Multiplicity Considerations in Confirmatory Subgroup Analyses

Frank Bretz

European Statistical Meeting on Subgroup Analyses
Brussels, November 30, 2012

Subgroup analyses

- **Exploratory** subgroup analyses are often used to:
  - assess internal consistency of study results
  - rescue a failed trial by assessing the expected risk-benefit compared to the whole trial population in a post-hoc manner
- **Confirmatory** subgroup analyses
  - pre-specify one (or more subgroups) in the trial protocol (based on demographic, genomic or disease characteristics)
  - control Type I error rate for the pre-specified multiple hypothesis test problem and fulfill other standard requirements for confirmatory trials

Why the concern about multiplicity?

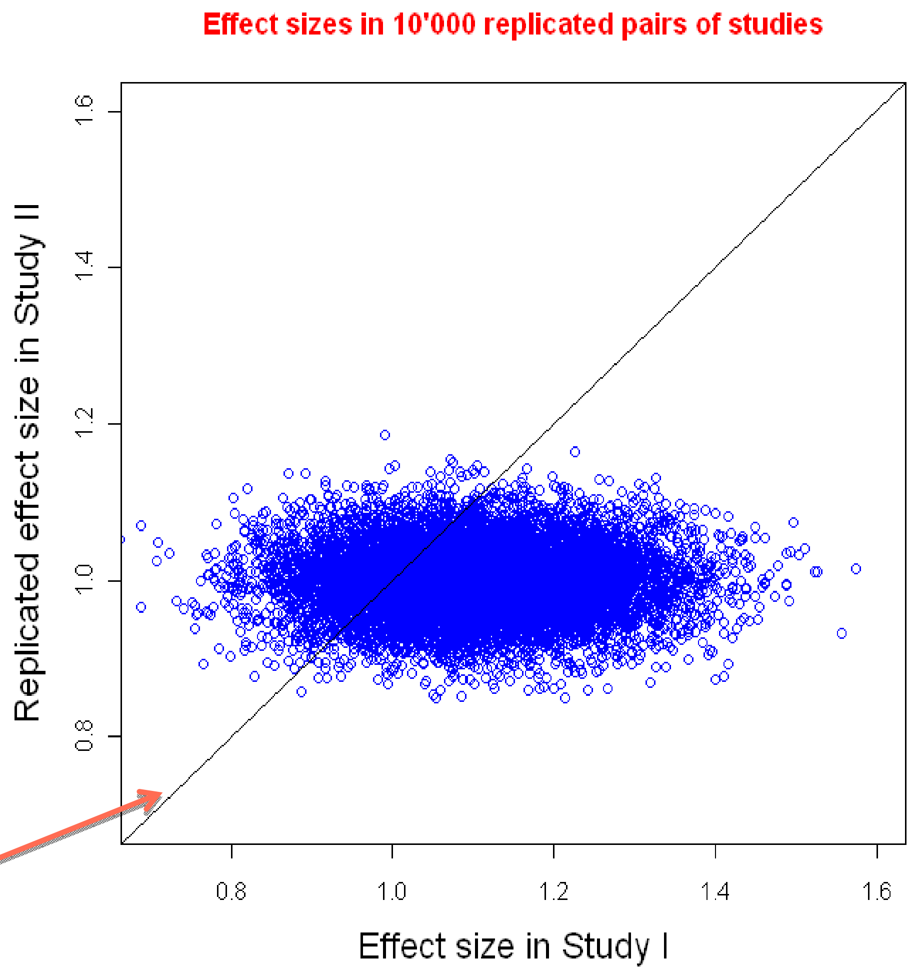
The Scientific Concern: **Reproducibility**

Assume 2 independent studies comparing treatment vs. placebo:

1. Study I with 4 disjoint subgroups

2. Study II only with the “best” subgroup from Study I

How do the observed effect sizes in the selected subgroup compare across both studies?



perfect reproducibility

(adapted from Westfall, Bretz and Tobias, 2012)

Confirmatory subgroup analyses

- Require tailored multiple test procedures for confirmatory inference
- Selected references:
  - Song and Chi (2007)
  - Wang et al. (2007)
  - Alosh and Huque (2009)
  - Spiessens and Debois (2010)
  - Zhao et al. (2010)
  - Bretz et al. (2011)
  - Dmitrienko and Tamhane (2011)
  - Alosh and Huque (2012)
  - Tang, Liu, Hsu (2012)
  - Tu, Hsu (2012)

Case Study 1

New treatment as add-on to **background therapy**

Primary objective:

To demonstrate efficacy of at least one of two regimen as add-on therapy despite stable **treatment with X**

Secondary objective:

To demonstrate efficacy of at least one of two regimen as add-on despite stable **treatment with X or other drugs of the same class (ALL)**

Design:

Randomization to be **stratified** by **X** or **not X**, enrollment such that 100p% of patients are on X.

