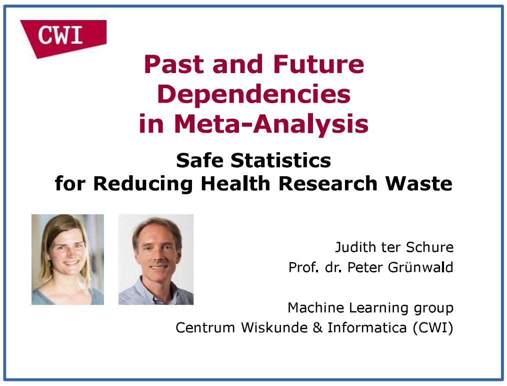


Past and Future Dependencies in Meta-Analysis

Safe Statistics for Reducing Health Research Waste

Judith ter Schure
Prof. dr. Peter Grünwald

Machine Learning group
Centrum Wiskunde & Informatica (CWI)

From the point of view of Machine Learning, the replication crisis is partly a failure to learn from accumulating data. This vision is shared with a group of health science advocates - among which those that founded the Cochrane Collaboration - that addressed the problem of 'Health Research Waste'. In this context, I would like to introduce methods that have been developed in our group, that I think are essential to reduce health research waste.

85% of Health Research Investment is wasted

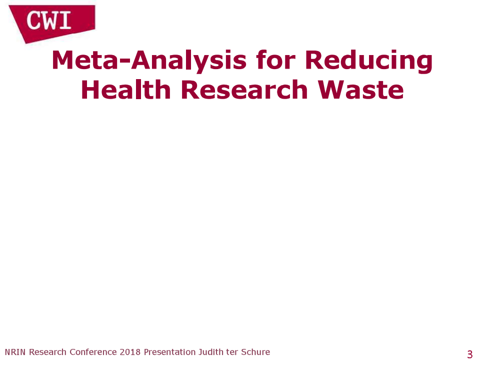
Viewpoint

Avoidable waste in the production and reporting of research evidence

Iain Chalmers, Paul Glasziou

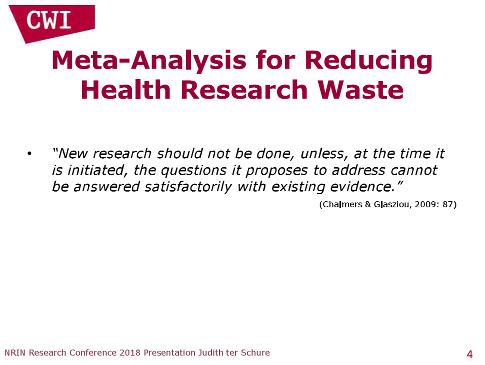
Lancet 2009; 374: 86–89

The term 'Health Research Waste' was coined in 2009 in a paper published in The Lancet claiming that 85% of Health Research Investment is wasted.



Meta-Analysis for Reducing Health Research Waste

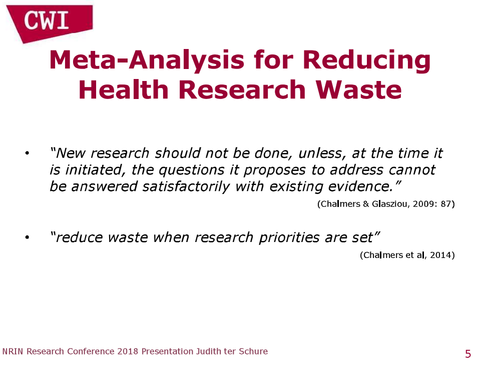
One reason for this research waste is a lack of systematic reviews and meta-analyses. Therefore, one of the paper's main recommendations was to rely more on systematic reviews and meta-analysis in the health research process.



Meta-Analysis for Reducing Health Research Waste

- *"New research should not be done, unless, at the time it is initiated, the questions it proposes to address cannot be answered satisfactorily with existing evidence."* (Chalmers & Glasziou, 2009: 87)

So what these authors propose is cumulative meta-analysis: Meta-analysis on all available evidence before designing a new study means meta-analysis after each previous study.



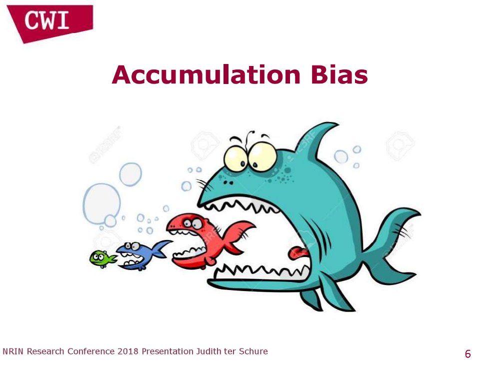
CWI

Meta-Analysis for Reducing Health Research Waste

- *"New research should not be done, unless, at the time it is initiated, the questions it proposes to address cannot be answered satisfactorily with existing evidence."* (Chalmers & Glasziou, 2009: 87)
- *"reduce waste when research priorities are set"* (Chalmers et al, 2014)

This recommendation was still quite urgent in 2014 when they published an entire series of papers on it. In 2014, the same recommendation was phrased as "setting research priorities" based on systematic reviews and meta-analyses.

