藥品臨床試驗相關統計學

Chin-Fu Hsiao
Institute of Population Health Sciences
National Health Research Institutes

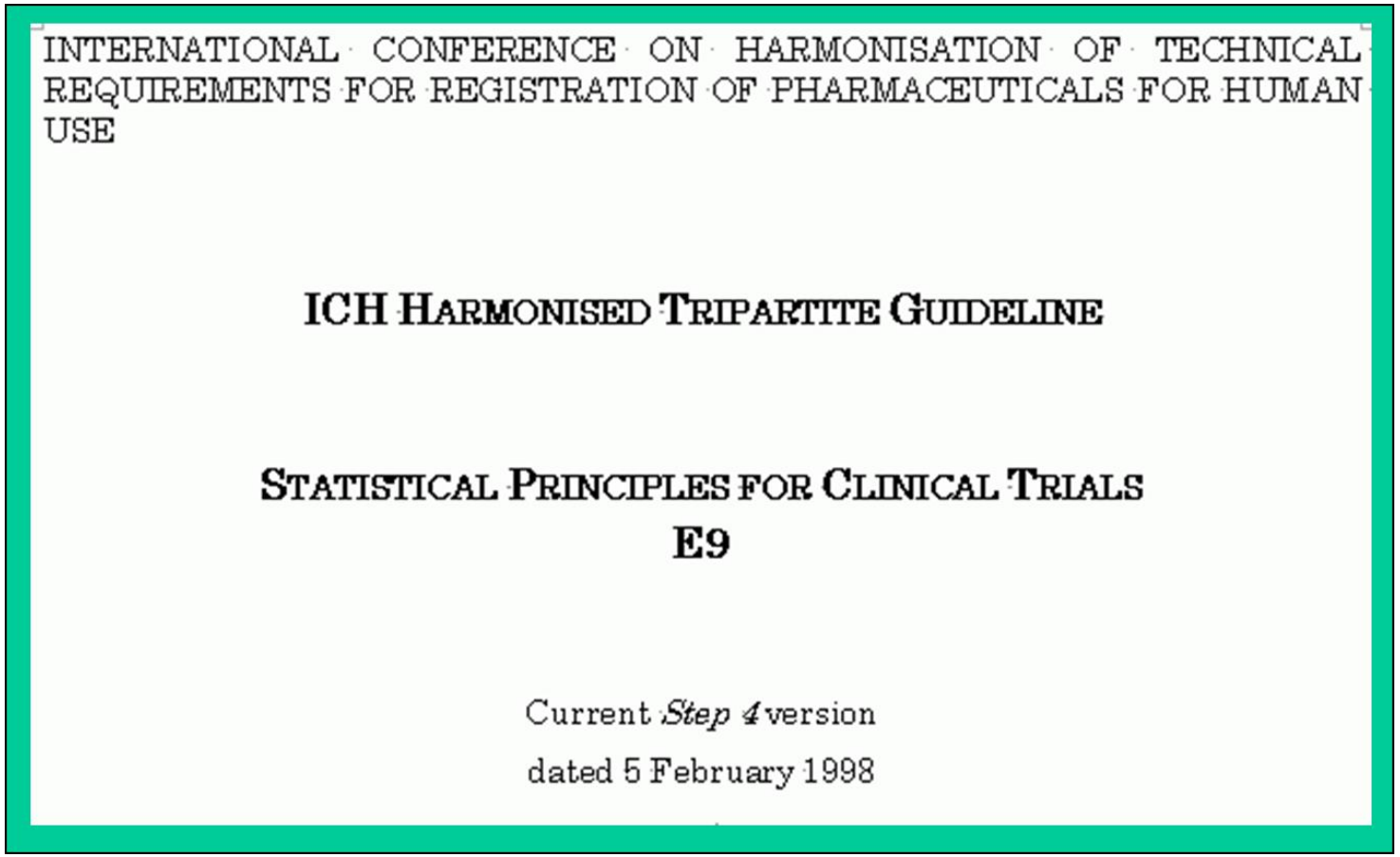
Outline

- Introduction
- Questions & Multiple Testing
- Randomization & Blinding
- Sample Size Calculation
- Study Population

Statistics in Clinical Trials

- Study objectives/hypotheses
- Study design
- Sample size calculation
- Randomization/blinding
- Study endpoint selection
- Protocol amendments
- Independent Data Monitoring Committee (IDMC)
- Statistical analysis plan
  - Statistical methods for data analysis

ICH E9 Statistical Principles



INTERNATIONAL CONFERENCE ON HARMONISATION OF TECHNICAL REQUIREMENTS FOR REGISTRATION OF PHARMACEUTICALS FOR HUMAN USE

ICH HARMONISED TRIPARTITE GUIDELINE

STATISTICAL PRINCIPLES FOR CLINICAL TRIALS

E9

Current *Step 4* version

dated 5 February 1998

Descriptive Statistics (Continuous Endpoint)

All statistics are estimates with sampling errors

- Continuous Data
  - Central tendency
    - Mean: arithmetic average of all observations y
    - Median: the middle observation
  - Dispersion
    - Standard deviation: s
    - Minimum: the smallest observation
    - Maximum: the largest observation
    - Range: maximum minus minimum

Endpoint PEFR	Placebo	Immunotherapy	Mean Difference
N	60	61	
Baseline	84.8±8.6	81.9±10.8	-2.9±9.8
Change	-1.4±11.1	2.5±11.1	3.8±11.1

Categorical Data

- Proportion of the patients with a certain attribute: the number of the patients with the attribute divided the total number of the patients in the group
- Presenting both of counts and proportions m, p

