ATLAS
ASIAN CLINICAL TRIALS NETWORK FOR CANCERS PROJECT

National Cancer Center Japan

Sample Size Calculation

National Cancer Center, Japan

Biostatistics Division/JCOG Data Center

Junki Mizusawa

JCOG
Japan Clinical Oncology Group

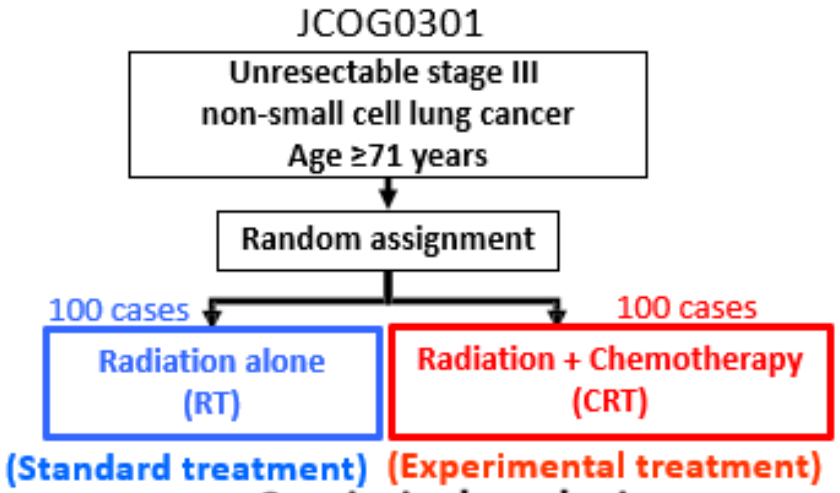
Objective of this lecture

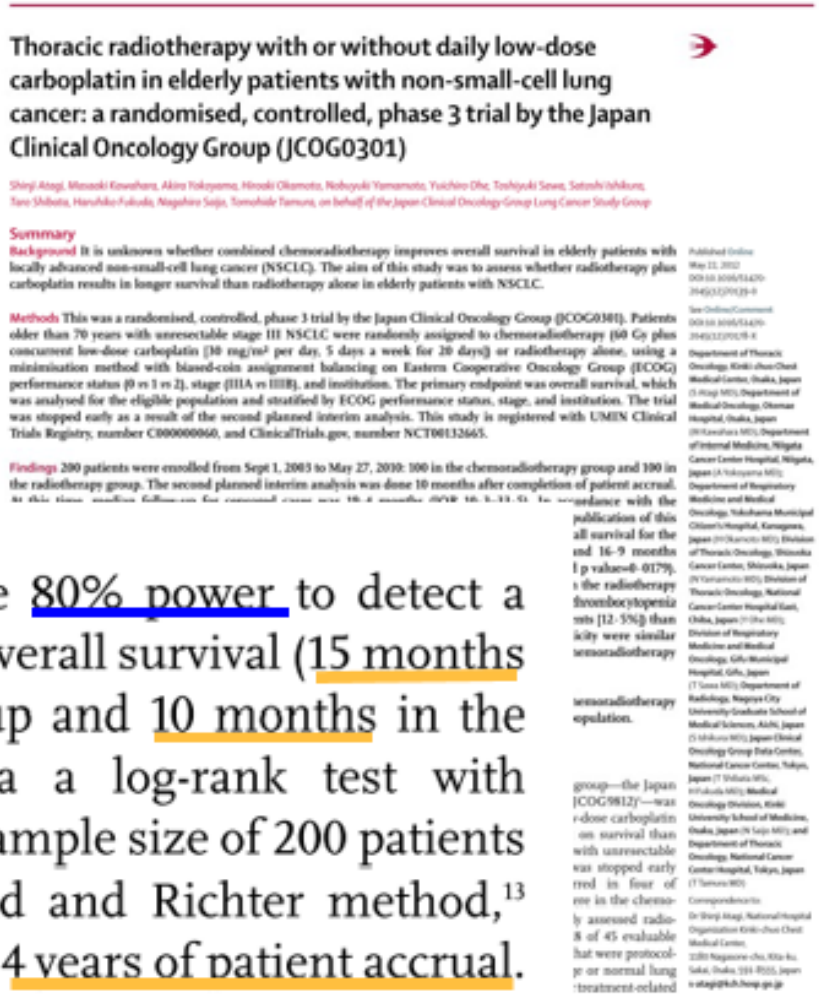
- As a clinician, learn what you need to know about sample size calculations
  - What to consider when consulting a statistician?
    - What information is needed for the sample size calculation?
  - What do statisticians think?

- Understand sample size calculations to better understand study design
  - Superiority and noninferiority tests
  - Risk-benefit balance

Directions for Sample Size Calculation

Example of Sample Size Description in an Article: Superiority Test





Thoracic radiotherapy with or without daily low-dose carboplatin in elderly patients with non-small-cell lung cancer: a randomised, controlled, phase 3 trial by the Japan Clinical Oncology Group (JCOG0301)

Summary

Background It is unknown whether combined chemoradiotherapy improves overall survival in elderly patients with locally advanced non-small-cell lung cancer (NSCLC). The aim of this study was to assess whether radiotherapy plus carboplatin results in longer survival than radiotherapy alone in elderly patients with NSCLC.

Methods This was a randomised, controlled, phase 3 trial by the Japan Clinical Oncology Group (JCOG0301). Patients older than 70 years with unresectable stage III NSCLC were randomly assigned to chemoradiotherapy (60 Gy plus concurrent low-dose carboplatin [30 mg/m² per day, 5 days a week for 20 days]) or radiotherapy alone, using a minimisation method with biased-coin assignment balancing on Eastern Cooperative Oncology Group (ECOG) performance status (0 vs 1 vs 2), stage (IIIA vs IIIB), and institution. The primary endpoint was overall survival, which was analysed for the eligible population and stratified by ECOG performance status, stage, and institution. The trial was stopped early as a result of the second planned interim analysis. This study is registered with UMIN Clinical Trials Registry, number C000000060, and ClinicalTrials.gov, number NCT00132665.

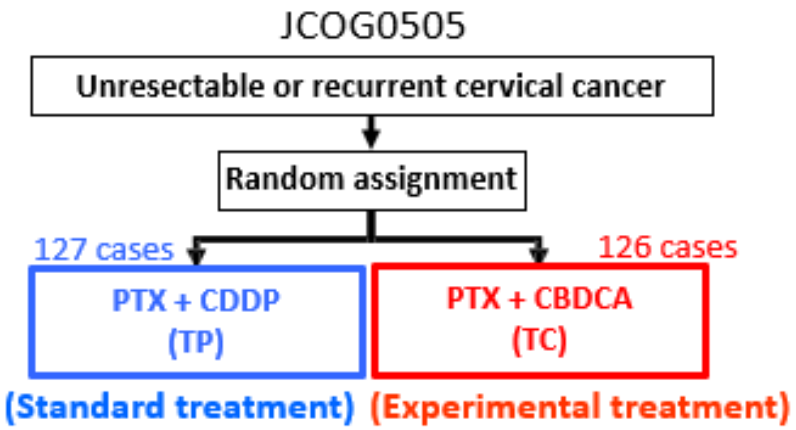
Findings 200 patients were enrolled from Sept 1, 2003 to May 27, 2010: 100 in the chemoradiotherapy group and 100 in the radiotherapy group. The second planned interim analysis was done 10 months after completion of patient accrual.

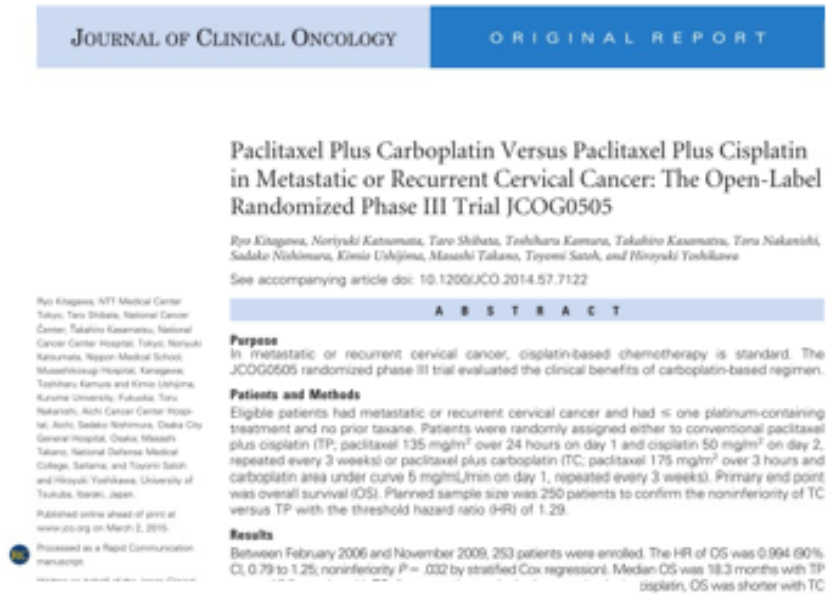
Statistical analysis

The trial was designed to have 80% power to detect a 5-month difference in median overall survival (15 months in the chemoradiotherapy group and 10 months in the radiotherapy alone group) via a log-rank test with one-sided alpha of 0·05.[7,10–12] A sample size of 200 patients was planned by the Schoenfeld and Richter method,[13] with 1·5 years of follow-up after 4 years of patient accrual.

Atagi et al. (2012) Lancet Oncol. 13 (7): 671-8.

Example of Sample Size Description in an Article: Noninferiority Test




JOURNAL OF CLINICAL ONCOLOGY ORIGINAL REPORT

Paclitaxel Plus Carboplatin Versus Paclitaxel Plus Cisplatin in Metastatic or Recurrent Cervical Cancer: The Open-Label Randomized Phase III Trial JCOG0505

Ryo Kitagawa, Noriyuki Katsumata, Taro Shibata, Toshiharu Kamura, Takahiro Kasumatsu, Toru Nakanishi, Sadako Nishimura, Kimio Ushijima, Masashi Takano, Toyomi Satoh, and Hiroyuki Toshikawa

See accompanying article doi: 10.1200/JCO.2014.57.7122

ABSTRACT

Purpose
In metastatic or recurrent cervical cancer, cisplatin-based chemotherapy is standard. The JCOG0505 randomized phase III trial evaluated the clinical benefits of carboplatin-based regimen.

Patients and Methods
Eligible patients had metastatic or recurrent cervical cancer and had ≤ one platinum-containing treatment and no prior taxane. Patients were randomly assigned either to conventional paclitaxel plus cisplatin (TP; paclitaxel 135 mg/m^2 over 24 hours on day 1 and cisplatin 50 mg/m^2 on day 2, repeated every 3 weeks) or paclitaxel plus carboplatin (TC; paclitaxel 175 mg/m^2 over 3 hours and carboplatin area under curve 5 mg/mL/min on day 1, repeated every 3 weeks). Primary end point was overall survival (OS). Planned sample size was 250 patients to confirm the noninferiority of TC versus TP with the threshold hazard ratio (HR) of 1.29.

Results
Between February 2006 and November 2009, 253 patients were enrolled. The HR of OS was 0.994 (90% CI, 0.79 to 1.25; noninferiority $P = .032$ by stratified Cox regression). Median OS was 18.3 months with TP

Statistical Design and Analysis

Initially, with an accrual time of 2.5 years and minimum follow-up period of 1 year, the required number of OS events was 209 and the planned sample size was 250 according to the Schoenfeld and Richter method[17] to confirm the noninferiority of TC compared with TP, with a one-sided α level of 0.05 and power of at least 70%, with noninferiority margin of 1.29, corresponding to 2 months in OS, assuming a median OS in the TP group of 9 months based on outcome in the GOG 169 trial (SCC only).

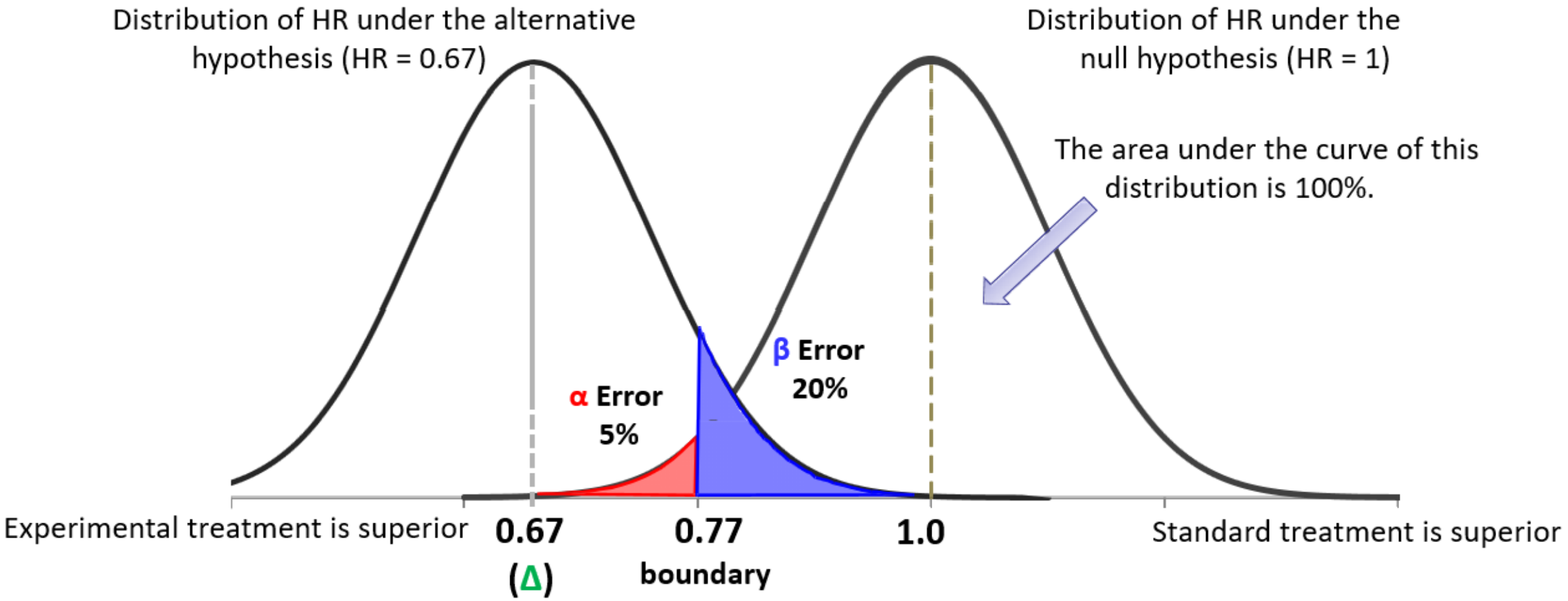
Kitagawa et al. (2015) J Clin Oncol.

