Scientific integrity, (un)ethical conduct,. . .
& Journal publishing

Geert Molenberghs

Acknowledging materials provided by Geert Verbeke

geert.molenberghs@uhasselt.be & geert.molenberghs@kuleuven.be

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

UHasselt & KU Leuven, Belgium

www.ibiostat.be

EFSPI, November 22, 2019

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Theme 1
Statistical / Scientific Evidence

1.1 Example: Captopril Data

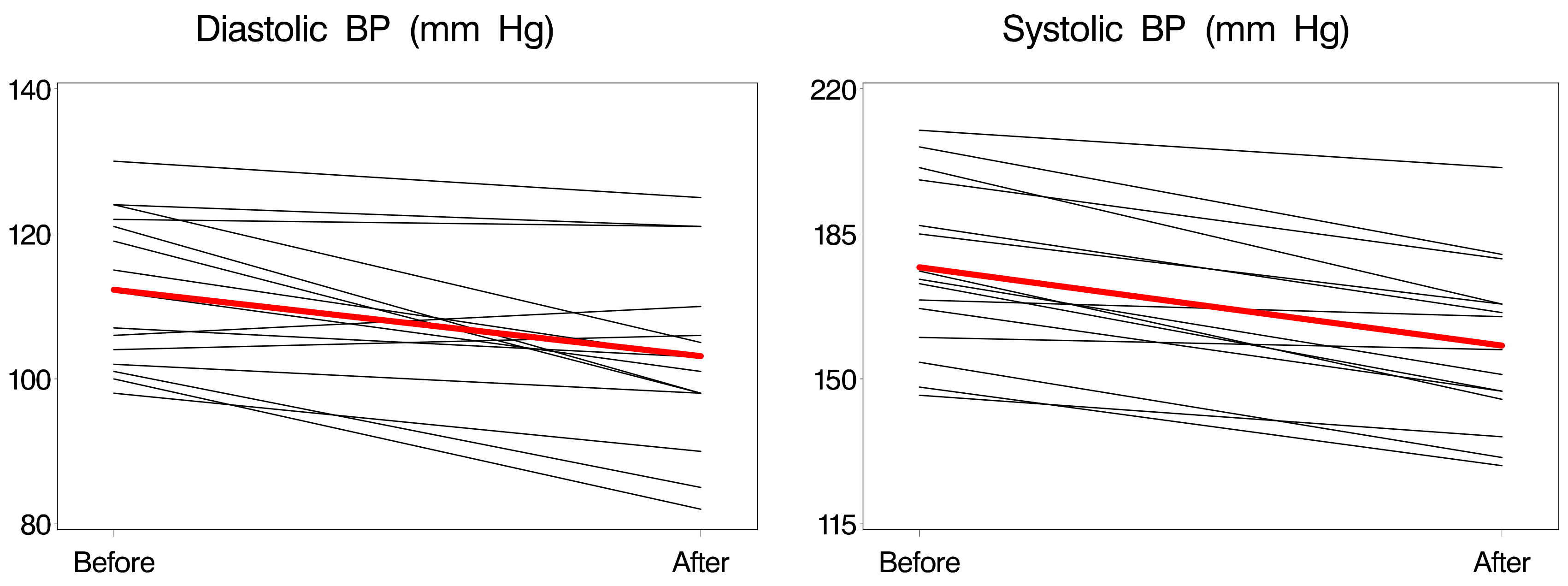
- 15 patients with hypertension
- The response of interest is the supine blood pressure, before and after treatment with CAPTOPRIL
- Research question:

How does treatment affect BP ?

- Dataset ‘Captopril’

	Before		After	
Patiënt	SBP	DBP	SBP	DBP
1	210	130	201	125
2	169	122	165	121
3	187	124	166	121
4	160	104	157	106
5	167	112	147	101
6	176	101	145	85
7	185	121	168	98
8	206	124	180	105
9	173	115	147	103
10	146	102	136	98
11	174	98	151	90
12	201	119	168	98
13	198	106	179	110
14	148	107	129	103
15	154	100	131	82

	Average (mm Hg)
Diastolic before:	112.3
Diastolic after:	103.1
Systolic before:	176.9
Systolic after:	158.0



- “The result is 9.27 (4.91;13.63) with P =0.001.”

