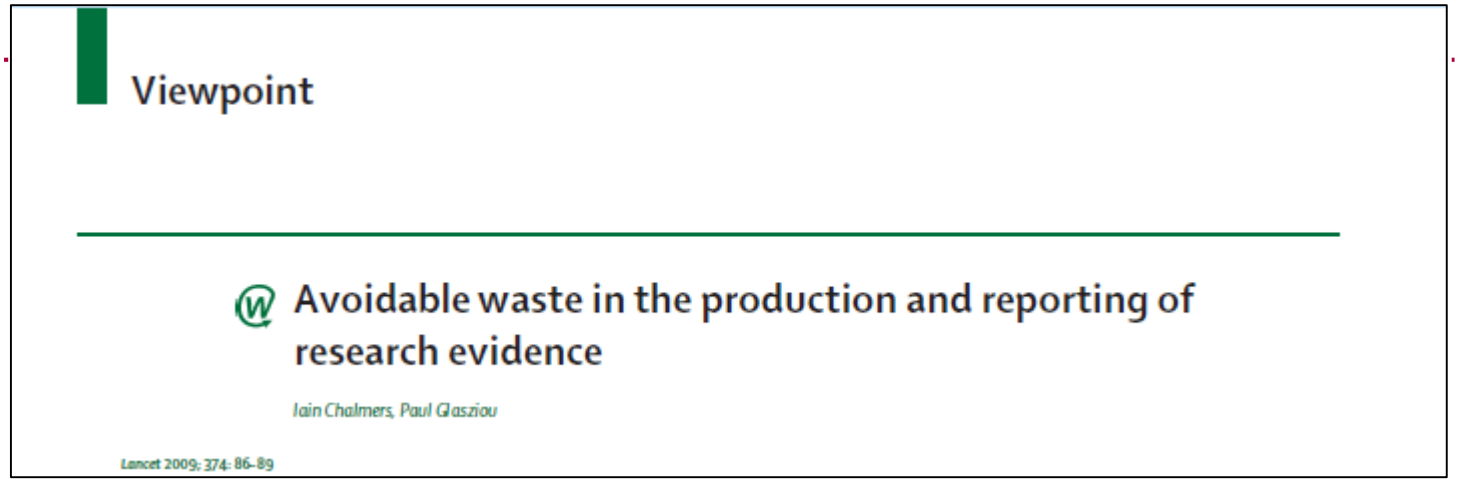
CWI Centrum Wiskunde & Informatica

# The rules of the game called significance testing

**Judith ter Schure**

**CWI, Machine Learning group**



Viewpoint

# Avoidable waste in the production and reporting of research evidence

Iain Chalmers, Paul Glasziou

Lancet 2009; 374: 86–89

Viewpoint

## Avoidable waste in the production and reporting of research evidence

Iain Chalmers, Paul Glasziou

Lancet 2009; 374: 86–89

Series

Research: increasi

How to increase v
priorities are set

Iain Chalmers, Michael B Bracken, Ben Djulbegovic, Silvio Garattini, Jonathan Grant, A Metin Gülmezoglu, David W Howells, John P A Ioannidis, Sandy Oliver