Characteristics	Prednisone	Placebo
N	48	45
Smoking status		
Current	21 (50.0%)	13 (31.0%)
Former	5 (11.9%)	6 (14.3%)
Never	16 (38.1%)	24 (54.8%)

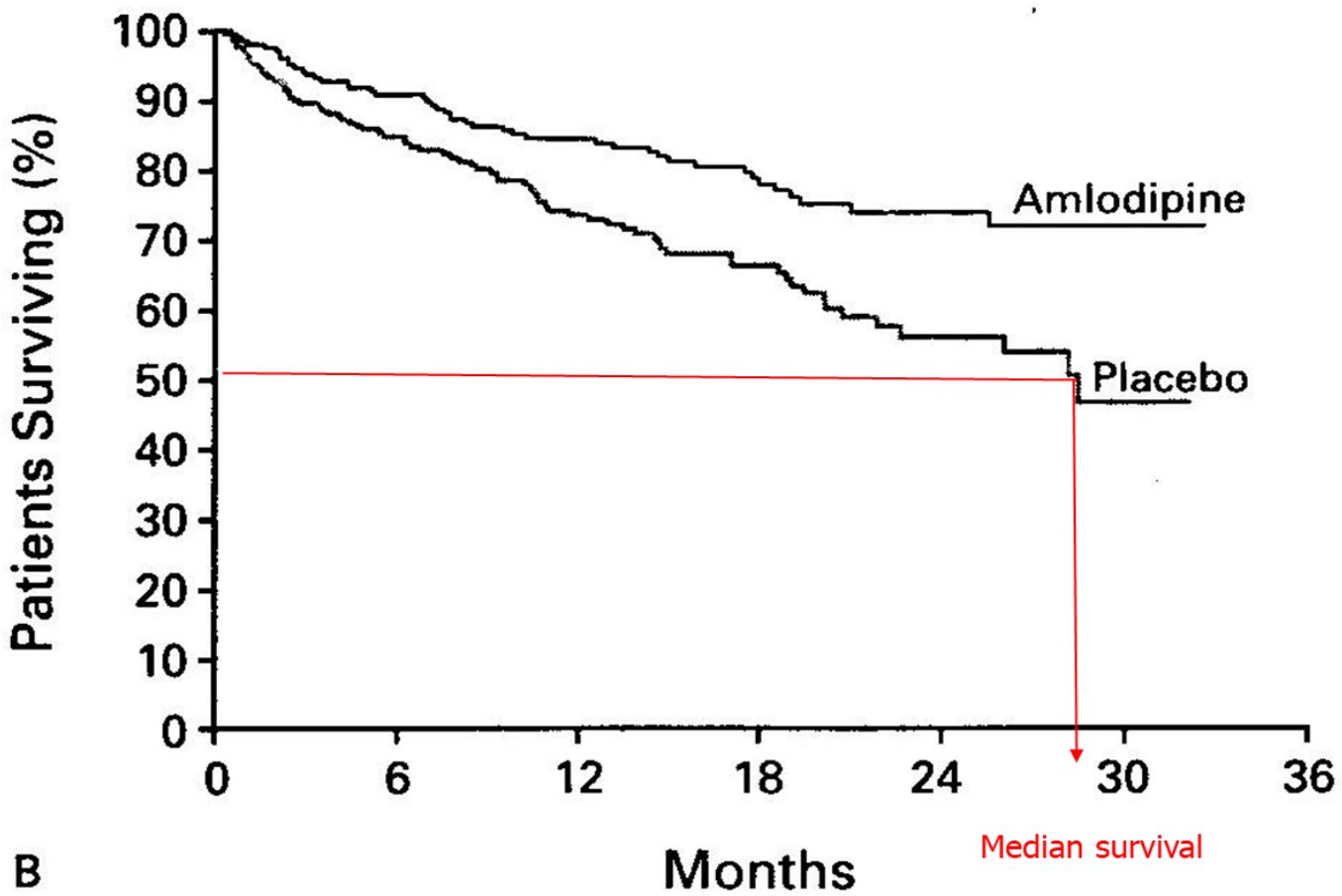
Endpoints for Cancer Trials (Time to Event)

- Overall survival (OS) (event: death)
- Progression free survival (PFS) (event: progression or death)

Censored Data

- Kaplan-Meier curve (actuarial probabilities)
  - The proportions of the patients with occurrence of a pre-defined event over a period of time
- Median survival
  - The time to the pre-defined event (e.g. death) occurring in 50% of the patients
- Hazard ratio
  - The hazard of the occurrence of a pre-defined event of the test group to the control group

PRAISE I – Non- Ischemic



Example: Crawford, et al
(NEJM 1989; 321: 419-24)

- A controlled trial of leuprolide with and without flutamide in prostatic carcinoma
- Randomized, double-blind, 2 parallel groups
- Primary endpoint: overall survival

Treatment	Median Survival
Leuprolide + flutamide	35.6 Months (95% CI)
Leuprolide + placebo	28.3 Months (95% CI)

Primary vs. Secondary Question

- Primary
  - most important, central question
  - ideally, only one (or at least <=2)
  - stated in advance
  - basis for design and sample size
- Secondary
  - related to primary
  - stated in advance
  - limited in number

Example

- Eastern Cooperative Oncology Group (ECOG - 1178)
  - tamoxifen vs. placebo
  - primary: tumor recurrence/relapse, disease-free survival
  - secondary: total mortality

Statistical Considerations

Null Hypothesis (H_0):

No difference in the response exists between treatment and control groups

Alternative Hypothesis (H_A):

A difference of a specified amount (δ) exists between treatment and control

Significance Level (α): Type I Error

The probability of rejecting H_0 given that H_0 is true

Power = (1 - β): (β = Type II Error)

