PREMIER Symposium
Berlin, Germany
March 16th, 2018

UNIVERSITÄT
BERN

The standardization fallacy

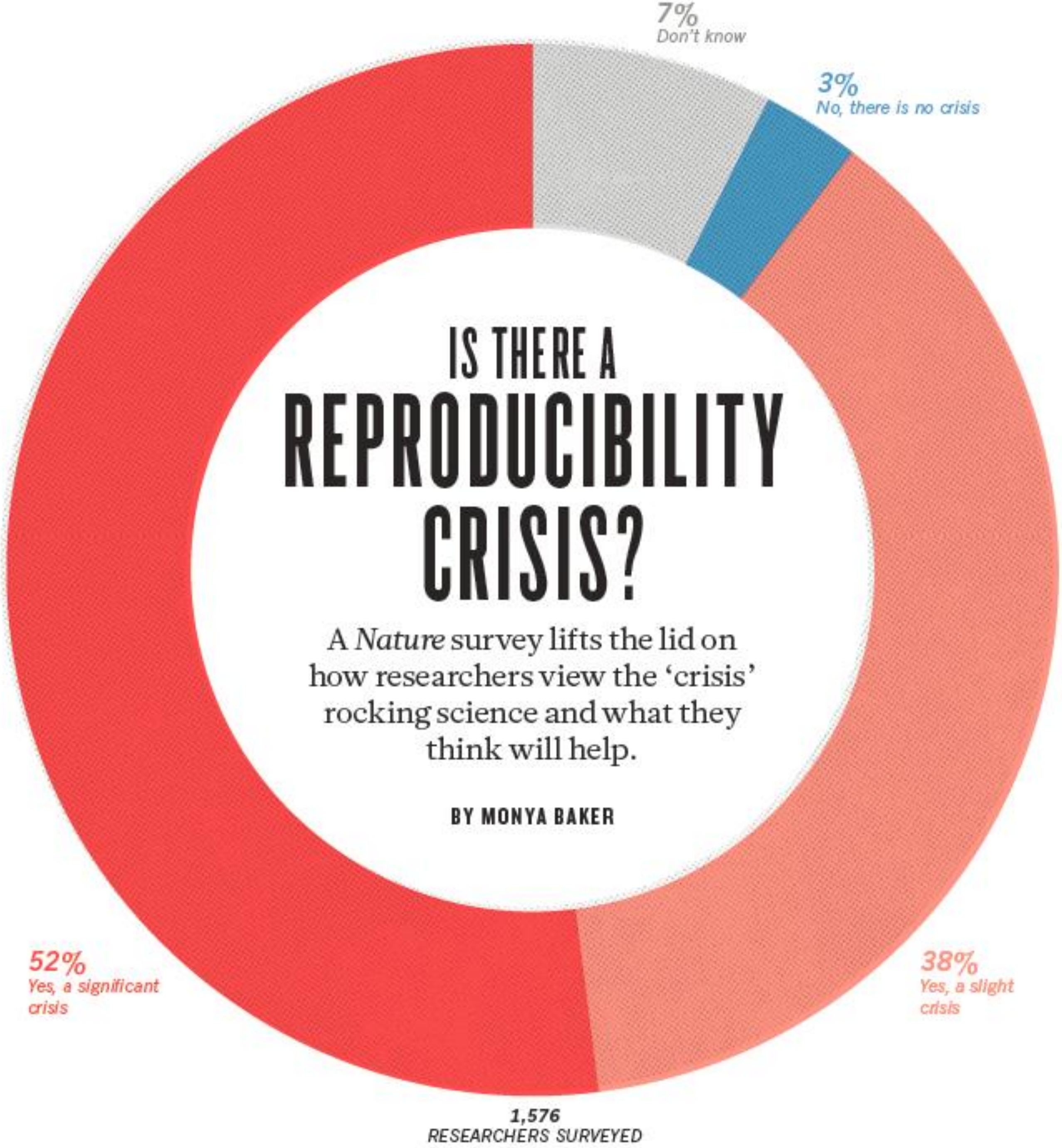
Hanno Würbel

Division of Animal Welfare
University of Bern
Switzerland

IS THERE A REPRODUCIBILITY CRISIS?

