>Improvement Over Current OS That Would Be Clinically Meaningful (months)</th>
<th>Target HRs</th>
<th>Improvement in 1-Year Survival Rate (%)</th>
<th>Improvement in PFS (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pancreatic cancer</td>
<td>FOLFIRINOX-eligible patients</td>
<td>10 to 11</td>
<td>4 to 5</td>
<td>0.67 to 0.69</td>
<td>48 → 63</td>
<td>4 to 5</td>
</tr>
<tr>
<td>Pancreatic cancer</td>
<td>Gemcitabine or gemcitabine/nab-paclitaxel-eligible patients</td>
<td>8 to 9^20,21</td>
<td>3 to 4</td>
<td>0.6 to 0.75</td>
<td>35 → 60</td>
<td>3 to 4</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>Nonsquamous cell carcinoma</td>
<td>13^22</td>
<td>3.25 to 4</td>
<td>0.76 to 0.8</td>
<td>53 → 61</td>
<td>4</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>Squamous cell carcinoma</td>
<td>10^23</td>
<td>2.5 to 3</td>
<td>0.77 to 0.8</td>
<td>44 → 61</td>
<td>3</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>Metastatic triple negative, previously untreated for metastatic disease</td>
<td>18^24,25</td>
<td>4.5 to 8</td>
<td>0.75 to 0.8</td>
<td>63 → 71</td>
<td>4</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>Disease progression with all prior therapies (or not a candidate for standard second- or third-line options)</td>
<td>4 to 8^28</td>
<td>3 to 5</td>
<td>0.67 to 0.67</td>
<td>25 → 35</td>
<td>3 to 5</td>
</tr>
</tbody>
</table>

Abbreviations: FOLFIRINOX, 5-fluorouracil, leucovorin, and oxaliplatin; HR, hazard ratio; OS, overall survival; PFS, progression-free survival.
*Current → target.

HR ≤ 0.8, with an improvement in median OS 2.5 to 6 months, is the minimum incremental improvement over standard therapy that would define a clinically meaningful outcome.

Ellis et al. 2013. JCO. American Society of Clinical Oncology Perspective: Raising the Bar for Clinical Trials by Defining Clinically Meaningful Outcomes
Surgery + (neo)adjuvant CT ± RT

Centrally confirmed IHC 3+ or FISH+

and LVEF ≥ 55

Randomization

Women with locally determined HER2-positive invasive early breast cancer

Observation

2 years trastuzumab
8 mg/kg ⇒ 6 mg/kg
3 weekly schedule

1 year trastuzumab
8 mg/kg ⇒ 6 mg/kg
3 weekly schedule

HERA, HERceptin Adjuvant; CT, chemotherapy; RT, radiation

HERA trial design
Statistical analysis

Enrollment of 4482 patients was planned to detect a 23 percent relative reduction in the risk of a disease-free-survival event with 80 percent power, with the use of a two-sided significance level of 2.5 percent for each comparison: two years of trastuzumab versus observation and one year of trastuzumab versus observation.

Clinical significance: 23%
Representation of the Hypotheses

- \( \lambda_T \) = Risk of a disease event for the trastuzumab group
- \( \lambda_S \) = Risk of a disease event for the observation group

Hazard Ratio \( HR = \lambda_T / \lambda_S \)

- Then we write
  - Null Hypothesis: \( H_0: HR=1 \) (\( \lambda_T = \lambda_S \))
  - Alternative Hypothesis: \( H_a: HR\neq 1, \) \( HR<1 \) (OR \( \lambda_T < \lambda_S \))

Clinical significance: 23%

If in fact HR is less by 23%, i.e., \( HR\leq0.77 \)
there needs to be HIGH PROBABILITY to detect this difference
This probability is the \textit{Statistical Power} = 1-\( \beta \) = 80%
Representation of the Hypotheses

One should be able to detect with high probability this reduction by 23% if in fact it is true, i.e. with HIGH STATISTICAL POWER = 80%

Statistical analysis

Enrollment of 4482 patients was planned to detect a 23 percent relative reduction in the risk of a disease-free–survival event with 80 percent power, with the use of a two-sided significance level of 2.5 percent for each comparison: two years of trastuzumab versus observation and one year of trastuzumab versus observation.
Representation of the Hypotheses

• We work with logarithms of the HR.
• Estimated log hazard ratio \([\ln(HR)]\) for disease event between the trastuzumab group and the observation group:

\[
\Delta = \ln(HR) = \ln(\lambda_T) - \ln(\lambda_S)
\]

where in the logarithmic scale, \(\Delta\) is the true relative reduction in the risk of a disease-free–survival event treated with one year of trastuzumab versus observation.