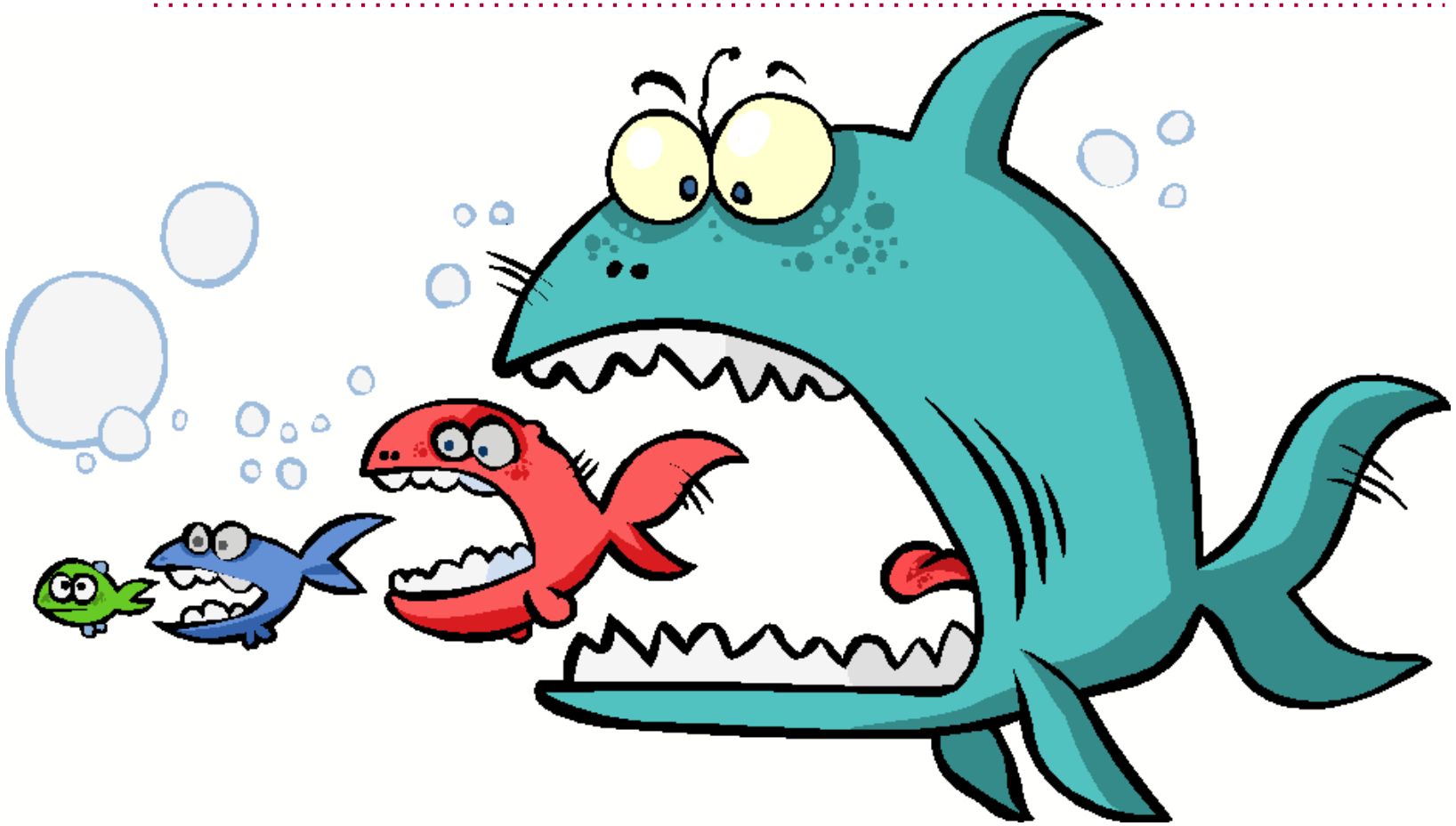
Lancet 2014; 383: 156–65

Research: increasing value, reducing waste 5

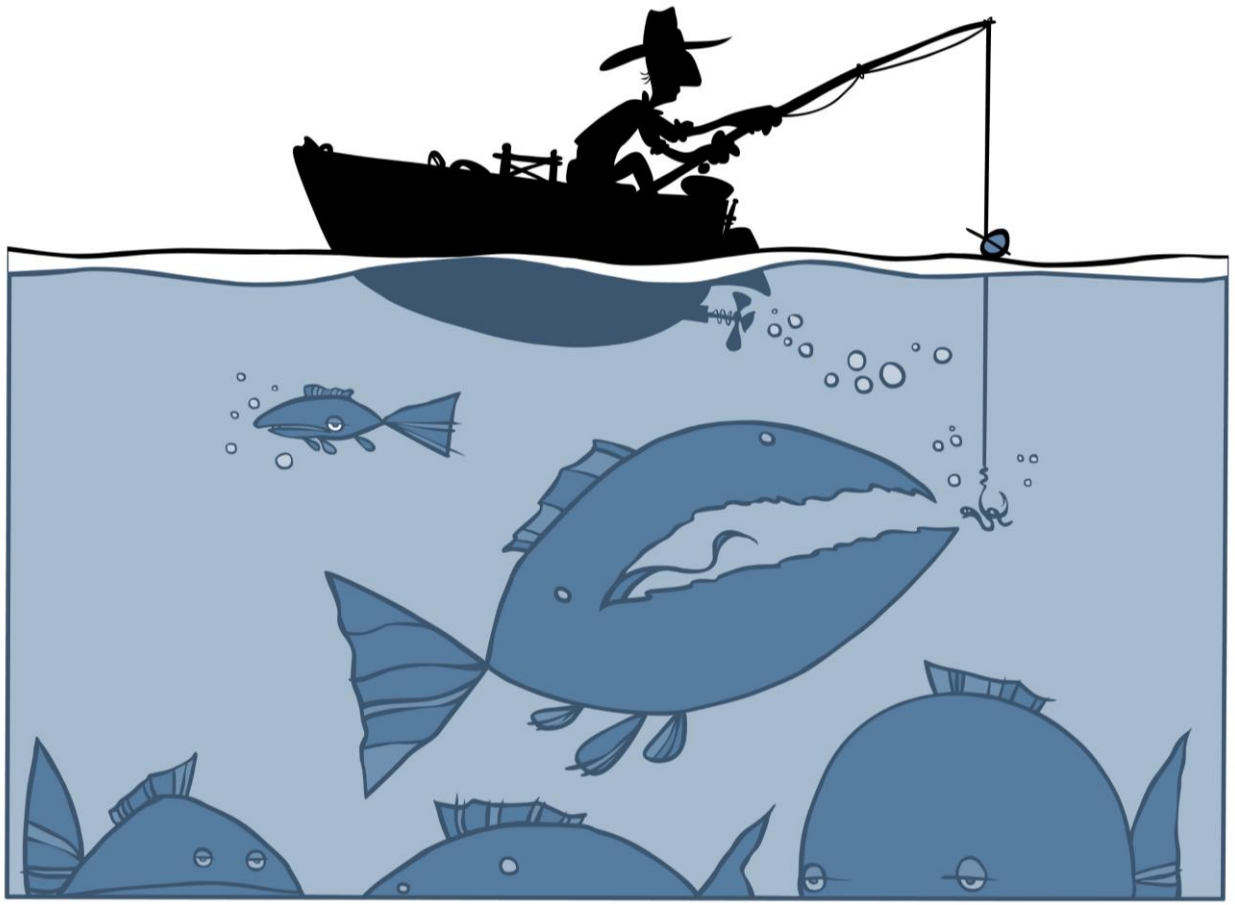
## Reducing waste from incomplete or unusable reports of biomedical research

