In Hochberg's tradition: Selective Inference for Clinical Trials

Yoav Benjamini

Tel Aviv University

with

Rami Cohen

The Association for Public Health, Jerusalem

2014 IMPACT Symposium III, NC

Research supported by a European Research Council grant (PSARPS)

Yosi Hochberg
1945-2013

Origin and Childhood

- Born in Russia 1943 after his parents fled from Poland at the beginning of world war two
- After the war, passing through Germany, the family arrived at Israel.

- Bright
  yet adventurous
- Leadership:
  Nasser Mission

High-school

- Tsietling religious high school Talmudic studies
- throughout his life he moved back and forth between the two worlds

The way to Statistics

- Economics and Stat HUJI
- Reuven Gabriel
- Sent to Department of Biostatistics, University of North Carolina by TAU

"... I wish to express my gratitude to Professor P. K. Sen who, as my adviser, efficiently guided this work."

Some generalizations of the *T*-method in simultaneous inference (1974)JMA☆

Work on Car Safety

- NYU-TLV
- Head of Statistical Lab at TAU (1982)
- Book with Ajit Tamhane

False Discovery Rate

- Hochberg's procedure (1988, back to MCP)
- Sandoz study on reducing high blood pressure
- "Plots of p-value to evaluate many tests simultaneously" Schweder & Spjotvoll (1982)

  Hochberg and YB (1989): Estimate m_0 algorithmically, plug in the Bonferroni and Hochberg procedures
- Soriç (1989) argued forcefully against the usual way in medical research of using level 0.05 testing:

"There is danger that a large part of science is false"

FDR The background

Reading Soriç's paper, Yosi Hochberg got convinced that "there is more to the paper than what is written in it"
With this insight he told me I should read the paper …
--> JW Tukey & J Shaffer
None said it was a well known approach…
--> JASA October 1989

Breadth of knowledge in the field,
Deep intuition
Always seeking to go beyond what is currently known

The first MCP conference 1996

- Yosi's initiative

The Temple Workshop

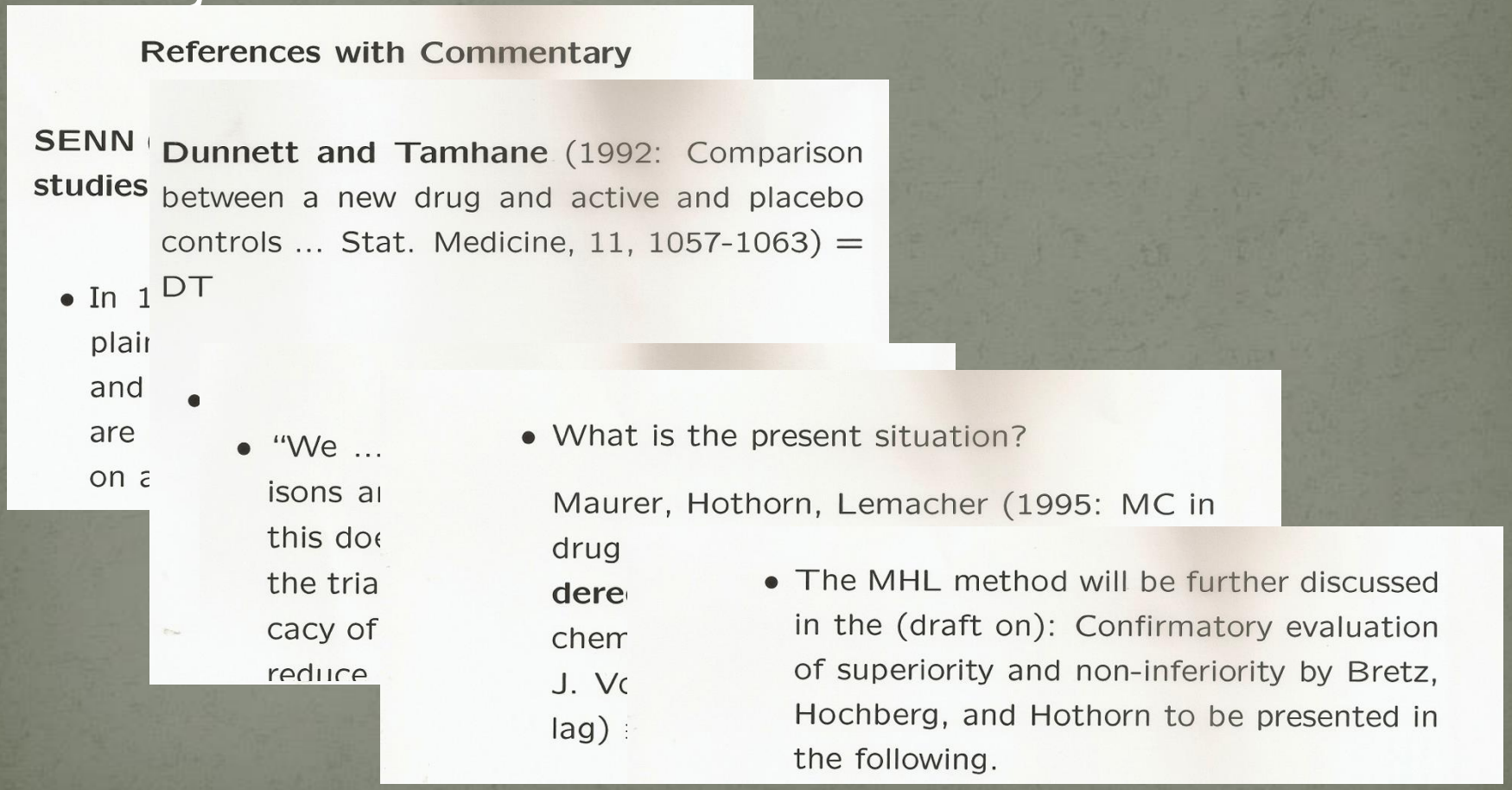
New Horizons (NH)
in Multiple Comparison (MC)
Procedures (MCP)