But there is a problem..



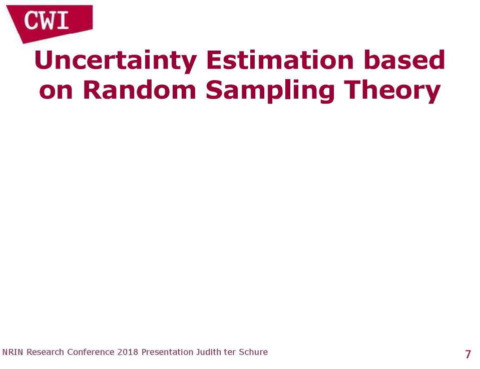
CWI

Accumulation Bias

This is a phenomenon in meta-analysis that hasn't been described before and that I gave the name 'Accumulation Bias'. It occurs when meta-analyses play a role in accumulating science, and that's exactly what causes it:

This is accumulating science clashing with the fundamentals of statistics.

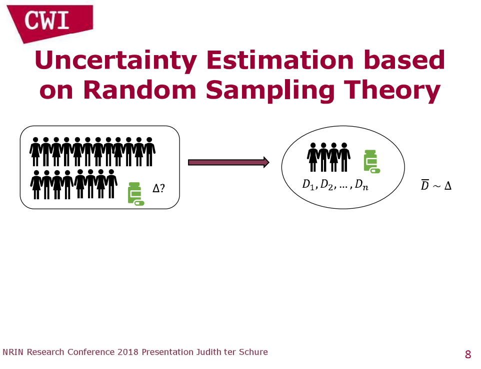
So let's have a look at the fundamental statistical machinery…



CWI

Uncertainty Estimation based on Random Sampling Theory

… which is random sampling theory.

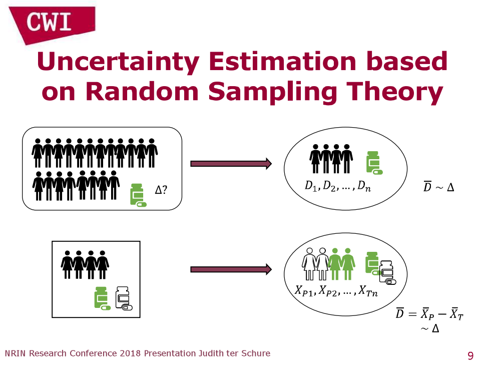


CWI

Uncertainty Estimation based on Random Sampling Theory

In the application of health research, this theory describes how to think about populations and samples of patients:

There is a population with all possible patients with a certain disease and you as a researcher wonder how these patients change if they were given a certain drug. Because you cannot observe that change directly, you sample a view patients, give them the drug, measure the change and then assume that the average change in your sample is somewhat related to the expected change in the population as a whole.

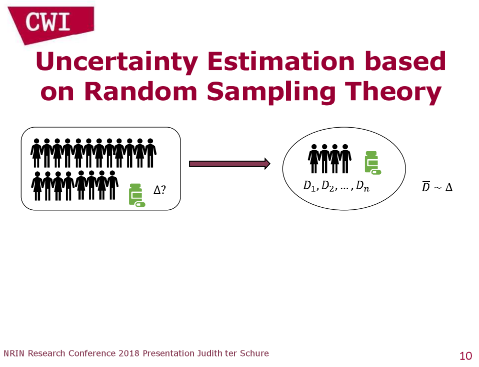


Uncertainty Estimation based on Random Sampling Theory

Δ?
$D_1, D_2, \ldots, D_n$
$\bar{D} \sim \Delta$
$X_{P1}, X_{P2}, \ldots, X_{Tn}$
$\bar{D} = \bar{X}_P - \bar{X}_T \sim \Delta$

NRIN Research Conference 2018 Presentation Judith ter Schure 9

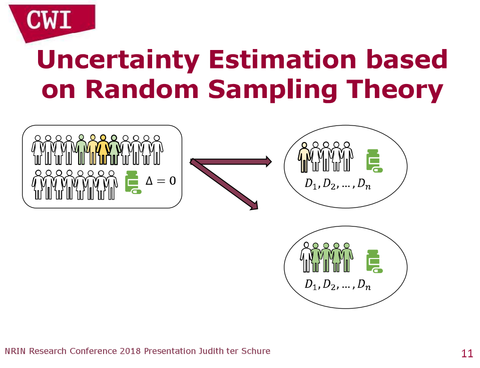
Of course in practice we do not randomly sample patients, but we randomize either placebo or real treatment.



Uncertainty Estimation based on Random Sampling Theory

NRIN Research Conference 2018 Presentation Judith ter Schure 10

But for now we can think about this in this upper simplified way.