The probability of not rejecting H_A given that H_A is true

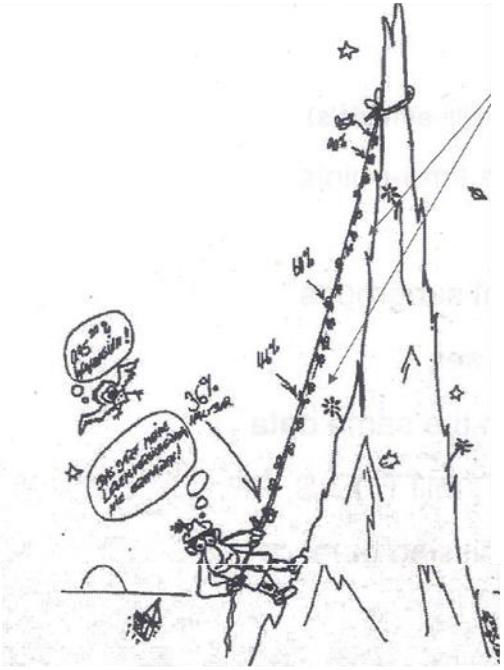
Controlling the Type I Error

- Ensuring that the chance of declaring a treatment efficacious when it in fact does not work is low (e.g., $\alpha \leq 0.05$)
- "Multiplicity" refers to having more than one opportunity to detect a difference between drugs (e.g., interim analyses, multiple endpoints of interest)

Multiple Testing

- Multiple treatment groups
- Multiple endpoints
- Multiple time points
- Multiple sub-group analyses
- Multiple interim analyses

Multiplicity



Each knot can cut with a probability of 5%.

Guess the probability of falling down the mountain!

k	1	2	3	4	5	50
	.05	.1	.14	.19	.23	.92

Multiple Treatment Groups

- Three groups, Trt1, Trt2, and Control
- Two comparisons: Trt1 vs. Control and Trt2 vs. Control
- Conduct 2 statistical tests
- The type I error will be inflated to 0.083 if p-value<0.05 for each test is used
- Need to adjust the significant level, 0.05/2=0.025 (Bonferroni adjustment)
- Once any one of three tests is significant , the trial can be claimed to be successful

Multiple Co-Primary Endpoints

- The clinical efficacy of a new treatment may be characterized by a set of possibly correlated endpoints
- Effects of Olanzapine Combined With Samidorphan on Weight Gain in Schizophrenia: A 24-Week Phase 3 Study
- Co-Primary Endpoints
  - Percent change from baseline at week 24 in body weight
  - The proportion of patients with >=10% weight gain from baseline at week 24
- Both endpoints should have p-value<0.05

Subgroup Questions

- Questions about effect of therapy in a sub-population of subjects entered into the trial
- Assess internal consistency of results
- Confirm previous hypothesis
- Generate new hypotheses
- Interpreted cautiously, qualitatively (次群體分析是無法視為結論)

MERIT-HF Study Design

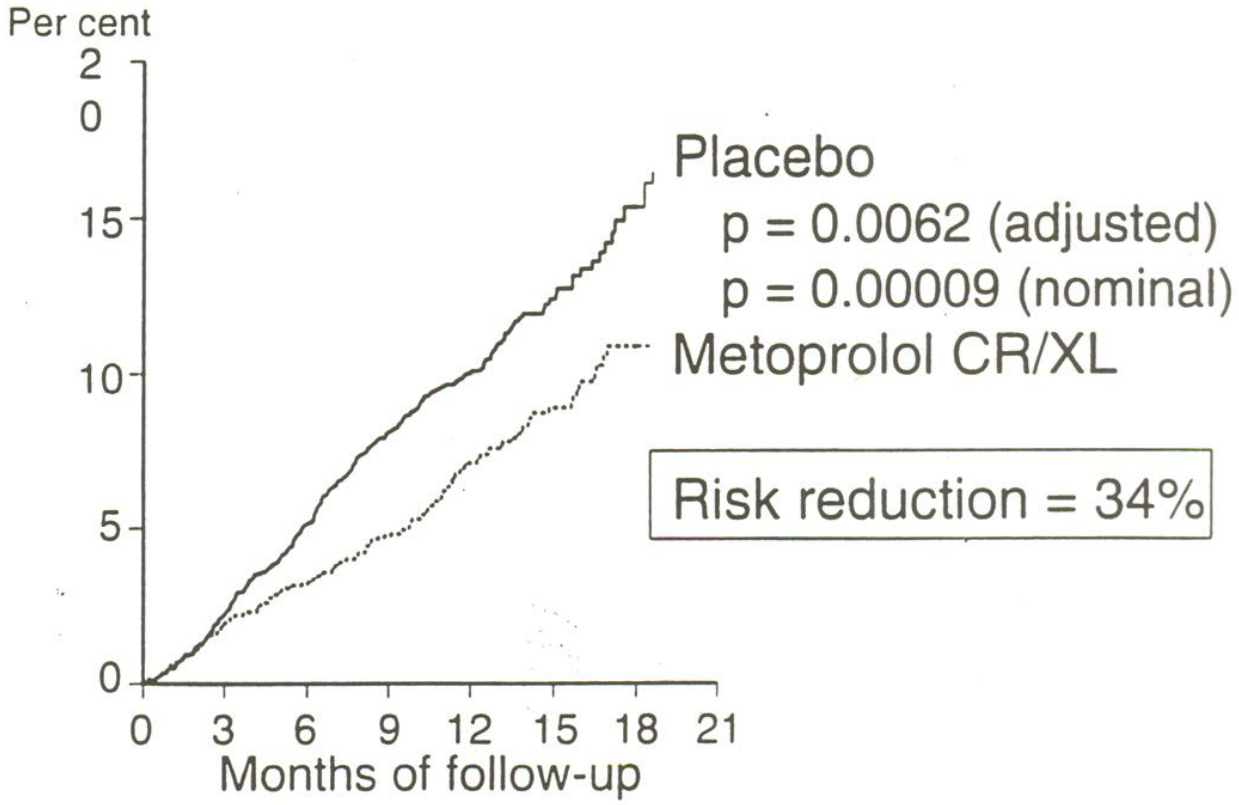
(Metoprolol Randomized Intervention Trial in congestive Heart Failure)