Lectures by Yosef Hochberg* (Yosi, PL)
NSF-CBMS Regional Research Conference
Temple University, Philadelphia
August 13-17, 2001

- **Other frequent abbreviations are:**
  More on This ≡ MOT
  With Reference to ≡ WRT
  With Special Reference to ≡ WSRT
  Scope of our Subject ≡ SOS

Examples

- of style



Examples

- of style
- of the concern about selective inference

13 **Selective Inference (SEI)**

131 Putter (1982) introduced SEI as follows:
A statistical inference procedure may be called selective if the identity of the object of inference (the parameter to be estimated, the hypothesis to be tested, etc.) is selected on the basis of the same sample data that are to be used in the procedure.

Examples

- of style
- of the concern about selective inference

- Yosi has not recognized that concern about selective inference was actually answered with the FDR. Neither did I at the time.
- In the following year, when Dani and I were working on False Coverage-statement Rate, we suddenly realized: we had a framework to address this concern which is different from simultaneous inference.
- Clearly we were strongly influenced by the Yosi'd talks

Examples

- The posterior p-value (ppv) with
  - publication ppv
  - discovery ppv
- Investigator with 0.06, 0.07, and 0.08.

- Unfortunately the Book never went beyond Preface
- Bretz and Hothorn and I editted others' lectures and Yosi's Outline in an IMS publication

- But every version of the Preface ended with the page…

Love and dedication

Dedicated to Miriam Hochberg

Seminar on serendipity

- His last departmental seminar
- Back to the issue of selective inference/serendipity/ (miracle?)

- The question he posed:

Can we assess the “significance of this happening” after it happened?

A selective inference question.

Relevant to the Open Access Clinical Trials Data panel

Last years

- Had serious health problem (Miriam's hospital at home)
- Withdrew entirely from professional life (no email)
- Enjoyed knowing about success – last week the FDR paper was discussed in Nature as 59th most cited paper across science - refused to participate

- Ending with personal remarks:

He attracted me into the field of MCP;

Passed on the conviction this was an important area of Stat;

Was an enormous source of knowledge.

And was fun to talk with.

- As a result of his life long contributions,

Simultaneous & Selective Inference is a very active area of research

at Tel Aviv University

as well as worldwide.

Let his memory be blessed יהי זכרו ברוך

Inference on the selected

- Inference on a selected subset of the parameters that turned out to be of interest

  after viewing the data!

- Worry about the effect of selection on properties of inference

How is selection manifested?

Selection by the Abstract

Selection by a table

Selection by highlighting

Selection by modeling: AIC, C_p, BIC, FDR, LASSO,…

P-values-free selection into the abstract

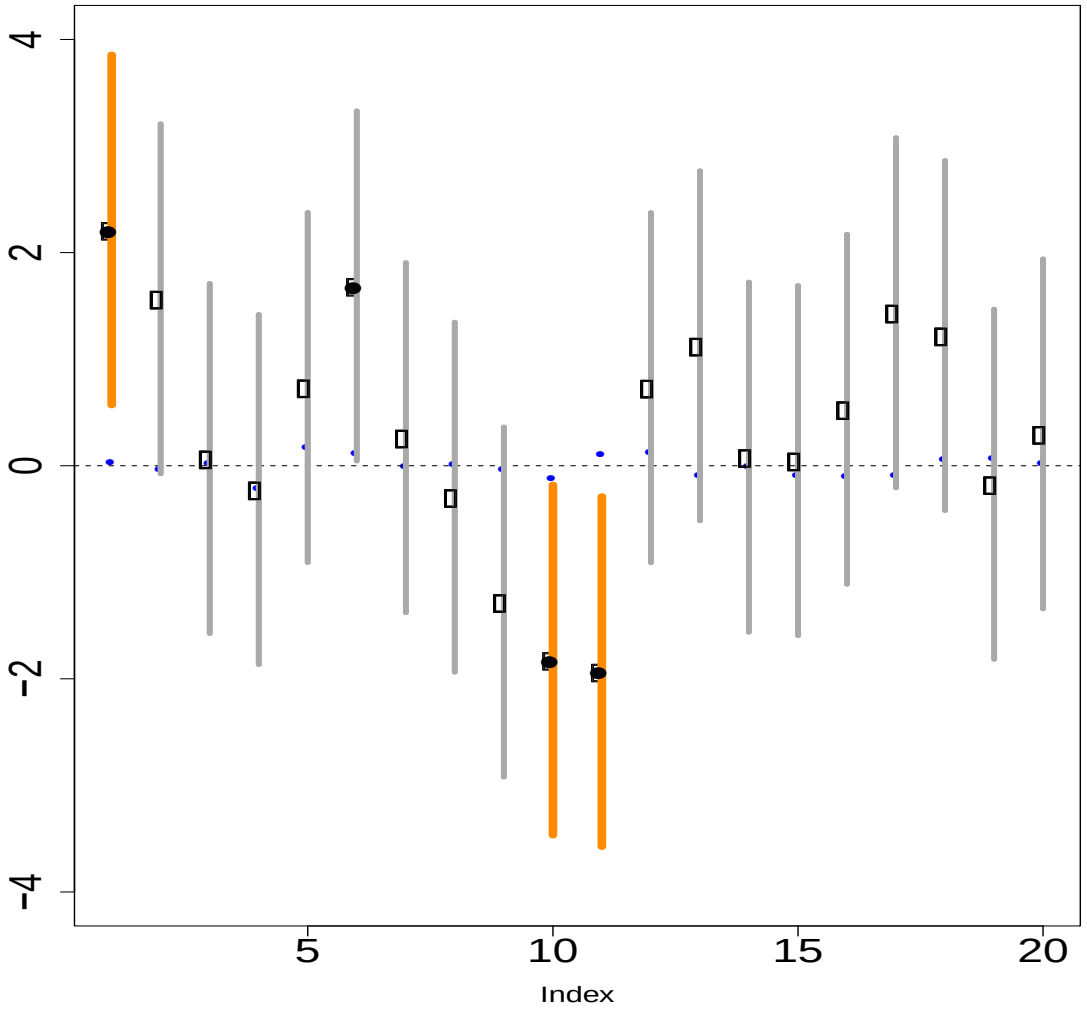
- Giovannucci et al. (1995) look for relationships between more than a hundred types of food intakes and the risk of prostate cancer
- The abstract reports only three (marginal) 95% confidence intervals (CIs), apparently only for those relative risks whose CIs do not cover 1.