Basic Principles of Sample Size Calculation

The required sample size determined in sample size calculations is established to satisfy the following conditions:

1. Suppress the probability (**α**) of mistakenly judging that a new treatment is effective when the null hypothesis is correct below a specified threshold
2. Maintain the probability of correctly judging that a new treatment is effective (**power [1 – β]**) when the true effect of the new treatment is **Δ** (when the alternative hypothesis is correct)

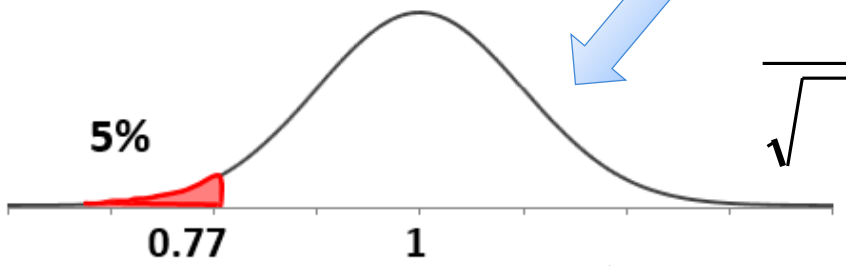
Distribution of HR Obtained as α, β, and Δ



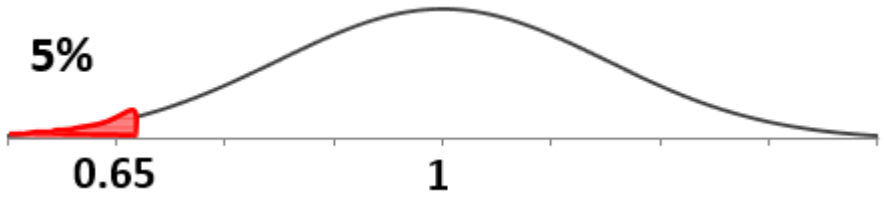
Relationship between Sample Size and Distribution of the Test Statistic

N is included in the formula defining the distribution of the test statistic

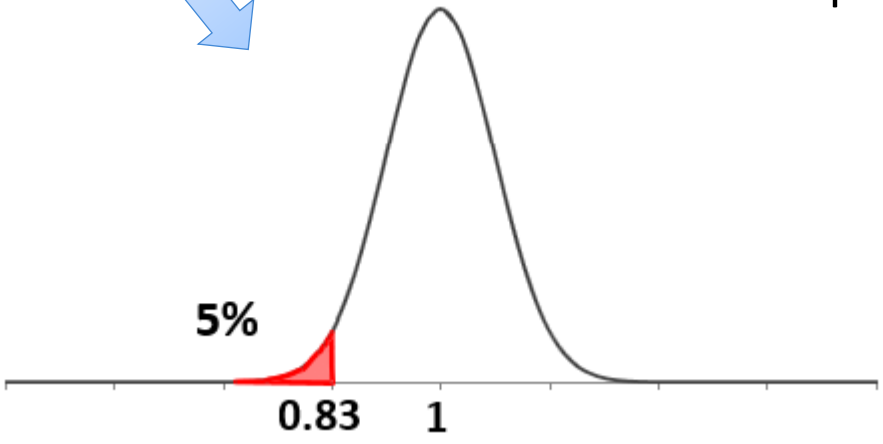
$$\frac{\Delta}{\sqrt{(\text{Data Variability})^2/N}}$$



As N decreases, the distribution becomes more gradual.



As N increases, the distribution becomes sharp.



Framework for Sample Size Calculation

Define N so that the distribution of the test statistic satisfies α, β, and Δ

Substitute the value corresponding to the significance level α
(1.96 if one-sided α = 2.5% and two-sided α = 5% and
1.64 if one-sided α = 5%)

$$N = \frac{\left(Z_\alpha + Z_\beta\right)^2 \cdot \text{variability}^2}{\Delta^2}$$

The format in which the data is entered depends on the outcome type
(continuous variable, binary data, time-to event) and test method

Enter the value corresponding to power
(0.84 if power = 80%and 1.28 if power = 90%)

Framework of the Sample Size Formula for Each Test Method

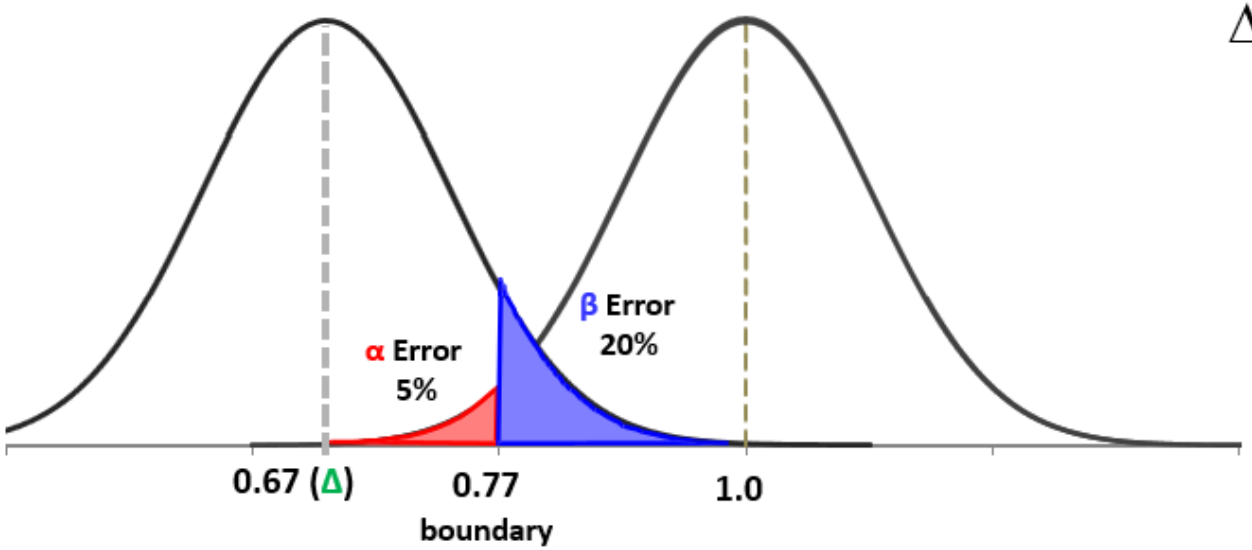
$$t\text{-test: } N = \frac{(Z_\alpha + Z_\beta)^2 \cdot (\text{Data Variability})^2}{(\text{Difference of mean})^2}$$

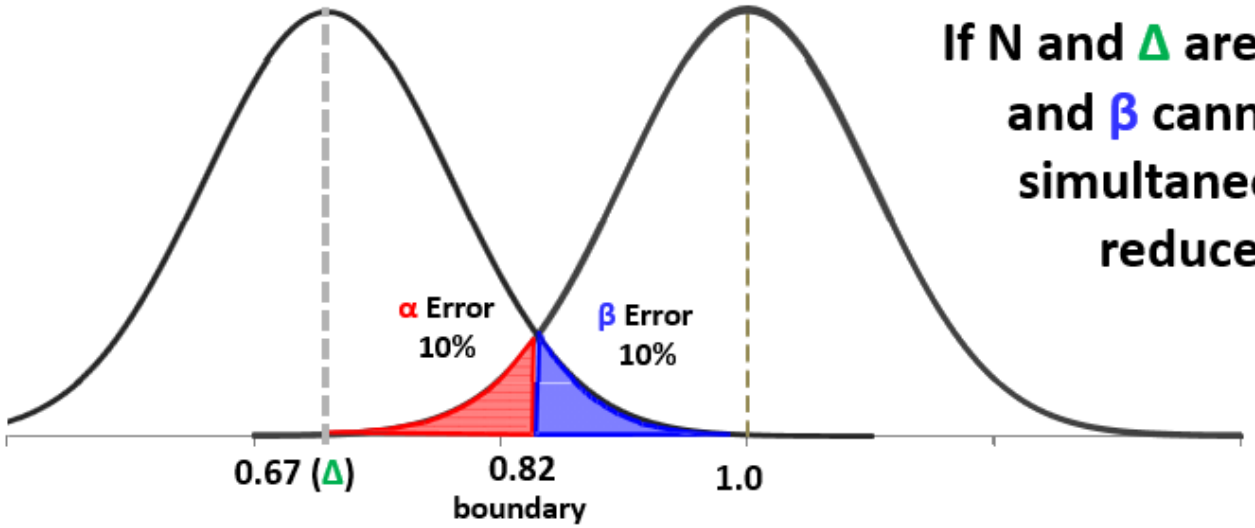
$$\chi^2 \text{ Test: } N = \frac{(Z_\alpha + Z_\beta)^2 \cdot \left[\begin{array}{l}\text{Proportion of standard group (1 – Proportion of standard group)} \\ + \text{Proportion of study group (1 – Proportion of study groups)}\end{array}\right]^2}{(\text{Proportion difference})^2}$$

$$\text{log rank test: } N = \frac{(Z_\alpha + Z_\beta)^2 \cdot \left[\begin{array}{l}\text{The value considering the prognosis} \\ \text{and enrollment/follow-up period.}\end{array}\right]}{\ln(HR)^2}$$

α and β are Trade-offs

$$N = \frac{(Z_\alpha + Z_\beta)^2 \cdot \text{Variability}^2}{\Delta^2}$$

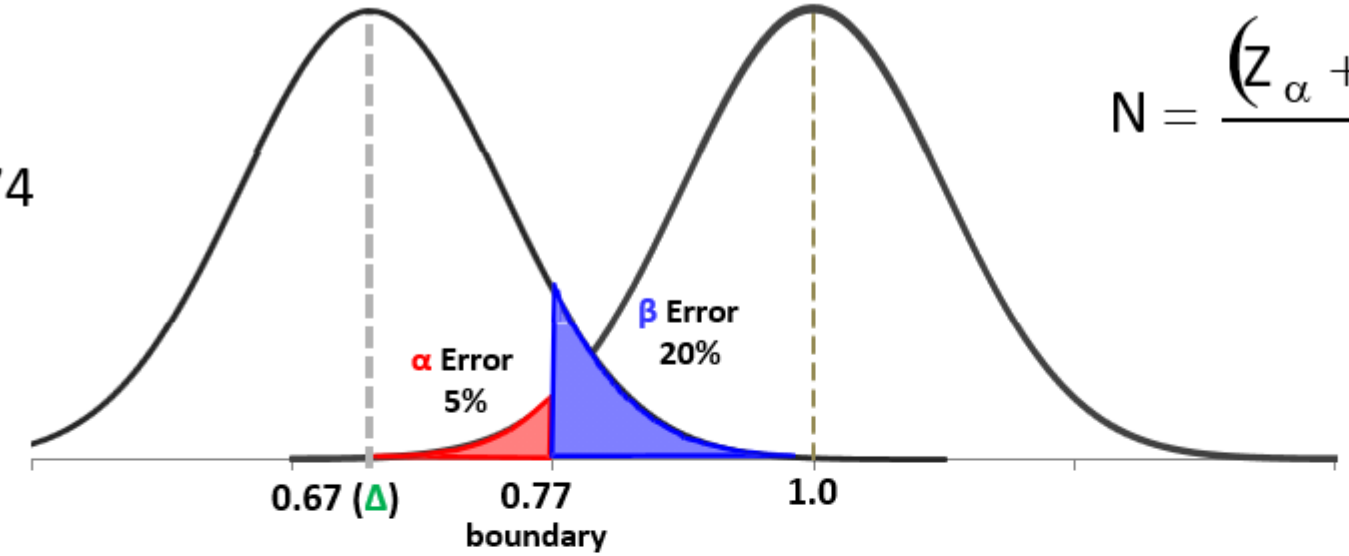




If N and Δ are fixed, α and β cannot be simultaneously reduced

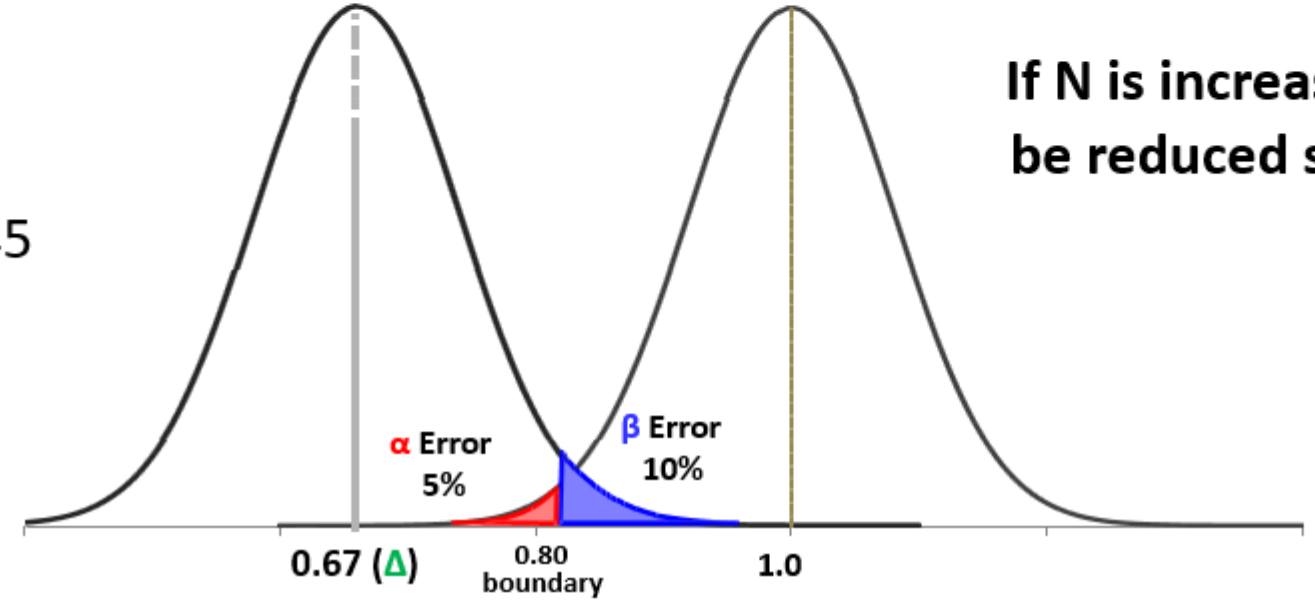
If You Increase N, Power can be Increased by Fixing α