Paul Glasziou, Douglas G Altman, Patrick Bossuyt, Isabelle Boutron, Mike Clarke, Steven Julious, Susan Michie, David Moher, Elizabeth Wager

# Size is dependent on intermediate results



# Timing is dependent on intermediate results



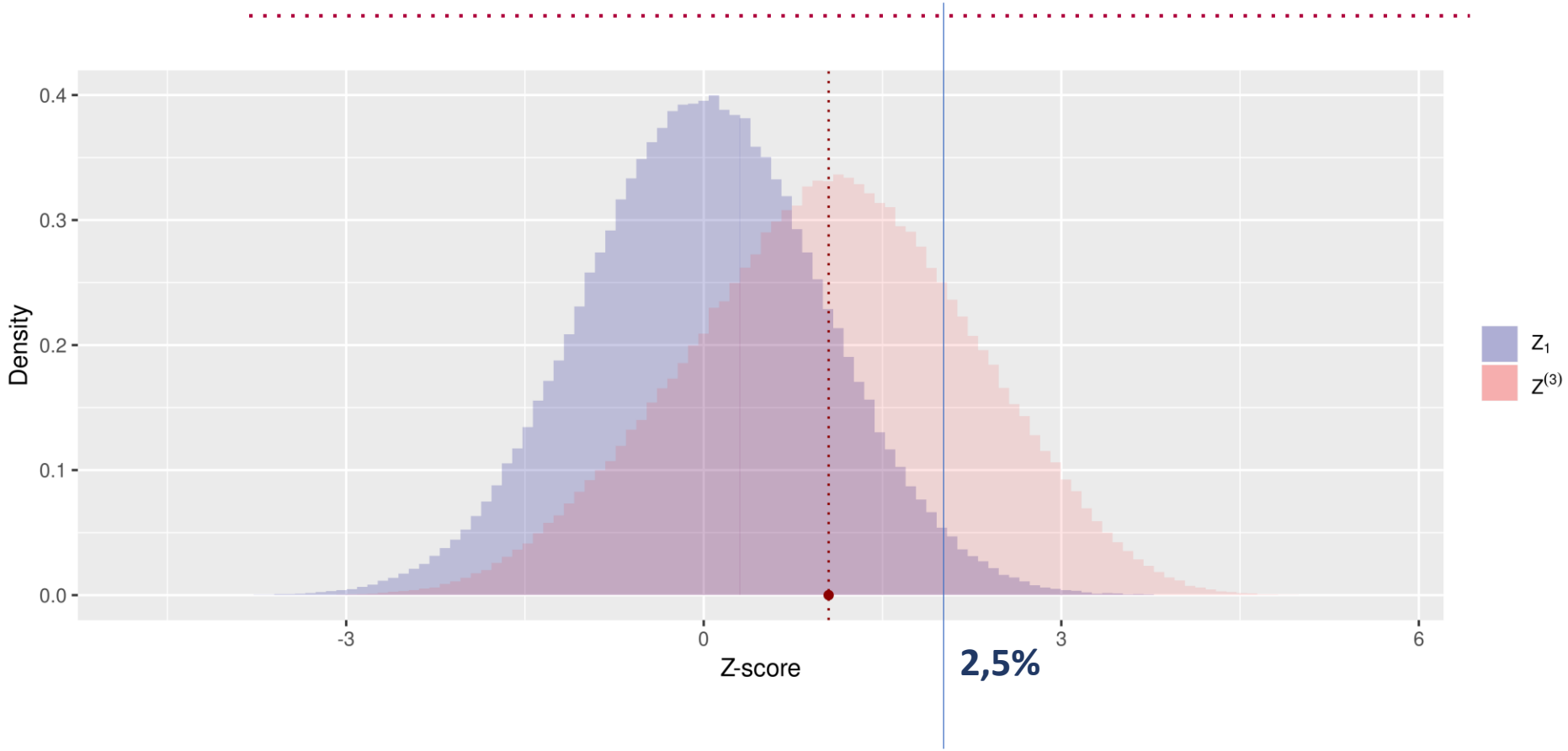
CWI Centrum Wiskunde & Informatica

# Accumulation Bias

*Judith ter Schure - 1 October 2019*

# Accumulation Bias

# Inflated error rates



BMC Medical Research Methodology

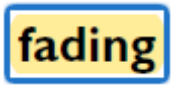
BioMed Central

Research article

Open Access

# The fading of reported effectiveness. A meta-analysis of randomised controlled trials

Bernhard T Gehr[1], Christel Weiss[2] and Franz Porzsolt*[3]


fading

# Correspondence

fading

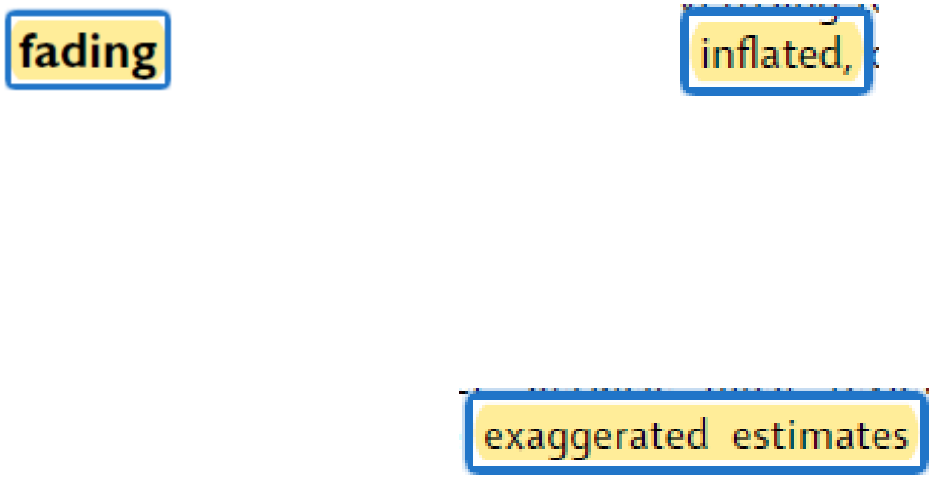
## How systematic reviews cause research waste