20 parameters to be estimated with 90% CIs

3/20 do not cover

3/4 CI do not cover when selected

These so selected 4 will tend to fail, or shrink back, when replicated



Inference on the selected

- When addressing a family of inferences (tests, estimates, CIs) we wish at least to assure that the property of the individual inference will still hold on the average over the selected
- For Confidence Intervals

  The False Coverage-statement Rate (FCR) of a selective CIs procedure is the expected proportion of coverage statements made that fail to cover their respective parameters

  (Schooler's complaint, Nature)

- This is also the essence of the FDR
- Simultaneous inference: a stronger desirable requirement

What I (we?) know about clinical trials

These enjoy complete transparency and clearly specified

population of study,

treatments' regimes,

endpoints for efficacy,

endpoints for safety,

statistical methods for the analysis and reporting,

and in particular methods that control

the probability of making even one false discovery claim

(the familywise error-rate) offering simultaneous inference

Netalizumab Study

Natalizumab, was examined by Ghosh et al (NEJM, 2003) for the treatment of Crohn's disease.

Comparing 3 regimes with placebo; 4 measures of success;

at 5 time points; Total 51 endpoints

1 primary endpoint: Treatment by 2 infusions of 6mg/kg dose
remission measured at week 6

Other 50 described as secondary endpoints

The result for the primary endpoint was not significant ($p = 0.533$);

27 secondary endpoints at $p \leq 0.05$ were considered as discoveries

Study reported as a success

Would not have been reported as such using FWER control

What is really going on in medical research

We conducted an in deep analysis of 100 papers from the NEJM 2002-2010. All had multiple endpoints

- # of endpoints in a paper 4-167 ; mean=27
- In 80% the issue of multiplicity was entirely ignore (in none fully addressed)
- All studies designated primary endpoints (in 84% a single one)

What we see is the difference between the phases of clinical research at the regulatory stage and earlier ones. The latter constitute most of the medical research

But even in phase III studies

- Posaconazole was tested against Fluconazole for the prevention of invasive fungal infections in severe graft-vs-host disease.

  Ulmann et al (NEJM, '07). Phase III bioequivalence study.

- Primary endpoint: incidence of invasive fungal infections
- Secondary endpoints:

  Mortality (p=0.07);

  Time to first breakthrough of fungal infection (p=0.048);

  Mortality from fungal invasion (p=0.046) ;

  Number of cases with fungal infections (p=0.006)

  Number of breakthroughs during exposure time (p=0.004);
  Same but of aspergillosis fungal infection (p=0.001)

- Primary endpoint significant in equivalence test, but > 0.5 for superiority.
- Abstract's conclusion: "It was superior in preventing invasive aspergillosis and reducing the rate of death related to fungal infections"

Estimating the science-wise false discovery rate

- Ioanidis in "Why most research findings are false" wrote about what may happen (same as Soric)
- Leah & Leek ('14) tried to estimate:

  Mined the Abstracts of 4 medical and 1 epid. journals over 10 years

  Collected all p-values < 0.05; Estimated FDR at ~15%
- Analyzing a sample of 25 papers:

  Problems seems (i) more severe (ii) different.

  # p-value ≤ 0.05 in the paper >> in the abstract, yet in 19 / 25

  the smallest p-value in the paper appeared in the abstract.

  Again, evidence of selection. Hechtlinger & YB ('14)

Weighted FDR

- For each H_i assign a weight $w_i \geq 0$, $i=1,2,\ldots m$. w.l.o.g. $\Sigma w_i = m$.
- Define $R_i = 1$ if H_i is rejected, otherwise $R_i = 0$;
- Define $V_i = 1$ if H_i is erroneously rejected, otherwise $V_i = 0$. Let

$$Q(w) = \begin{cases} \dfrac{\sum_{i=1}^{m} w_i \cdot V_i}{\sum_{i=1}^{m} w_i \cdot R_i} & \sum_{i=1}^{m} w_i \cdot R_i > 0 \\ 0 & otherwise \end{cases} \qquad (2.1)$$

wFDR=E(Q(w)) is the *weighted FDR*

which we wish to control it at level *q*. Hochberg &YB('97)