- **“The difference in diastolic blood pressure is estimated as 9.27 mmHg, with 95% confidence limits [4.91,13.63] and p-value based on a two-sided t-test of 0.001.”**

1.2 Possible Errors in Decision Making

- In our example about the Captopril treatment, we obtained $p = 0.001$ leading to the rejection of the null hypothesis of no treatment effect.
- This should not be considered as formal proof that there is a treatment effect
- Even if the treatment has no effect at all, a sample like ours would occur once every 1000 times.
- Maybe, our sample was indeed the extreme one that happens once every thousand experiments.
- Alternatively, suppose we would have obtained $p = 0.9812$. We then would not have rejected the null hypothesis, and concluded that there is no evidence for any treatment effect.

- This should not have been considered as formal proof that any treatment effect would be absent.

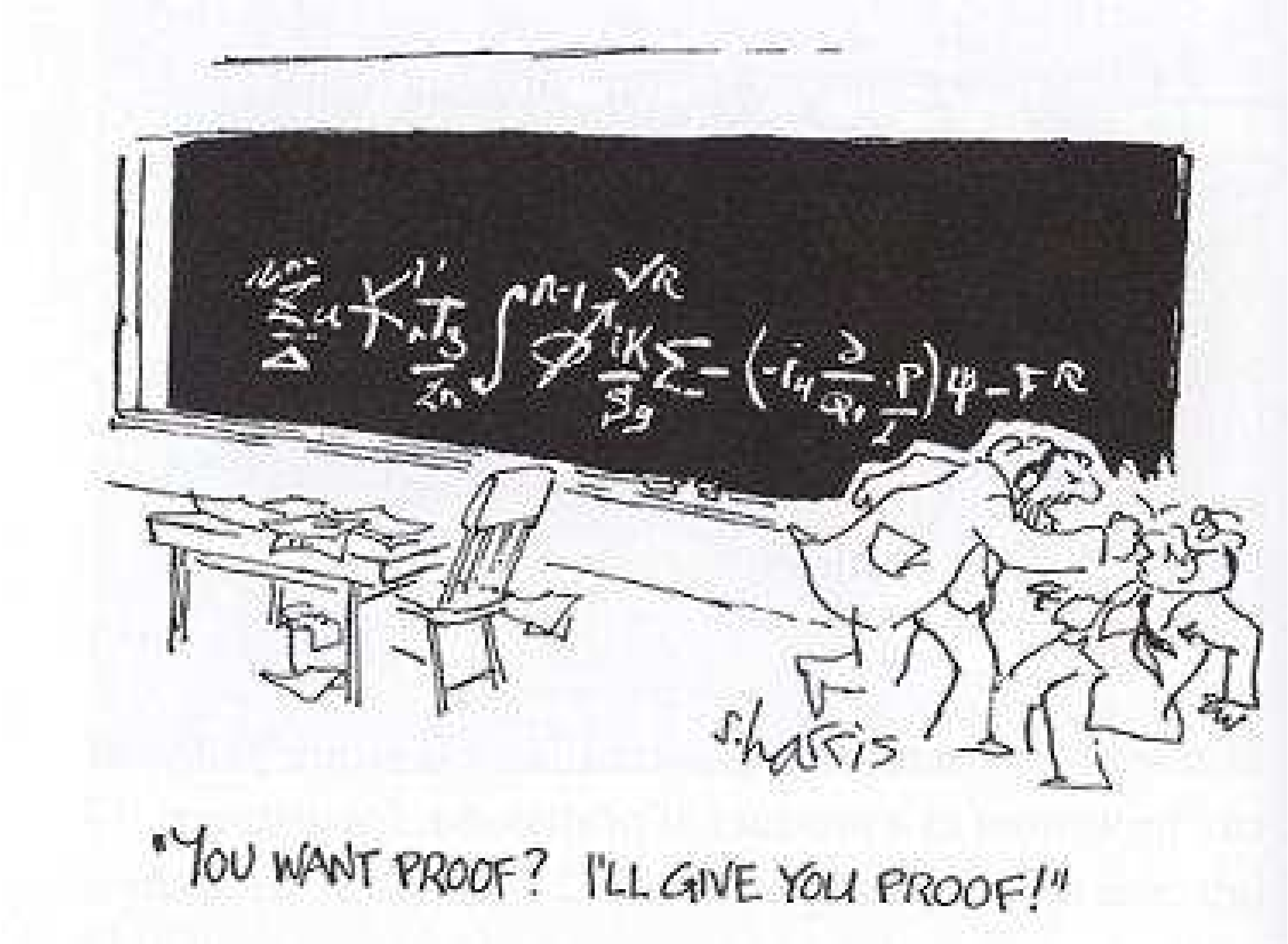
- Maybe, the treatment effect μ is not 0, but very close to 0. The data one then would observe would look very similar to data that would be observed if $\mu = 0$, such that the data do not allow to detect that $\mu \neq 0$