- Chronic heart failure patients
- Randomized placebo controlled
- Metoprolol (beta-blockers) vs. placebo
- Two-week placebo run in (compliance)
- Entered 3991 patients
- Terminated early
- Mean follow-up approximately one year

Primary Endpoints

- All-cause mortality
- All-cause mortality + All-cause hospitalization
- The type I error of 0.05 (two-sided) was allocated to these endpoints as 0.04 and 0.01, respectively

MERIT Total Mortality



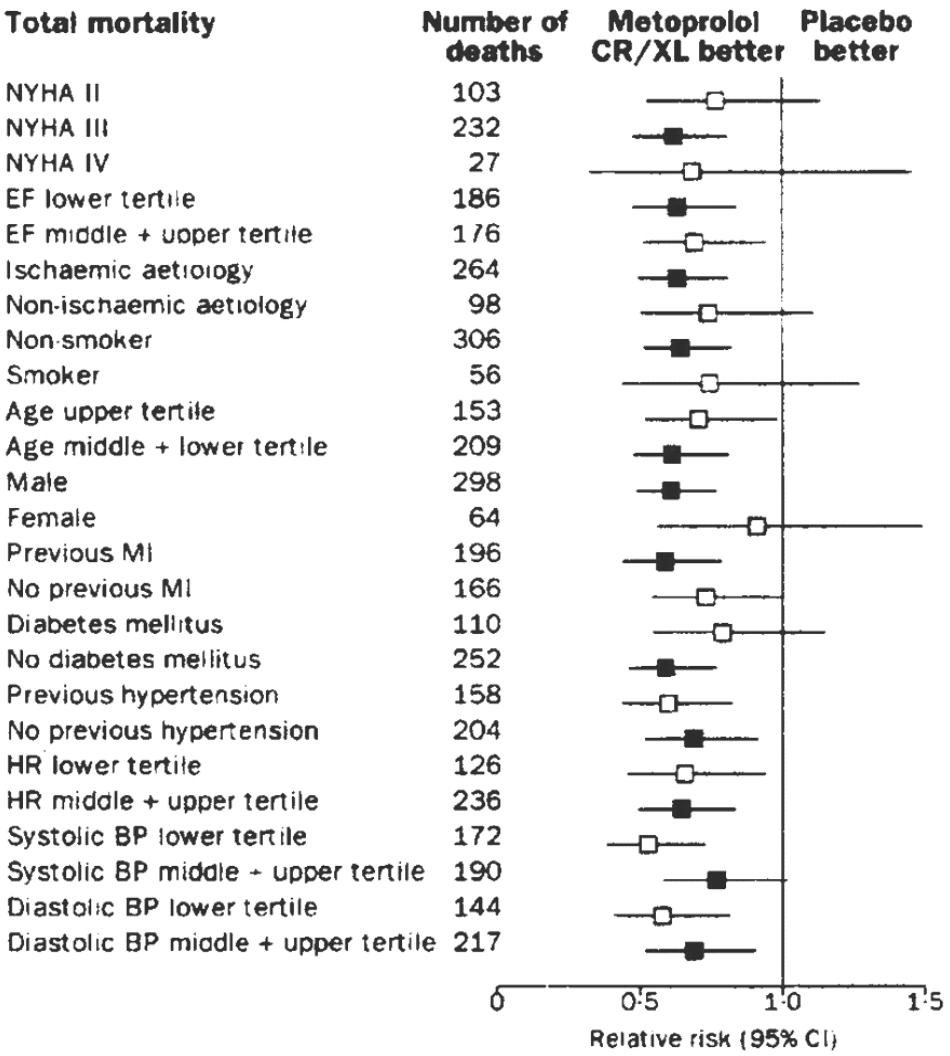
Data unblinded by ISaC

The MERIT-HF Study Group, ACC, March 1999

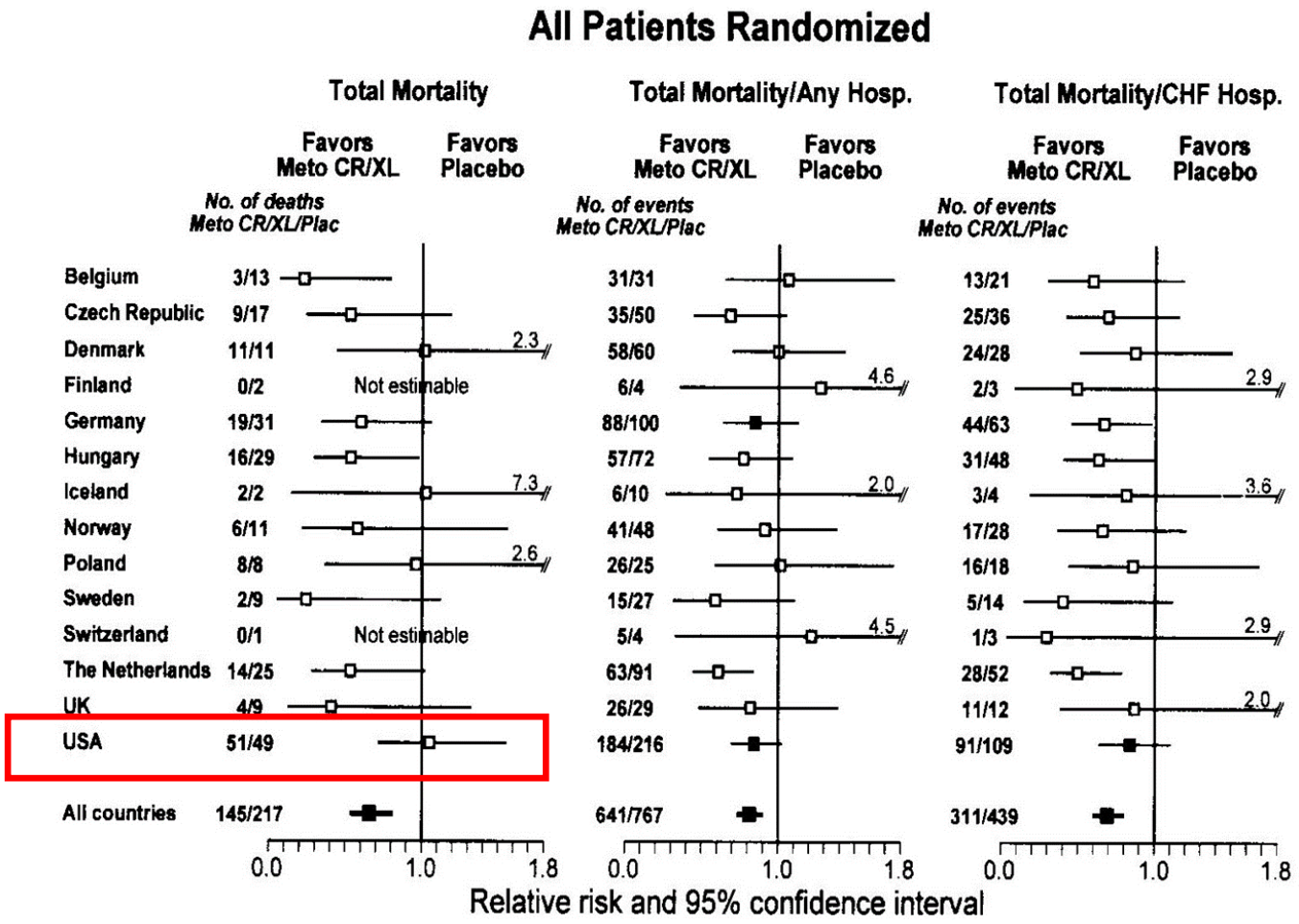
Results

- For all-cause mortality, the HR=0.66 ($p<0.0001$)
- For mortality-hospitalization composite, the HR=0.81 ($p=0.00012$)

MERIT


Total mortality
Number of deaths
Metoprolol CR/XL better
Placebo better
NYHA II 103
NYHA III 232
NYHA IV 27
EF lower tertile 186
EF middle + upper tertile 176
Ischaemic aetiology 264
Non-ischaemic aetiology 98
Non-smoker 306
Smoker 56
Age upper tertile 153
Age middle + lower tertile 209
Male 298
Female 64
Previous MI 196
No previous MI 166
Diabetes mellitus 110
No diabetes mellitus 252
Previous hypertension 158
No previous hypertension 204
HR lower tertile 126
HR middle + upper tertile 236
Systolic BP lower tertile 172
Systolic BP middle + upper tertile 190
Diastolic BP lower tertile 144
Diastolic BP middle + upper tertile 217
0
0·5