Weights for primary and secondary endpoints

- *P* primary endpoints with corresponding p-values

  $\mathbf{p_p}=(p_{(p1)}, p_{(p2),\ldots,} p_{(sP)})$,

- *S* secondary endpoints with corresponding p-values

  $\mathbf{p_S}=(p_{(s1)}, p_{(s2),\ldots,} p_{(sS)})$,

- $\mathbf{w_P}$ and $\mathbf{w_S}$ be the corresponding vectors of weights, with w_{pi} and w_{sj} denoting the weights assigned to the *i*-th primary and the *j*-th secondary endpoints respectively.
- Obviously we assign $w_{pi} \geq w_{sj}$ for all *i* and *j*
- and sometimes further $R' = \sum w_{P_i} / \sum w_{s_i} \geq 1.$
- For ease of notation we'll require $\sum w_{s_i} + \sum w_{P_i} = S + P.$

The Weighted FDR controlling procedure

- Sort all p-values $p_{(i)}$ corresponds to $H_{(i)}$

$$p_{(1)} \le \ldots p_{(i)} \le \ldots \le p_{(m)}$$

- Let $$k = \max\{j : p_{(j)} \le \sum_{i=1}^{j} w_{(i)} \cdot q / m\}$$

- Reject $$H_{(1)} \ldots H_{(k)}$$

- When applied in the multiple endpoint problem

$wBH_q(\boldsymbol{w},\boldsymbol{p})$

wBH in the extreme

w_{si}= *0* to every one of m secondary endpoint

w_p= *m+1* to the primary endpoint

Using the wBH at 0.05:

- If $p_p > 0.05$ no secondary can be rejected

  because even $p_{si} << 0.05$ it is compared with *.05*0/(m+1)*

- If $p_p \leq 0.05$ the primary is rejected

  because $p_p \leq 0.05 * (0+..+0+m+1)/(m+1)$

  and then any secondary whose $p_{si} \leq 0.05$ is rejected

So this reflects the ongoing unfortunate practice

Hierarchical Weighted FDR controlling procedure

- Calculate the p-value for the intersection hypothesis of the secondary hypotheses using the weighted Simes test

$$p^* = \min_i \; p_{s_{(i)}} \cdot \sum_{j=1}^{S} w_{s_j} \Big/ \sum_{j=1}^{i} w_{s_{(j)}}$$

- Assign it the sum of secondary weights $w^* = \sum w_{s_i}$.
- Pool p^* and its weight w^* with the p-values and weights of the primary hypotheses, to get

$$\boldsymbol{p'}=(p^*,\boldsymbol{p_P}) \text{ and } \boldsymbol{w'}=(w^*,\boldsymbol{w_P}).$$

- Apply the $wBH_\alpha(\boldsymbol{p'},\boldsymbol{w'})$; reject primary hypotheses accordingly
- If the intersection hypotheses is among the rejected

  test the secondary endpoints with $wBH_\alpha(\boldsymbol{p_s},\boldsymbol{w_s})$

- The HWF procedure should be applied at level

  $\alpha(q,P,S,R') \leq \alpha$

  in order to assure $wFDR \leq q$

- We treat only the case P=1 (recall 84% in NEJM sample);

  and $w_{si}=w_s$

- A close bound is calculated under independence
- A somewhat looser bound under Positive Regression Dependency, (a very reasonable assumption for efficacy endpoints in clinical trials)
- Implemented in an easy to use website http://spark.rstudio.com/shayy/HWBH/

  accessed via http://www.replicability.tau.ac.il

HWF in the extreme

$w_{si} = 0$ to every one of m secondary endpoint

$w_p = m+1$ to the primary endpoint

Using the HWF at $\alpha \sim 0.05$:

- If $p_p > 0.05$ no secondary can be rejected, as

  the intersection's p-value is compared with $.05*0/(m+1)$

- If $p_p \leq 0.05$ the primary is rejected

The secondary endpoints are also tested at level 0.05

first their intersection,

then the secondary themselves – all at level $q = 0.05$

Let’s leave out the math

Back to Netalizumab case

Recall 51 endpoints: *P=1, S=50. Choose R'=2,* ($R=100=w_p/w_s$)

- HWF: The list of sorted secondary endpoints p-values:

$$p_{S_{(1)}} = 4.76 \cdot 10^{-5}, p_{S_{(2)}} = 6.29 \cdot 10^{-5}, p_{S_{(3)}} = 1.44 \cdot 10^{-4}, \ldots, p_{S_{(50)}} = .992$$

Simes p-value for intersection 0.00157 < 1/3*0.0257

The p-value of the primary 0.533 > 2/3*0.0257

$\alpha(0.05,1,50,2) = 0.0257$

12 secondary p-values ≤ 0.0257*12/50 rejected by HWF

- wBH: The p-value of the primary 0.533 > 0.05*102/150
- 9 secondary p-values ≤ 0.05*9/(150) rejected by wBH

In both cases secondary endpoints were rejected while controlling error rate. Study's general conclusion should be reported as it was .