- Conclusion:

“~~Statistics can prove everything~~”

© Mike Baldwin / Cornered

- **Intuitively:** Absolute certainty about population characteristics cannot be attained based on a finite sample of observations

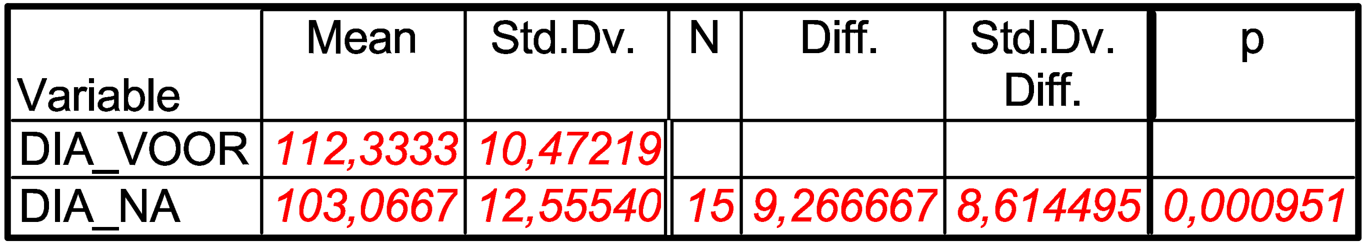

S. Harris
"YOU WANT PROOF? I'LL GIVE YOU PROOF!"

1.3 Significance versus Relevance

- Of course, the power to detect some effect Δ increases with the sample size

- This implies that any effect Δ, no matter how small, will, sooner or later, be detected, if the sample is sufficiently large.

- For example, consider the Captopril data, where the observed difference of 9.27 mmHg was found significantly different from zero $(p < 0.001)$, based on data from 15 patients only:



Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	p
DIA_VOOR	112,3333	10,47219				
DIA_NA	103,0667	12,55540	15	9,266667	8,614495	0,000951

- Suppose that the observed difference would have been 0.1 mmHg.

- A p-value as small as 0.001 would be likely to be obtained, provided that the sample would be sufficiently large.

- Obviously, an average change in BP as small as 0.1 mmHg is not relevant from a clinical point of view.

- Conclusion:

Statistical significance $\neq$ **Clinical relevance**

- The p-value cannot distinguish between both situations

- It is therefore important not to blindly overinterpret significant results without knowing the size of the effect

1.4 Didn't We Know All That?

- Yes, but the discussion continues

> **Scientists rise up against statistical significance**
>
> **Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.**

- Relevance $\simeq$ Meaningful effect size

- One needs good understanding, by a statistician, to disentangle the issues:
  - ▷ Proper interpretation of effects, statistically and scientifically
  - ▷ Improper behavior (cf. **non-uniformity of p-values**)

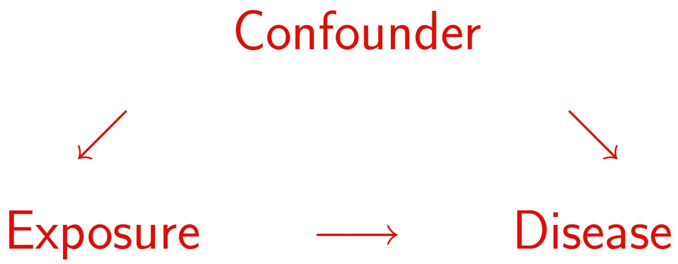
- Cf. Bolland *et al.* paper: Fujii case & Sato cases & . . . case

- Implausible similarity of numbers discussed:
  - ▷ Implausibly similar p-values
  - ▷ Implausibly similar baseline characteristics
  - ▷ “Implausibility” of. . .