1·0
1·5
Relative risk (95% CI)

MERIT


All Patients Randomized
Total Mortality
Total Mortality/Any Hosp.
Total Mortality/CHF Hosp.
Favors Meto CR/XL
Favors Placebo
No. of deaths Meto CR/XL/Plac
No. of events Meto CR/XL/Plac
No. of events Meto CR/XL/Plac
Belgium 3/13 31/31 13/21
Czech Republic 9/17 35/50 25/36
Denmark 11/11 2.3 58/60 24/28
Finland 0/2 Not estimable 6/4 4.6 2/3 2.9
Germany 19/31 88/100 44/63
Hungary 16/29 57/72 31/48
Iceland 2/2 7.3 6/10 2.0 3/4 3.6
Norway 6/11 41/48 17/28
Poland 8/8 2.6 26/25 16/18
Sweden 2/9 15/27 5/14
Switzerland 0/1 Not estimable 5/4 4.5 1/3 2.9
The Netherlands 14/25 63/91 28/52
UK 4/9 26/29 11/12 2.0
USA 51/49 184/216 91/109
All countries 145/217 641/767 311/439
0.0 1.0 1.8
Relative risk and 95% confidence interval

FDA NDA Review

- The statistical reviewer conducted a subgroup analysis by region (USA vs. EU) and noted a strong suggestion of a treatment-by-region interaction (p=0.006) for all-cause mortality
- In EU, the HR=0.55 (p=0.0001), while in US, HR=1.05 (p=0.80)
- For the composite, effects were similar (0.84 vs. 0.81) across both regions

FDA NDA Review

- Extensive exploratory analyses were conducted
- Females and blacks were more heavily represented in the US. The HR in the US excluded non-whites was higher than 1.05
- The HRs in women vs. man was 0.92 vs. 0.61 for overall population. In US, 1.45 vs. 0.95