N = 174



$$N = \frac{(Z_\alpha + Z_\beta)^2 \cdot \text{Variability}^2}{\Delta^2}$$

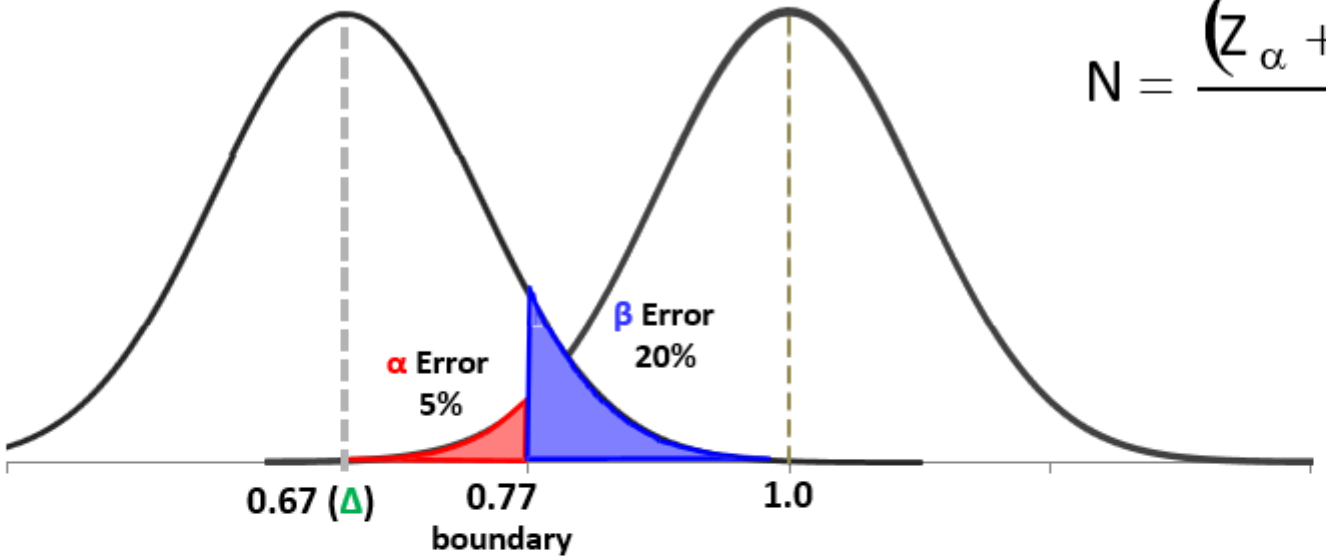
N = 245

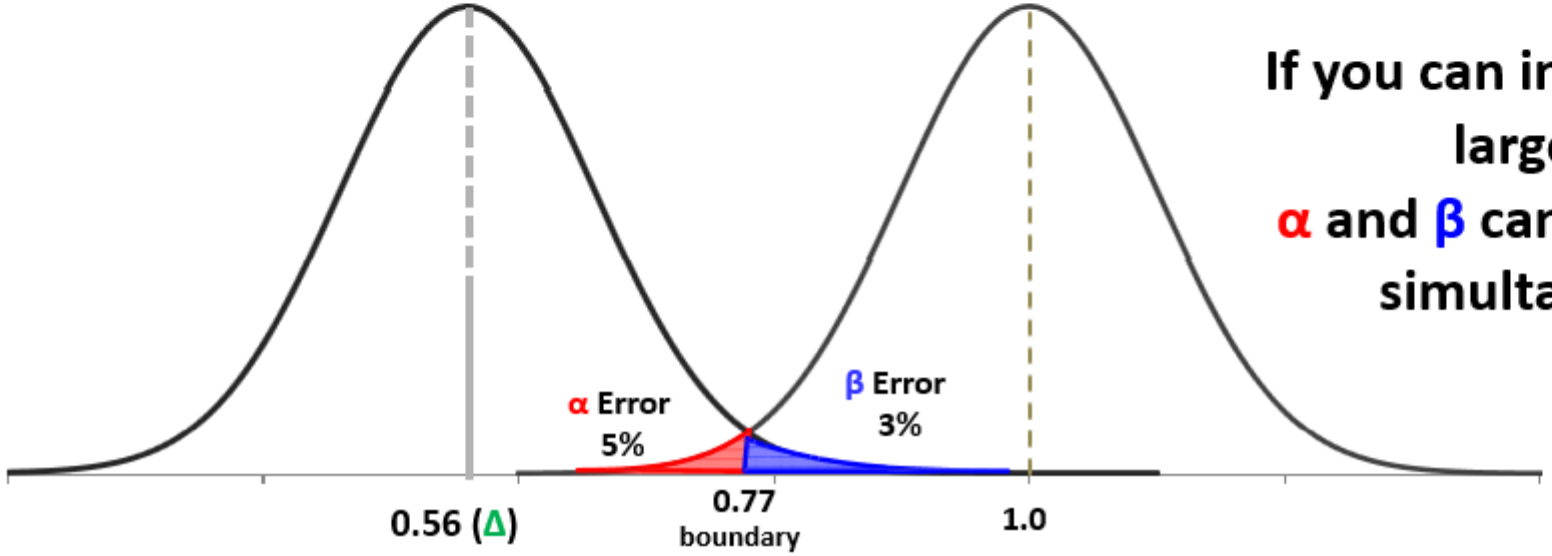


If N is increased, α and β can be reduced simultaneously

If You Can Increase Δ, You Can Fix α and Increase Power

$$N = \frac{(Z_\alpha + Z_\beta)^2 \cdot \text{Variability}^2}{\Delta^2}$$





If you can increase Δ to a large size, α and β can be reduced simultaneously

Memorize!

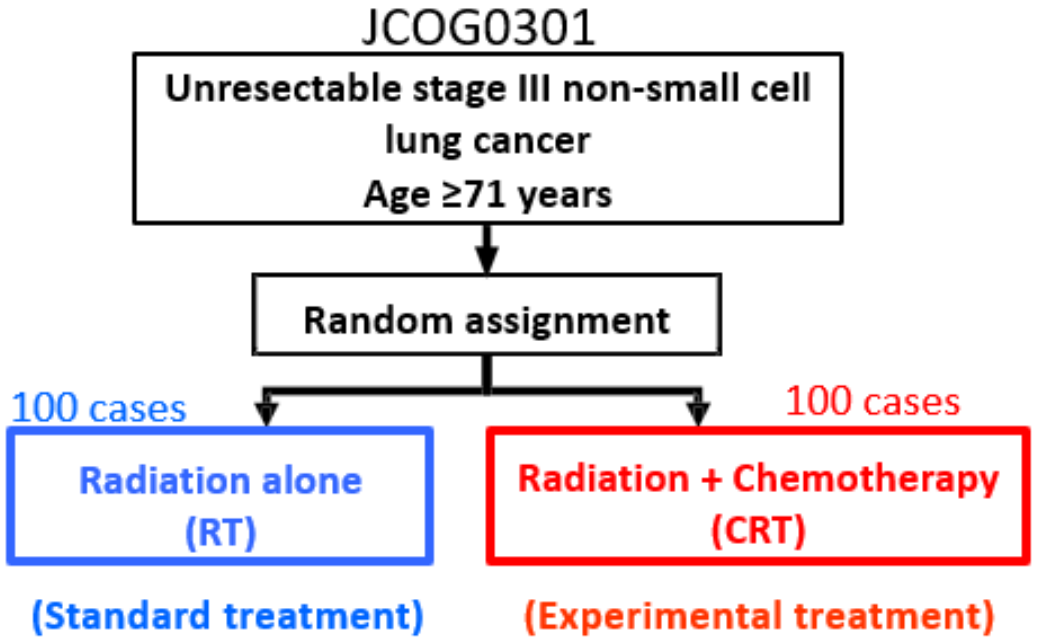
Minimum Parameters Required for Sample Size Calculation

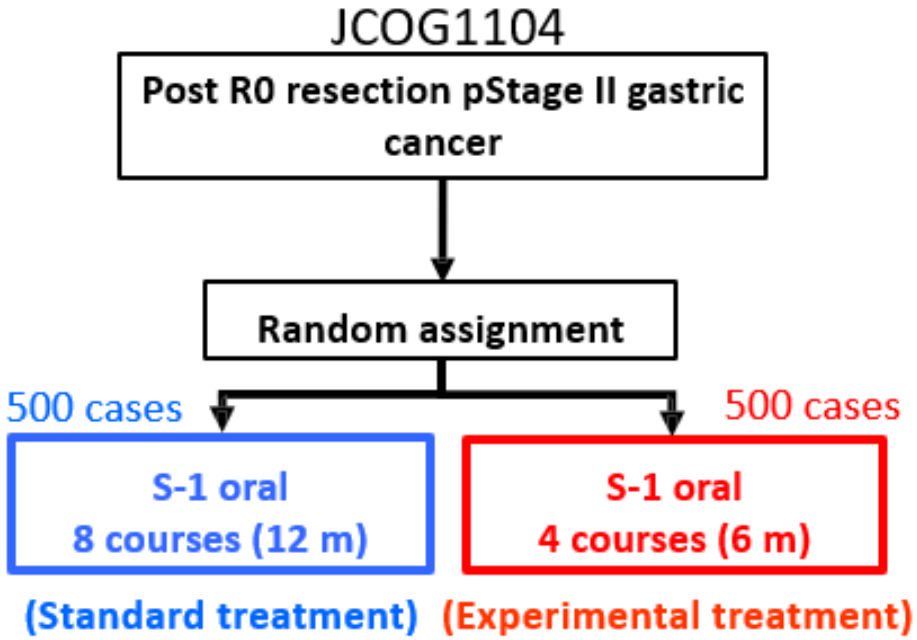
- **Size of treatment effect: Δ (delta)**
- **Variability**
- **Significance level: α (alpha)**
  - The value below which the *P* value is considered to indicate a “significant difference”
- **Power: 1–β (1–beta, power)**
  - The probability of correctly determining that a truly effective treatment is effective

Things to Consider When Setting Parameters

- Design: superiority or noninferiority
- Significance level: one-sided or two-sided, always 5%?
- Power
- Significance of treatment effect (clinical significance or expectation)

Superiority and Non-inferiority Tests





Superiority test

- Experimental treatment must be superior in efficacy
- Experimental treatment is inferior to standard treatment in terms of safety and other factors (**Toxic new**)

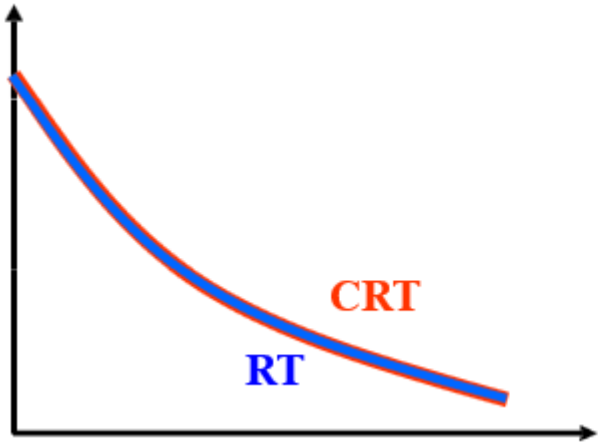
Non-inferiority test

- Experimental treatment must be no less effective than the standard treatment
- Experimental treatment is superior to standard treatment in terms of safety and other factors (**Less toxic new**)

Superiority/Non-inferiority Determination

- Decide on the situation in which **efficacy endpoints** (survival curves) overlap.
  - Select **standard treatment** → **Superiority trials**
  - Select **experimental treatment** → **Non-inferiority trials**

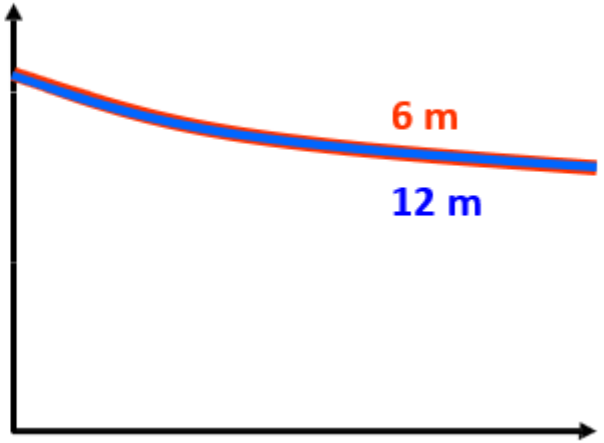
For JCOG0301



If the OS of the **CRT group** (**Toxic new**) and **RT group** are equivalent, the standard treatment is the **RT group**

→ **Superiority test**

For JCOG1104



If the RFS of the **6 m group** (**Less toxic new**) and the **12 m group** are equivalent, the standard treatment is the **6 m group**