Systematic reviews of small trials increase waste by ac[illegible] to the scientific community inflated, often significant treatment effects that become smaller or absent when large, high-quality trials are done. Effect estimates from systematic reviews often inform sample size calculations.[4] However, because most reviews provide exaggerated estimates of treatment effects due to inclusion

We declare no competing interests.

*Ian Roberts, Katharine Ker
**Ian.roberts@lshtm.ac.uk**

Clinical Trials Unit, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK


fading
inflated,
exaggerated estimates

Journal of Cardiac Failure Vol. 9 No. 5 2003

Perspective

# Why do Phase III Trials of Promising Heart Failure Drugs Often Fail? The Contribution of 'Regression to the Truth'

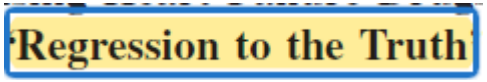
HENRY KRUM, MBBS, PHD, FRACP, ANDREW TONKIN, MD, FRACP

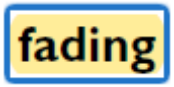
*Victoria, Australia*

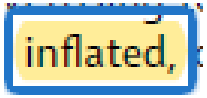
fading

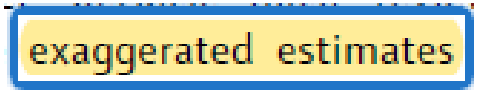
inflated,

exaggerated estimates


Regression to the Truth


fading


inflated,


exaggerated estimates

ELSEVIER

Journal of Clinical Epidemiology 58 (2005) 543–549

Journal of Clinical Epidemiology

REVIEW ARTICLES

# Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials

John P.A. Ioannidis[a,b,c,*], Thomas A. Trikalinos[a,b]

'Regression to the Truth'

fading

inflated,

exaggerated estimates

# Correspondence

## How systematic reviews cause research waste

Systematic reviews of small trials increase waste by advertising to the scientific community inflated, often significant treatment effects that become smaller or absent when large, high-quality trials are done. Effect estimates from systematic reviews often inform sample size calculations.[4] However because most reviews provide exaggerated estimates of treatment effects due to inclusion

We declare no competing interests.

*Ian Roberts, Katharine Ker
**Ian.roberts@lshtm.ac.uk**

Clinical Trials Unit, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK

# Correspondence

## Systematic reviews and research waste

We share Ian Roberts' and Katharine Ker's frustration with the poor quality of much research (Oct 17,

their suggestion scientifically flawed, it is also unrealistic. Funders and regulators cannot be expected to support and endorse large studies without some reassurance from the results of smaller existing studies that the substantial investment needed is justified. We hope that, instead of promoting their

*Iain Chalmers, Paul Glasziou
ichalmers@jameslind.net

## Correspondence

Early extreme contradictory estimates

The Proteus phenomenon

'Regression to the Truth'

fading

**ystematic reviews and research waste**

We share Ian Roberts' and Katharine Ker's frustration with the poor quality of much research (Oct 17,

inflated,

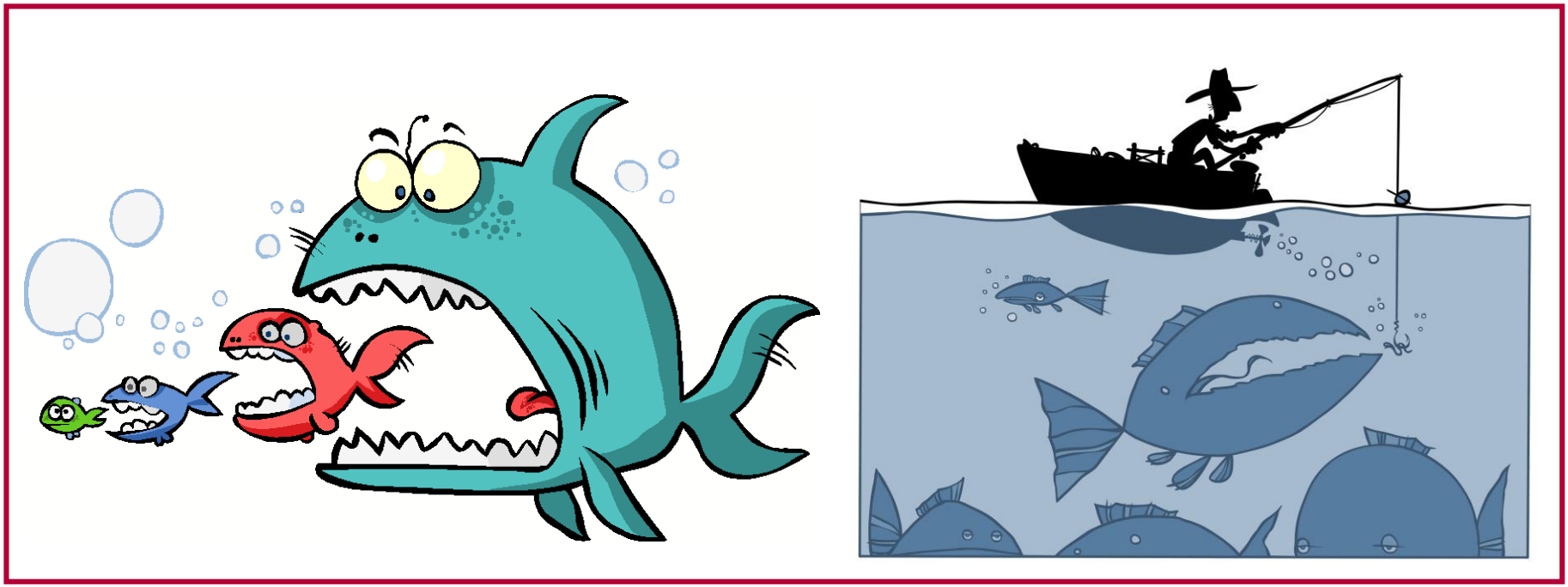
eir suggestion scientifically flawed, it also unrealistic