But with less exaggerated claims about all 27 differences

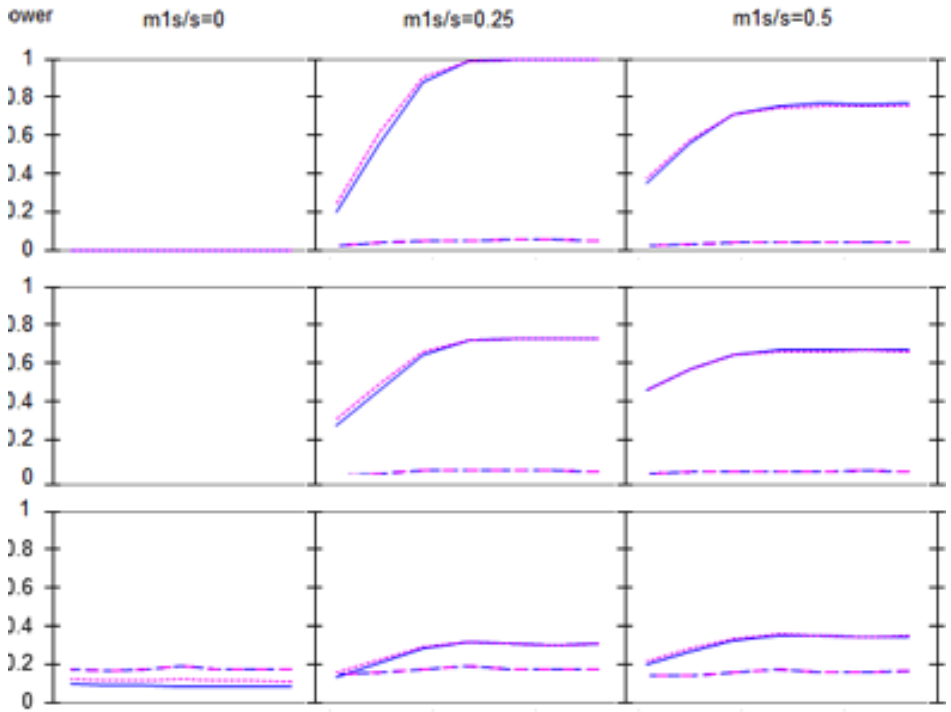
Concepts of power

- Overall (weighted) power $\Pi_G = E\left(\sum_{i\in I_1} w_i \cdot R_i \Big/ \sum_{i\in I_1} w_i\right)$
- Primary's power $\Pi_P = E(R_0)$
- Secondary's power $\Pi_S = E\left(\sum_{i\in Is_1} w_i \cdot R_i \Big/ \sum_{i\in Is_1} w_i\right)$
- Weighted primary and secondary

$$\Pi_\delta = \frac{\delta}{\delta+1}\Pi_p + \frac{1}{\delta+1}\Pi_s$$

- which will serve us in practical recommendations

Simulations analysis of overall power



μ_p=0

---- HWF

---- wBH

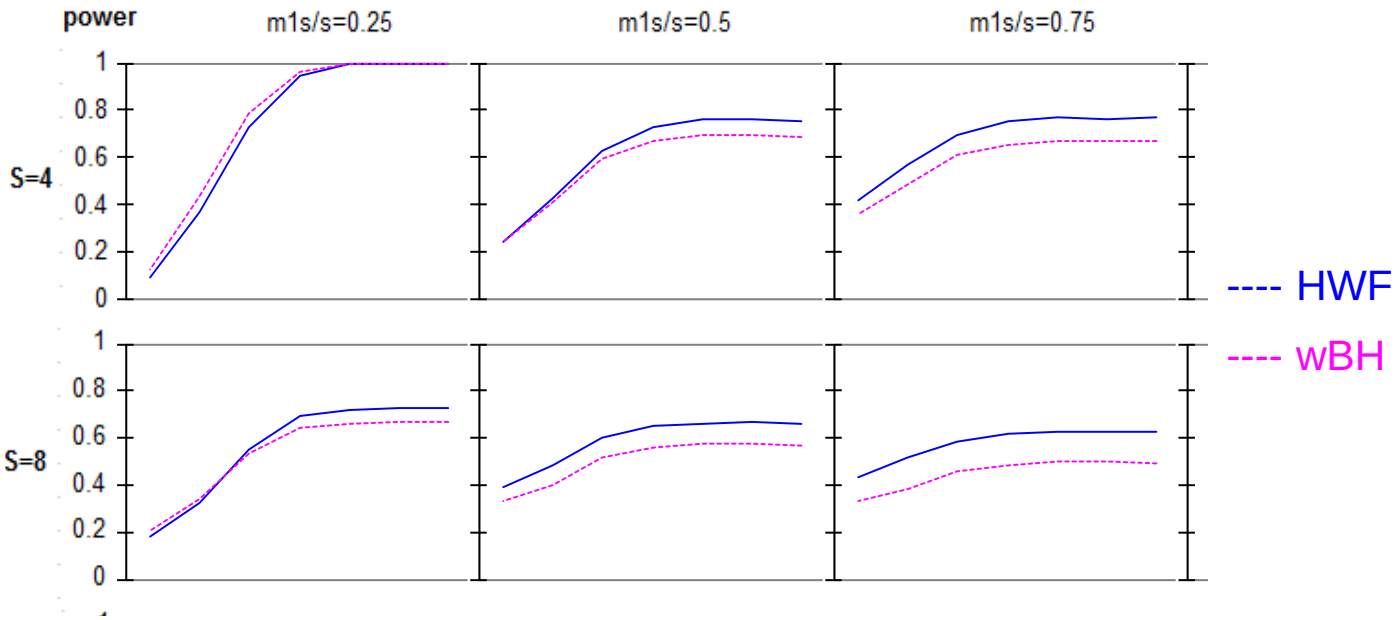
- - Simes GK

- - Bonferroni GK

μ_p=1

Overall power vs secondary endpoint's parmeter μ_{S_1} : The overall power of the GK procedure is bounded by α while the HWF and wBH procedures have much higher power for all different parameter values. This remains also true even when the primary endpoint has but a small effect (last 2 rows) where $\mu_p = 1$.