• Then we write
  – Null Hypothesis: \(H_0: \Delta = 0\) (for \(HR=1\))
  – Alternative Hypothesis: \(H_a: \Delta < 0\) (for \(HR<1\))

Here \(\Delta = \ln(1-0.23) = \ln(0.77) = -0.26\)
\[ \Delta = \ln(HR) = \ln(1 - 0.23) = -0.26 \]
Representation of the Hypotheses

Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer

Martine J. Piccart-Gebhart, M.D., Ph.D., Marion Proctor, M.Sci., Brian Leyland-Jones, M.D., Ph.D., Aron Goldhirsch, M.D., Michael Untch, M.D., Ian Smith, M.D., Luca Gianni, M.D., Jose Baselga, M.D., Richard Bell, M.D., Christian Jackisch, M.D., David Cameron, M.D., Mitch Dowsett, Ph.D., Carlos H. Barrios, M.D., Günther Steger, M.D., Chiun-Shen Huang, M.D., Ph.D., M.P.H., Michael Andersson, M.D., Dr.Med.Sci., Moshe Inbar, M.D., Mikhail Lichinitser, M.D., István Láng, M.D., Ulrike Nitz, M.D., Hiroji Iwata, M.D., Christoph Thomssen, M.D., Caroline Lohrisch, M.D., Thomas M. Suter, M.D., Josef Rüschhoff, M.D., Tamás Sütő, M.D., Ph.D., Victoria Greuloch, M.Sc., Carol Ward, M.Sc., Carolyn Straehle, Ph.D., Eleanor McFadden, M.A., M. Stella Dolci, and Richard D. Gelber, Ph.D., for the Herceptin Adjuvant (HERA) Trial Study Team

Statistical analysis

Enrollment of 4482 patients was planned to detect a 23 percent relative reduction in the risk of a disease-free–survival event with 80 percent power, with the use of a two-sided significance level of 2.5 percent for each comparison: two years of trastuzumab versus observation and one year of trastuzumab versus observation.
Hypothesis Testing: General Principles

We reject $H_0$ if:

- when assuming that $H_0$ is true (i.e., HR=1),
- the probability (p-value) to observe the OBSERVED HR from the clinical trial is small enough

We discriminate between rare (unexpected) and common (expected) outcomes.

- But what is small enough probability – rare event?
  - Reject $H_0$ if p-value $\leq \alpha$
  - Fail to reject if p-value $> \alpha$

In HERA $\alpha = 2.5\%$ for each comparison.
\[ \Delta = \ln(HR) = \ln(1 - 0.23) = -0.26 \]
Hypothesis Testing: Errors

We can enumerate the possible outcomes of a hypothesis testing in a table:

<table>
<thead>
<tr>
<th>Test Result</th>
<th>True State</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not reject $H_0$</td>
<td>$H_0$ True</td>
<td>$H_0$ False</td>
</tr>
<tr>
<td>$\Delta = 0$</td>
<td>✓</td>
<td>FALSE NEGATIVE</td>
</tr>
<tr>
<td>$\ln(\text{HR}) = 0$, $\text{HR}=1$</td>
<td>(1-$\alpha=97.5%$)</td>
<td>Type II error ($\beta=20%$)</td>
</tr>
<tr>
<td>$\ln(\lambda_T) - \ln(\lambda_S)=0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>FALSE POSITIVE</td>
<td>✓</td>
</tr>
<tr>
<td>$\Delta &lt; 0$, $\text{HR}&lt;1$</td>
<td></td>
<td>Power</td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>(1-$\beta=80%$)</td>
</tr>
</tbody>
</table>
Representation of the Hypotheses

Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer

Martine J. Piccart-Gebhart, M.D., Ph.D., Marion Proctor, M.Sci., Brian Leyland-Jones, M.D., Ph.D., Aron Goldhirsch, M.D., Michael Untch, M.D., Ian Smith, M.D., Luca Gianni, M.D., Jose Baselga, M.D., Richard Bell, M.D., Christian Jackisch, M.D., David Cameron, M.D., Mitch Dowsett, Ph.D., Carlos H. Barrios, M.D., Günther Steger, M.D., Chiun-Shen Huang, M.D., Ph.D., M.P.H., Michael Andersson, M.D., Dr.Med.Sci., Moshe Inbar, M.D., Mikhail Lichinitser, M.D., István Láng, M.D., Ulrike Nitz, M.D., Hiroji Iwata, M.D., Christoph Thomssen, M.D., Caroline Lohrisch, M.D., Thomas M. Suter, M.D., Josef Rüschoff, M.D., Tamás Sütő, M.D., Ph.D., Victoria Greatorex, M.Sc., Carol Ward, M.Sc., Carolyn Straehle, Ph.D., Eleanor McFadden, M.A., M. Stella Dolci, and Richard D. Gelber, Ph.D., for the Herceptin Adjuvant (HERA) Trial Study Team

Statistical analysis

Enrollment of 4482 patients was planned to detect a 23 percent relative reduction in the risk of a disease-free–survival event with 80 percent power, with the use of a two-sided significance level of 2.5 percent for each comparison: two years of trastuzumab versus observation and one year of trastuzumab versus observation.
HERA: DFS

**Patients(%)**

- **Observation**
- **1 year Herceptin**

**Events** | **2-year DFS** | **HR** | **95% CI** | **p value**
--- | --- | --- | --- | ---
127 | 85.8 | 0.54 | 0.43, 0.67 | <0.0001
220 | 77.4 |

**No. at risk**
- Observation: 1694
- Herceptin: 1693

**Median follow-up:** 1 year; DFS, disease-free survival; HR, hazard ratio; CI, confidence interval
Results

The unadjusted hazard ratio for an event in the trastuzumab group, as compared with the observation group, was 0.54 (95 percent confidence interval, 0.43 to 0.67; P<0.0001 by the log-rank test, crossing the interim analysis boundary), representing an absolute benefit in terms of disease-free survival at two years of 8.4 percentage points.
Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer

Discussion

This study shows that trastuzumab can benefit women with HER2-positive breast cancer when given after completion of adjuvant chemotherapy. As compared with observation after primary therapy (including surgery with or without radiotherapy and neoadjuvant or adjuvant chemotherapy), trastuzumab given after primary therapy reduced the rate of recurrence, particularly distant recurrence, by approximately 50 percent. This degree of benefit in early breast cancer is the largest to be reported since the introduction of tamoxifen in hormone-receptor–positive disease.
Final conclusion in a superiority trial

- Reject the null hypothesis:
  reject equivalence
  accept significant difference
- Fail to reject the null hypothesis:

  FAILURE TO REJECT the Null hypothesis should not be confused with PROOF OF EQUIVALENCE
Objective for equivalence or non-inferiority trials

• The objective for equivalence trials is to demonstrate that a new treatment is equivalent to a standard therapy with regard to a specific clinical end point and has an intrinsic benefit for other clinical end points.

• A non-inferiority trial refers to a study in which the primary objective is to evaluate whether the new treatment is not inferior to or as effective as the standard therapy for a particular end point.

Non-Inferiority Margin
Predefined tolerance: $\delta$

Zee, B. C.-Y. J Clin Oncol; 24:1026-1028 2006
Equivalence and non-inferiority trial

Prespecified quantity $\delta$: equivalence limit or non-inferiority margin

$95\%$ CI for $\ln(\theta)$

$\ln(\theta) - 1.96 \ SE$  $\ln(\theta)$  $\ln(\theta) + 1.96 \ SE$
Non-inferiority trial example

Randomized Trial Comparing Two Fractionation Schedules for Patients With Localized Prostate Cancer

Himu Lukka, Charles Hayter, Jim A. Julian, Padraig Warde, W. James Morris, Mary Gospodarowicz, Mark Levine, Jinka Sathya, Richard Choo, Hugh Prichard, Michael Brundage, and Winkle Kwan

Abstract

Purpose
The optimal radiation dose fractionation schedule for localized prostate cancer is unclear. This study was designed to compare two dose fractionation schemes (a shorter 4-week radiation schedule vs a longer 6.5-week schedule).