- What is required:
  - ▷ Detection systems:
  - ▷ Legal and other procedures
  - ▷ Reflection on the systems that lead to all this

- **Multiple testing** has related problems

1.5 Observational Studies: Environment and Health



- **Smoking & lung cancer:** tobacco industry versus the states of the US

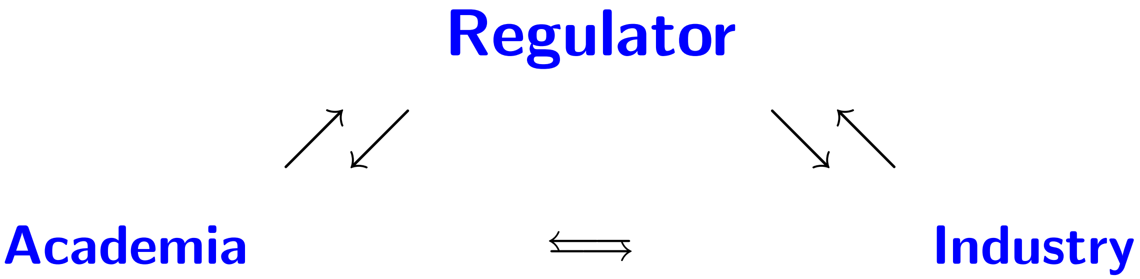
1.6 Further Problems with Experiments: Psychology, Sociology, Economy, Medicine, Exact Science...

! Not done: invention of studies/study results

? Not done: removal of study subjects that do not fit well with the rest

- 67,138 BEF (= €1664.31 = $ 1886.26)

1.7 An Experimental Setting: Randomized Clinical Trials



Regulator

Academia ⟺ **Industry**

- Phases of clinical research
- Randomization
- Blinding
- Various committees:
  - **Institutional Review Board**
  - **Data Monitoring Committee**
- Informed consent

Theme 2
Checking Data

2.1 Cleverly Looking At Data

- A key point: **Variability** structure
- Single variables ⟷ multiple variables
- Subsets reproducible via multiple imputation — if not, maybe mechanism is “fishy”
- Fractions between quantiles

Theme 3
Journals & Measurements, a Good Idea?

3.1 Peer Review is Watching You

	Who or what?	By whom or what?
1	**Scientific texts**	journal, publisher
2	**Grant applications**	granting agency
3	**Assessment of scientist**	Employer
4	**Assessment of entity**	Government

3.2 The Impact Factor

$$\text{I.F.} = \frac{\#\ \text{citations in 2018 of articles published in 2016–2017}}{\#\ \text{articles published in 2016–2017}}$$

- **Well defined?**

- **Even if well defined. . .**

3.3 Well defined?

- Scientific area with n papers, where everyone cites everyone:

$$2\binom{n}{2} = \frac{2n(n-1)}{2} = n(n-1)$$

- Corrected yes, but only linearly!

$\Longrightarrow$ **advantage for large fields**

$\Longrightarrow$ **advantage for areas that self-cite a lot**

- Half life very different!

  In spite of there being a 5-year version.

3.4 “I, Poor Journal Editor”

“It is less useful to compare areas, but very useful within a given area!”

- “Oh, really?”
  - ▷ Mathematics ⟷ statistics
  - ▷ Theoretical statistics ⟷ applied statistics
  - ▷ *Biometrics* (1.86) ⟷ *Statistics in Medicine* (1.99)
- Statistics in Medicine has got a high shelf life!

“At every rate, the I.F. must go up!”

- “Rather genetic statistics than classical medical statistics.”
- “Rather data science than statistics.”

- “Could you please cite three more articles from our journal before we accept your manuscript?”

- *“We especially welcome review papers.”*

3.5 "I, Poor Research Coordinator"

- . . . "How can we score well on all of these measures?"

- "We preferably make (professorial) appointments in high I.F. areas."

$\Longrightarrow$ State of equilibrium:

University with solely a Faculty of Medicine!

- **Exercise:** Translate this to industry

3.6 “I, Poor Researcher”

- **No go:**
  - ▷ Risky and/or time-consuming research
  - ▷ Controversial papers
  - ▷ Books: *book with 1700 citations, top article with 200 citations. . .*
  - ▷ Bye Monseigneur Lemaître
  - ▷ Bye Andrew Wiles:

    Fermat’s Last Theorem really was the last. . .

- **Yes:**
  - ▷ ε incremental research (ε)
  - ▷ “I don’t want to be a reviewer, but want my papers reviewed quickly.”