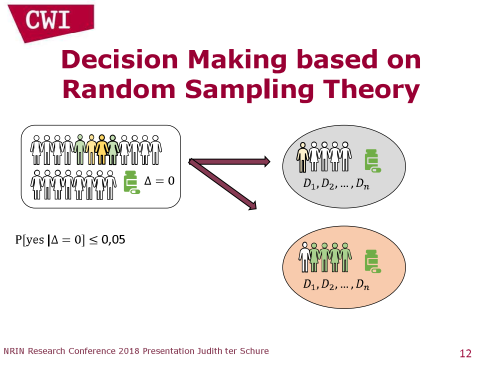


Uncertainty Estimation based on Random Sampling Theory

NRIN Research Conference 2018 Presentation Judith ter Schure 11

So how do we estimate uncertainty?

The uncertainty is in how the sample measurements relate to the population. We assume that the drug does not do anything to the patients in the population: Most do not change, some might improve a little bit, some might get a little bit worse due to background variation, but in expectation they don't change. Then either that is exactly what we observe in our sample, and we are not able tot conclude that the drug works, or we see, just by chance, very atypical measurements in our sample (a majority of improving patients) that incorrectly suggests that the drug does work.

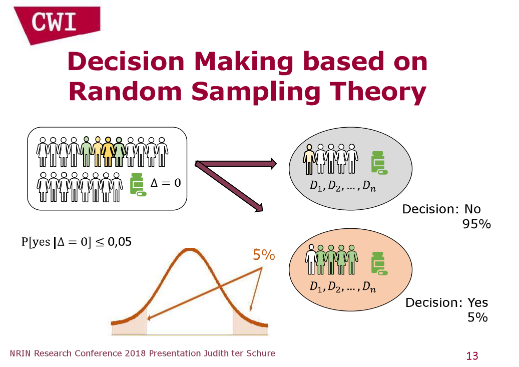


Decision Making based on Random Sampling Theory

$\mathrm{P}[\text{yes} \mid \Delta = 0] \leq 0{,}05$

NRIN Research Conference 2018 Presentation Judith ter Schure 12

When we want to make decisions under this uncertainty, we establish a threshold that decides how often we allow the atypical measurements to fool us.



CWI

Decision Making based on Random Sampling Theory

$P[\text{yes} \mid \Delta = 0] \leq 0{,}05$

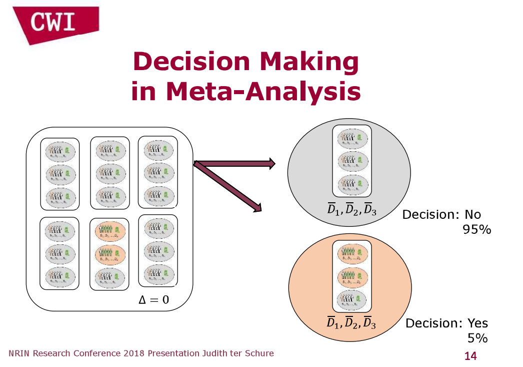
Decision: No
95%

Decision: Yes
5%

NRIN Research Conference 2018 Presentation Judith ter Schure

13

This threshold describes the tails of the sampling distribution of the test-statistic that we can calculate for each sample. If for a given sample of measurements the test statistic is inside this critical region, we're allowed to say 'Yes' to our data and reject the null hypothesis.



CWI

Decision Making in Meta-Analysis

$\Delta = 0$
$\bar{D}_1, \bar{D}_2, \bar{D}_3$
$\bar{D}_1, \bar{D}_2, \bar{D}_3$

Decision: No
95%

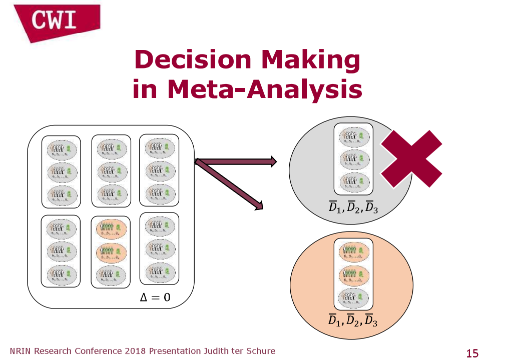
Decision: Yes
5%

NRIN Research Conference 2018 Presentation Judith ter Schure

14

So let's carry over this idea to meta-analysis. If we perform a meta-analysis on a sequence of three trials, we implicitly assume a population of three trial sequences, of which our sequence is a sample. If again, in the population the drug doesn't do anything, then most trials inside those sequences will show just that. But some samples will have seen atypical measurements and have found a significant improvement.

So we can use the random sampling machinery to distinguish the typical from the atypical sequences right?