Patients and Methods
Patients with early-stage (T1 or T2) prostate cancer were randomly assigned to 66 Gy in 33 fractions over 45 days (long arm) or 52.5 Gy in 20 fractions over 28 days (short arm). The study was designed as a noninferiority investigation with a predefined tolerance of -7.5%. The primary outcome was a composite of biochemical or clinical failure (BCF). Secondary outcomes included presence of tumor on prostate biopsy at 2 years, survival, and toxicity.

Results
From March 1995 to December 1998, 936 men were randomly assigned to treatment; 470 were assigned to the long arm, and 466 were assigned to the short arm. The median follow-up time was 5.7 years. At 5 years, the BCF probability was 52.96% in the long arm and 59.95% in the short arm (difference = -7.0%; 90% CI, -12.6% to -1.4%), favoring the long arm. No difference in 2-year postradiotherapy biopsy or in overall survival was detected between the arms. Acute toxicity was found to be slightly higher in the short arm (11.4%) compared with the long arm (7%; difference = -4.4%; 95% CI, -8.1% to -0.6%); however, late toxicity was similarly low in both arms (3.2%).
Non-inferiority trial example

- The study was designed as a non-inferiority investigation with a predefined tolerance of 7.5%, comparing a 4-week to a 6.5 week radiation schedule.
- The primary outcome was a composite endpoint: time to biochemical or clinical failure (BCF); 5-year BCF.
- BCF was defined as a cluster of any one of the following events (whichever occurred first):
  - three consecutive increases in PSA,
  - clinical evidence of failure (local or distant),
  - commencement of hormonal therapy, or
  - death as a result of prostate cancer.
Non-inferiority Trial example

Fig 1. Biochemical or clinical failure (BCF) by randomized treatment arm

7.0% at 5 years
Non-inferiority Trial example

At 5 years, the Kaplan-Meier estimates of BCF in the long arm and short arm were **52.95% and 59.95%**, respectively.

The difference was **7.0% (90% CI, 1.42% to 12.58%)**.

Because the upper bound is more than the predefined tolerance of 7.5%, we could not exclude the possibility of the short arm being inferior.

**Conclusion**

Given the results, we cannot exclude the possibility that the chosen hypofractionated radiation regimen may be inferior to the standard regimen. Further evaluation involving higher dose hypofractionated radiation regimens in contemporary radiation settings is necessary.

Final conclusion

In a superiority trial

• Reject the null hypothesis:
  reject equivalence
  accept significant difference

• Fail to reject null hypothesis: not to be confused with
  PROOF OF EQUIVALENCE or non-inferiority

In a non-inferiority trial

• Reject the null hypothesis:
  reject inferiority
  accept non-inferiority
Phase II trials in oncology

• One-stage designs
• Multi-stage designs -> Simon’s two stage design
• Two-endpoint designs
• Multinomial designs
• Three-outcome designs
Testing a new drug in phase II

Question: does new drug have sufficient activity to warrant further study? i.e. is response rate of new drug large enough?

Response rate = proportion of patients who have a “tumor response” (CR or PR):

CR = tumor disappears
PR = tumor size is reduced (by at least 50% in area or by 30% in largest diameter)
SD = tumor size remains about the same
PD = tumor size increases
Purposes of phase II trials in oncology

• Select active drugs for further testing (reject inactive drugs)
• Document toxicity
  not
• Provide *definite* estimate of response rate or efficacy of new drugs
Simon’s two stage design (1)