Simulations analysis of power



Overall power vs secondary endpoint's effect μ_{S_1} where $R' = 10$, $\mu_p = 0$,. This shows similar trends to those in the upper two rows of Figure 1, but in this setting the difference in power between the two weighted procedures becomes evident, and the HWF has higher power. Continuous blue line = HWF; Dotted pink = wB-H; Dashed blue = Simes GK; Dashed pink = Bonferroni GK.

Simulations analysis of power

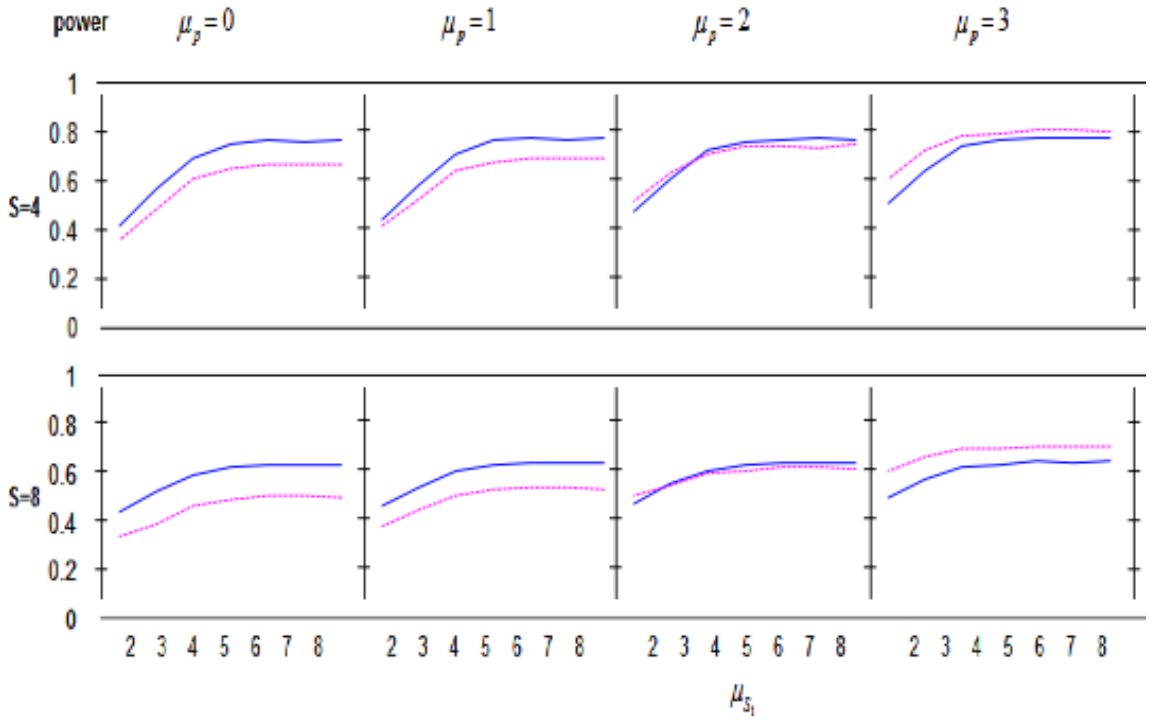


Figure 3 Power for secondary endpoints vs secondary endpoint's effect μ_{S_1} $R' = 10$, $\frac{m_1 S}{s} = 0.75, \mu_p = 0,1,2,3,4$. When the primary effect is low the HWF procedure is more powerful than the wBH in discovering secondary endpoints. When it is high the opposite is true. Continuous blue line= HWF; Dotted pink = wBH.