A *Nature* survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

BY MONYA BAKER



Things proposed to cause poor reproducibility

- Lack of scientific rigor (risks of bias)
- Too small sample sizes (lack of statistical power)
- “Analytical flexibility” (p-hacking, HARKing, selective reporting)
- Publishing “positive” results only (publication bias)

Things proposed to cause poor reproducibility

- Lack of scientific rigor (risks of bias)
- Too small sample sizes (lack of statistical power)
- “Analytical flexibility” (p-hacking, HARKing, selective reporting)
- Publishing “positive” results only (publication bias)
- **Rigorous standardization**

Things proposed to cause poor reproducibility

- Lack of scientific rigor (risks of bias)
- Too small sample sizes (lack of statistical power)
- “Analytical flexibility” (p-hacking, HARKing, selective reporting)
- Publishing “positive” results only (publication bias)
- **Rigorous standardization**

► **Reproducibility is a function of external validity**

External validity and reproducibility

Birth of *reproducibility* as key principle to establish “matters of fact”

RITA
GREER
2007

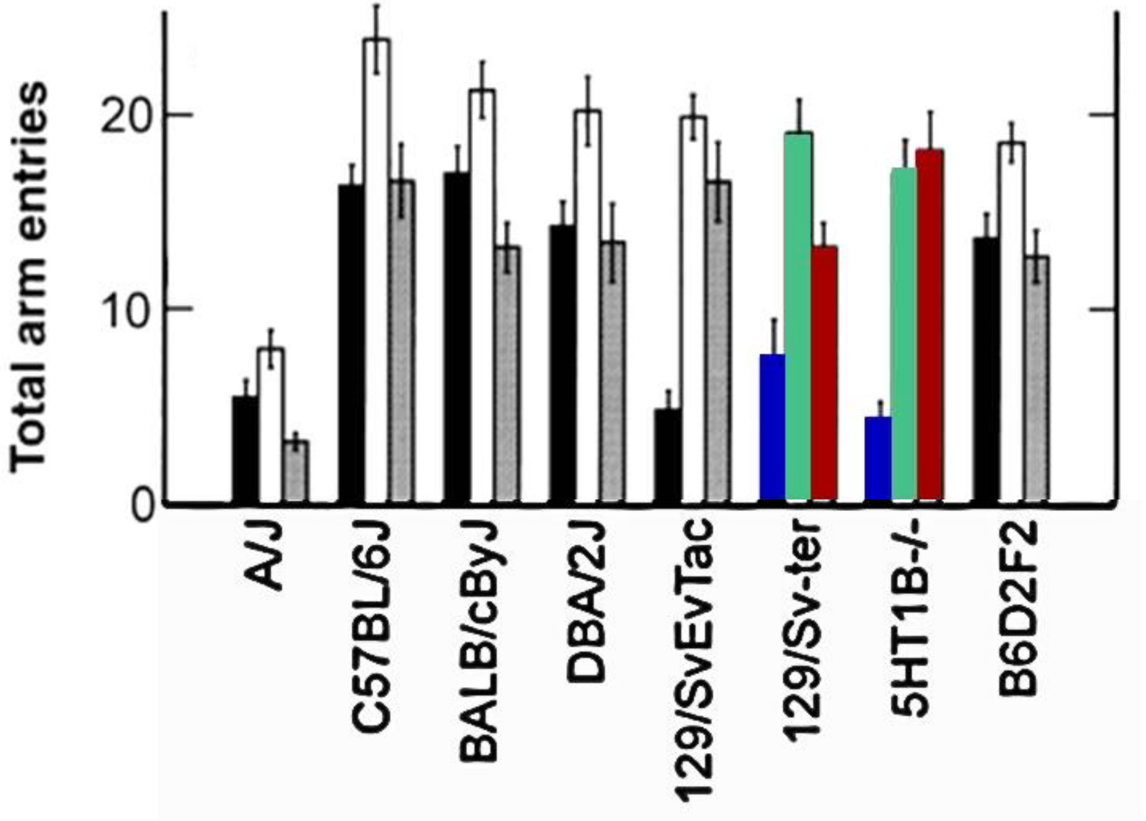
Open Science
in the 17th century

Under the eyes of Royal Society members, Robert Hooke replicates an observation reported by a Dutch scientist.