Possible Designs For P0=0.300, P1=0.500, Alpha=0.017, Beta=0.200

<table>
<thead>
<tr>
<th>N1</th>
<th>R1</th>
<th>PET</th>
<th>N</th>
<th>R</th>
<th>Ave N</th>
<th>Alpha</th>
<th>Beta Satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>24</td>
<td>0.000</td>
<td>56</td>
<td>24</td>
<td>56.00</td>
<td>0.175</td>
<td>Single Stage</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>0.725</td>
<td>54</td>
<td>23</td>
<td>32.25</td>
<td>0.197</td>
<td>Minimax</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>0.814</td>
<td>63</td>
<td>26</td>
<td>29.64</td>
<td>0.197</td>
<td>Optimum</td>
</tr>
</tbody>
</table>

Report Definitions
- N1 is the sample size in the first stage.
- R1 is the drug rejection number in the first stage.
- PET is the probability of early termination of the study.
- N is the combined sample size of both stages.
- P0 is the response proportion of a poor drug.
- P1 is the response proportion of a good drug.
- R is the combined drug rejection number after both stages.
- Ave N is the average sample size if this design is repeated many times.
- Alpha is the probability of rejecting that P<=P0 when this is true.
- Beta is the probability of rejecting that P>=P1 when this is true.
Simon’s two stage design (2)

• The optimal two-stage design to test the null hypothesis that \( P \leq 0.300 \) versus the alternative that \( P \geq 0.500 \) has an expected sample size of 29.64 and a probability of early termination of 0.814.

• If the drug is actually not effective, there is a 0.015 probability of concluding that it is (the target for this value was 0.017).

• If the drug is actually effective, there is a 0.197 probability of concluding that it is not (the target for this value was 0.200).

• After testing the drug on 22 patients in the first stage, the trial will be terminated if 8 or fewer respond.

• If the trial goes on to the second stage, a total of 63 patients will be studied.

• If the total number responding is less than or equal to 26, the drug is rejected.
Limitations of non-randomized phase II trials in oncology

Non-randomized phase II trials may be seriously misleading because

- the impact of prognostic factors is usually far larger than that of treatment (hence response rate depends on patient selection more than on treatment efficacy)
- known prognostic factors explain little variance (hence describing the patient characteristics does not help much)
Randomized phase II trials in oncology

• Need for randomization
• Screening designs
• Selection designs
• Randomized discontinuation designs
Objectives of randomized phase II trials in oncology

• Control of selection bias
• Simultaneous testing of several new treatments, analogs, combinations, doses, schedules, ...

NOT

• A statistical comparison between randomized groups
Randomized phase II trials

*DO NOT* replace phase III trials

- are useful to relate results of a new treatment to those obtained with a known treatment, or to test several treatments at once
- can be misleading because of the small sample sizes
- are generally preferable to non-randomized phase II trials, because some degree of control is better than none!
Statistical Challenges

• Multiple testing
  – Interim analysis
  – Subset analysis
    – *Posthoc examination of subgroups will commonly reveal certain ones in which the treatment seems to work better or worse than overall.*
  – Interaction
    – *Investigators sometimes seek to claim that the effect of the factor of interest differs significantly between patient groups.*
Question 3

• Let's suppose that we examine the difference of the impact of a new treatment to the standard for 5 factors, each with two levels [e.g., smoking (yes-no), age group (>50 vs. <=50 years old), gender, presence or absence of mutation].
• This would correspond to conducting 10 statistical tests at a 5% significance level.
• If we find a statistically significant result for the mutation subgroup, what is the probability that this is not a real finding?
  A. 5%
  B. 10%
  C. 40%
  D. 80%
Curse of multiplicity

• Common practice to perform multiple subgroup analyses
• The probability of a false positive finding (type-I error) increases as the number of subgroup analyses increases
• 10 analyses conducted → 40% chance at least 1 yields $p \leq 0.05$
• 14 analyses conducted → 50% chance at least 1 yields $p \leq 0.05$
• Easy to find one subgroup in which the treatment appears to work
Probability That Multiple Subgroup Analyses Will Yield at Least One (Red), Two (Blue), or Three (Yellow) False Positive Results.
Interim Monitoring

• Comparative clinical trials can take several years to complete
• Because of the availability of ‘updated’ information, it is appropriate and sometimes ethically necessary to conduct **interim analyses** to determine:
  1. Is it still ethical to randomly assign the study treatments and to allow randomized patients to continue to take their assigned treatments?
  2. Does the study still have the potential to achieve its scientific goals?
Group Sequential Methods

• The data is analyzed after groups of pre-specified size have been collected and compare the test statistic to stopping boundaries at each interim analysis (e.g., analyze data after 50, 100, 150, … patients are enrolled).