Funders and regulators cannot be expected to support and endorse large studies without some reassurance from the results of smaller existing studies that the substantial investment needed is justified. We hope that, instead of promoting their

exaggerated estimates

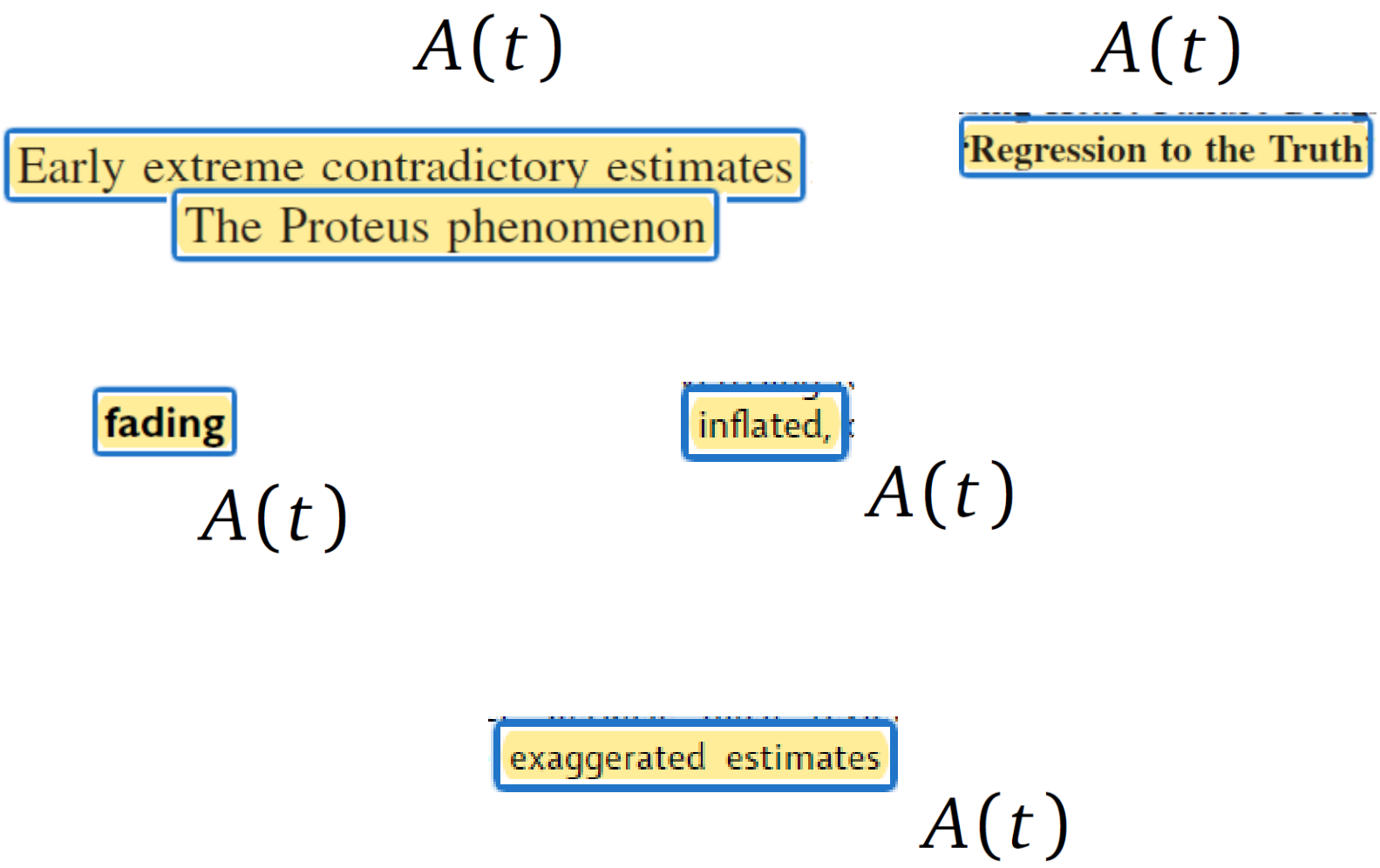
Iain Chalmers, Paul Glasziou
ichalmers@jameslind.net