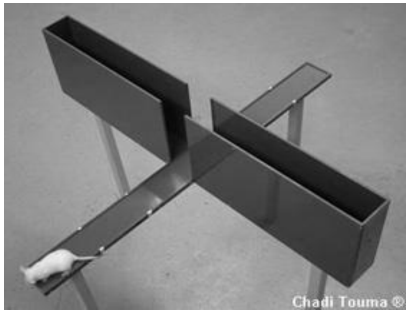
Rita Greer 2007 The Scientists
Source: **Wikimedia Commons**

External validity and reproducibility

Poor reproducibility “despite” rigorous standardization



🟥 Portland ↗
🟩 Edmonton →
🟦 Albany ↘



Crabbe et al. 1999 **Science**

External validity and reproducibility

The standardization fallacy

«Standardization is the attempt to increase reproducibility at the expense of external validity.»

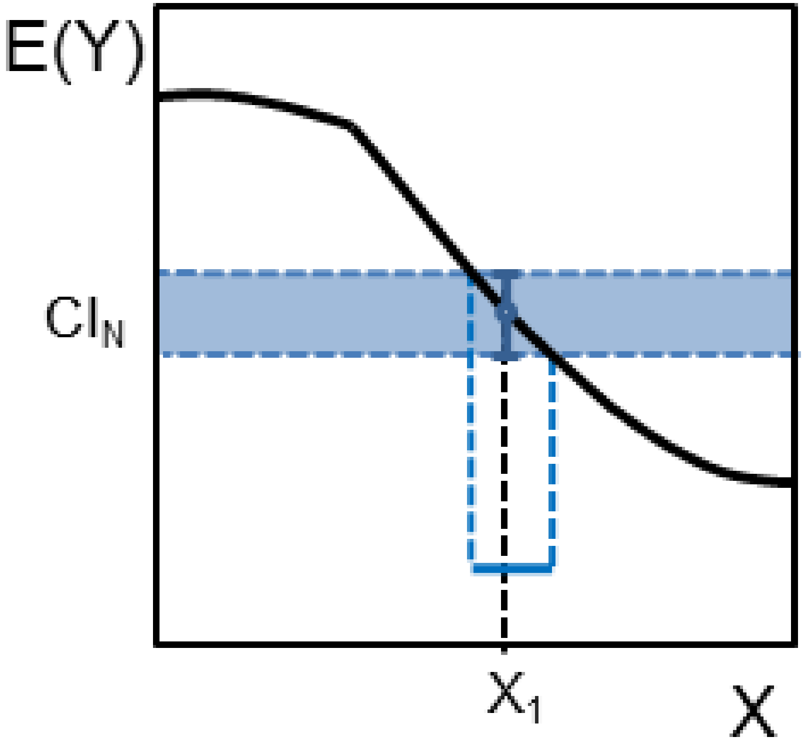
Würbel 2000 **Nature Genetics**

«A highly standardized experiment supplies direct information only in respect of the narrow range of conditions achieved by standardization. ***Standardization, therefore, weakens rather than strengthens our ground for inferring a result****, when, as is the case in practice, these conditions are somewhat varied.»*