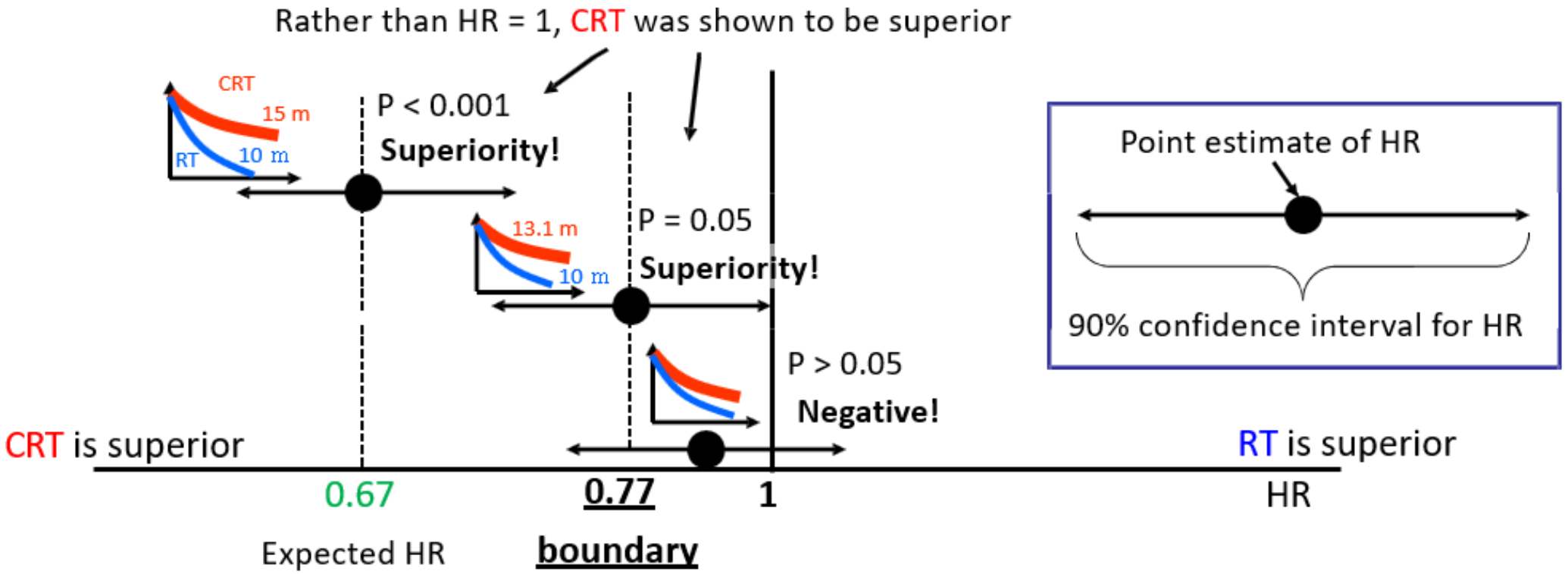
→ **Non-inferiority test**

Statistical Significance and Boundary: The Case of Superiority Tests

- One-sided *P* value < 0.05 = 90% CI of two-sided hazard ratio (HR) is less than or equal to the null hypothesis (**HR = 1**)

Sample size design of JCOG0301

- One-sided α = 5%
- Power = 80
- MST for RT vs. CRT: 10 m vs. 15 m (HR = 0.67)
- Required number of participants for analysis (expected number of events): 174 cases (151) in total

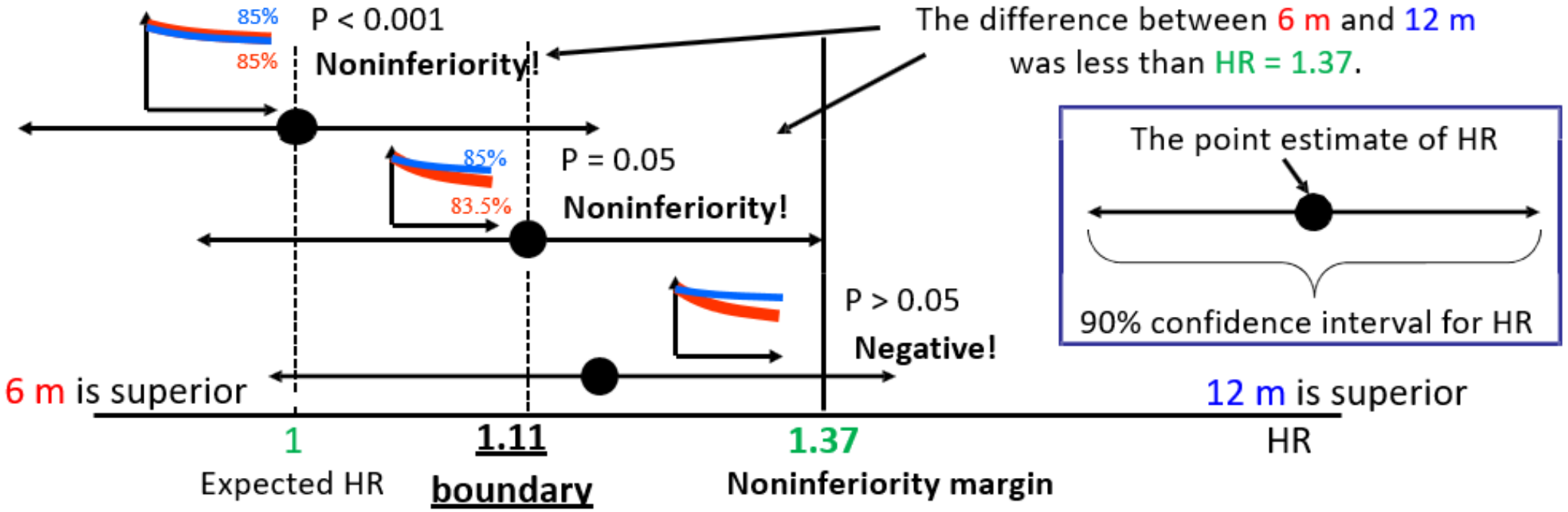


Statistical Significance and Boundary: The Case of Non-inferiority Trials

- Non-inferiority one-sided *P* value < 0.05 = two-sided 90% CI of HR is less than the null hypothesis (**non-inferiority margin)**

Sample size design of JCOG1104

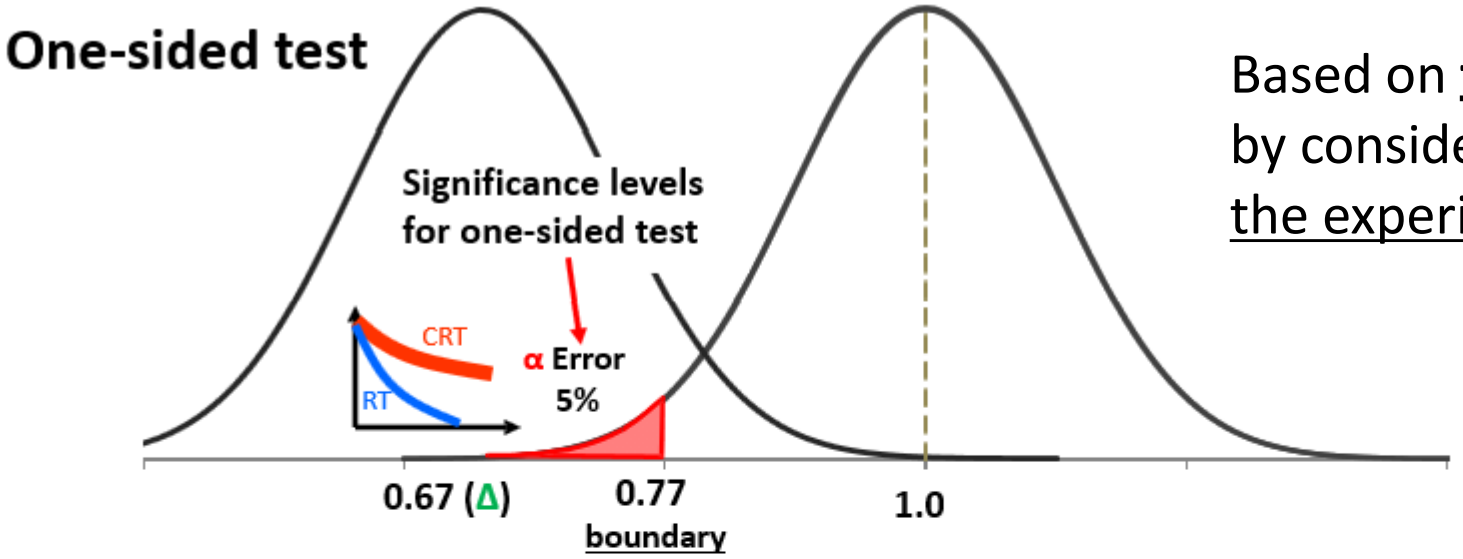
- One-sided α = 5%
- Power = 80
- Three-year RFS (3yRFS) at 12 m vs. 6 m: 85% vs. 85% (HR = 1)
- Non-inferiority margin for HR: 1.37 (equivalent to 5% for 3yRFS)
- Required number of participants for analysis (expected number of events): 964 cases (246) in total



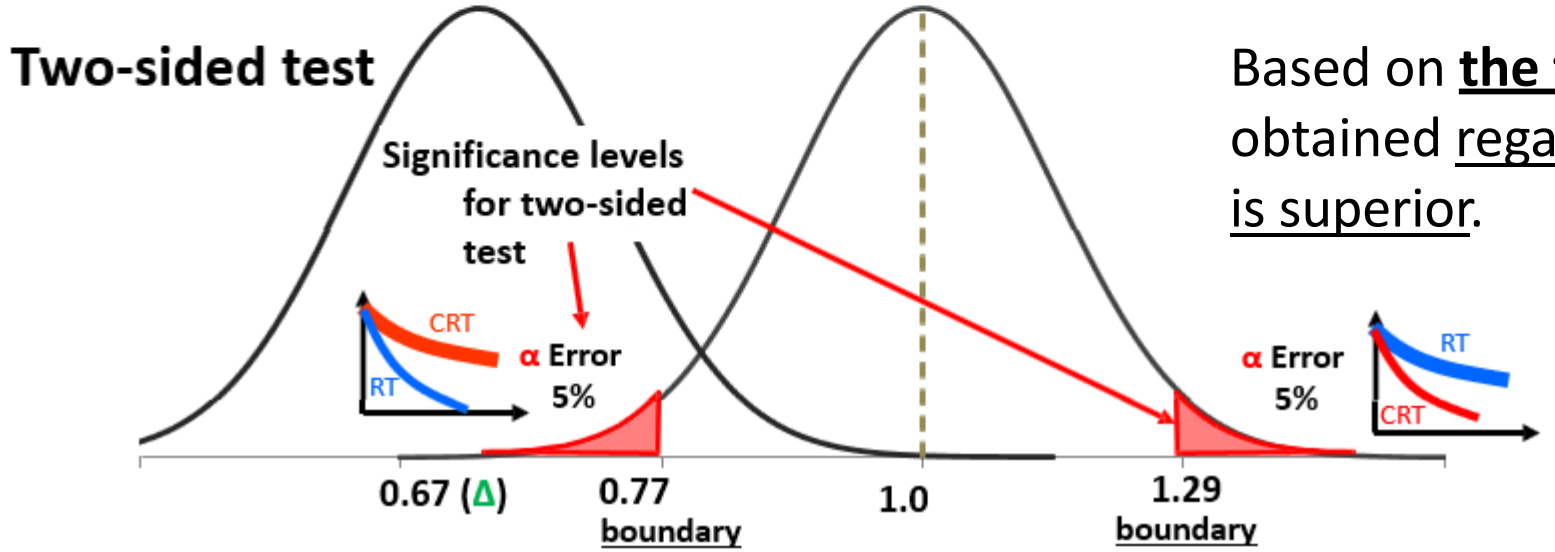
Things to Consider When Setting Parameters

- Design: superiority or noninferiority
- Significance level: one-sided or two-sided, always 5%?
- Power
- Significance of treatment effect (clinical significance or expectation)

Significance Level α: One-sided and Two-sided Tests



Based on **the one-sided *P* value** obtained by considering only the direction in which the experimental treatment is superior.



Based on **the two-sided *P* values** obtained regardless of which treatment is superior.

Two ways to decide whether to use a one-sided or two-sided test

Recommended!

A) Decide **according to the decision-making process**, considering the risks/benefits of the treatment

- JCOG0301 (superiority): in case of a comparison of RT vs. CRT: **one-sided test**
  - CRT that is Toxic new must be superior to RT
  - Not interested in CRT being significantly inferior to RT; even if CRT is not significantly inferior to RT in the interim analysis, the study will be terminated.
- JCOG1104 (non-inferiority): In the case of a comparison of 12 m vs. 6 m: **one-sided test**
  - The less toxic 6 m is acceptable unless it is inferior to 12 m (it can be superior)
    - In cases where superiority is not allowed, equivalence testing, rather than non-inferiority testing, is required (e.g., for generic antihypertensive drugs).
- If the risk-benefit ratio of the treatments is comparable (superiority [a test to prove a difference]): **two-sided test**
  - Either treatment may be superior, and **if there is no significant difference, either treatment is the standard treatment (JCOG1213)**

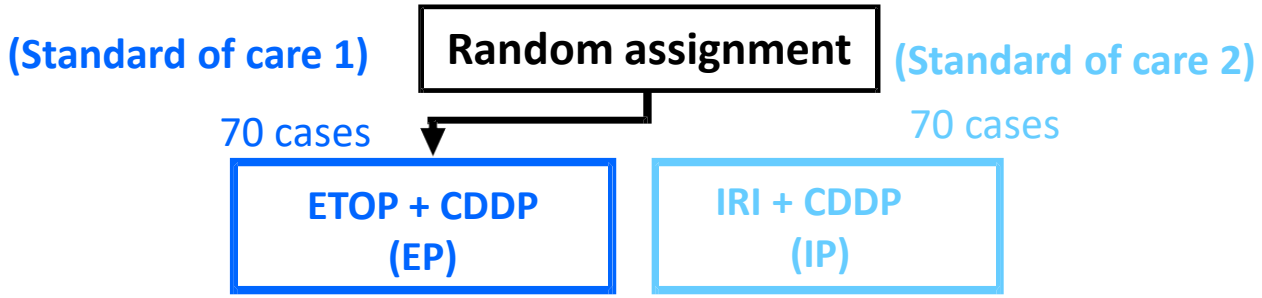
B) Should **always be a two-sided test if there is a possibility that** the experimental treatment is **inferior to** the standard treatment

- Two-sided test to also prove that the experimental treatment is inferior
- Even if highly toxic experimental treatments are found to be inferior in interim analysis, the trial will not be discontinued until they are significantly inferior.