Case Study 1

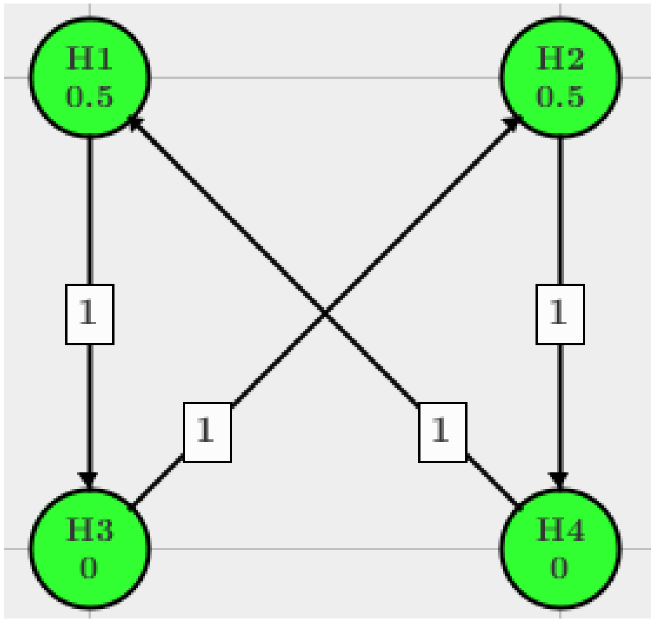
New treatment as add-on to **background therapy**

Results in **4 hypotheses**, where after discussions with clinical team:

- for each regimen, ALL is tested only if X is significant
- both regimens are considered equally important
- only if X and ALL significant for a same regimen, its significance level is propagated to competing regimen

X

ALL



Regimen 1 **Regimen 2**

Case Study 2

New treatment for **targeted therapy**

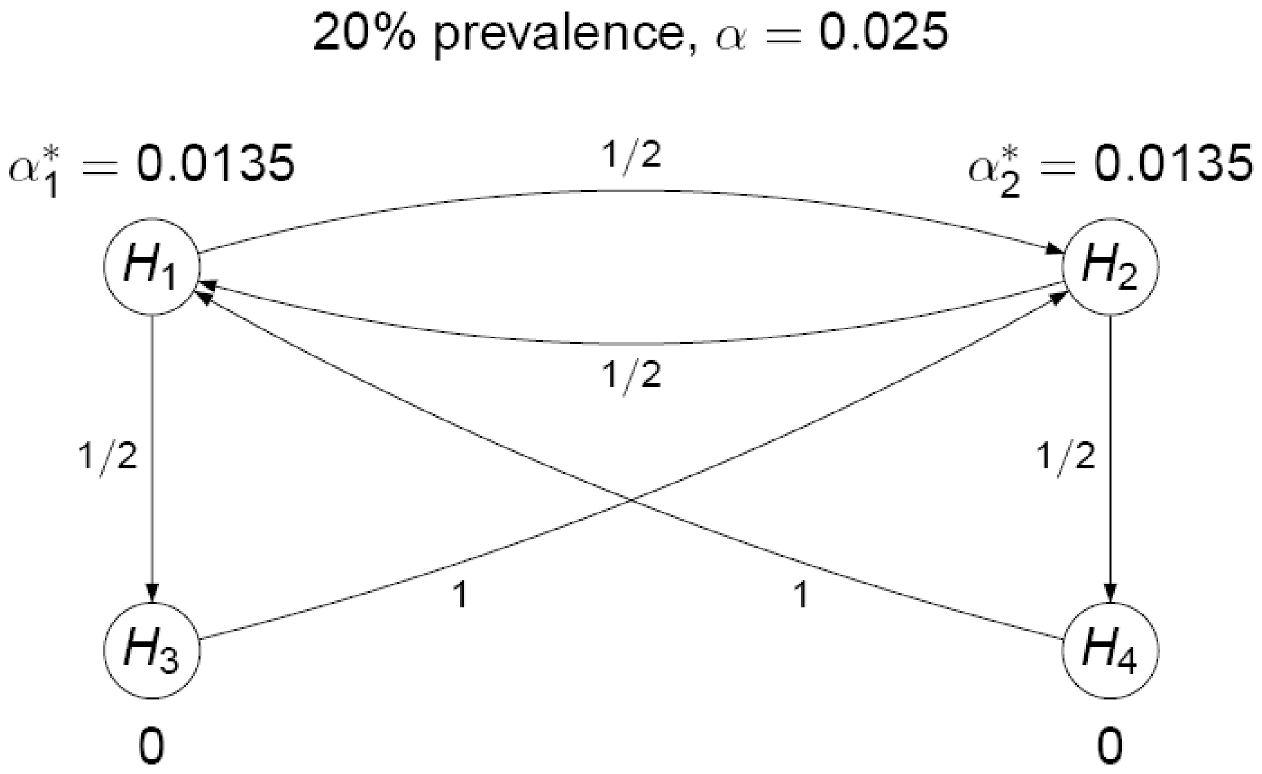
1. Targeted therapy of benefit in a **subpopulation S**
   - If beneficial in S, test for efficacy in **full population F**
   - Compare **two doses** (low / high) of new treatment against Standard-of-Care
2. Clinical considerations:
   - For each dose, F is tested only if S is significant
   - Both doses are considered equally important
   - As soon as S is significant for one dose, propagate some of the significance level to the other dose (safety considerations)
3. Sequentially rejective graphical procedure based on weighted Dunnett and t tests (Bretz et al., 2011; Millen and Dmitrienko, 2011)
   - Correlation between all 4 test statistics fully known and determined through sample sizes
   - In the balanced case and with $p = n_S / n_F$