Fisher 1935 **The Design of Experiments**

Standardization and poor reproducibility

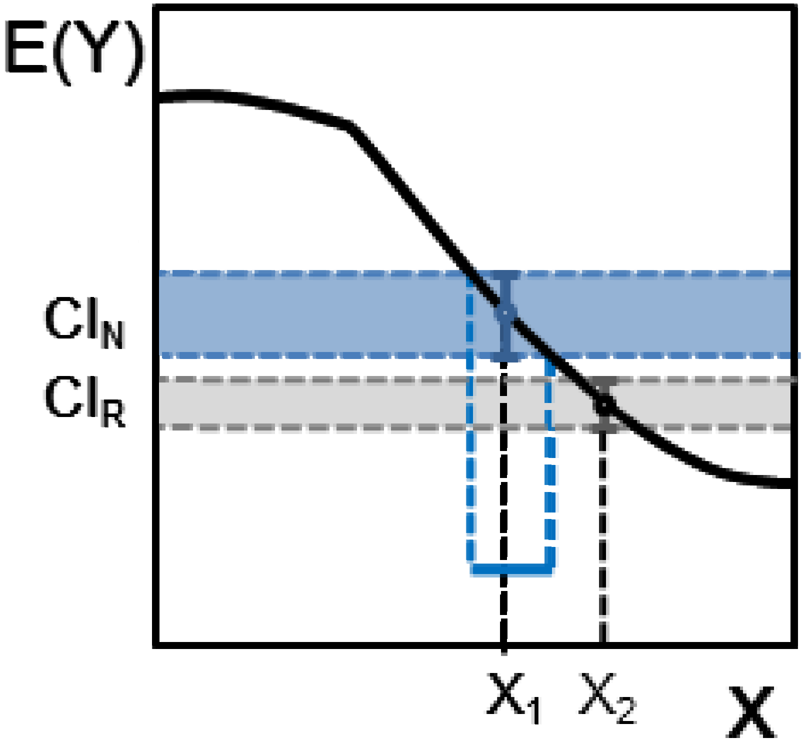
The reproducibility paradox



Voelkl & Würbel 2016 **Trends Pharmacol Sci**

Standardization and poor reproducibility

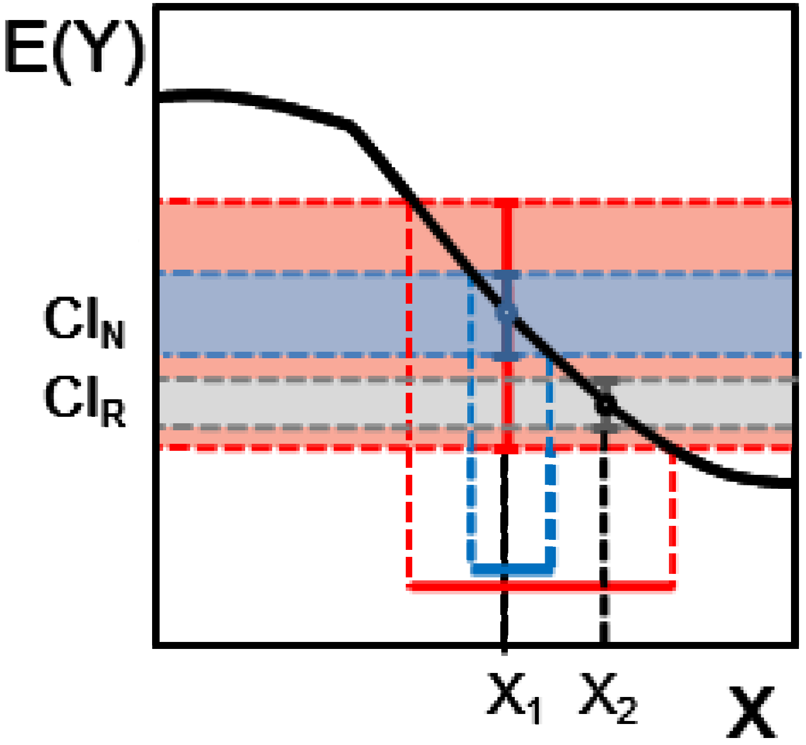
The reproducibility paradox