CWI

Decision Making in Meta-Analysis

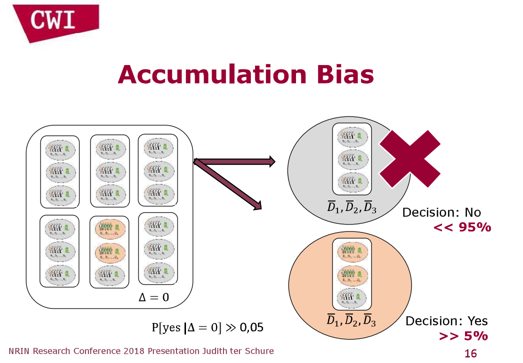
$\Delta = 0$
$\bar{D}_1, \bar{D}_2, \bar{D}_3$
$\bar{D}_1, \bar{D}_2, \bar{D}_3$

NRIN Research Conference 2018 Presentation Judith ter Schure

15

Except that we can't.

Because think about what such a sequence is in real science: It is a third trial being designed and performed, probably knowing what the first two trial results were. In that case, a third trial being performed when the first two showed no effect is less likely than random sampling theory suggests, and a third trial following two significant previous trials is more likely.



CWI

Accumulation Bias

$\Delta = 0$
$\bar{D}_1, \bar{D}_2, \bar{D}_3$
$\bar{D}_1, \bar{D}_2, \bar{D}_3$

$P[\text{yes} \mid \Delta = 0] \gg 0{,}05$

Decision: No
<< 95%

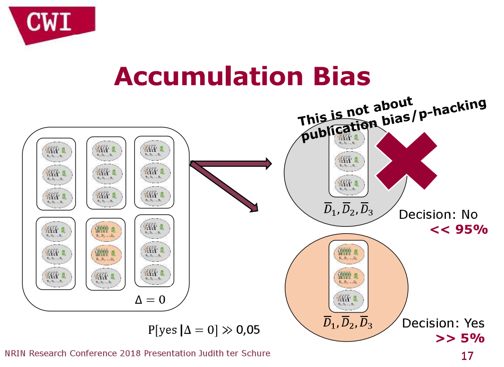
Decision: Yes
>> 5%

NRIN Research Conference 2018 Presentation Judith ter Schure

16

This results in bias in the expectation of the test statistic under the null hypothesis: Accumulation Bias.

And this bias inflates Type-I errors.



Accumulation Bias

This is not about publication bias/p-hacking

$\bar{D}_1, \bar{D}_2, \bar{D}_3$

Decision: No

<< 95%

$\Delta = 0$

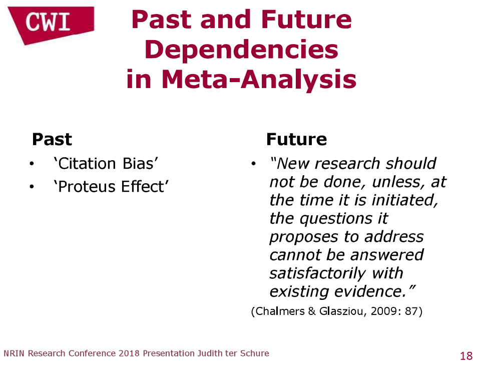
$\mathrm{P}[\text{yes} \mid \Delta = 0] \gg 0{,}05$

$\bar{D}_1, \bar{D}_2, \bar{D}_3$

Decision: Yes

>> 5%

I have to stress: This is not about publication bias or p-hacking. Even if we would be able to get rid of those practices entirely, this problem would still exist. Because this is not about trials being performed and the results stashed away in a file drawer or p-hacked, it is about trials not being performed at all. So given that a large sequence exists, the results in those sequence are no random sample from the population of all possible results, because some sequences never come into existence.



Past and Future Dependencies in Meta-Analysis

Past

- ‘Citation Bias’
- ‘Proteus Effect’

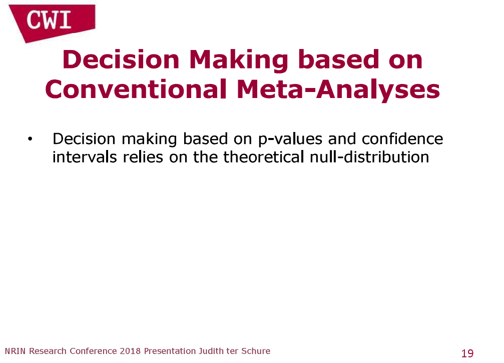
Future

- *"New research should not be done, unless, at the time it is initiated, the questions it proposes to address cannot be answered satisfactorily with existing evidence."*

(Chalmers & Glasziou, 2009: 87)

Also, this is not something we want to get rid of. If you look at the right hand side of this slide, it is exactly what the research waste advocates are proposing: To make the existence of trials dependent on previous trial results.