$$\mathrm{corr}(T_i, T_j) = 0.5, \sqrt{p}, \text{ or } \sqrt{p}/2$$

Case Study 2

New treatment for **targeted therapy**

- Resulting graphical test procedure reflecting the clinical considerations
- Dunnett-adjusted significance levels 0.0135 (> 0.0125 = α/2 from Bonferroni)



Sub

Full

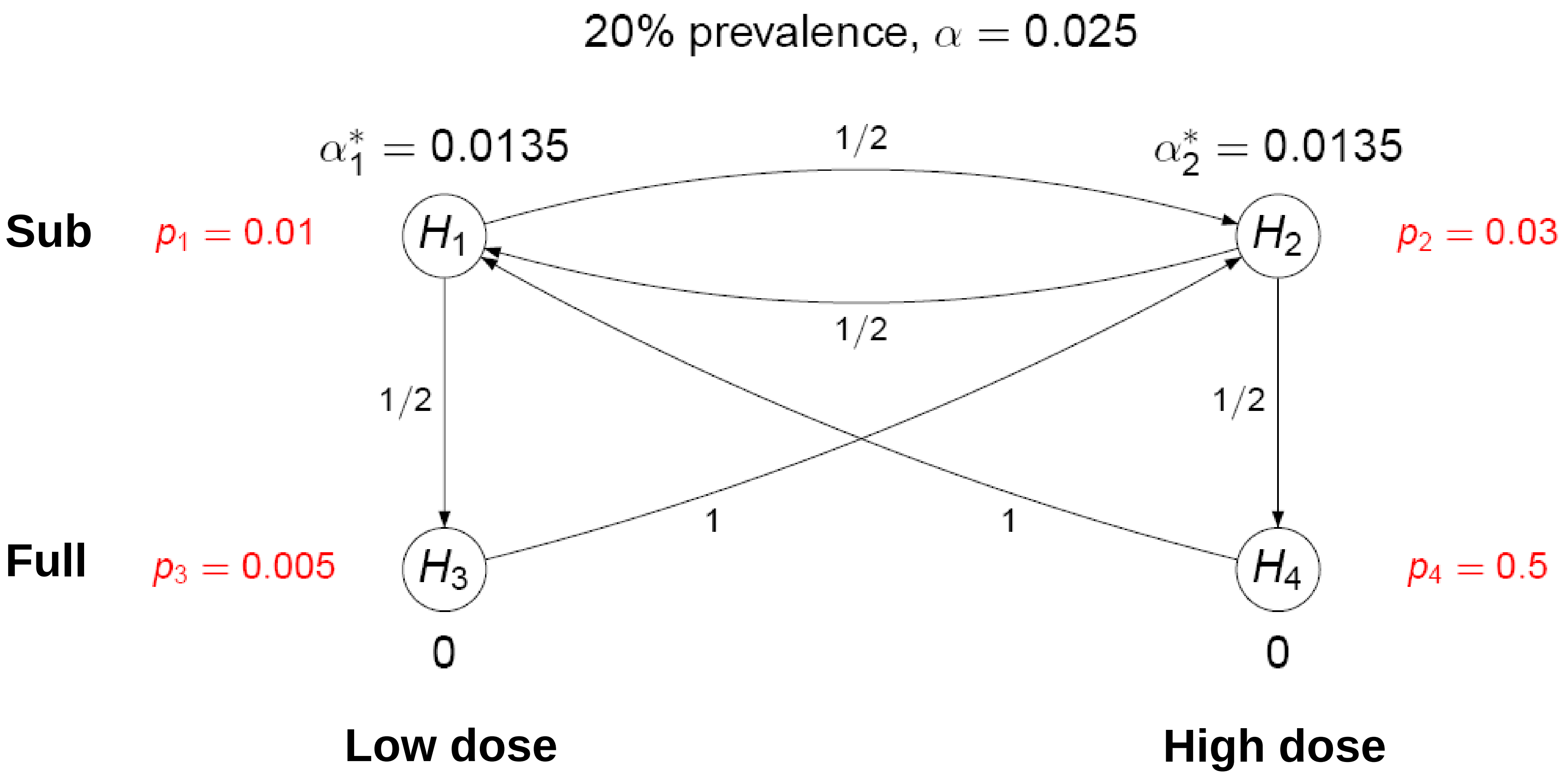
Low dose

High dose

Case Study 2

New treatment for **targeted therapy**

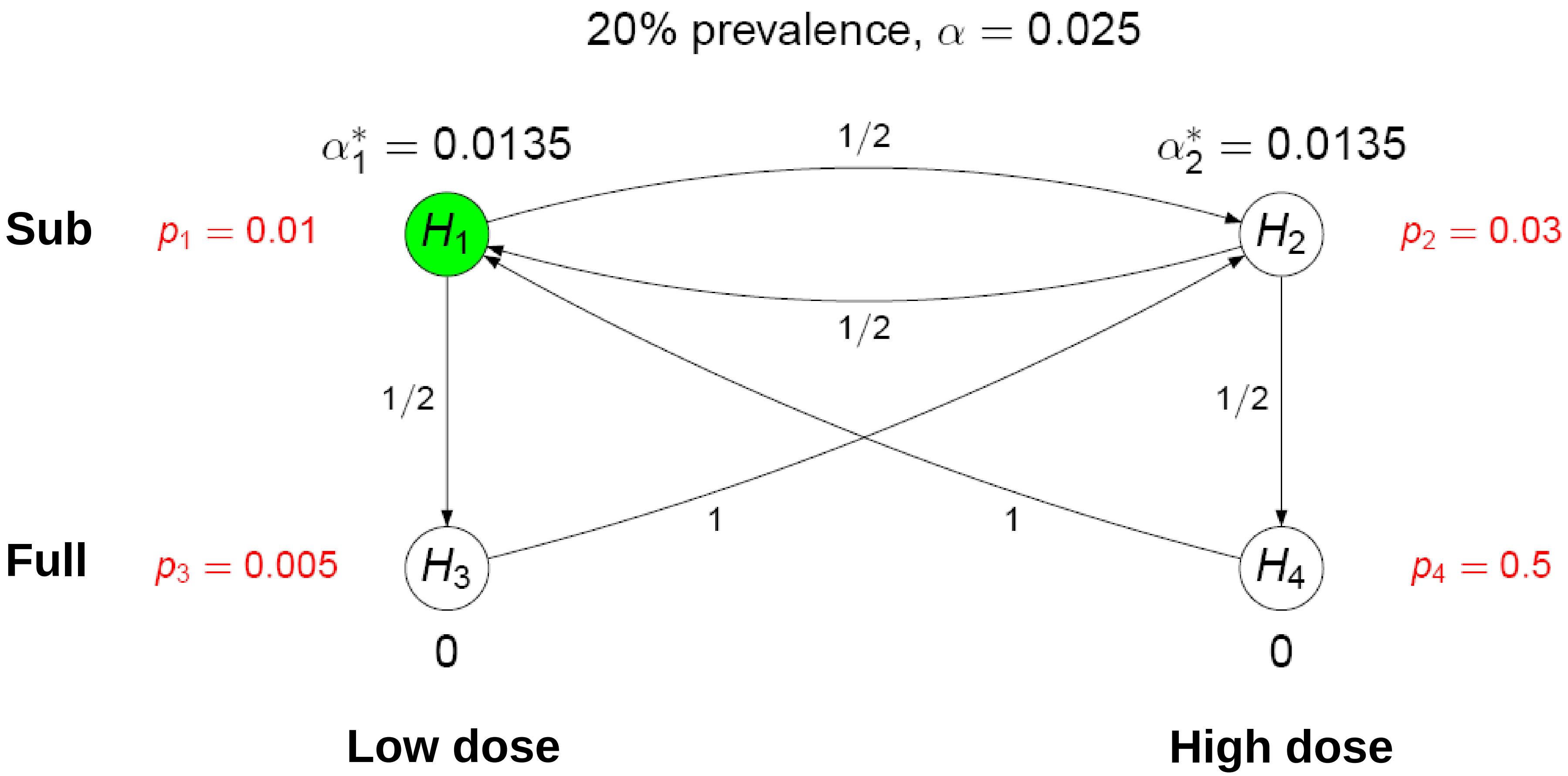
- Numerical example with 4 unadjusted p-values



