# Mutiple Testing in Clinical Trials

Livio Finos

# Weeks

- Week 1:
  One test, Many test,
  Lab and simulation
- Week 2:
  Multiple testing procedures (FamilyWise Error Rate)
  Lab multiple testing procedures
  Discussion with dr Glauco Cappellini (Quantiles Italia)
- Week 3:
  False Discovery Rate and other measures
  Sequential Rejecton Principle for FamilyWise Error Rate
- Week 4:
  Univariate and Multivariate Permutation Tests

# American Statistical Association's Ethical Guidelines for Statistical Practice

Recognize that any frequentist statistical test has a random chance of indicating significance when it is not really present.

Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion.

Failure to disclose the full extent of tests and their results in such a case would be highly misleading

# American Statistical Association's Ethical Guidelines for Statistical Practice

Recognize that any frequentist statistical test has a random chance of indicating significance when it is not really present.

Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion.

Failure to disclose the full extent of tests and their results in such a case would be highly misleading

e.g. VaxGen's AIDSVAX trial . . .

# VaxGen’s AIDSVAX trial?

VaxGen announced the results of the first-ever efficacy trial of an AIDS vaccine on 24 February 2003:

the vaccine prevent HIV infection?

| | Total | Infected | | | |
|---|---|---|---|---|---|
| All subjects | 1679 | 96 | 🟩 | 5.8% | PLACEBO |
| | 3330 | 191 | 🟥 | 5.7% | VACCINE |

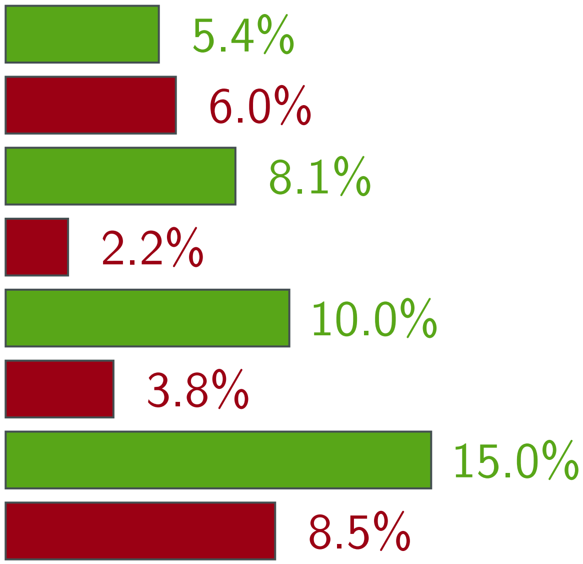
*”We saw absolutely no difference between the vaccine and placebo groups. Everyone was pretty depressed.”*

but the next day...

# VaxGen’s AIDSVAX trial

...by broking the data down into racial groups – which they say was part of the original design – the vaccine appeared to have worked in blacks:

| | Total | Infected | Fisher’s exact test |
|---|---|---|---|
| White | 1508 | 81 | $p_W = 0.898$ |
| | 3003 | 179 | |
| Black | 111 | 9 | $p_B = 0.015$ |
| | 203 | 4 | |
| Asian | 20 | 2 | $p_A = 0.301$ |
| | 53 | 4 | |
| Other | 40 | 6 | $p_O = 0.345$ |
| | 71 | 6 | |



*”The numbers were small, which concerned us, but the result was highly statistically significant. They were pretty incredible results.”*

# Criticisms

1. **failure to account for multiplicity**

   *"The p-values were not adjusted."*

2. **selective reporting (data snooping)**

   *"It's all murky because it's all post hoc analysis. They might as well do a subgroup analysis based on signs of the zodiac."*

   If you torture your data long enough, they will confess you whatever you want to hear!

# Revived interest in multiple testing

“-omics”

e.g. genomics experiments with microarray data: which genes are differentially expressed?

model selection

e.g. multiple regression: which coefficients matter?

econometric

e.g. comparing several strategies with a benchmark: any better? which ones?

. . .

## clinical trials

### sources of multiplicity

- multiple endpoints
- several treatments
- multiple time points
- subgroup analysis
- interim analysis
- . . .

### regulatory guidelines

- statistical principles for clinical trials (ICH E9)
- points to consider on multiplicity issues in clinical trials (EMEA)
- . . .