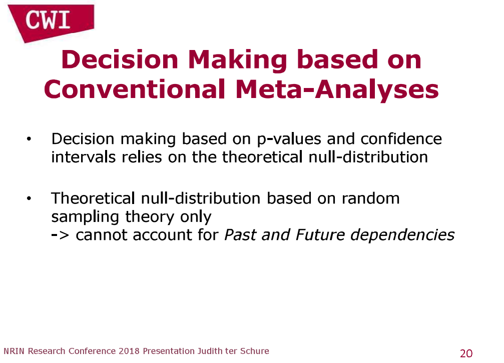
And this is probably already affecting our current meta-analyses, since empirical research under names as ‘Citation Bias’ and ‘Proteus Effect’ has shown that very promising initial trials are more likely to be replicated and cited in a replication as the reason for new research –and thus are able to end up in a sequence of trials- than not so promising initial trials.



Decision Making based on Conventional Meta-Analyses

- Decision making based on p-values and confidence intervals relies on the theoretical null-distribution

So here we sum up the problem, before we discuss the solution.



Decision Making based on Conventional Meta-Analyses

- Decision making based on p-values and confidence intervals relies on the theoretical null-distribution
- Theoretical null-distribution based on random sampling theory only
  -> cannot account for *Past and Future dependencies*

CWI

Decision Making based on Conventional Meta-Analyses

- Decision making based on p-values and confidence intervals relies on the theoretical null-distribution
- Theoretical null-distribution based on random sampling theory only
  -> cannot account for *Past and Future dependencies*
- Conventional meta-analyses cannot optimally reduce health research waste

CWI

Accumulating Science
needs
Accumulating Tests
to avoid
Accumulation Bias

CWI

Safe Tests

- **$$$**

And that is exactly what has been developed in our Machine Learning group at CWI. We call these accumulating tests 'Safe Tests', and they have the nice property that their test statistics can be interpreted in terms of gambling profits. So our intuitions about them can rely on our ideas about gambling: Both 'luck and skill' are involved and larger profits are less likely than smaller profits to be based on pure luck. There is no such intuitive interpretation for p-values.

CWI

Safe Tests

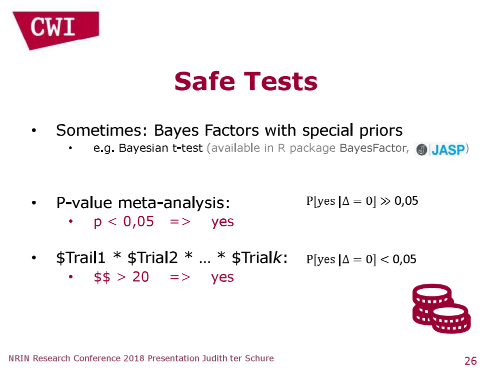
- **$$$**
- Accumulating Meta-Analysis: Reinvesting after each trial

The gambling profit interpretation also intuitively incorporates dependencies in accumulating science. In meta-analysis, this can be seen as reinvesting the results of previous trials in new ones.

Safe Tests

- Sometimes: Bayes Factors with special priors
  - e.g. Bayesian t-test (available in R package BayesFactor, JASP)

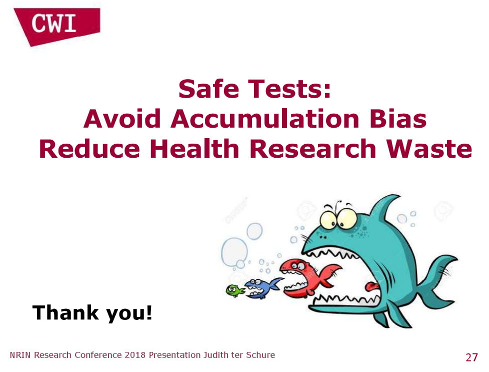
Some Bayes Factor tests are Safe Tests, such as the Bayesian t-test that is already available in software.



Safe Tests

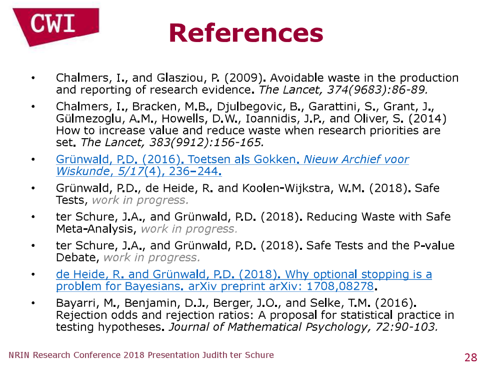
- Sometimes: Bayes Factors with special priors
  - e.g. Bayesian t-test (available in R package BayesFactor, JASP)
- P-value meta-analysis: $\mathrm{P}[\text{yes} \mid \Delta = 0] \gg 0{,}05$
  - p < 0,05 => yes
- $Trail1 * $Trial2 * … * $Trial*k*: $\mathrm{P}[\text{yes} \mid \Delta = 0] < 0{,}05$
  - $$ > 20 => yes

And Safe Tests are able to keep the Type-I errors under control as scientific data accumulate. In contrast to p-value tests, as I have just shown.



Safe Tests: Avoid Accumulation Bias Reduce Health Research Waste

Thank you!