3.7 Where Do We Go From Here?

- I.F. not well defined

$\longrightarrow$ almost surely problems with whatever **summary** measure

- Even if well defined: perverse behavior:

$\Delta p \cdot \Delta r \geq \overline{h}$: **The measure influences what is being measured**

The Heisenberg Uncertainty Principle of Peer Review

Theme 4
Case Study: The Toenail Data

- **T**oenail **D**ermatophyte **O**nychomycosis: Common toenail infection, difficult to treat, affecting more than 2% of population.

- Classical treatments with antifungal compounds need to be administered until the whole nail has grown out healthy.

- New compounds have been developed which reduce treatment to 3 months

- Randomized, double-blind, parallel group, multicenter study for the comparison of two such new compounds (A and B) for oral treatment.

- Research question:

Severity relative to treatment of TDO ?

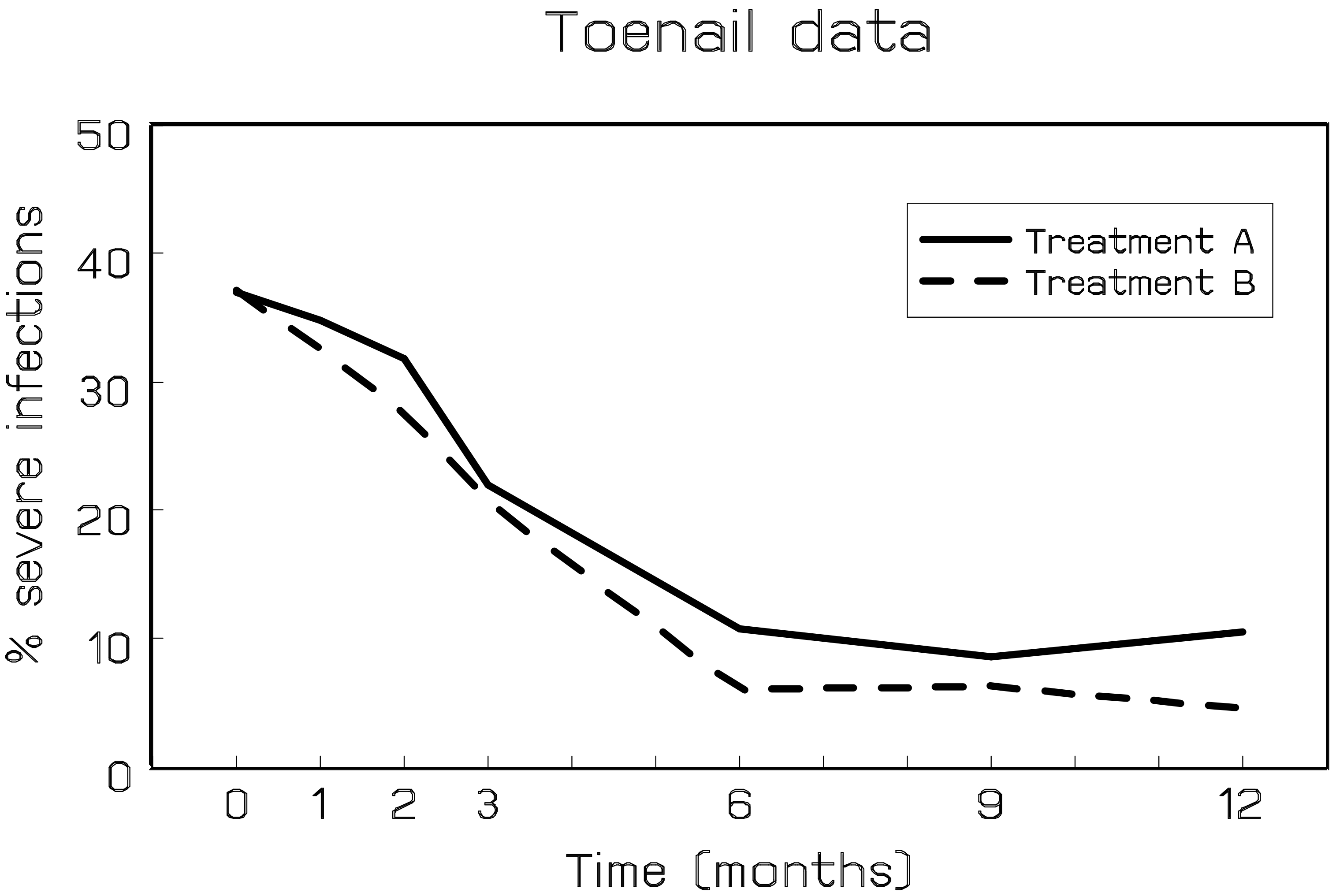
- 2×189 patients randomized

- 48 weeks of total follow up (12 months)

- 12 weeks of treatment (3 months)

- measurements at months 0, 1, 2, 3, 6, 9, 12.