Voelkl & Würbel 2016 **Trends Pharmacol Sci**

Standardization and poor reproducibility

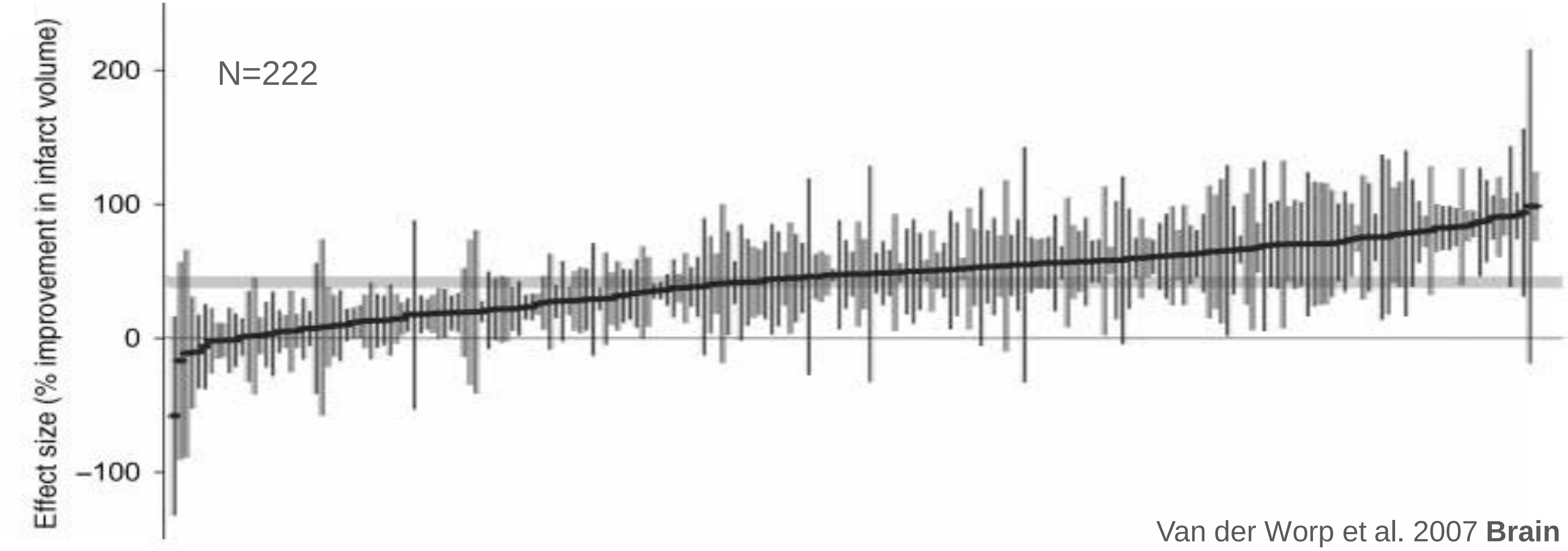
The reproducibility paradox



Voelkl & Würbel 2016 **Trends Pharmacol Sci**

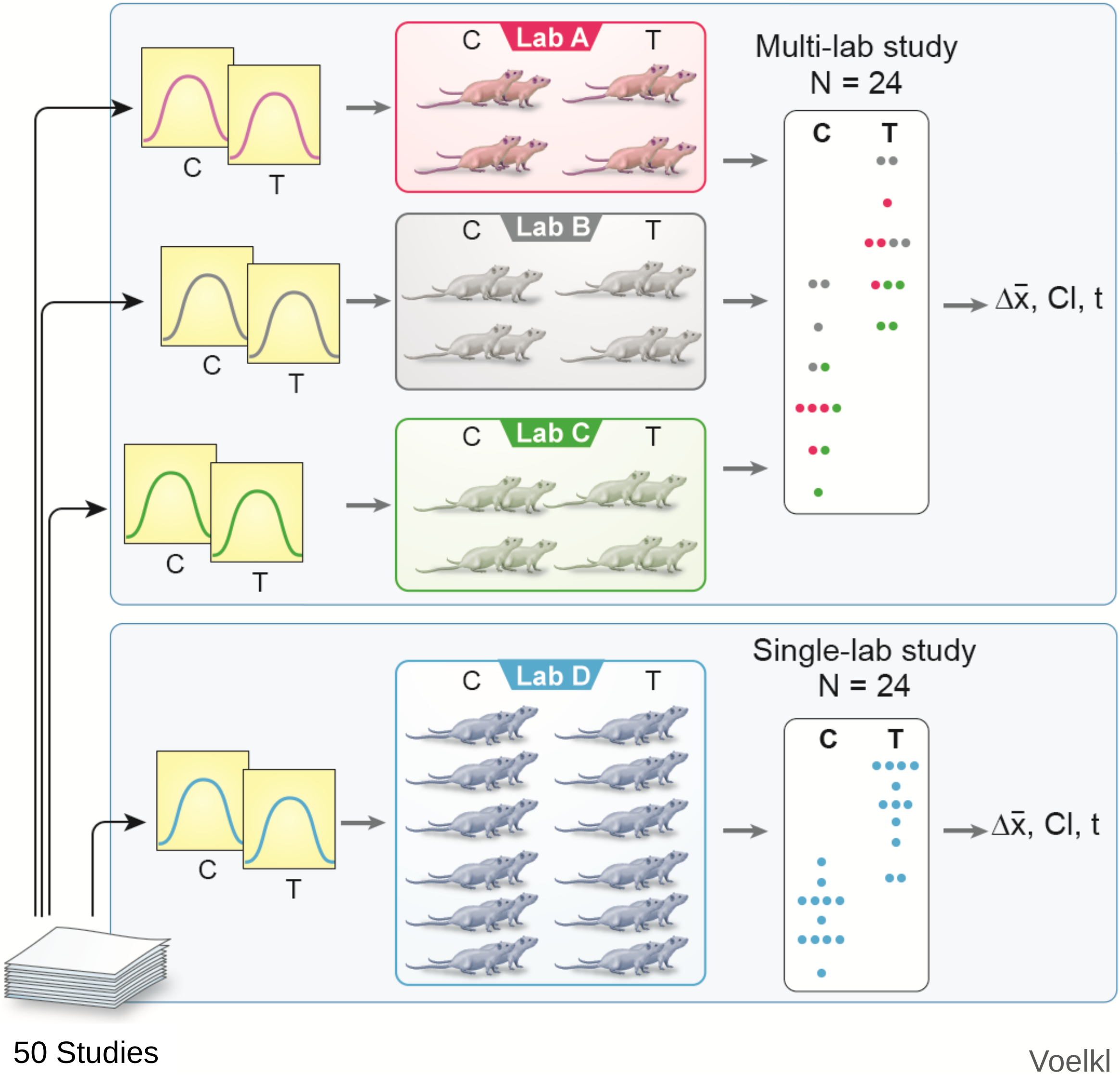
Simulation of multi-lab studies using real data