Power considerations

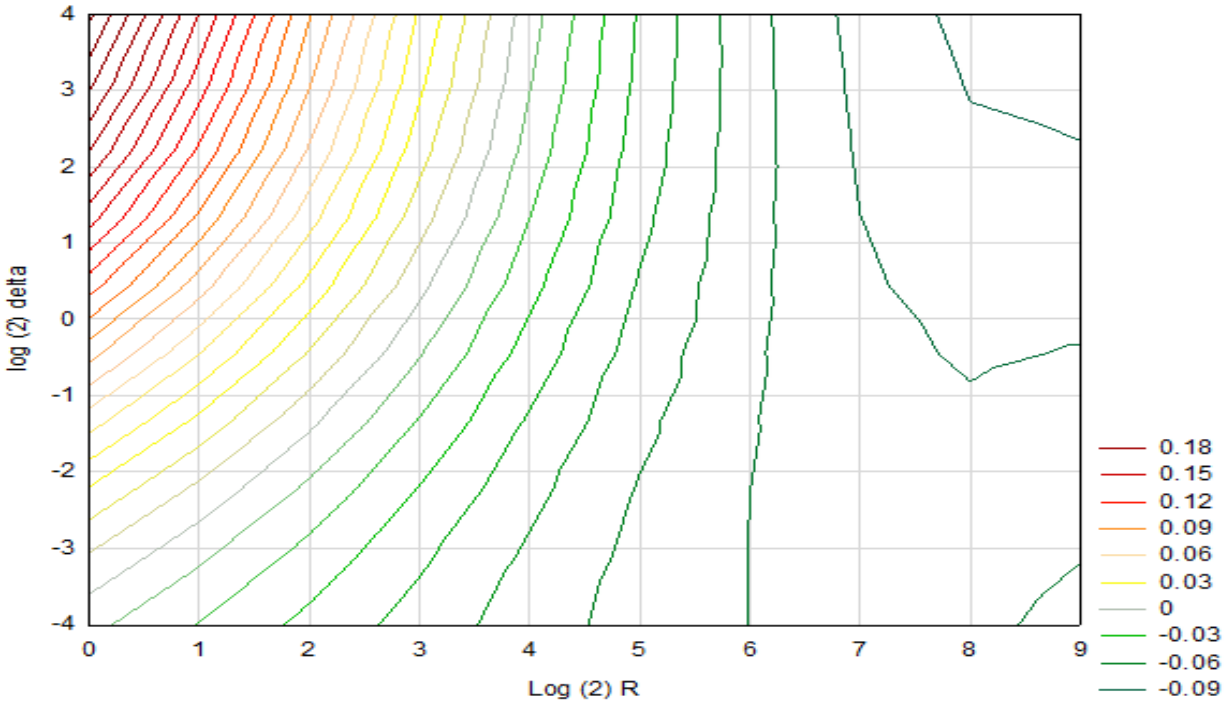
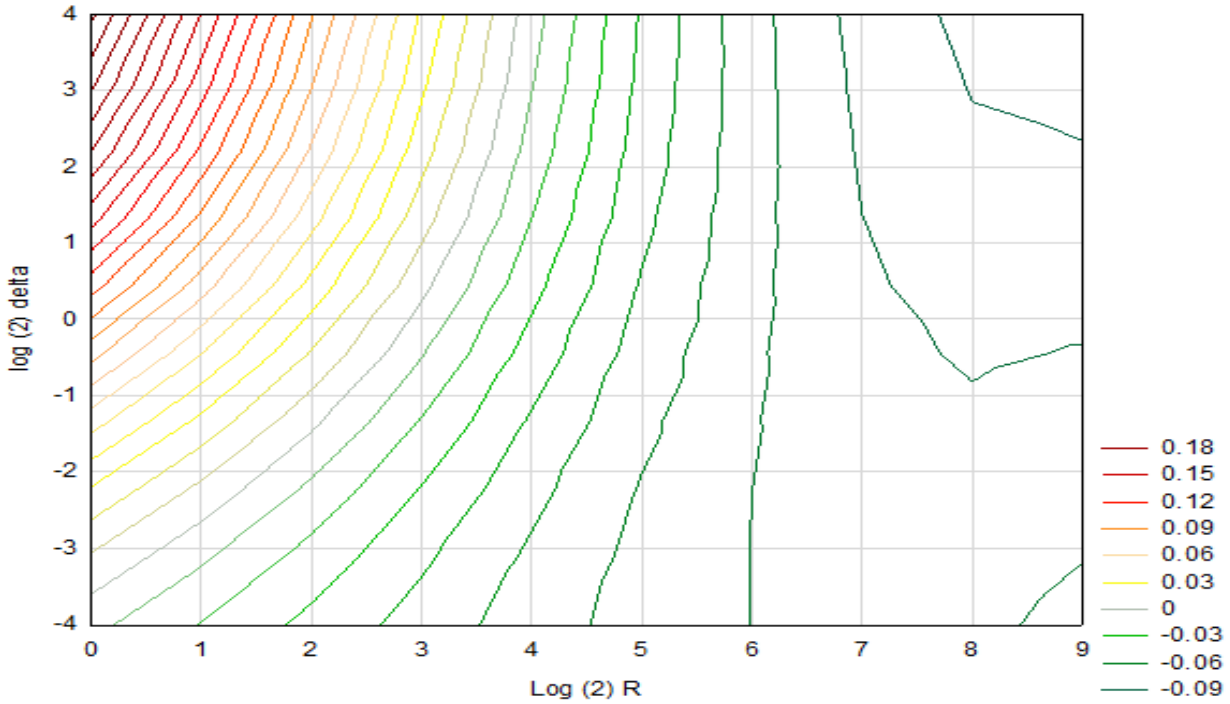


Figure 4. The difference in power as a function of R and δ: HWF weighted power – wBH weighted power, for $\mu_p = 2$, S=16, $m_1S/S=.25$, $\mu_S = 1.5$, $\mu_{S_1} = 6$. For R between 1 and 10 and δ assigning more weight to the power for the primary endpoint, HWF is superior. This remains true for R closer to 1 even when the weight given to the secondary is more than 8 times more than to the primary. For R bigger than 10 the wBH is more powerful

Practical considerations (i)

- Any of the secondary endpoint at this phase can replace the primary in later phases.
- Hence a type I error for a secondary endpoint is almost as crucial as for the primary. Choose R close to 1, (≤ 2.)
- Assessing that the primary endpoint reflects the superiority of the treatment better than any secondary, the higher R.
- Power consideration should be similar for the primary and secondary ones so we choose δ~1, in evaluation .
- Use of HWF is recommended if we assess that at few of the secondary endpoints might be affected and most of them not substantially so. Otherwise use wBH but…