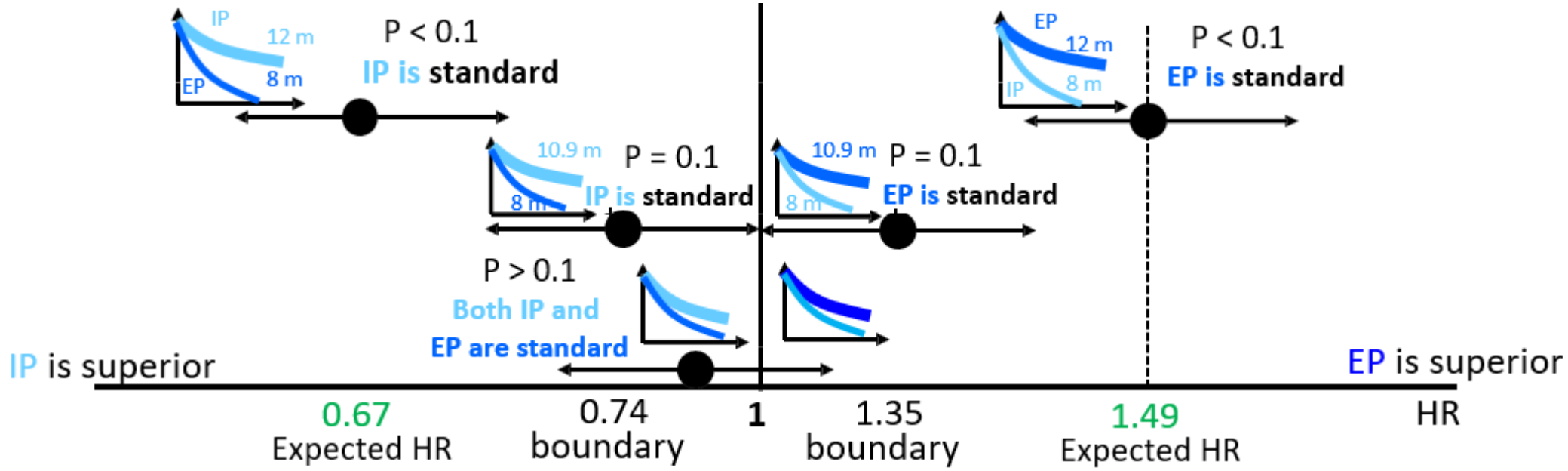
Example of JCOG Study With Two-Sided α

JCOG1213

Gastrointestinal tract, hepatobiliary, and pancreatic primary unresectable or recurrent neuroendocrine cancer



- Efficacy and safety are hard to choose Comparison of two standard treatments
- We are also interested in determining which one is superior to the other.
- **Two-sided** α = 10%, power = 80%
- MST 8 m vs. 12 m (HR = 0.67)



α Is Not Always 5%: Set α According to the Situation

- Two-sided α = 0.05 or one-sided α = 0.025
  - Global standard (**Phase III standard [ICH E9]**): There is no reason to comply
  - Changing from two-sided to one-sided does not reduce the sample size
    - Two-sided 5% corresponds to one-sided 2.5% (same sample size)
  - α error is a risk for the patient, so they should be kept as small as possible
- One-sided α = 0.05–0.2
  - **Phase II**: Even if it mistakenly shows significance, it can be validated next in Phase III
- Two-sided α = 0.1–0.2
  - Evaluation of bias when comparing **background factors** between groups
    - P > 0.05 can affect endpoint → it is wrong to uniformly say "N.S."
  - Test for differences in treatment effects between subgroups (**interaction effects**)
- Two-sided or one-sided α = 0.0001–0.005
  - **Phase III**: Significance levels in **interim analysis** considering the multiplicity
  - Unless there is a significant difference, the trial will not be discontinued for efficacy during its course

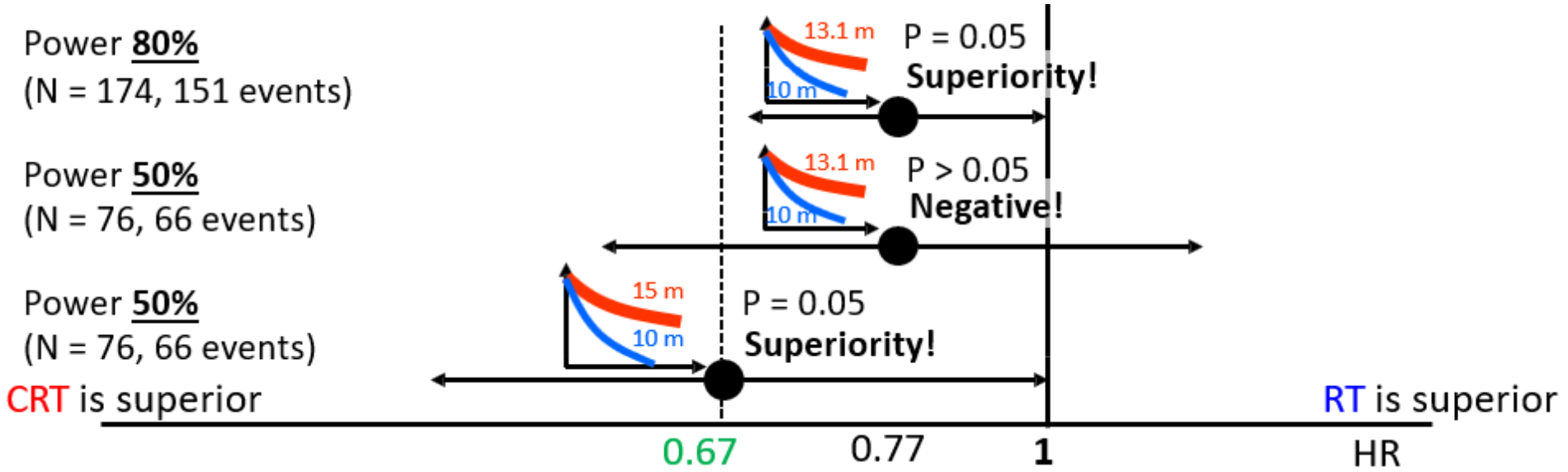
Things to Consider When Setting Parameters

- Design: superiority or noninferiority
- Significance level: one-sided or two-sided, always 5%?
- Power
- Significance of treatment effect (clinical significance or expectation)

Power Setting: 80% or more is the default

- Because beta errors are a risk for the researchers, they do not need to be as small as α errors.
  - Even if it is not significant, it will not be worse than the current situation.
- Power 50% = expected Δ is the boundary test
  - It is unethical because it is like flipping a coin to decide
  - There is no consensus on how much the statistical power can be reduced.

JCOG0301 example (one-sided α = 5%, MST 10 m vs. 15 m [HR = 0.67])



Things to consider when setting Parameters

- Design: superiority or noninferiority
- Significance level: one-sided or two-sided, always 5%?
- Power
- Significance of treatment effect (clinical significance or expectation)

Direction for Determining Δ

Common to superiority and noninferiority studies

A) Decide on the **risk-benefit** of the treatment to be compared

- Set the clinically meaningful difference to Δ
  - If the difference in toxicity is large, the Δ is large. If the difference is small, the Δ is small
- Clinically meaningful differences are determined by consensus among researchers
  - To what extent of difference would be acceptable for adopting the new treatment?

B) Decide based on the **expected** additional effect

- If the likelihood of success is high, set a large Δ

In the case of a noninferiority test (how to determine the noninferiority margin)

C) Decide to ensure that the **product is superior to the placebo to a certain degree**

Hypothetical Example of the planning Phase of JCOG0301 (Superiority Study)

RT group

- Assuming MST = **10 m**
- RT is 2 Gy/day, 5 days/week, 60 Gy

CRT group [low-dose daily CBDCA + RT]

- Phase II results, MST = **20 m**
- RT is 2 Gy/day, 5 days/week, 60 Gy
- Intravenous infusion of CBDCA is 30 mg/m^2 once daily **within 1 h before the start of RT**

- AnyGrade3-4 ≤ 10%
- Grade 4 AE is 0%

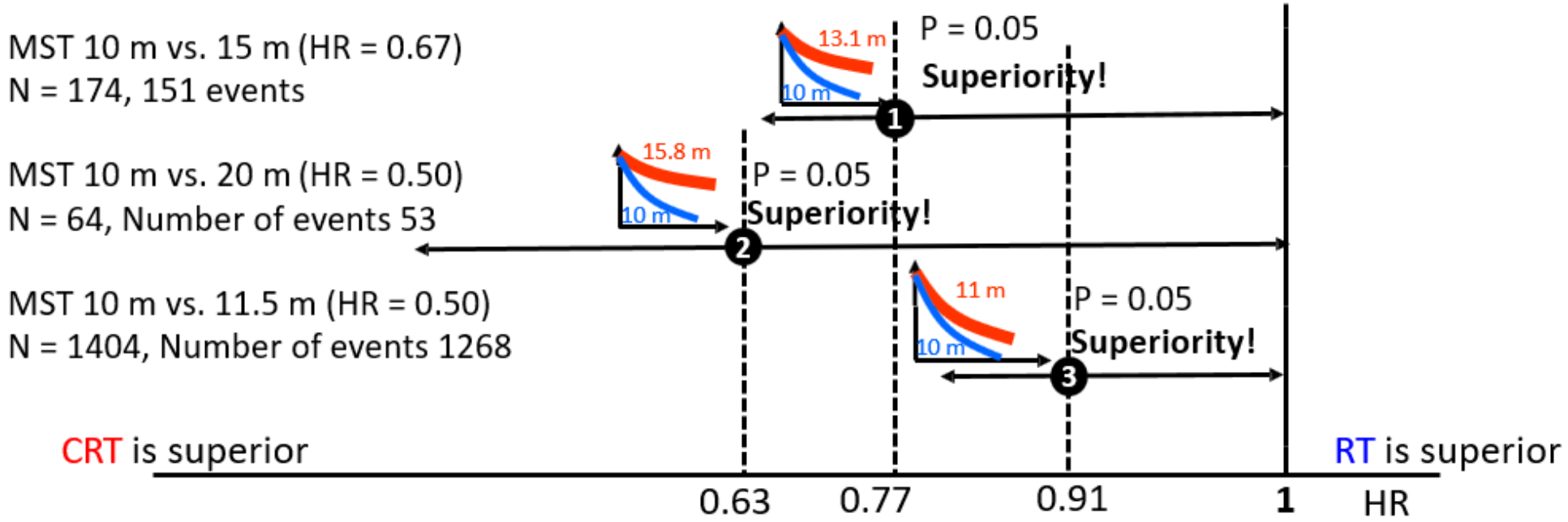
- G4 neutropenia = 25%,
  FN of G3 = 12.5%

- If there is an additional 5 m, it justifies the toxicity and effort involved

(= the least clinically meaningful difference)

Setting of Δ and Statistical Significance

Example of JCOG0301 (one-sided α = 5%, 80% power)



Recommended!

(1) Detect clinically meaningful differences

(2) The number of required analyses is reduced, but even if a clinically meaningful difference is obtained, it can no longer be considered statistically significant

(3) Even clinically meaningless differences are judged to be statistically significant

Background of the JCOG1104 Study Plan

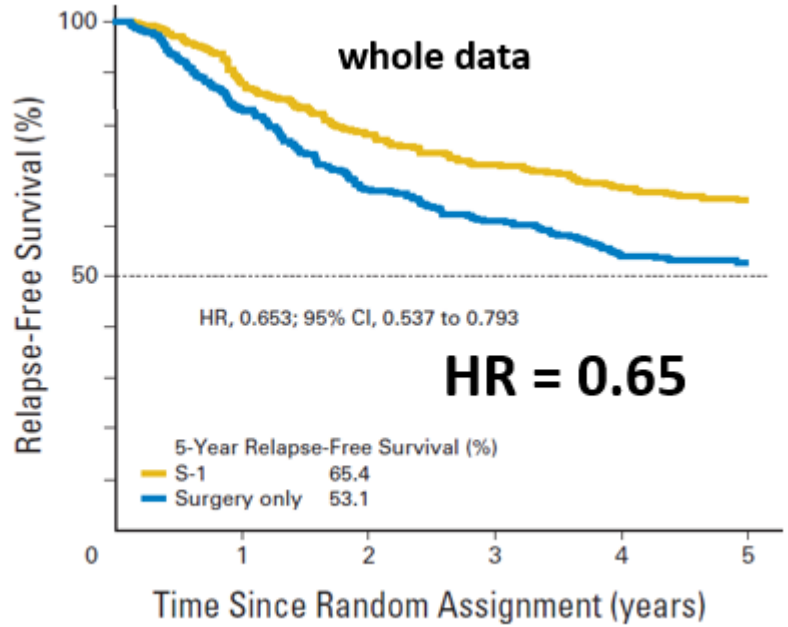
ACTS-GC

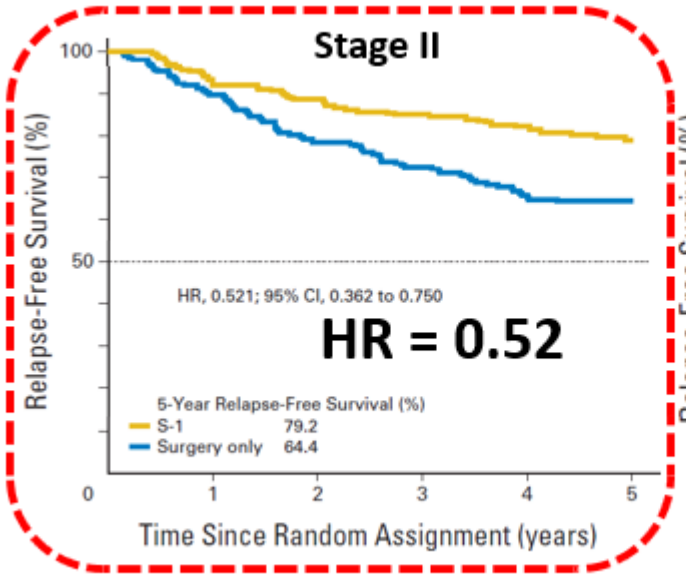
Post-gastrectomy stage II-IIIB gastric cancer
20–80 years old

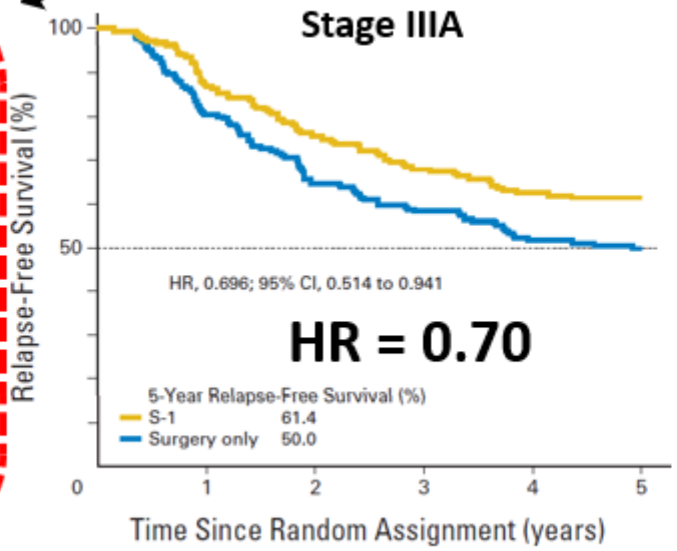
Random assignment

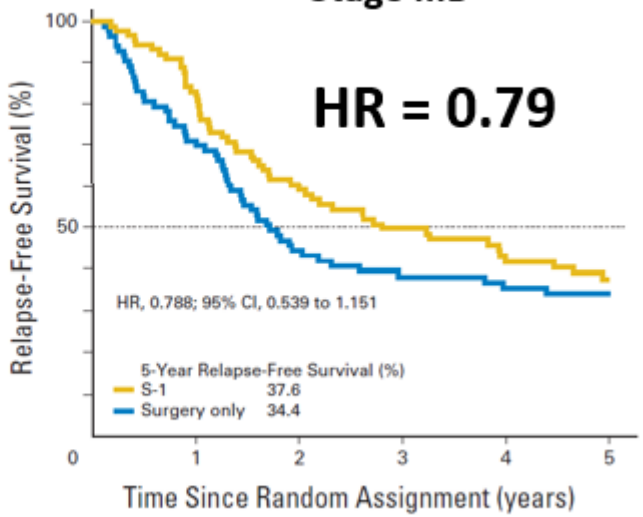
529 cases → **Follow-up (surgery only)** **(Standard treatment)**