Case Study 2

New treatment for **targeted therapy**

- Reject H_1 because $p_1 = 0.01 < 0.0135 = \alpha_1^*$

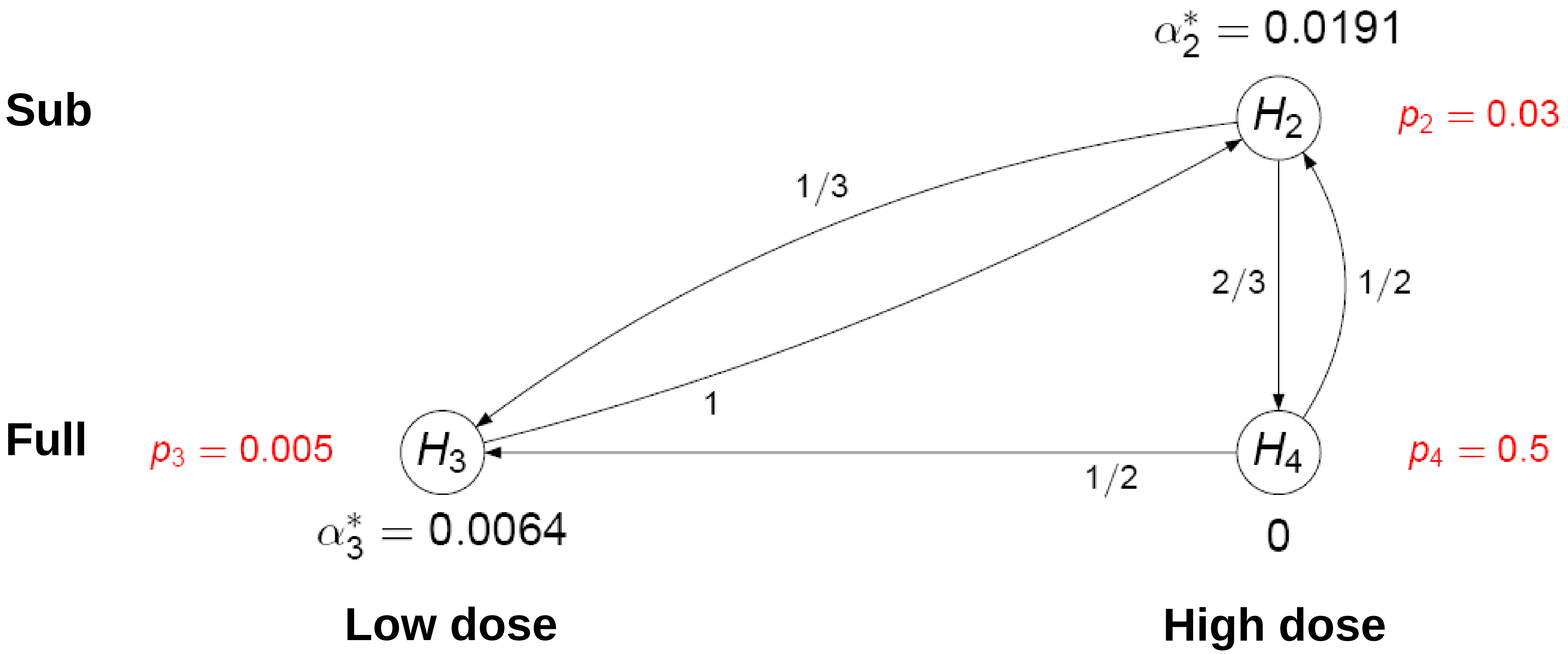


Case Study 2

New treatment for **targeted therapy**

- Update graph to complete α-propagation after first rejection

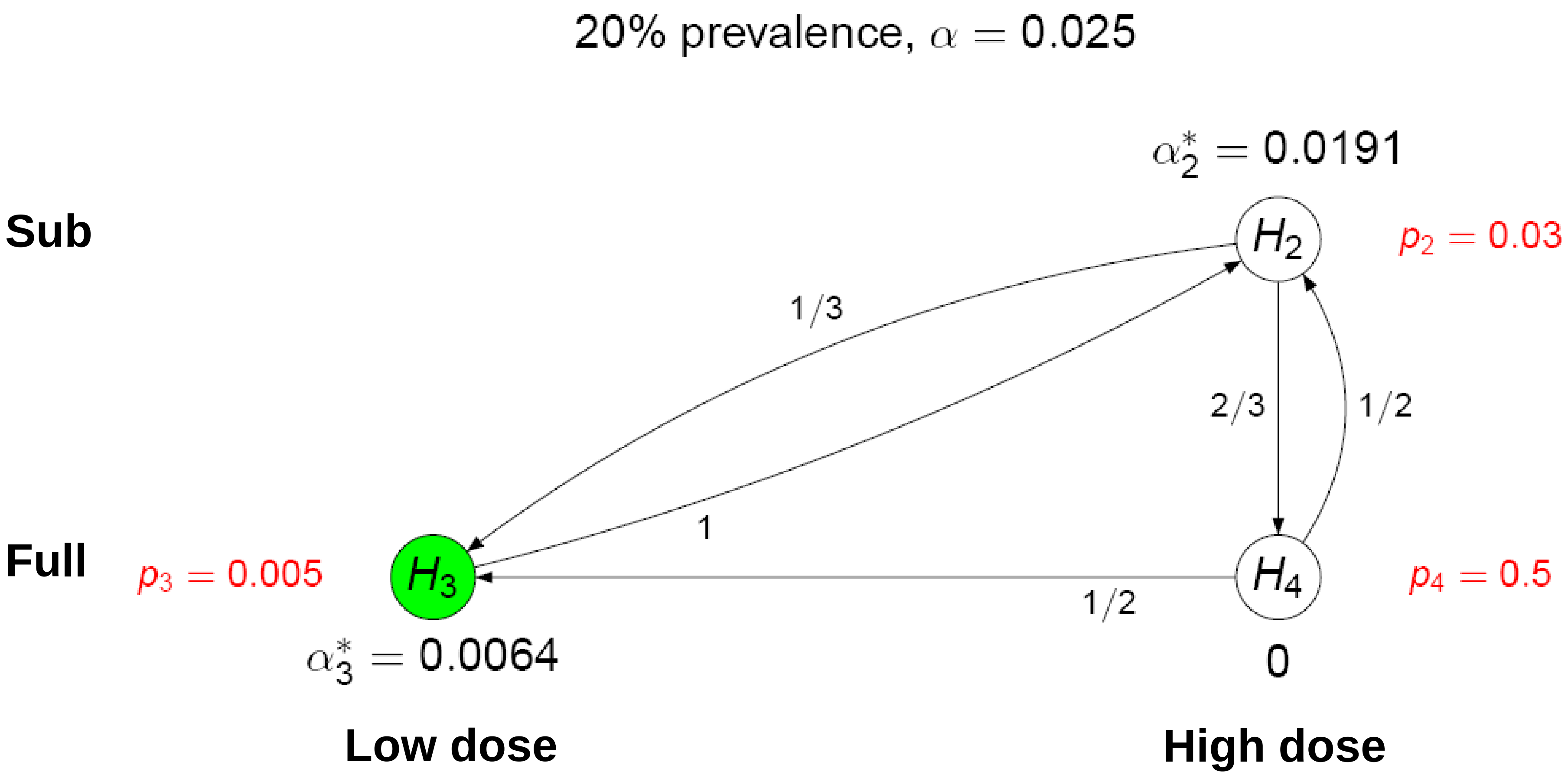
20% prevalence, $\alpha = 0.025$