FDA NDA Review

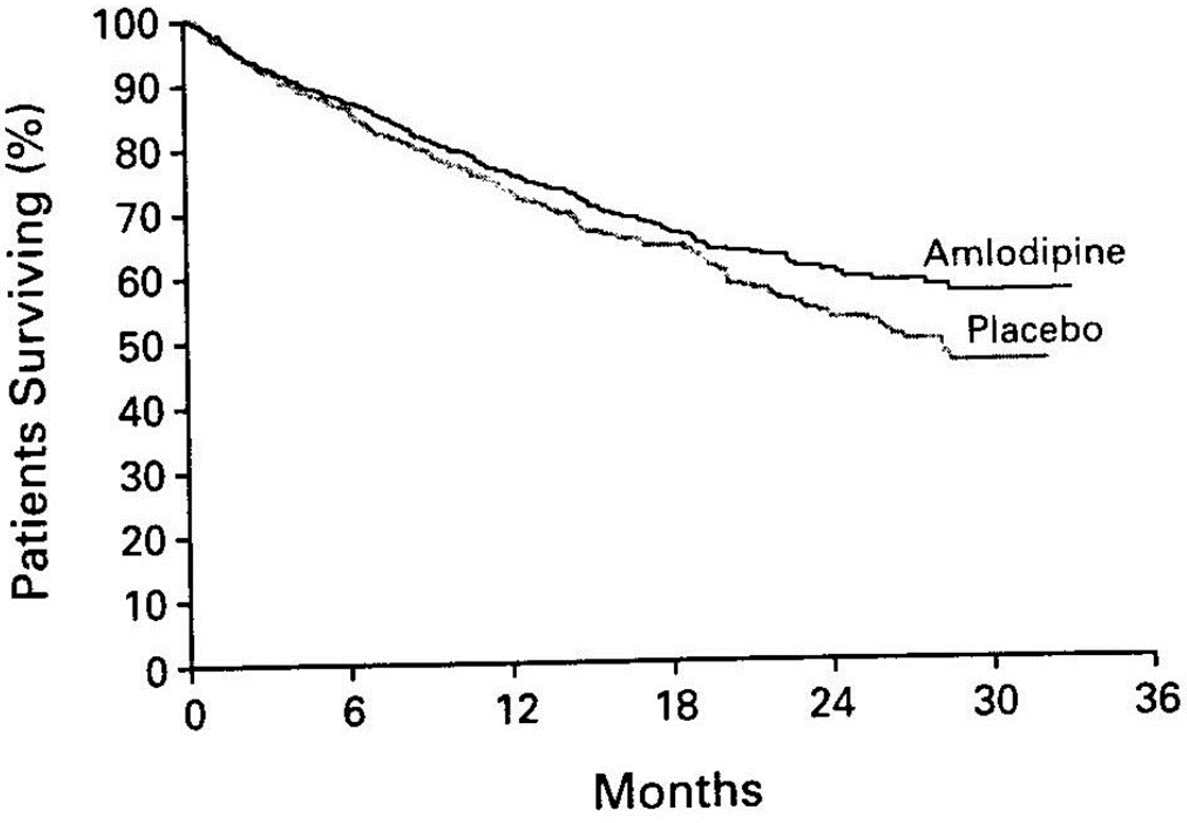
- No plausible covariates stood out to explain the US mortality result
- The absence of observed mortality benefit in the US was perhaps due to chance, but was more likely caused by inter-country differences in gender distribution of randomized patients, cause of deaths in CHF, or other demographic or medical practice differences
- The US patient population had not been intended to be used to evaluate the overall efficacy, which had been established by the pre-specificed primary analysis in the overall population
- In 2009, FDA-approved metoprolol for treatment of CHF based on the MERIT-HF study, but regional differences were described in the product label

PRAISE I

(Prospective Randomized Amlodipine Survival Evaluation) Ref: NEJM, 1996

- Amlodipine (calcium channel blocker) vs. placebo
- NYHA (New York Heart Association) class II-III
- Randomized double-blind
- Mortality/hospitalization outcomes
- Stratified by etiology (ischemic / non-ischemic)
- 1153 patients