• The trial is stopped prematurely if superiority (or noninferiority) of a treatment has been established

Repeated significance testing
Statistical Stopping Rules: The Group Sequential Approach

Address the high **Probability of False Positive:**

– The fundamental statistical problem raised by trial monitoring is that the more analyses that are undertaken, the greater is the probability of false positive (Type I error).

– Planning a limited number of IM at pre-specified times using formal rules that address this problem is the key.
Cumulative Probability of Rejecting H0 for $K = 5$ and $\alpha = 0.05$

<table>
<thead>
<tr>
<th>$k$</th>
<th>Pocock</th>
<th>O'Brien &amp; Fleming</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0158</td>
<td>0.00000504</td>
</tr>
<tr>
<td>2</td>
<td>0.0275</td>
<td>0.00126</td>
</tr>
<tr>
<td>3</td>
<td>0.0365</td>
<td>0.00891</td>
</tr>
<tr>
<td>4</td>
<td>0.0439</td>
<td>0.0256</td>
</tr>
<tr>
<td>5</td>
<td>0.0500</td>
<td>0.0500</td>
</tr>
</tbody>
</table>
O’Brien-Fleming boundaries Ho
O’Brien-Fleming boundaries $H_0 / H_1$
Subgroup analysis

• Multiplicity issues
• Prognostic & Predictive Factors
• Interaction

• Comparison of two or more treatment regimens restricted to a subgroup of the study sample, defined according to a baseline characteristic
• Examples: age, sex, mutation, smoking

• Prognostic & Predictive factors
Prognostic - Predictive

• Correlation between biomarker and true clinical endpoint:
  – makes a **prognostic marker**.
  – does not make a **predictive biomarker**.

• **Predictive biomarker**: difference in treatment effect between its levels
Heterogeneity and interaction

Heterogeneity of treatment effect:
varying effect across levels of a baseline characteristic

- Quantitative heterogeneity: one treatment is always better, but by varying degrees
- Qualitative heterogeneity: one treatment is better in a particular subgroup but worse in another subgroup

→ Predictive
Estimating the magnitude of treatment effect
Heterogeneity: varying effect across levels of a baseline characteristic
Estimating the magnitude of treatment effect:
Qualitative heterogeneity
Estimating the magnitude of treatment effect: Quantitative heterogeneity
Prognostic & predictive factors

**Prognostic**
- Std vs Exp: Factor present
- Std vs Exp: Factor absent

**Predictive**
- Std vs Exp: Factor present
- Std vs Exp: Factor absent

**Prognostic and predictive**
- Std vs Exp: Factor present
- Std vs Exp: Factor absent

---

ESMO Clinical Trials Course

27-28 March 2015; Vienna, Austria
Heterogeneity and interaction

• Evaluating heterogeneity requires a test of interaction
• Separate tests within subgroups cannot be used to determine heterogeneity
• Interaction tests are often underpowered; leading to false negative conclusions

• Predictive Factor: Only if treatment by group interaction is significant
Is Group predictive for differential survival benefit from treatment vs. control?
Prognostic vs. Predictive factors

Survival of patients by histology (Clark, Molecular Oncology 2008)

<table>
<thead>
<tr>
<th></th>
<th>Erlotinib arm</th>
<th>Placebo arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR (95% C.I.)</td>
<td>0.66 (0.52-0.83)</td>
<td>0.65 (0.48-0.88)</td>
</tr>
</tbody>
</table>
Prognostic vs. Predictive factors

Survival of patients by treatment arm (Clark, Molecular Oncology 2008)

Adenocarcinoma
HR (95% C.I.) = 0.71 (0.56-0.92)

Squamous cell
HR (95% C.I.) = 0.67 (0.50-0.90)
Prognostic vs. Predictive factors

Survival of patients by histology & treatment arm
(Clark, Molecular Oncology 2008)

- Test for interaction between histology and treatment benefit: $P=0.97$
Question 4.

• What do the results on histology indicate?