# Accumulation Bias framework



# Accumulation Bias framework

$$A(t)$$


$A(t)$
$A(t)$
Early extreme contradictory estimates
The Proteus phenomenon
Regression to the Truth
fading
$A(t)$
inflated,
$A(t)$
exaggerated estimates
$A(t)$

# Correspondence

$$A(t)$$

## Systematic reviews and research waste

We share Ian Roberts' and Katharine Ker's frustration with the poor quality of much research (Oct 17,

their suggestion scientifically flawed, it is also unrealistic. Funders and regulators cannot be expected to support and endorse large studies without some reassurance from the results of smaller existing studies that the substantial investment needed is justified. We hope that, instead of promoting their

*Iain Chalmers, Paul Glasziou
ichalmers@jameslind.net

# So, bias right?

# How come we can still do inference?

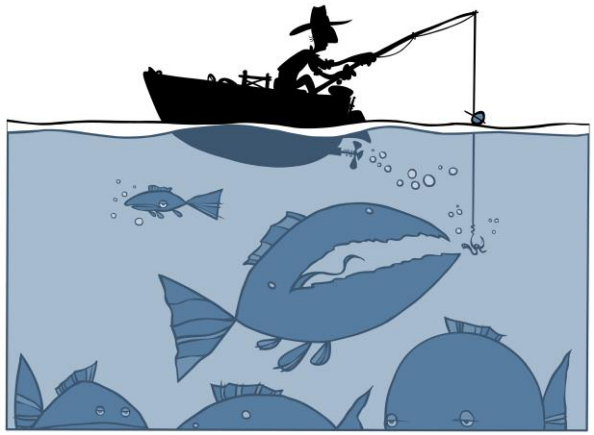
$$\mathbf{LR}_{10}{}^{(t)}$$

$$\mathbf{LR}_{10}^{(t)} = \frac{f_1(z_1, \ldots, z_t) \cdot A(t \mid z_1, \ldots, z_t)}{f_0(z_1, \ldots, z_t) \cdot A(t \mid z_1, \ldots, z_t)}$$

$$\mathbf{LR}_{10}^{(t)} = \frac{f_1(z_1,\ldots,z_t)\cdot A(t\,|\,z_1,\ldots,z_t)}{f_0(z_1,\ldots,z_t)\cdot A(t\,|\,z_1,\ldots,z_t)} = \frac{f_1(z_1,\ldots,z_t)}{f_0(z_1,\ldots,z_t)}$$