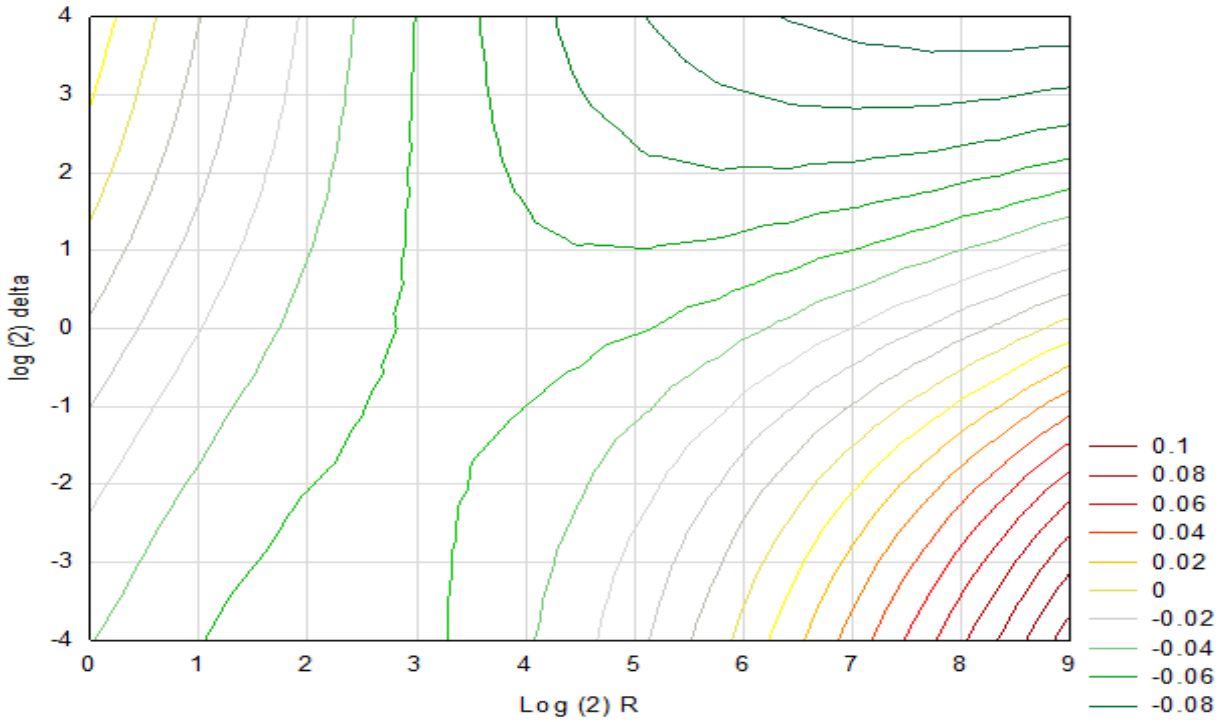
Power considerations



1≤R≤2; log(d)~0: Use of HWF is recommended if we assess that at few of the secondary endpoints might be affected and most of them not substantially so. Otherwise use wBH but…

the secondary is more than 8 times more than to the primary. For R bigger than 10 the wBH is more powerful

Power consideration



1≤R≤2; log(d)~0: Still Use of HWF is recommended if we assess that at few of the secondary endpoints might be affected and most of them not substantially so. Otherwise use wBH but…

the secondary, an unlikely setting, the wBH is superior.

Practical considerations (ii)

- Any of the secondary endpoint at this phase can replace the primary in later phases.
- Still, there is regulatory/economic advantage in rejecting the primary endpoint.
- As in previous case, R close to 1, say in the range 1-2.

  But choose $\delta>1$, in order to evaluate:
- Use of HWF is recommended as before with less reservations.

Practical considerations (iii)

- The primary endpoint indeed determines the success of the trial, and secondary endpoints can strengthen it.
- Now, R >> 1, say in the range R>S R/S=R'>1.
- (i) Choose δ>1, in order to evaluate:
- small advantage to wBH.
- (ii) Sure about success of primary so choose δ<1 to discover new benefits: Use of HWF

It may be that wBH is preferable but..

the statements about the secondary ones are hardly protected!

Related work: Hierarchical FDR testing

When no weights are involves

a) BY with Bogomolov (‘13) The primary endpoints as one family, secondary as a second family

1st stage: test families using intersection p-values

2nd stage test within family using Bonferroni or BH

at level α(# significant families)/2

Controls* FWER or FDR on the average over the selected

b) Guo & Sarkar et al (+12) for families of equal size shows that the over-all FDR is controlled when the second stage uses adaptive Bonferroni method.

Related work: Other weighing schemes

Genovese Roeder and Wasserman ('06):

- Hypotheses that get larger weights have increased probability of being rejected, at the (small) expense of reducing the power for the low weighted hypotheses.
- Such weights may reflect prior probabilities about correctness of hypotheses - can play a role in Bayesian analysis.
- They control the regular FDR rather than a weighted one.

Sometimes a virtue - but not in this case:

Weigh differently for the primary and a secondary

the benefit and cost from a rejection (both correct and false)

Are we heading this way?

- Kaplan (2008, with FDA) reveals

  increase in Phase III failure rates from 30% to 50% in recent years.
- One of the reasons offered is the lack of attention to the multiplicity issue.
- In 2009, the FDA 's issued a directive concerning necessary multiplicity correction required for significance levels of secondary endpoints in studies concerning surgical ablation devices for treatment of atrial fibrillation. "If you intend to present comparisons between groups for a secondary effectiveness endpoint in your labeling, …".

- Control of the some error criterion for the secondary endpoints is getting attention by the regulatory agencies. (?)

Where are we heading

- Recent concern of Medical journals about replicability (reproducibility) should be reflected in attention to the multiplicity of secondary endpoints
- wFDR control is an appropriate answer to this concern.
- It balances control with the more flexible way medical research is done at earlier stages.
- It will also protect researchers (investors and treatment developers) from over confidence in their initial results.
- Better methods may indeed be developed for controlling hierarchicaly the wFDR.
- Choice of R should be agreed upon collaborative research into it needed (and of course the choice should be stated in the protocol).

Visit us at

www.replicability.tau.ac.il

Thanks