References

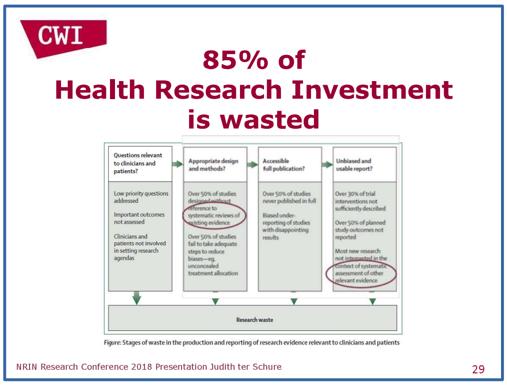
- Chalmers, I., and Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet, 374(9683):86-89.*
- Chalmers, I., Bracken, M.B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A.M., Howells, D.W., Ioannidis, J.P., and Oliver, S. (2014) How to increase value and reduce waste when research priorities are set. *The Lancet, 383(9912):156-165.*
- Grünwald, P.D. (2016). Toetsen als Gokken. *Nieuw Archief voor Wiskunde, 5/17*(4), 236–244.
- Grünwald, P.D., de Heide, R. and Koolen-Wijkstra, W.M. (2018). Safe Tests, *work in progress.*
- ter Schure, J.A., and Grünwald, P.D. (2018). Reducing Waste with Safe Meta-Analysis, *work in progress.*
- ter Schure, J.A., and Grünwald, P.D. (2018). Safe Tests and the P-value Debate, *work in progress.*
- de Heide, R. and Grünwald, P.D. (2018). Why optional stopping is a problem for Bayesians. arXiv preprint arXiv: 1708,08278.
- Bayarri, M., Benjamin, D.J., Berger, J.O., and Selke, T.M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology, 72:90-103.*

Grünwald, P.D. (2016). Toetsen als Gokken. *Nieuw Archief voor Wiskunde, 5/17*(4), 236–244.

https://ir.cwi.nl/pub/25373

de Heide, R. and Grünwald, P.D. (2018). Why optional stopping is a problem for Bayesians. arXiv preprint arXiv: 1708,08278.

https://arxiv.org/abs/1708.08278v2



CWI

85% of Health Research Investment is wasted

Questions relevant to clinicians and patients?

Low priority questions addressed

Important outcomes not assessed

Clinicians and patients not involved in setting research agendas

Appropriate design and methods?

Over 50% of studies designed without reference to systematic reviews of existing evidence

Over 50% of studies fail to take adequate steps to reduce biases—eg, unconcealed treatment allocation

Accessible full publication?

Over 50% of studies never published in full

Biased under-reporting of studies with disappointing results

Unbiased and usable report?

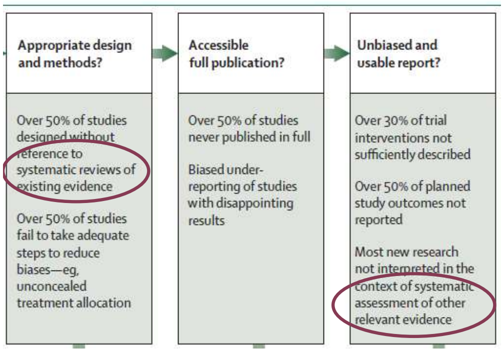
Over 30% of trial interventions not sufficiently described

Over 50% of planned study outcomes not reported

Most new research not interpreted in the context of systematic assessment of other relevant evidence

Research waste

Figure: Stages of waste in the production and reporting of research evidence relevant to clinicians and patients



Appropriate design and methods?

Over 50% of studies designed without reference to systematic reviews of existing evidence

Over 50% of studies fail to take adequate steps to reduce biases—eg, unconcealed treatment allocation

Accessible full publication?

Over 50% of studies never published in full

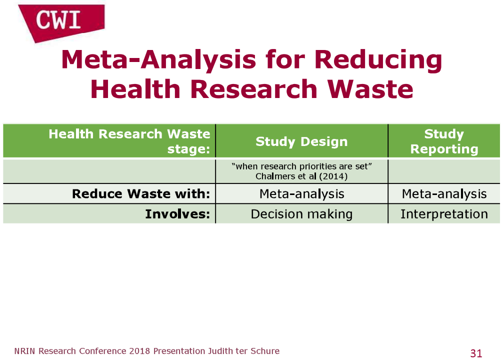
Biased under-reporting of studies with disappointing results

Unbiased and usable report?