Effect of hypothermia on infarct volume in animal models of stroke



Van der Worp et al. 2007 **Brain**

Simulation of multi-lab studies using real data

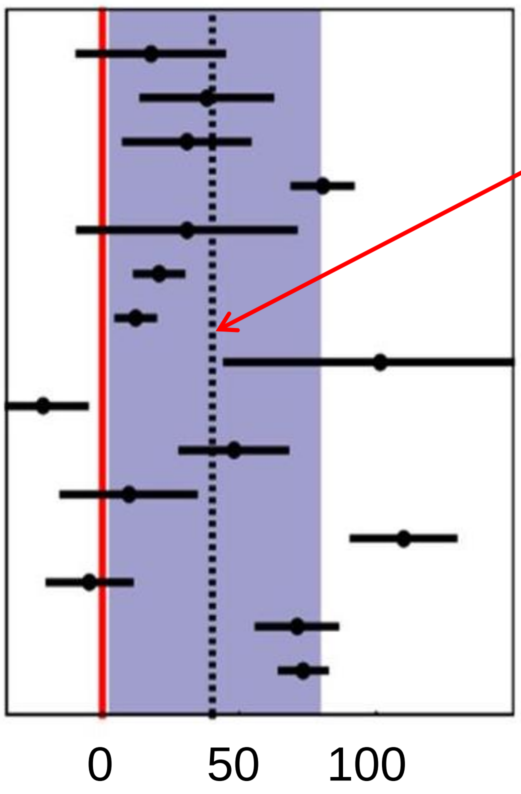


Voelkl et al. 2018 **PLOS Biol**

Simulation of single-lab versus multi-lab studies

Effect of hypothermia on infarct volume in rodent models of stroke

1-lab studies



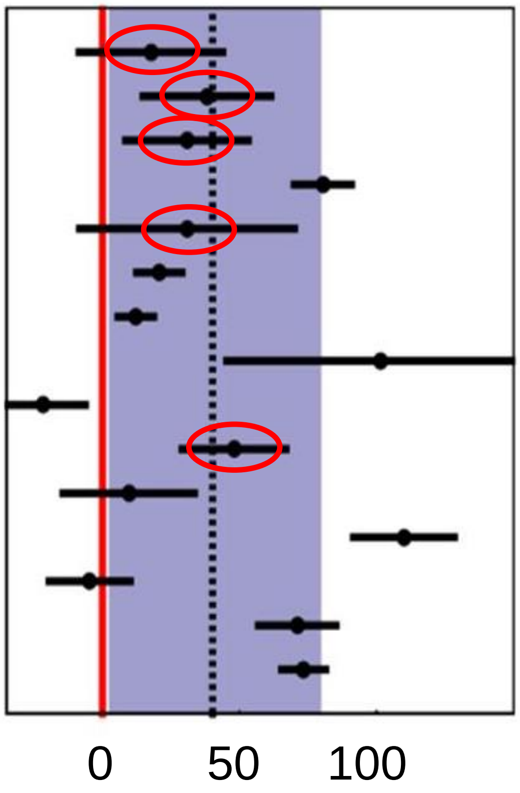
„true" effect size:
summary effect size across all 50 studies