Case Study 2

New treatment for **targeted therapy**

- Reject H_3 because $p_3 = 0.005 < 0.0064 = \alpha_3^*$



20% prevalence, $\alpha = 0.025$

$\alpha_2^* = 0.0191$

Sub

$p_2 = 0.03$

Full

$p_3 = 0.005$

$p_4 = 0.5$

$\alpha_3^* = 0.0064$

0

Low dose

High dose

Case Study 2

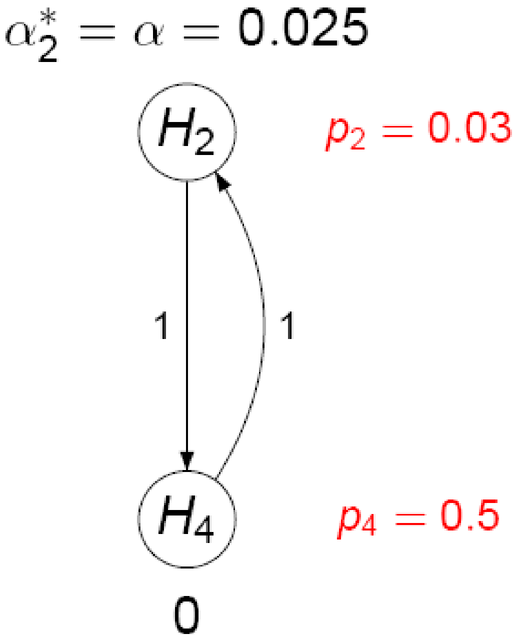
New treatment for **targeted therapy**