PRAISE I

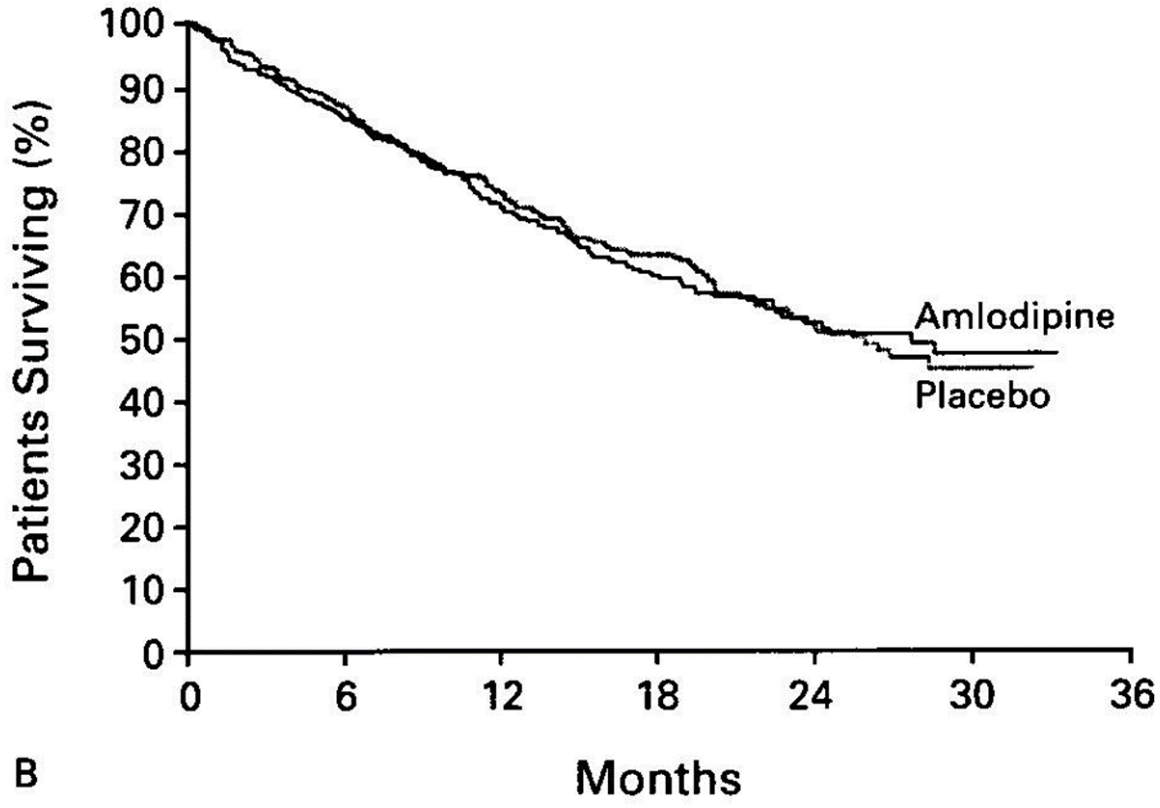

Patients Surviving (%)
100
90
80
70
60
50
40
30
20
10
0
0
6
12
18
24
30
36
Months
Amlodipine
Placebo

PRAISE I - Interaction

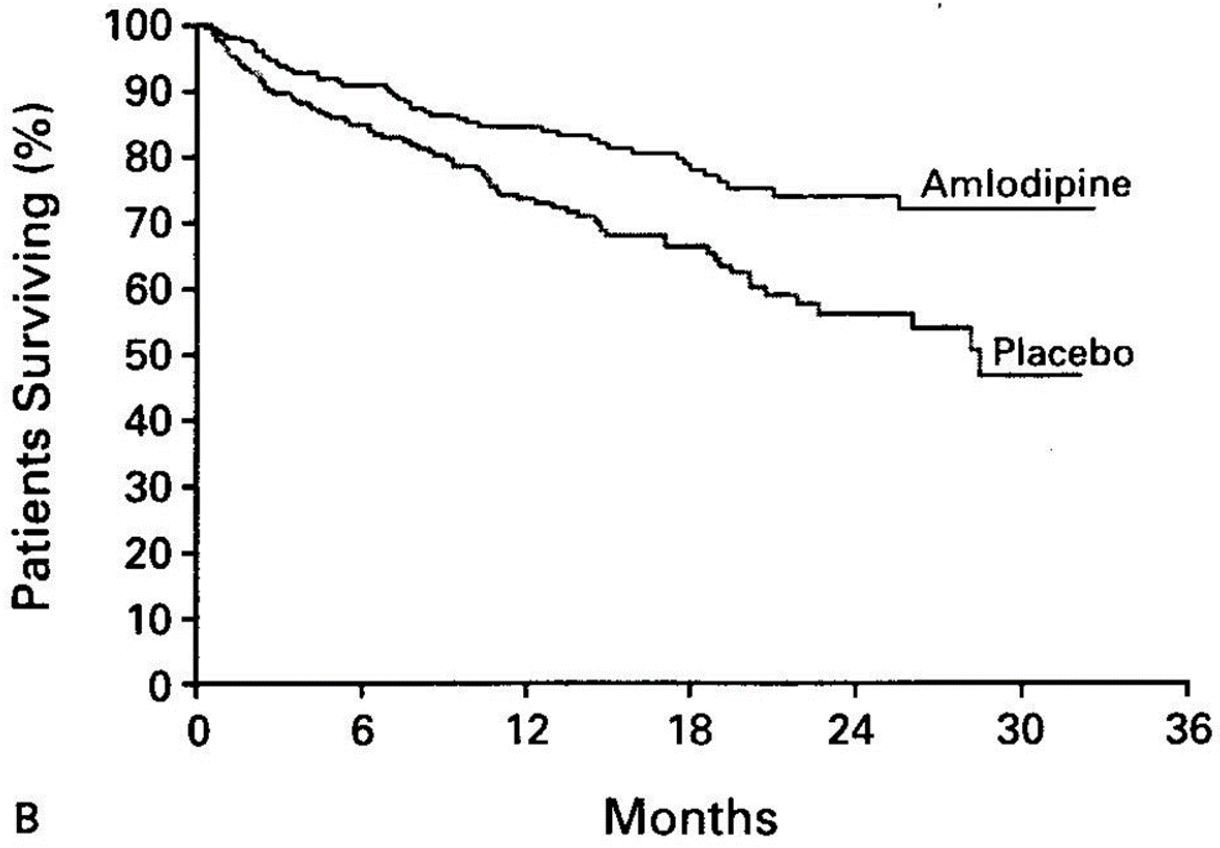
- Overall P = 0.07
- Etiology by Trt Interaction

  P = 0.004
- Ischemic P = NS
- Non-Ischemic P < 0.001

PRAISE I - Ischemic



PRAISE I – Non- Ischemic



PRAISE II

- Repeated non-ischemic strata
- Amlodipine vs. placebo
- Randomized double-blind
- 1653 patients
- Mortality outcome
- RR≅1.0

Example - Subgroup Concern

- Second International Study of Infarct Survival (ISIS 2)
  - 2 x 2 factorial design
  - (aspirin vs. placebo and streptokinase (抗凝血劑) vs. placebo)
  - vascular and total mortality in patients with an acute myocardial infarction (MI)
- Gemini or Libra astrological birth signs did somewhat worse on aspirin while all other signs and overall results impressive and highly significant benefit from aspirin

Randomized Control Studies

- Comparative studies with an intervention group and a control group
- Assign the subject to a group followed by the formal procedure of randomization
- The gold standard design

Advantages

- randomization removes the allocation bias
- randomization tends to produce comparable groups
- randomization guarantees the validity of statistical tests