530 cases → **S-1 1-year oral (S-1)** **(New treatment)**









Sasako M. et al. J Clin Oncol. 29(33): 4387-93, 2011

Hypothetical Example of the Planning Phase of JCOG1104 (Non-inferiority Study)

S-1 12 m dose group

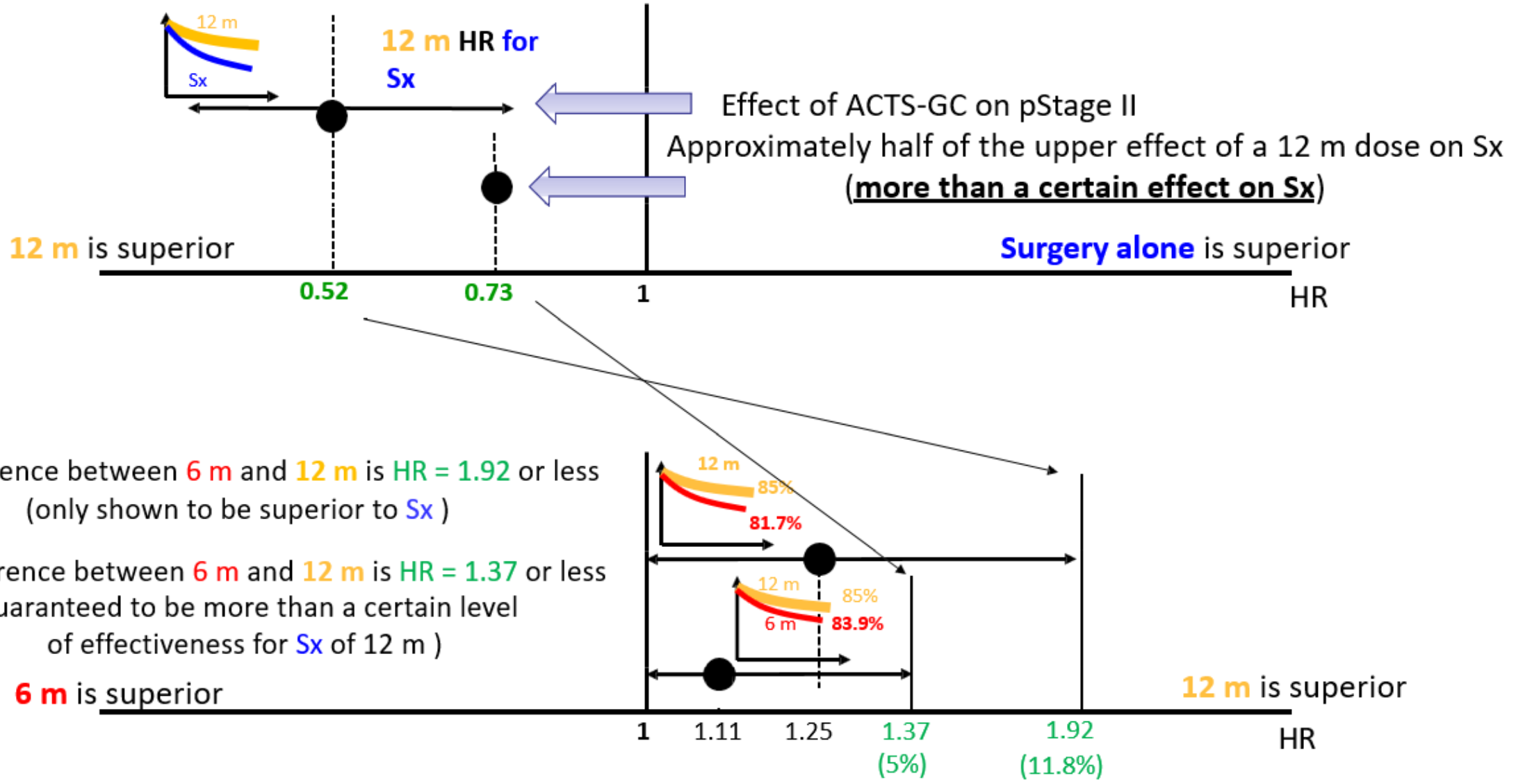
- Assumed 3yRFS = **85%**
- Gastrointestinal toxicity for 1 year
  - Grade 2 or higher anorexia, nausea, diarrhea, fatigue, dermatitis ≥ 20
- Cost per year

S-1 6 m dose group

- Assumed 3yRFS = **85%**
- AEs that occur after 6 m disappear
  - Reduces the burden on patients, such as hospital visits
- Half the cost

- If the difference in 3yRFS is ≤5% (non-inferiority margin of HR = 1.37), there is an advantage to administer for 6 m
- Is more than a certain level of effectiveness guaranteed for “Surgery alone”?

What Is More Than a Certain Level of Effectiveness for Surgery Alone?



Non-inferiority Trial Design with a Reasonable Chance of Success

VOLUME 25 · NUMBER 31 · NOVEMBER 1 2007

JOURNAL OF CLINICAL ONCOLOGY | SPECIAL ARTICLE

Randomized Clinical Trial Design for Assessing Noninferiority When Superiority Is Expected

Boris Freidlin, Edward L. Korn, Stephen L. George, and Robert Gray

From the Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD; Department of Biostatistics and Bioinformatics and the *Cancer and Leukemia Group B Statistical* Center, Duke University School of Medicine, Durham, NC; and Eastern Cooperative Oncology Group, Harvard School of Public Health, Boston, MA.

Submitted March 23, 2007; accepted July 30, 2007.

Supported in part by Grant No. U10CA033601 from the National Cancer Institute (S.L.G.).

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Address reprint requests to Boris Freidlin, PhD, Biometric Research Branch, 6130 Executive Blvd, EPN-8122, MSC 7434 National Cancer Institute, Bethesda, MD 20892; e-mail: freidlinb@ctep.nci.nih.gov.

0732-183X/07/2531-5019/$20.00

DOI: 10.1200/JCO.2007.11.8711

A B S T R A C T

The randomized clinical trial (RCT) is the gold standard for definitive evaluation of new therapies. RCTs designed to show that the therapeutic efficacy of a new therapy is not unacceptably inferior *to that of standard therapy are called noninferiority trials. Traditionally, noninferiority trials have* required very large sample sizes. Sometimes, a new treatment regimen with a favorable toxicity and/or tolerability profile is also expected to have some modest improvement in efficacy. In such specialized settings we describe a hybrid trial-design approach that requires a dramatically smaller sample size than that of a standard noninferiority design. This hybrid design can naturally incorporate a formal test of superiority as well as noninferiority.

J Clin Oncol 25:5019-5023. © 2007 by American Society of Clinical Oncology

INTRODUCTION

The majority of randomized clinical trials (RCTs) are superiority trials, designed to show that one treatment arm (experimental arm) is superior to another (standard arm) with respect to some clinically relevant outcome. The objectives of these trials are often contrasted with noninferiority trials designed to show that a new treatment arm is noninferior to the standard therapy with respect to some clinical outcome. The latter approach may be justified by a favorable toxicity profile, convenience, or the cost associated with the new regimen relative to the standard arm.

The goal of RCTs is to provide definitive evaluation of new regimens. When the new regimen is incremental in nature (eg, addition of a new drug to the standard backbone), a superiority trial is appropriate. A definitive evaluation must provide sufficiently convincing evidence to distinguish between the conclusion that the new regimen is superior to the standard versus the conclusion that new regimen does not provide clinically meaningful benefit. When the new regimen was developed to provide a better safety or tolerability profile while preserving most of the efficacy of the standard regimen (eg, by reducing a dose of a toxic component), a noninferiority trial is appropriate. A definitive evaluation must be able to provide compelling evidence to distinguish between the conclusion that the new regimen preserves the benefit of the standard regimen (noninferiority) versus the conclusion that the new regimen does not preserve a clinically important fraction of the standard arm benefit (eg, the dose-reduced regimen produces only one half of the complete remissions induced by the full-dose regimen).

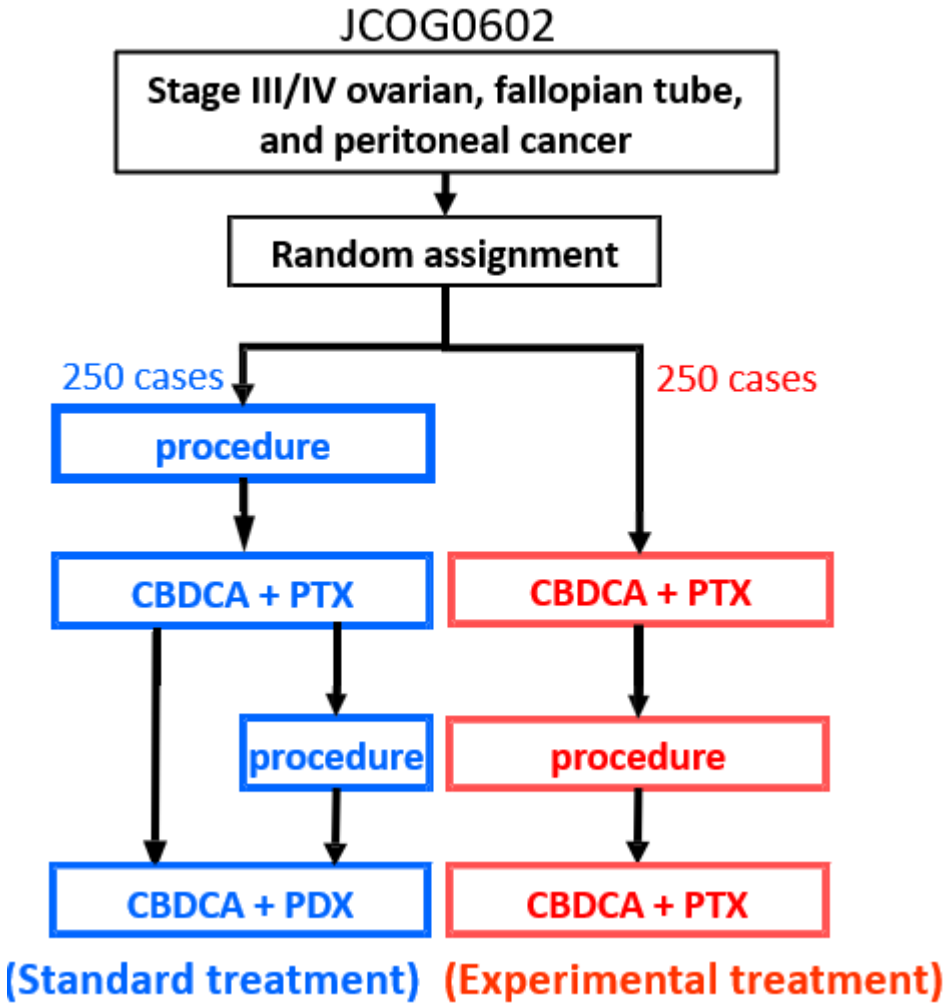
Whether the trial is a superiority trial or a noninferiority trial, the sample size needed to provide sufficiently compelling evidence to address its goal is determined mainly by the distance (on the treatment effect scale) between the negative and the positive conclusions (this difference is called the target treatment effect). In noninferiority trials, the target treatment effect generally is smaller than in superiority trials. Thus, larger sample sizes generally are required to achieve the necessary level of evidence. The increase in sample size can be quite dramatic. This often makes noninferiority designs infeasible because of the inability to complete the study in a reasonable time frame.

Sometimes the traditional classification of trials as either superiority trials or noninferiority trials is inadequate, and there are some features of each type of trial present. This occurs in the special case when a new treatment regimen with a favorable toxicity and/or tolerability profile is also expected to have some modest improvement in efficacy. For example, lenalidomide, an immunomodulatory analog of thalidomide, is believed to have a more potent antimyeloma effect than thalidomide, with lower systemic toxicity.[1] We argue that in this case the design should take into consideration the expected marginal benefit of the new regimen; this leads to a hybrid design with a dramatically smaller sample size than the standard noninferiority design. Similar designs have been mentioned in the statistical literature[2] and implemented in two recent studies.[3,4] Furthermore, we argue that a definitive evaluation

- Sample sizes for non-inferiority tests with a small inferiority margin and superiority tests with small differences are almost identical.
- Well-designed non-inferiority tests can be extensive.
- If the new treatment is expected to be less toxic and more effective, is it possible to design a feasible sample size?

Freidlin B, Korn EL, George SL, Gray R. J Clin Oncol. 2007; 25(31): 5019-23.