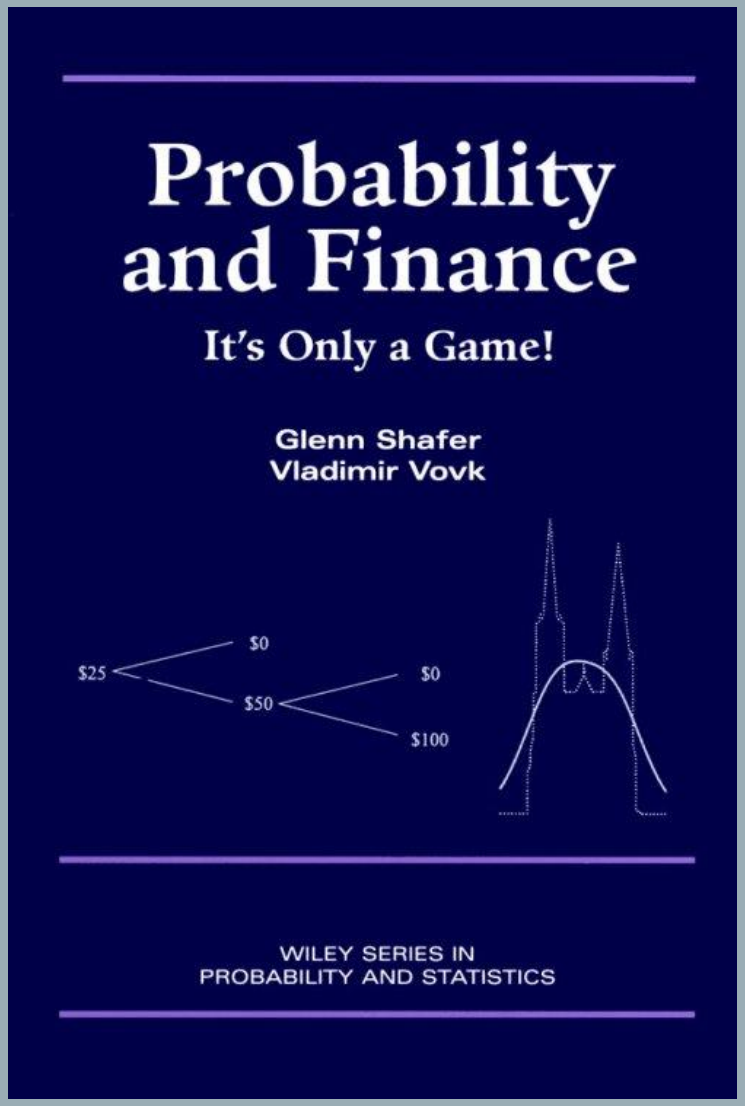
$$= \mathbf{LR}_{10}(z_1,\ldots,z_t).$$

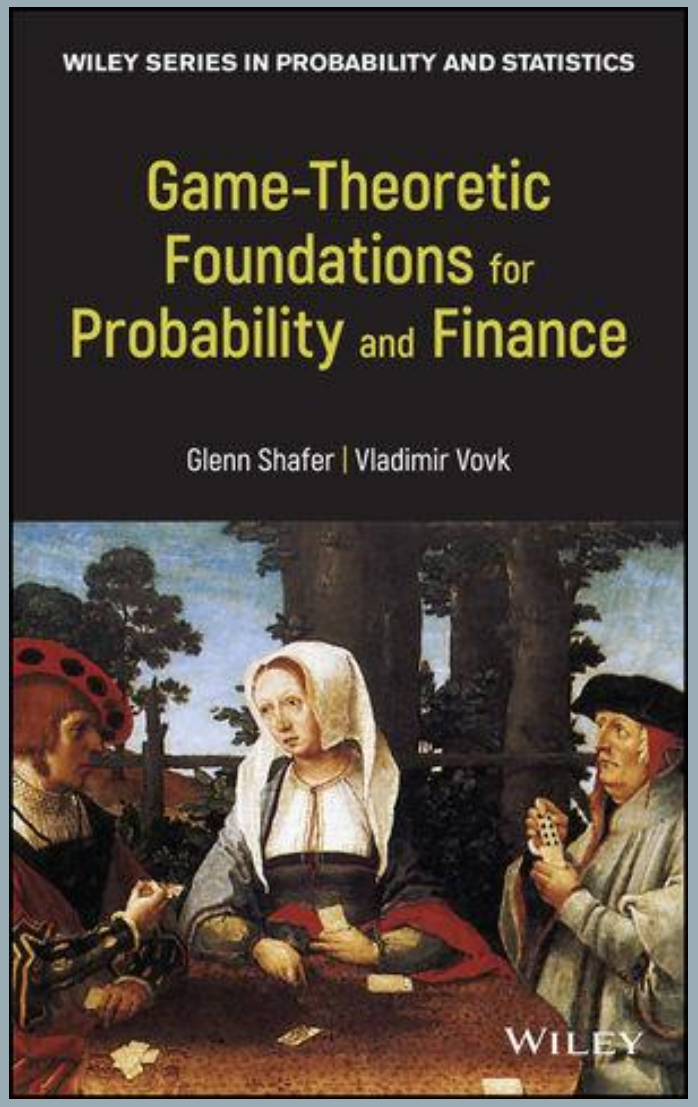
- ter Schure J and Grünwald P. Accumulation Bias in meta-analysis: the need to consider *time* in error control [version 1; peer review: 1 approved]. *F1000Research* 2019, **8**:962 (https://doi.org/10.12688/f1000research.19375.1)



# Accumulating Science needs Accumulating Tests

*Judith ter Schure - 1 October 2019*


Probability and Finance
It's Only a Game!
Glenn Shafer
Vladimir Vovk
$25
$0
$50
$0
$100
WILEY SERIES IN
PROBABILITY AND STATISTICS


WILEY SERIES IN PROBABILITY AND STATISTICS
Game-Theoretic Foundations for Probability and Finance
Glenn Shafer | Vladimir Vovk
WILEY

# Neyman Pearson *all-or-nothing* bet

$\$_{NP} = 0$ if *p-value* $> \alpha$

$\$_{NP} = \frac{1}{\alpha}$ if *p-value* $\leq \alpha$

see

- Grünwald, De Heide & Koolen (2019)
- Shafer (2019)
  and for the *all-or-nothing* strategy vs Kelly gambling, see:
- Tijms (2018)

# Accumulating Science needs Likelihood Ratios

*Judith ter Schure - 1 October 2019*

# How to define a bet

Fair bet: $\mathbf{E}_0[s(Z_t)] = \$1.$

Reinvesting $s(Z_1) \cdot s(Z_2) \cdot s(Z_3)$

# How to define a bet: likelihood ratio!

Fair bet: $\mathbf{E}_0[s(Z_t)] = \$1.$

$$\text{if we set } s(Z_t) = f_1(Z_t)/f_0(Z_t)$$

$$\mathbf{E}_0\left[\frac{f_1(Z_t)}{f_0(Z_t)}\right] = \int_z f_0(z)\frac{f_1(z)}{f_0(z)}dz = \int_z f_1(z)dz = 1.$$

Reinvesting (i.i.d. studies!): $s(Z_1) \cdot s(Z_2) \cdot s(Z_3)$

# How to turn the bet into a test

$$\mathbf{E}_0[s(Z_t)] = \$1.$$

Markov: $$\mathbf{P}_0\left[s(Z_t) \geq \frac{1}{\alpha}\right] \leq \frac{\mathbf{E}_0[s(Z_t)]}{\frac{1}{\alpha}} = \alpha$$

And with likelihood ratios you never go bankrupt!

CWI Centrum Wiskunde & Informatica

# Safe Tests generalize Likelihood Ratios to common test scenarios

- T-test (one-sided, two-sided)
- Test of two proportions (contingency tables)
- Meta-Analysis under heterogeneity