- Update graph to complete α-propagation after second rejection

20% prevalence, $\alpha = 0.025$

Sub

Full



Low dose

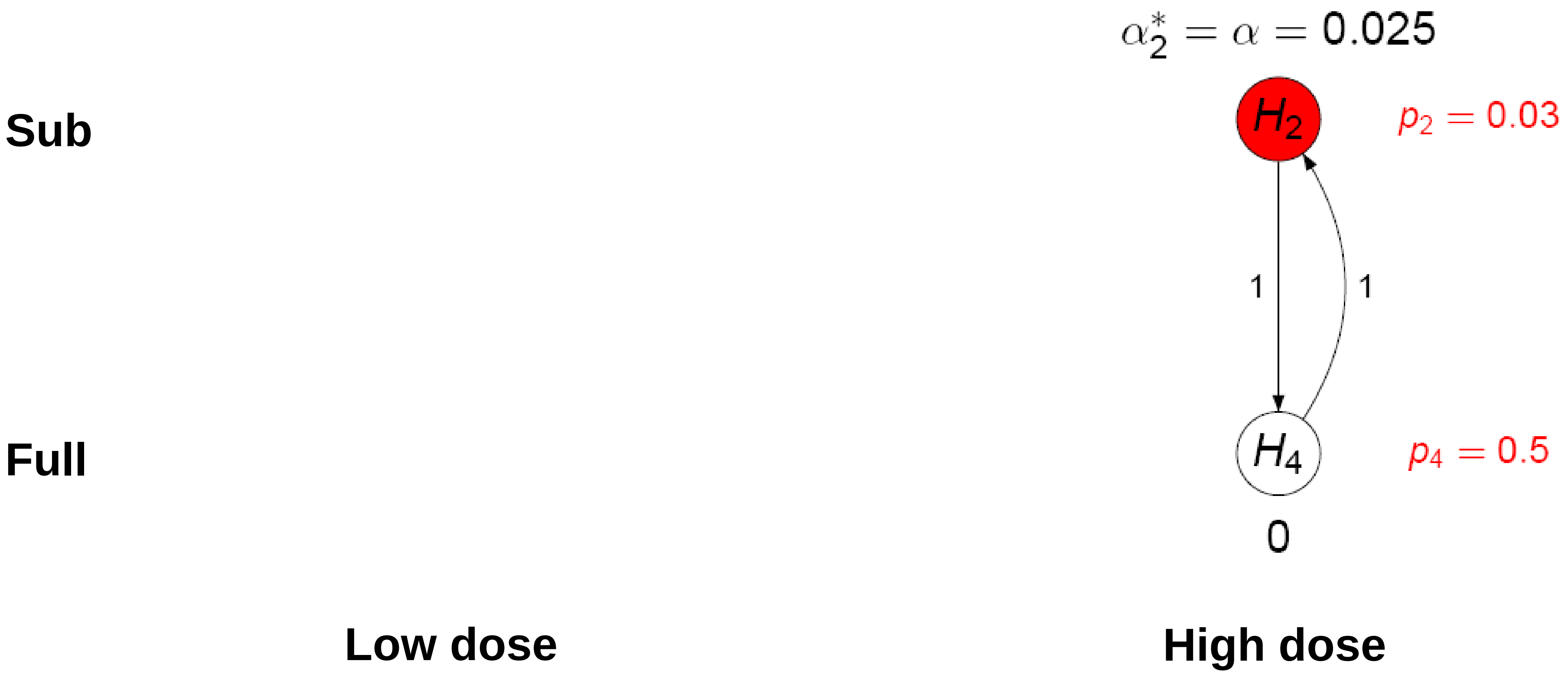
High dose

Case Study 2

New treatment for **targeted therapy**

- Stop the test procedure because $p_2 = 0.03 > 0.025 = \alpha_2^*$
- No further rejection possible