Goal: Achieve Comparable Groups to Allow Unbiased Estimate of Treatment

Beta-Blocker Heart Attack Trial
Baseline Comparisons

	Propranolol (N=1,916)	Placebo (N=1,921)
Average Age (yrs)	55.2	55.5
Male (%)	83.8	85.2
White (%)	89.3	88.4
Systolic BP	112.3	111.7
Diastolic BP	72.6	72.3
Heart rate	76.2	75.7
Cholesterol	212.7	213.6
Current smoker (%)	57.3	56.8

Permuted Blocked Randomization

- Block size $2m = 4$

  2 Trts A,B $\} \Rightarrow {}_4C_2 = 6$ possible

- Write down all possible assignments
- For each block, randomly choose one of the six possible arrangements

  {AABB, ABAB, BAAB, BABA, BBAA, ABBA}

	ABAB	BABA	……
Pts	1 2 3 4	5 6 7 8	9 10 11 12

Random Errors

- 受試者被隨機分派至A組
- 但PI卻給了B組藥品
- 此受試者療效分析應納入A組分析或是B組分析呢？(efficacy)
- 此受試者安全性分析應納入A組分析或是B組分析呢？(safety)

Patient Closeout

- ICH E9 Glossary
  - "Intention-to-treat principle - …It has the consequence that subjects allocated to a treatment group should be followed up, assessed, and analyzed as members of that group irrespective of their compliance with the planned course of treatment."

Blinding or Masking

- No Blind
  - All patients know treatment
- Single Blind
  - Patient does not know treatment
- Double Blind
  - Neither patient nor health care provider know treatment
- Triple Blind
  - Patient, physician and statistician/monitors do not know treatment
- Double blind recommended when possible

Unbiased Evaluation

Subject Bias (NIH Cold Study)

(Karlowski, 1975)

Duration of Cold (Days)

	Blinded Subjects	Unblinded Subjects
Placebo	6.3	8.6
Ascorbic Acid	6.5	4.8

Unbiased Evaluation

Investigator Bias - (Taste & Smell Study)
(Henkin et al, 1972 & 1976)

	Single Blind	Double Blind
Zinc	8/8*	5/8
Placebo	0/8	7/8

*Number of variables with significant improvement/Number of variables

Sample Size Determination

臨床試驗的樣本數愈多愈接近研究母群體，研究結果愈不會造成偏差；然而在試驗器材未確認其療效與安全前，冒然納入過多受試者，不符合倫理考量；但若是樣本數太少，試驗可能因為檢定力太低，浪費了寶貴醫療資源，卻無法得到任何有意義的結果，這也不符合倫理考量。

Primary Question

- Primary question maybe framed in the form of testing a hypothesis
  - difference between intervention and control (superiority)

Sample Size Issues

- Before computing sample size, the primary response variable used to judge the effectiveness of intervention must be identified.
  - Dichotomous response variables (success and failure)
  - Compare event rate P_T and P_C

Typical Design Assumptions

- Type I Error, α =0.05 (two-sided), 0.025 (one-sided)
- Power=0.80, 0.90

  Better be at least 0.80 for design
- δ =smallest difference wish to see

  e.g. $\delta = P_C - P_T = 0.4 - 0.3 = 0.1$

  25% reduction

Typical Design Assumptions

Two Sided

Significance Level		Power	
α	$Z_{\alpha/2}$	$1 - \beta$	Z_{β}
0.05	1.96	0.80	0.84
		0.90	1.282
		0.95	1.645

Binary Endpoint Sample Size Formula

$$N = \frac{\left[Z_{\alpha/2}\sqrt{2\bar{P}\left(1-\bar{P}\right)} + Z_{\beta}\sqrt{P_C\left(1-P_C\right)+P_T\left(1-P_T\right)} \right]^2}{\delta^2}$$

Per group

$$\bar{P} = \frac{P_C + P_T}{2}$$

Example

- H_0: $P_C = P_T$
- H_A: $P_C = 0.4$, $P_T = 0.3$
- Assume $\alpha = 0.05$, $1-\beta = 0.90$

  i.e. $Z_{\alpha/2} = 1.96$, $Z_{\beta} = 1.282$

  $\bar{P} = (0.3 + 0.4)/2 = 0.35$

$$N = \frac{\left[1.96\sqrt{2(.35)(.65)} + 1.282\sqrt{(.3)(.7)+(.4)(.6)}\right]^2}{(.4-.3)^2} = 476$$

per group

Analysis Sets

- Basic Intention-to-Treat Principle(Full Analysis Set, FAS)
  - Analyze what is randomized!
  - All subjects randomized, all events during follow-up
- Beware of “look alikes”
  - Modified ITT: Analyze subjects who get some intervention
  - Per Protocol: Analyze subjects who comply according to the protocol

Statistical Testing Procedures Categorical Data

- Within-group
  - McNemar test for two categories
  - Stuart-Maxwell-Bhapkar test for more than two categories
- Between-group
  - 2×2 Table
    - Fisher's exact test (small sample size)
    - Pearson's chi-square test
  - 2×2 tables for ordered categories
  - Mantel-Haenszel test
  - Logistic regression

Statistical Testing Procedures Continuous Data

- Within-group
  - Parametric methods: paired t-test
  - Nonparametric methods: Wilcoxon signed rank test
- Between-group
  - Parametric methods: unpaired t-test, analysis of variance
  - Nonparametric methods: Wilcoxon rank sum test, Kruskal-Wallis test

Statistical Testing Procedures Time to Event Data

- Between-group
  - Log-rank test: difference later in time (癌症)
  - Gehan's test: difference early in time (流感)
  - Cox's proportional hazard model

Thanks for Your Attention!