Over 30% of trial interventions not sufficiently described

Over 50% of planned study outcomes not reported

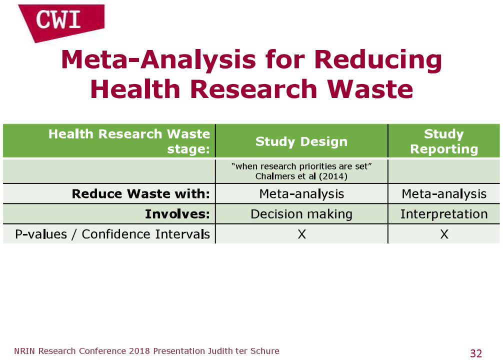
Most new research not interpreted in the context of systematic assessment of other relevant evidence



CWI

Meta-Analysis for Reducing Health Research Waste

Health Research Waste stage:	Study Design	Study Reporting
	“when research priorities are set” Chalmers et al (2014)	
Reduce Waste with:	Meta-analysis	Meta-analysis
Involves:	Decision making	Interpretation

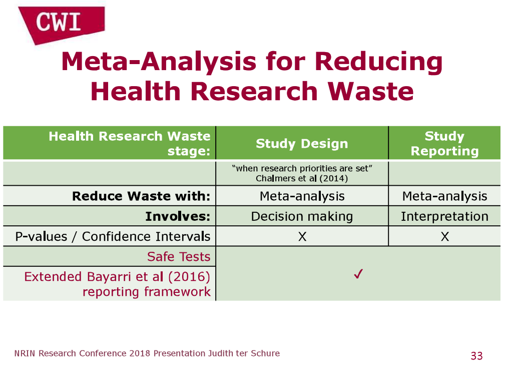


CWI

Meta-Analysis for Reducing Health Research Waste

Health Research Waste stage:	Study Design	Study Reporting
	“when research priorities are set” Chalmers et al (2014)	
Reduce Waste with:	Meta-analysis	Meta-analysis
Involves:	Decision making	Interpretation
P-values / Confidence Intervals	X	X

The ASA statement on p-values has shown that there are also severe interpretation issues with conventional meta-analysis reporting based on p-values and confidence intervals.

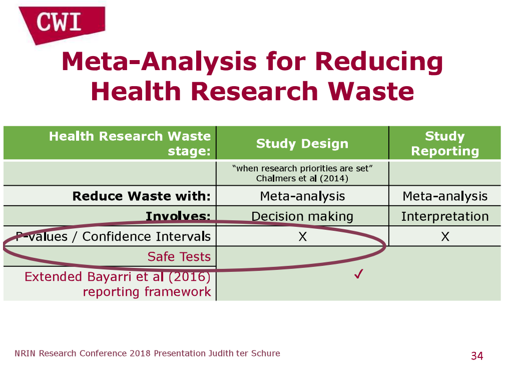


CWI

Meta-Analysis for Reducing Health Research Waste

Health Research Waste stage:	Study Design	Study Reporting
	"when research priorities are set" Chalmers et al (2014)	
Reduce Waste with:	Meta-analysis	Meta-analysis
Involves:	Decision making	Interpretation
P-values / Confidence Intervals	X	X
Safe Tests		
Extended Bayarri et al (2016) reporting framework	✓	

Safe Tests, together with a framework of reporting and study design based on earlier work by Bayarri et al (2016) can solve all problems with decision making and interpration based on meta-analysis in an accumulating science setting.

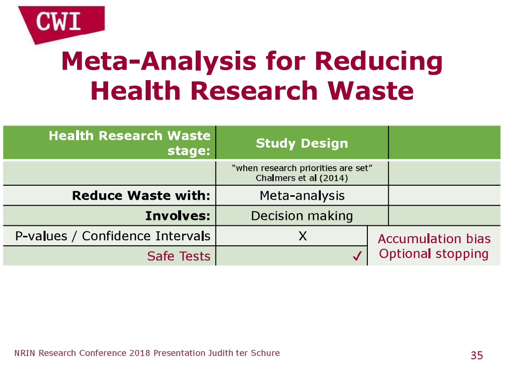


CWI

Meta-Analysis for Reducing Health Research Waste

Health Research Waste stage:	Study Design	Study Reporting
	"when research priorities are set" Chalmers et al (2014)	
Reduce Waste with:	Meta-analysis	Meta-analysis
Involves:	Decision making	Interpretation
P-values / Confidence Intervals	X	X
Safe Tests		
Extended Bayarri et al (2016) reporting framework	✓	

Today I only discussed the decision making setting.

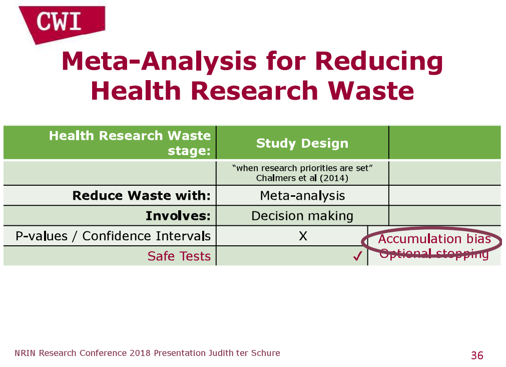


CWI

Meta-Analysis for Reducing Health Research Waste

Health Research Waste stage:	Study Design	
	"when research priorities are set" Chalmers et al (2014)	
Reduce Waste with:	Meta-analysis	
Involves:	Decision making	
P-values / Confidence Intervals	X	Accumulation bias Optional stopping
Safe Tests	✓	

The decision making setting involves two problems when scientific data accumulates that can both be summarized as a dependence of sample size (study sequence length) on previous results: Accumulation Bias and Optional Stopping.