Examples of Non-inferiority Trial Designs with a Reasonable Chance of Success



Rationale for non-inferiority design

- If OS is the same, choose **experimental treatment**
  - Only one surgery is required
  - Reduction of surgery-related complications by post-NAC surgery

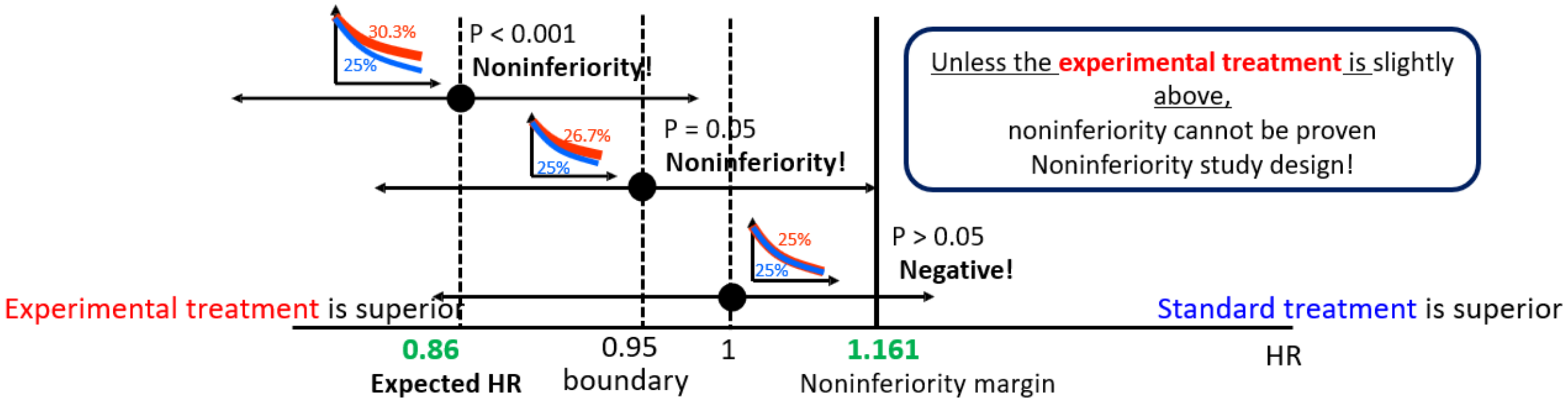
Rationale for a reasonable chance of success

- Early NAC
- Surgery of residual <1 cm tumor can be performed with improvement

Boundary of JCOG0602

Sample size design of JCOG0602

- One-sided α = 5%
- Power = 80
- 3yOS of standard vs. experimental treatment: 25% vs. 30.3% (HR = **0.86**)
- Non-inferiority margin for HR: **1.161** (equivalent to 5% in 3yOS)
- Required number of participants for analysis (expected number of events): 298 cases in total (278)
- When HR = 1 for standard vs. experimental, the number of required analyses (events): 1174 (1110)



Unless the **experimental treatment** is slightly above,
noninferiority cannot be proven
Noninferiority study design!

What Type of Test Design Would You Use?

For stage IIIB/IV, post-platinum therapy for non-squamous non-small cell lung cancer,
docetaxel vs. **nivolumab** (for simplicity, cost shall not be considered)

Some figures are hypothetical

Docetaxel

- Highly toxic
  - Any AE grade 3-4 50% or more
    - G3-4 Fatigue 10
    - G3-4 neutropenia 30%
    - FN 10%
  - Any AE grade 1-4 86%
- MST = approximately 8 m

Nivolumab

- Low toxicity
  - Any AE grade 3-4 7%
    - G3-4 Fatigue 1
    - G3-4 leukopenia 1%.
    - Others 0%
  - Any AE grade1-4 58%
- MST = about 12 m

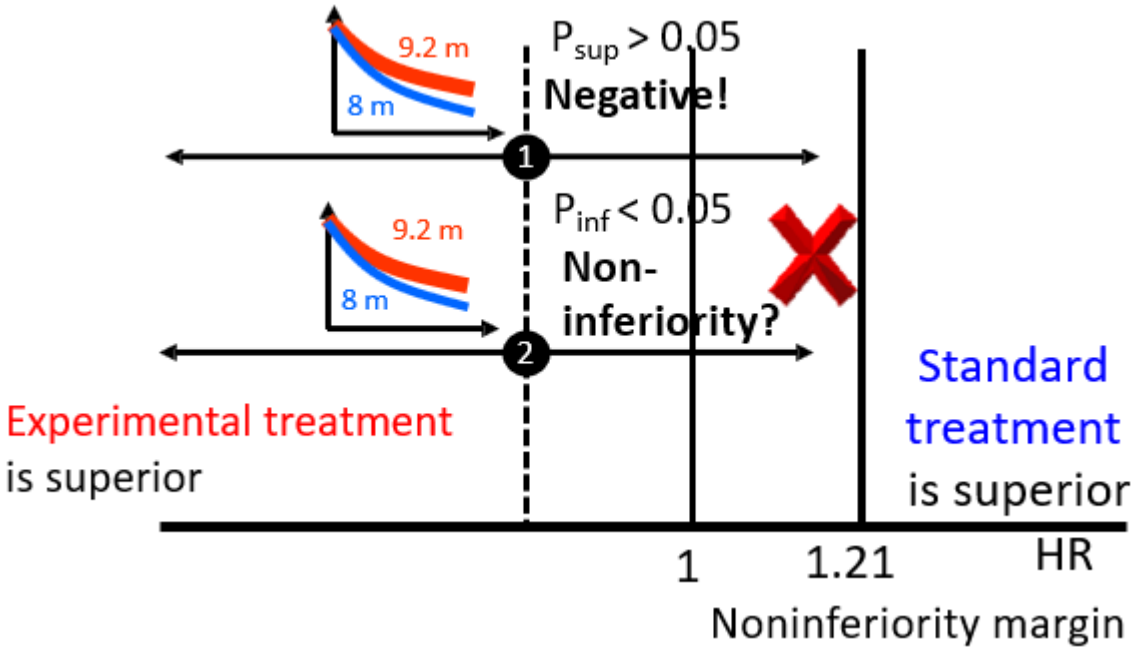
(1) Superiority test
(2) Non-inferiority test
(3) Non-inferiority test with a chance of winning

If superiority is achievable, it is hard to settle for non-inferiority...
In a non-inferiority trial, you cannot claim superiority, can you?

It Is Possible to Claim Superiority Even with a Non-inferiority Design Setup

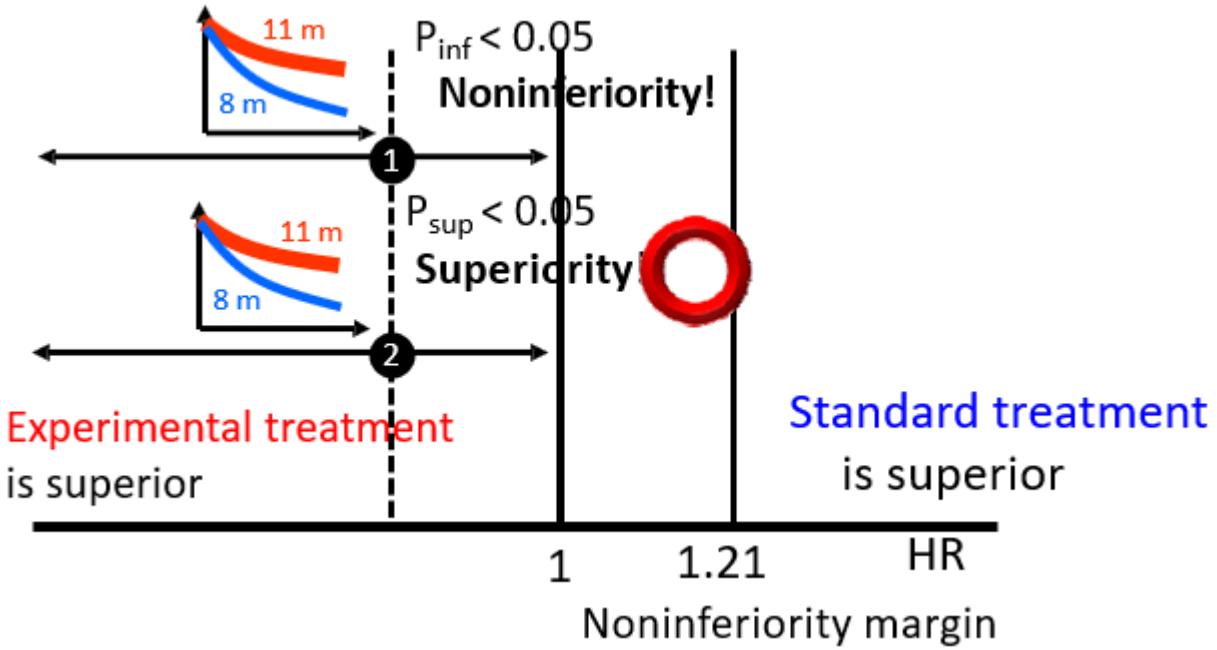
In the case of superiority design

When superiority is not displayed, the non-inferiority hypothesis cannot be retroactively analyzed.



In the case of non-inferiority design

It is acceptable to analyze the superiority hypothesis when non-inferiority has been demonstrated.



Sample Size Calculation

– In the Case of a Group Comparison of Survival

Required Information for Calculating Sample Size for Survival Analysis

- Significance level: **α**
- Power: **1−β (=power)**
- Size of treatment effect: **Δ**
  - **Δ** is a relative value **(HR)** rather than an absolute value (5yOS or MST)
  - Assuming an exponential curve, convert 5yOS and MST delta to HR

- Estimation of prognosis for the target population
- Enrollment period
- Follow-up period

Support the variability of sample size calculation formula

The Power of a Survival Analysis is Determined by the Number of Events

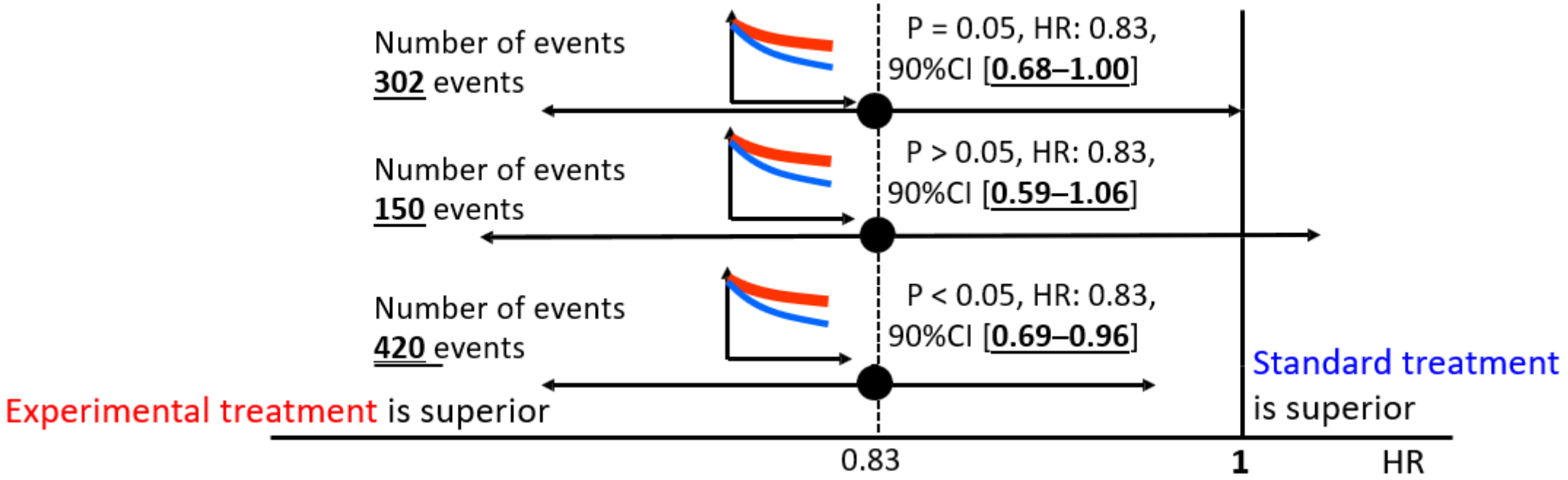
- 10000 cases enrolled and 100 events
- 1000 cases enrolled and 100 events
- 100 cases enrolled and 100 events

All have the same power

(If the point estimates of HR are the same, the confidence intervals are also the same)

"How many events occurred?" is more important than "How many cases have been enrolled?"

The number of events determines the width of the confidence interval



Sample Size Calculation Procedure for Survival Analysis

$$N = \frac{(Z_\alpha + Z_\beta)^2}{\ln(HR)^2} \cdot$$

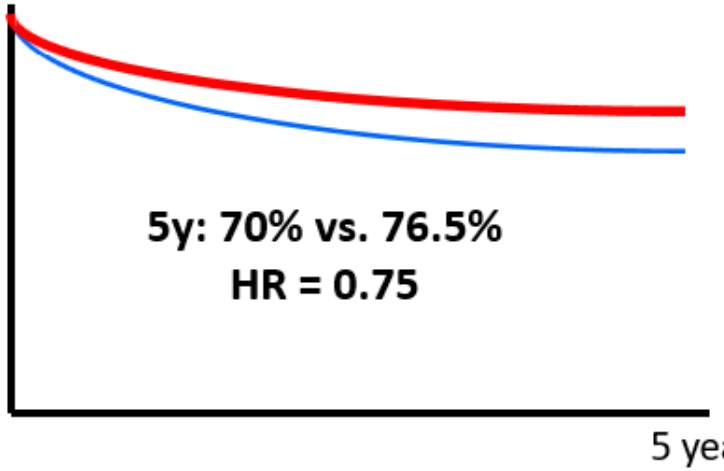
Values considering prognosis and the registration/follow-up period

(1) Calculate the number of events required by α, β, and Δ

(2) Calculate the N necessary to obtain the required number of events

Number of participants required to observe 380 events

For one-sided α = 2.5%, power 80%, HR = 0.75, about 380 events required

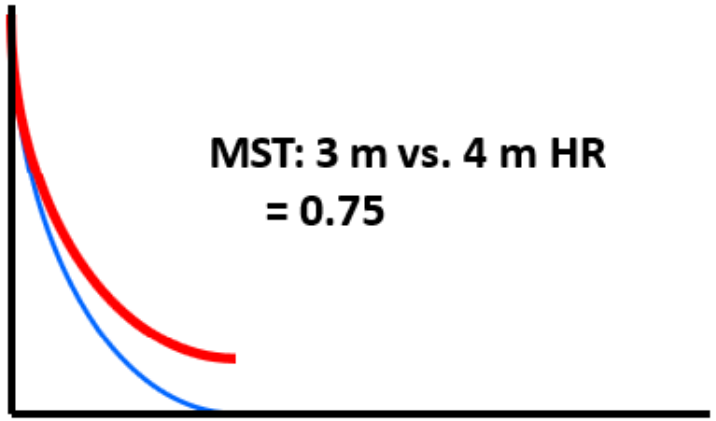


- Pooled groups 5y ≈ 73%

→ 27% of enrolled patients

If an event occurs, 380 events

→ Number of samples required = 380 ÷ 0.27 ≈ 1400 cases



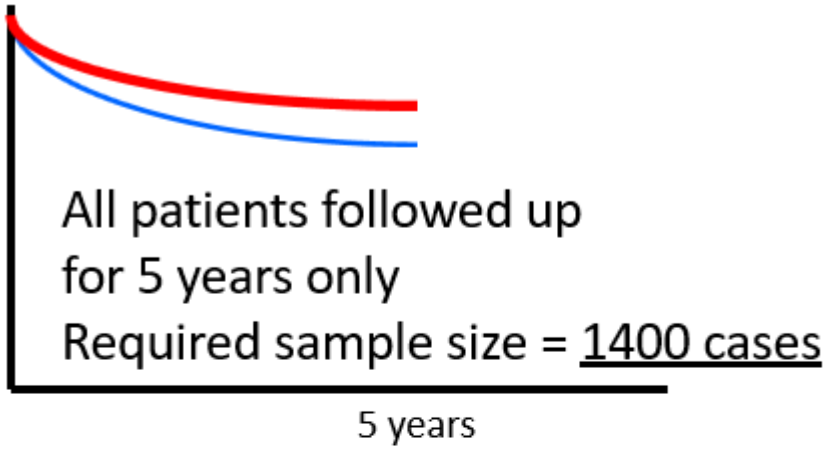
- Almost all enrolled patients have an event