Voelkl et al. 2018 **PLOS Biol**

Simulation of single-lab versus multi-lab studies

Effect of hypothermia on infarct volume in rodent models of stroke

1-lab studies



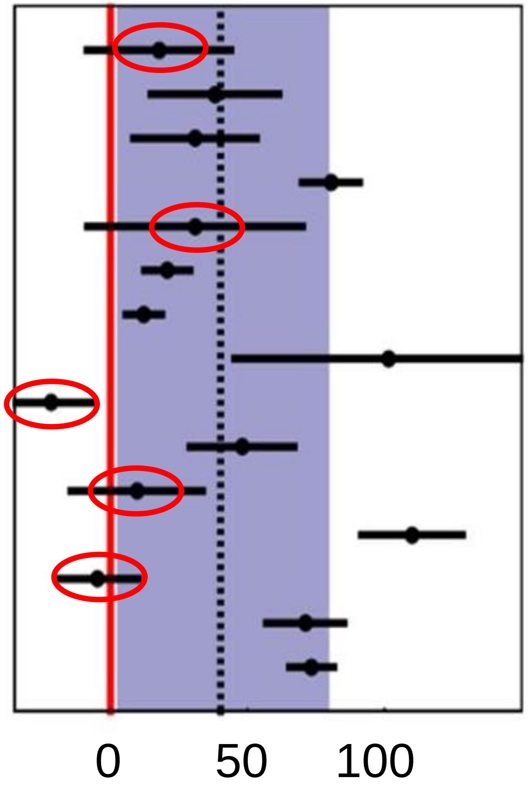
5 of 15 = accurate (coverage probability)