20% prevalence, $\alpha = 0.025$



Case Study 3

New treatment in **naive/pre-treated patients** for PFS and OS

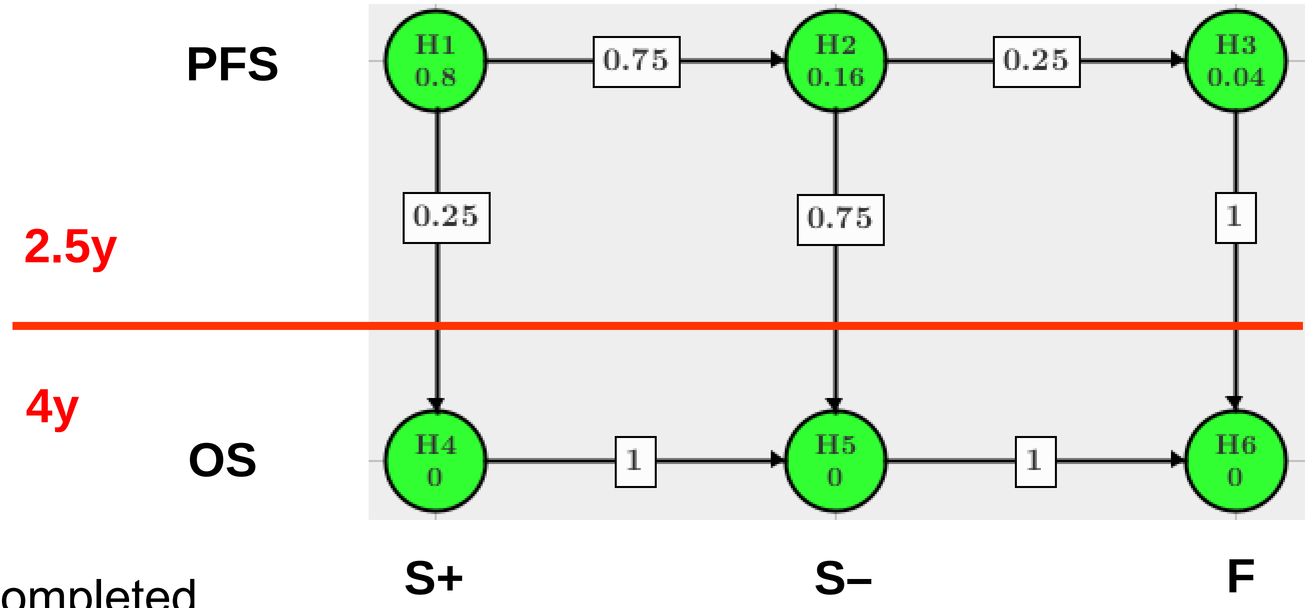
Structured hypotheses with **two levels of multiplicity**

1. Two-armed trial comparing novum vs. verum with six hypotheses:
   - three populations (S+ = naive, S– = pre-treated, F = full population)
   - two hierarchical endpoints: PFS (after 2.5 years) ➔ OS (after 4 years)
2. Important clinical considerations
   - conditional approval envisaged if PFS significant (study then continued until OS analysis)
   - avoid significance in S+ and F, but no significance in S– (otherwise difficulties with label)

How to construct decision strategy that reflects these requirements?

Case Study 3

New treatment in **naive/pre-treated patients** for PFS and OS



Remarks:

- After 2.5 years:
  a. Recruitment is completed
  b. No OS analysis is performed (otherwise extension to group-sequential setting mandatory)
- No edges from OS to PFS, as the PFS analysis is concluded by the time of the OS analysis
- Choice of α_3 :
  a. Very small α_3 = 0.04 * 0.025 = 0.001 ensures that PFS effect in F is declared significant only in case of an overwhelming effect
  b. Setting α_3 = 0 is an alternative possibility

Is strong FWER control always appropriate ?

- Consider two disjoint subgroups S_+ and S_- based on e.g. background therapy, predictive biomarker, disease status, or regions, with associated hypotheses H_+ and H_-

  → Applying strong FWER control, we have to adjust for multiplicty (e.g. test at $\alpha/2$)

  → However, if H_+ is rejected, drug is approved only for S_+

  – Risk of a false decision is strictly restricted to S_+ which can be controlled by testing H_+ at level α

  → Testing H_+ and H_- each at level α seems reasonable, although FWER can become almost 2α

  → FWER does not account for the relative risk that comes with false decisions

- Testing $\{H_1, H_2\}$ (e.g. two doses against placebo) and $\{H_+, H_-\}$ (e.g. disjoint subgroups) lead to different multiple testing problems

Summary

- Many **different applications** involving confirmatory subgroup analyses
  - Background therapy
  - Targeted therapy (e.g. based on a predictive biomarker)
  - Naive / pre-treated patients
  - Regional subgroups
  - ...
- **Lack of reproducibility** is a major concern, **even more in retrospective analyses** than in studies with prospectively defined subgroups
- Closer look at the subgroup hypotheses testing problem suggests that **strong FWER control may not always be appropriate** for clinical studies