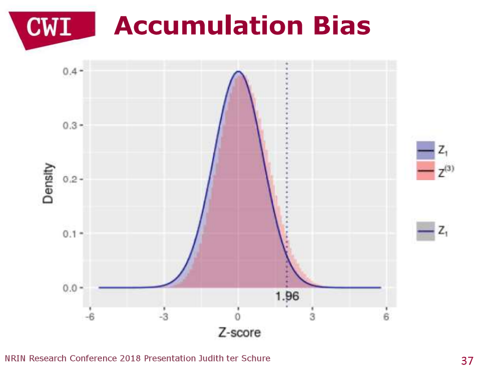


CWI

Meta-Analysis for Reducing Health Research Waste

Health Research Waste stage:	Study Design	
	"when research priorities are set" Chalmers et al (2014)	
Reduce Waste with:	Meta-analysis	
Involves:	Decision making	
P-values / Confidence Intervals	X	Accumulation bias Optional stopping
Safe Tests	✓	

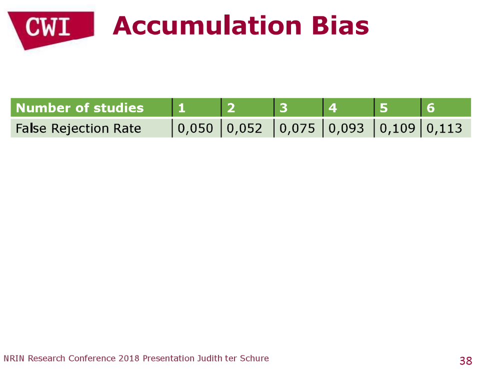
Today I only discussed Accumulation Bias.



Accumulation Bias

NRIN Research Conference 2018 Presentation Judith ter Schure 37

Accumulation Bias is a shift in the null-distribution as a result of the existence of a sequence of trials. Z^(3) displays the combined Z-score of a sequence of three studies, Z_1 displays the Z-score of an individual study. In this simulation I assumed that initial trials showing a significant negative effect (patients get worse), are not replicated at all, and that initial trials showing significant improvements are extra likely to be replicated and to end up in a three trial sequence.

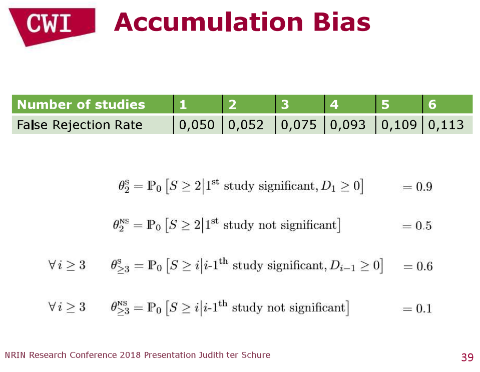


Accumulation Bias

Number of studies	1	2	3	4	5	6
False Rejection Rate	0,050	0,052	0,075	0,093	0,109	0,113

NRIN Research Conference 2018 Presentation Judith ter Schure 38

Under certain assumptions, the problem increases as studies accumulate: the larger the sequence, the larger the Type-I error rate (‘False Rejection Rate’). Not only does the expectation of the Z-score under the null-distribution show a shift, the distribution also gets skewed, which causes extra inflation of the Type-I error rate.



Accumulation Bias

Number of studies	1	2	3	4	5	6
False Rejection Rate	0,050	0,052	0,075	0,093	0,109	0,113

$$\theta_2^{\mathrm{S}} = \mathbb{P}_0\left[S \geq 2 \middle| 1^{\mathrm{st}} \text{ study significant}, D_1 \geq 0\right] \qquad = 0.9$$

$$\theta_2^{\mathrm{NS}} = \mathbb{P}_0\left[S \geq 2 \middle| 1^{\mathrm{st}} \text{ study not significant}\right] \qquad = 0.5$$

$$\forall\, i \geq 3 \qquad \theta_{\geq 3}^{\mathrm{S}} = \mathbb{P}_0\left[S \geq i \middle| i\text{-}1^{\mathrm{th}} \text{ study significant}, D_{i-1} \geq 0\right] \qquad = 0.6$$

$$\forall\, i \geq 3 \qquad \theta_{\geq 3}^{\mathrm{NS}} = \mathbb{P}_0\left[S \geq i \middle| i\text{-}1^{\mathrm{th}} \text{ study not significant}\right] \qquad = 0.1$$

NRIN Research Conference 2018 Presentation Judith ter Schure 39

These are the assumptions on which the previous graph and table were based.