- Frequencies at each visit (both treatments):



4.1 Application to the Toenail Data

- Consider the model:

$$
\begin{aligned}
Y_{ij} &\sim \text{Bernoulli}(\mu_{ij}) \\
\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) &= \beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \boxed{\beta_3 T_i t_{ij}} \\
\text{Corr}(Y_{ij}, Y_{ij'}) &= \alpha \qquad \text{(working correlation)}
\end{aligned}
$$

- Y_{ij}: severe infection (yes/no) at occasion j for patient i

- t_{ij}: measurement time for occasion j

- T_i: treatment group

4.1.1 Inference on Key Parameter: β_3. Story 1.

Model	Estimate (s.e.)	p-value
Initial model	-0.0783 (0.0394)	0.0469
Model-based (naive)	-0.0886 (0.0362)	0.0143
Empirically corrected (robust)	-0.0886 (0.0571)	0.1208

"The initial model is the most efficient estimator, because it assumes that each data point provides an independent piece of information. Based on this model, the treatment effect is marginally significant."

4.1.2 Inference on Key Parameter: β_3. Story 2.

Model	Estimate (s.e.)	p-value
Initial model	-0.0783 (0.0394)	0.0469
Model-based (naive)	-0.0886 (0.0362)	0.0143
Empirically corrected (robust)	-0.0886 (0.0571)	0.1208

“The model-based estimator assumes that the various pairs of measurements per patient exhibit a common correlation. This is estimated to be $\widehat{\alpha} = 0.42$, considered to be a plausible value. Therefore, inferences are based on the model-based estimator; this leads to a significant effect of treatment, with $p = 0.0143$.”

4.1.3 Inference on Key Parameter: β_3. Story 3.

Model	Estimate (s.e.)	p-value
Initial model	-0.0783 (0.0394)	0.0469
Model-based (naive)	-0.0886 (0.0362)	0.0143
Empirically corrected (robust)	-0.0886 (0.0571)	0.1208

"The empirically-corrected estimator assumes that the various pairs of measurements per patient exhibit a common correlation, but that, at the same time, this correlation assumption may be incorrect. In other words, it protects against misspecification. Inferences are based on this estimator. We conclude that there is no significant effect of treatment, with $p = 0.1208$*."*

4.1.4 Inference on Key Parameter: β_3. Story 4.

Model	Working corr. α	Estimate (s.e.)	p-value
Initial model		-0.078 (0.039)	0.0469
Model-based (naive)	exchangeable	-0.089 (0.036)	0.0143
Emp. corr. (robust)	independence	-0.078 (0.055)	0.1515
Emp. corr. (robust)	exchangeable	-0.089 (0.057)	0.1208
Emp. corr. (robust)	unstructured	-0.114 (0.052)	0.0275

"The empirically-corrected estimator assumes that the various pairs of measurements per patient exhibit a certain structure, but that, at the same time, this correlation assumption may be incorrect. The working correlation that is closest to the true structure is generally most efficient. Inferences are based on this estimator, with unstructured working correlation. We conclude that there is a significant effect of treatment, with $p = 0.0275$*."*

4.2 The Generalized Estimating Equations Case: Discussion

- Nice method to efficiently and correctly analyze non-Gaussian longitudinal data

- But: there are pitfalls

  ▷ Initial $\longleftrightarrow$ Model-based $\longleftrightarrow$ Empirically corrected

  ▷ Various working correlation structures possible: **how** and **when** to choose?

- **Know the method! Know the pitfalls! Stand firm on principles!**

Theme 5
Multiple Imputation

- Incompletely observed repeated measures

- A procedure gaining a lot of clout,. . .