A. It is prognostic but not predictive for survival
B. It is predictive but not prognostic for survival
C. It is neither predictive nor prognostic for survival
D. It is both predictive and prognostic for survival
Answer to Question 4.

• What do the results on histology indicate?
  
  A. It is prognostic but not predictive for survival  
  B. It is predictive but not prognostic for survival  
  C. It is neither predictive nor prognostic for survival  
  D. It is both predictive and prognostic for survival
Is Group predictive for differential survival benefit from treatment vs. control?
Prognostic vs. Predictive factors

Survival of patients by smoking history (Clark, Molecular Oncology 2008)

Erlotinib arm
HR (95% C.I.) = 0.54 (0.41-0.71)

Placebo arm
HR (95% C.I.) = 1.01 (0.71-1.45)
Prognostic vs. Predictive factors

Survival of patients by treatment arm (Clark, Molecular Oncology 2008)

Never smokers
HR (95% C.I.) = 0.42 (0.28-0.64)

Current / Former smokers
HR (95% C.I.) = 0.87 (0.71-1.05)
Prognostic vs. Predictive factors

Survival of patients by smoking history & treatment arm
(Clark, Molecular Oncology 2008)

Test for interaction between smoking history and treatment benefit: $P=0.006$
Question 5.

• What do the results on smoking history indicate?

  A. It is prognostic but not predictive for survival
  B. It is predictive but not prognostic for survival
  C. It is neither predictive nor prognostic for survival
  D. It is both predictive and prognostic for survival
Answer to Question 5.

• What do the results on smoking history indicate?

   A. It is prognostic but not predictive for survival
   B. It is predictive but not prognostic for survival
   C. It is neither predictive nor prognostic for survival
   D. It is both predictive and prognostic for survival
Definitive validation for a predictive marker: Phase III setting

- Retrospective validation may be acceptable as a marker validation strategy in certain cases, but
- the gold standard for predictive marker validation continues to be a prospective RCT.

Designs:

• Targeted or enrichment designs (HERA)
• Allcomers designs
  - Hybrid designs (TAILORx, MINDACT)
  - Marker by treatment interaction designs
  - Sequential testing strategy designs
• Adaptive designs
Randomized phase III designs

B. Enrichment design

Assess biomarker

Biomarker positive

Randomize

Treatment A

Treatment B

Biomarker negative

Off study

Randomized phase III designs

A. Biomarker-stratified design


ESMO Clinical Trials Course

27-28 March 2015; Vienna, Austria
Adaptive Trials Increasing

- Survey
  - 2003 to 2006: 3 adaptive trials/year
  - 2007: 13 adaptive trials
  (Quinlan J, Cytel survey)

- Acceptance of Bayesian statistical framework
  - 5-10% of the medical devices approved by FDA recently used Bayesian designs and analyses vs none, 10 years ago
  (Biswas S, Clinical Trials 2009;6:205-216)

Debated in the statistical community: Subjectivity in the design!

Adaptive trials are great for learning,
but are not a panacea (Lee, MD Anderson)

Nelson N. Adaptive Clinical Trial Design: Has its time come? JNCI 2010; 102(16):1217-8
Adaptive design methods

Clinical point of view:
• reflect real clinical practice in clinical development
• very attractive due to their flexibility and
• very useful especially in early clinical development

Statistical point of view:
• the use of adaptive methods in clinical trials makes current good statistics practice even more complicated
• The validity of the use of adaptive design methods is not well established and fully understood

Adaptive design methods

• **Guidelines** regarding the use of adaptive design methods must be developed so that
  – appropriate statistical methods and
  – statistical software packages can be developed accordingly

• **Regulatory guidelines** can
  – prevent possible misuse and/or abuse of adaptive design methods in clinical trials
  – maintain the validity and integrity of the trial

Concluding Remarks

• the flexibility afforded by adaptive designs might not be compensated for by the potential loss in credibility associated with their use.

• The use of adaptive design methods in clinical trials is motivated by its flexibility and efficiency, but there are still challenges in implementation.

• An independent data monitoring committee (IDMC) should be established.

• The future will tell which of these innovative designs are useful and when they constitute a definite improvement over classic approaches.