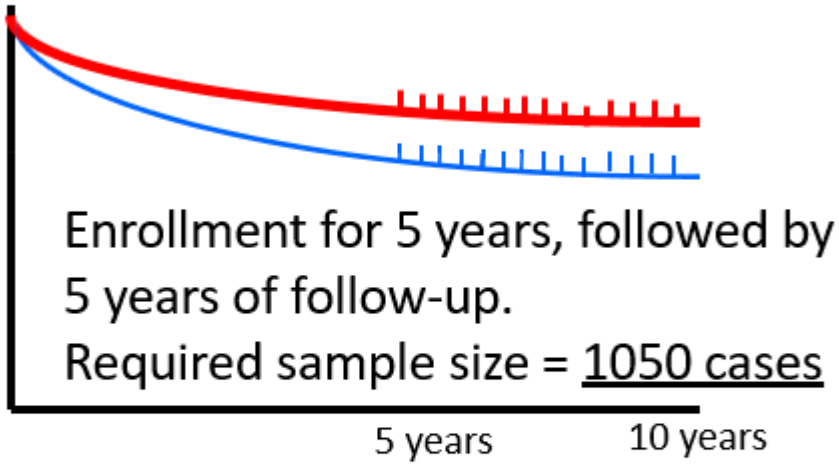
→ Number of required analyses ≈ Number of required events

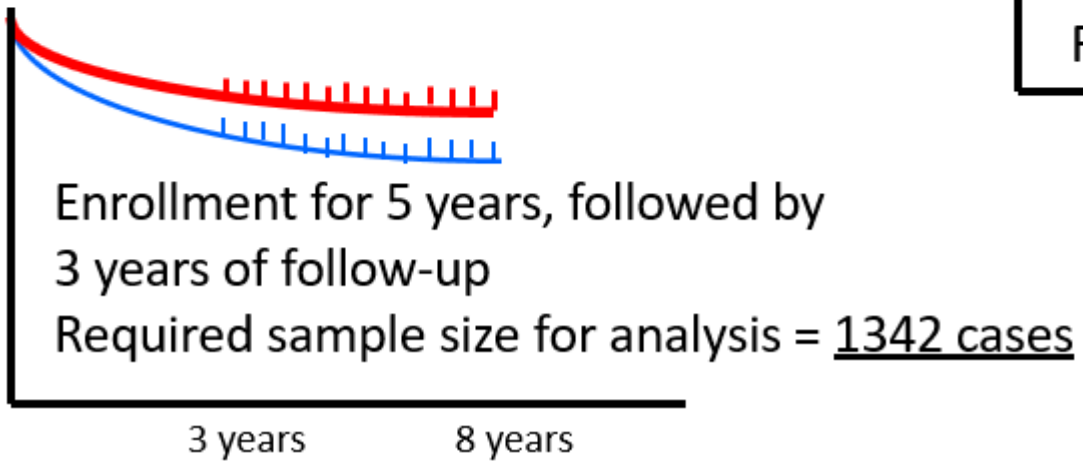
→ Number of cases required for analysis ≈ 380 cases

Consider Events That Occur during the Enrollment/Follow-up Period

For one-sided α = 2.5%, power 80%, HR = 0.75, approximately 380 events are required 5 years: 70% vs 76.5% (HR = 0.75)



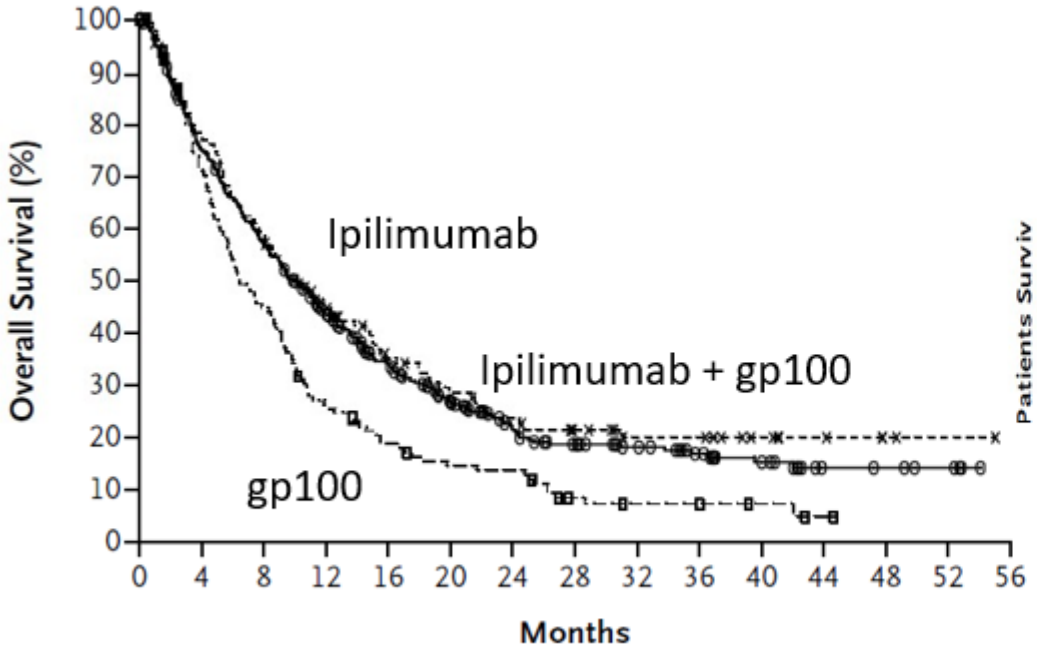




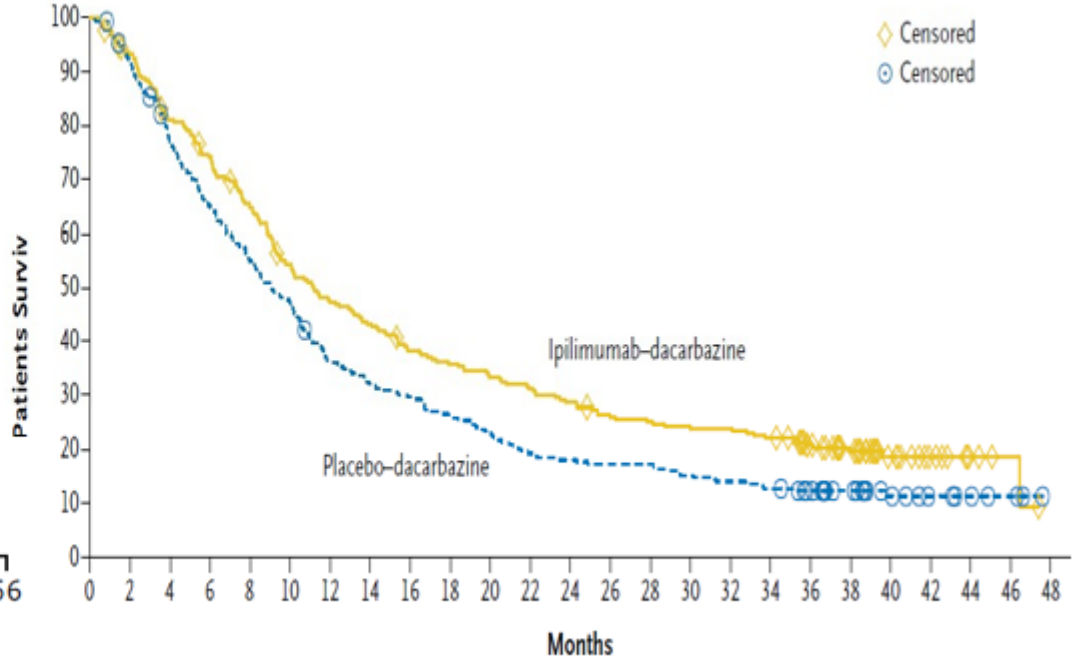
Assumptions of the Sample Size Calculation Method for Standard Survival Analysis

- Both groups assume an exponential curve
  - What if the exponential curve does not hold?
  - Participants with good prognosis plateau in the middle of the curve and are undetectable.
- The proportional hazard property is established
  - What if the proportional hazard property does not hold?
  - If the curves overlap up to a certain point and move apart after that point, that is insufficient power.

Application: Sample Size Calculation for Immunotherapy



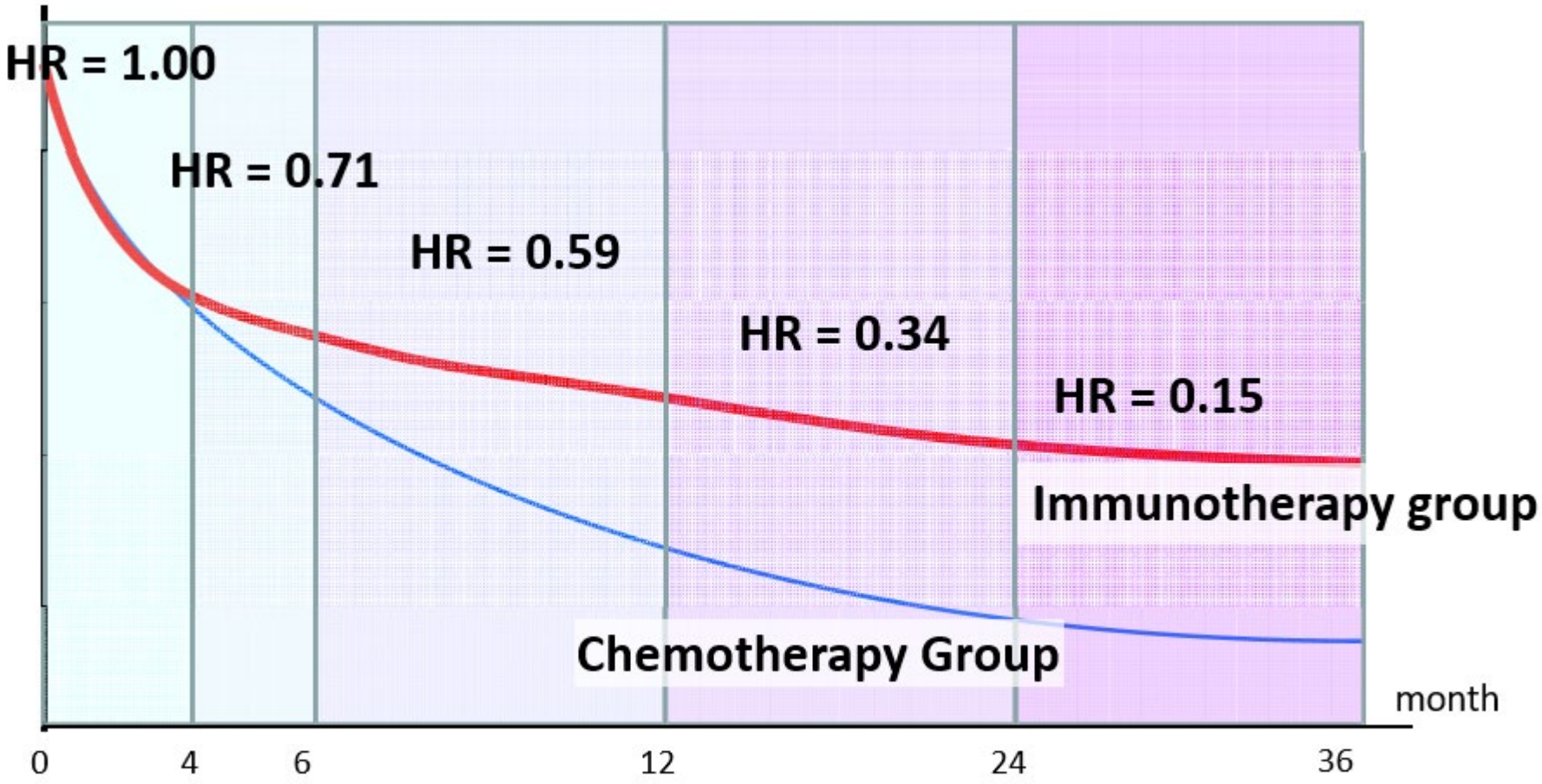
Hodi, F. Stephen, et al. N Engl J Med. 363.8 (2010): 711-723.



Robert, Caroline, et al. N Engl J Med. 364.26 (2011): 2517-2526.

- **The proportional hazards assumption is violated**
  - **No effect for the first 4 months**
  - **A certain number of patients are cured**
- **Power is insufficient at a sample size calculation assuming an exponential curve for both groups**

Assume Separate Exponential Curves for Each Section of the Curve



http://www.nejm.org/doi/full/10.1056/NEJMoa1507643#t=abstract
http://www.nejm.org/doi/suppl/10.1056/NEJMoa1507643/suppl_file/nejmoa1507643_protocol.pdf

Summary

- Parameters required for determining sample size are α, β, Δ, and variability
- Number of events, not N, is important in intergroup comparison of survival time
- Sample size calculation is a collaborative effort between statistician and clinician
  - In particular, the determination of Δ is a parameter for which the clinician is primarily responsible
  - Clarifying the clinical question is essential for determining Δ
    - Superiority or noninferiority
    - One-sided α or two-sided α
    - Clinically meaningful difference (Δ)
- There is no need to memorize detailed calculation methods
  - Calculations can be performed on software
  - Examine the validity of the design and parameters with a statistician