- Three steps:

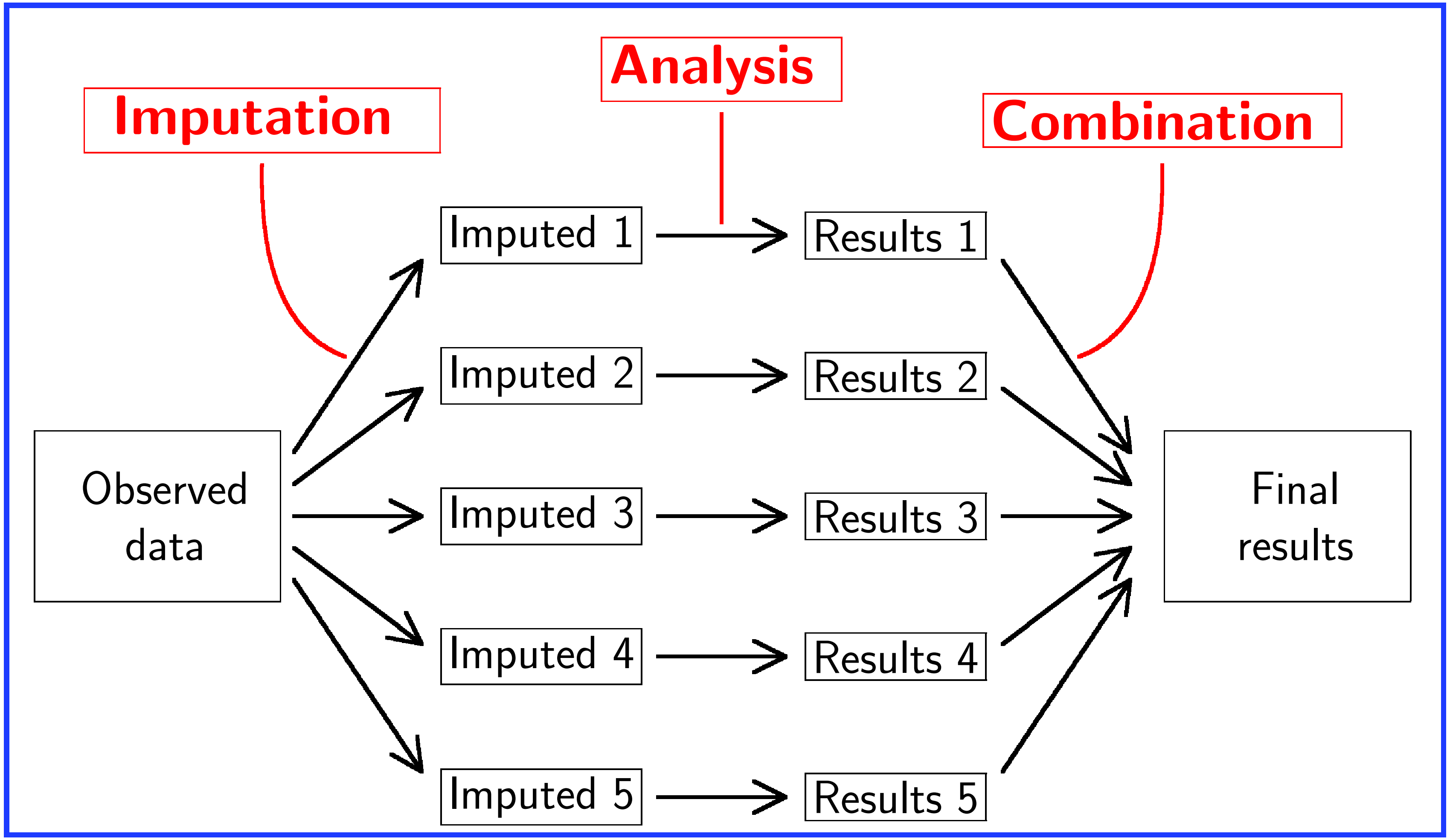
  1. The missing values are **sampled** M times $\Longrightarrow$ M complete data sets

  2. The M complete data sets are analyzed by using standard procedures

  3. The results from the M analyses are combined into a single inference

- Rubin (1987), Rubin and Schenker (1986), Little and Rubin (1987)

- Multiple imputation ($M = 5$ imputations):



5.1 Multiple Imputation: Ticket to Fraud?

- Code for imputations:

```
proc mi data=armd13 seed=486048 out=armd13a simple nimpute=10 round=0.1;
   var lesion diff4 diff12 diff24 diff52;
   by treat;
   run;
```

- ?

- **seed=486048**

- !