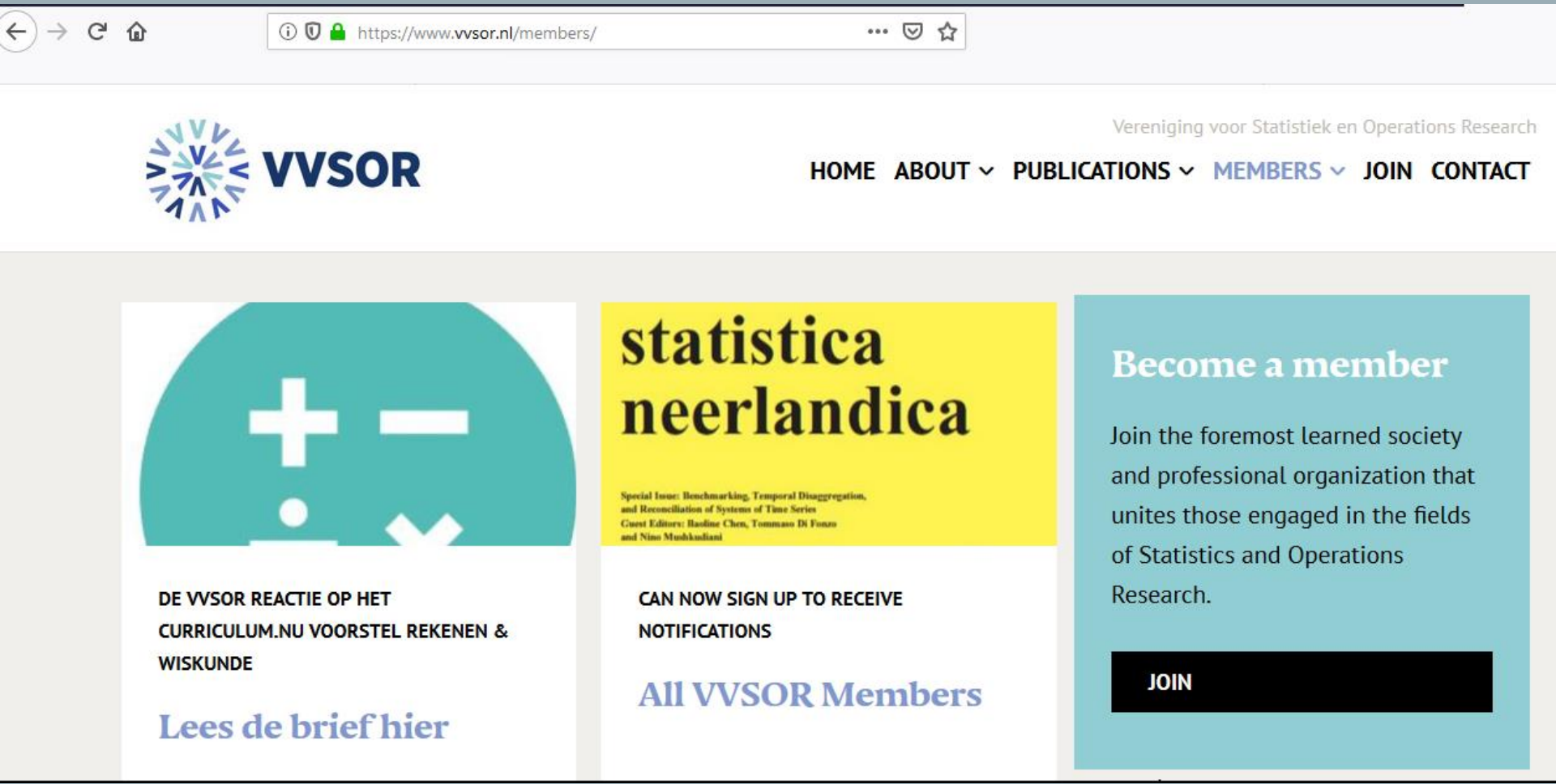
*Judith ter Schure - 1 October 2019*

# Would this change researcher's mindset?

# Background reading (*more papers coming up*)


- ter Schure J and Grünwald P. Accumulation Bias in meta-analysis: the need to consider *time* in error control [version 1; peer review: 1 approved]. *F1000Research* 2019, **8**:96
- Grünwald, P., de Heide, R., & Koolen, W. (2019). Safe testing. *arXiv preprint arXiv:1906.07801*.
- Shafer, G. (2019). The language of betting as a strategy for statistical and scientific communication. *arXiv preprint arXiv:1903.06991*. (most current version: http://probabilityandfinance.com/articles/54.pdf)
- Grünwald, P. D. (2015). Paranormale Statistiek: over de vele problemen met p-waarden, en een redelijk alternatief. StatOR (https://www.vvsor.nl/wp-content/uploads/2018/01/STAtOR_2015-3_totaal.pdf )
- Grünwald, P. D. (2016). Toetsen als gokken: een redelijk alternatief voor de p-waarde. *NAW*, *5*(17), 4. (https://homepages.cwi.nl/~pdg/ftp/naw.pdf )
- Tijms, H. (2018) Investeren en wedden met Kelly. StatOR (https://www.vvsor.nl/wp-content/uploads/2018/12/STAtOR-2018-4-1-40-LR-spreads.pdf)

# Become a member!



# Some conclusions

- All science is biased (if series > pilot study).
  BUT THAT'S OK.

# Some conclusions

- All science is biased (if series > pilot study).
  BUT THAT'S OK.

- Conventional meta-analysis is invalid

# Some conclusions

- All science is biased (if series > pilot study).
  BUT THAT'S OK.
- Conventional meta-analysis is invalid
- The language of betting + reinvesting might be a way out of p-hacking and filedrawering
  + restore meta-analysis and reduce waste

# Some conclusions

- All science is biased (if series > pilot study).
  BUT THAT'S OK.
- Conventional meta-analysis is invalid
- The language of betting + reinvesting might be a way out of p-hacking and filedrawering
  + restore meta-analysis and reduce waste

**Contact me at: schure@cwi.nl**