Voelkl et al. 2018 **PLOS Biol**

Simulation of single-lab versus multi-lab studies

Effect of hypothermia on infarct volume in rodent models of stroke

1-lab studies



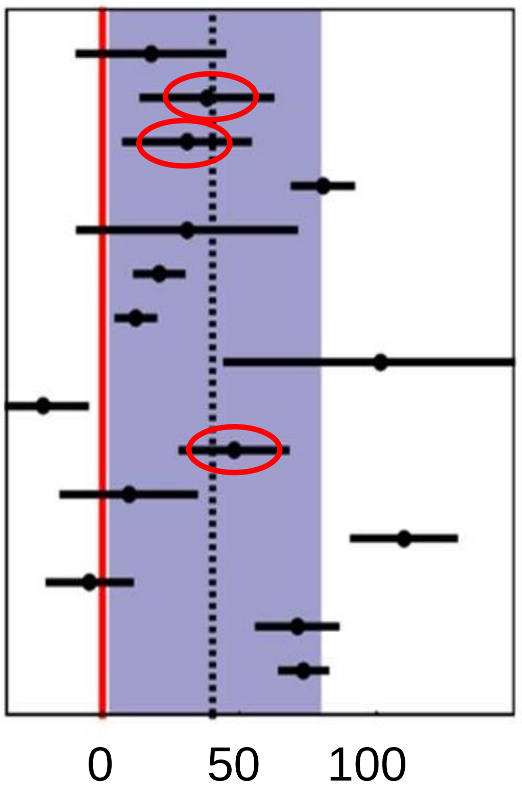
5 of 15 = false negatives

Voelkl et al. 2018 **PLOS Biol**

Simulation of single-lab versus multi-lab studies

Effect of hypothermia on infarct volume in rodent models of stroke

1-lab studies

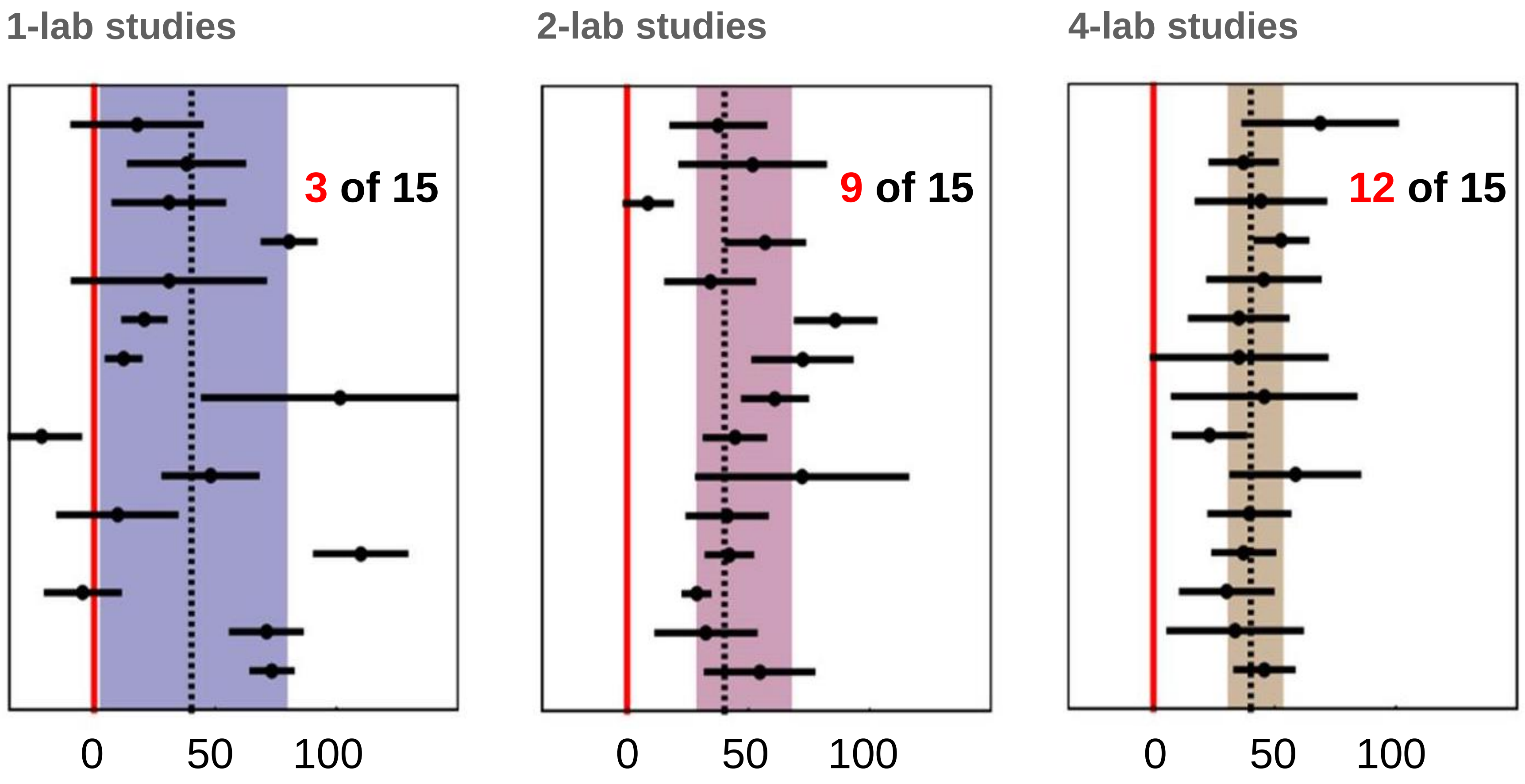


only 3 of 15 = significant and accurate

Voelkl et al. 2018 **PLOS Biol**

Simulation of single-lab versus multi-lab studies

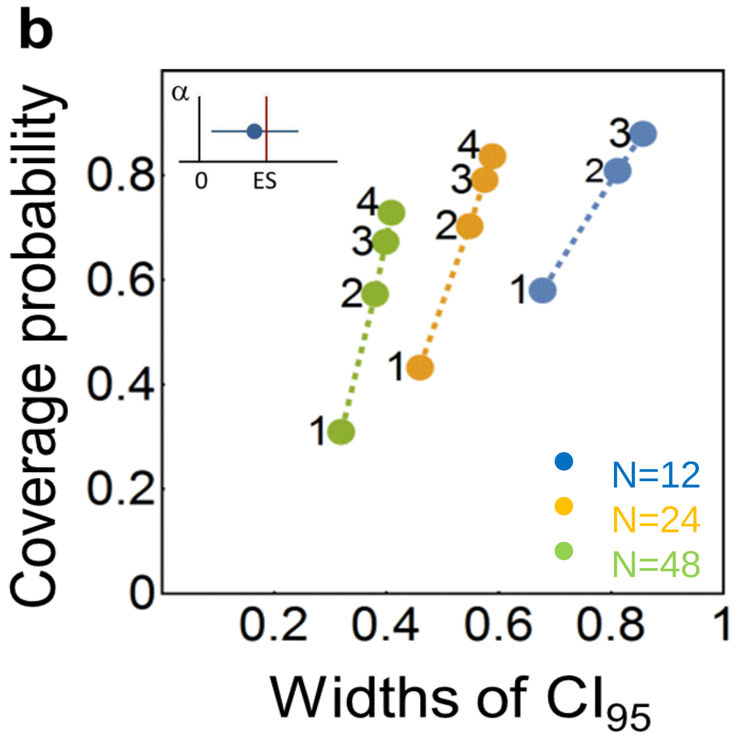
Effect of hypothermia on infarct volume in rodent models of stroke



Voelkl et al. 2018 **PLOS Biol**

Simulation of single-lab versus multi-lab studies

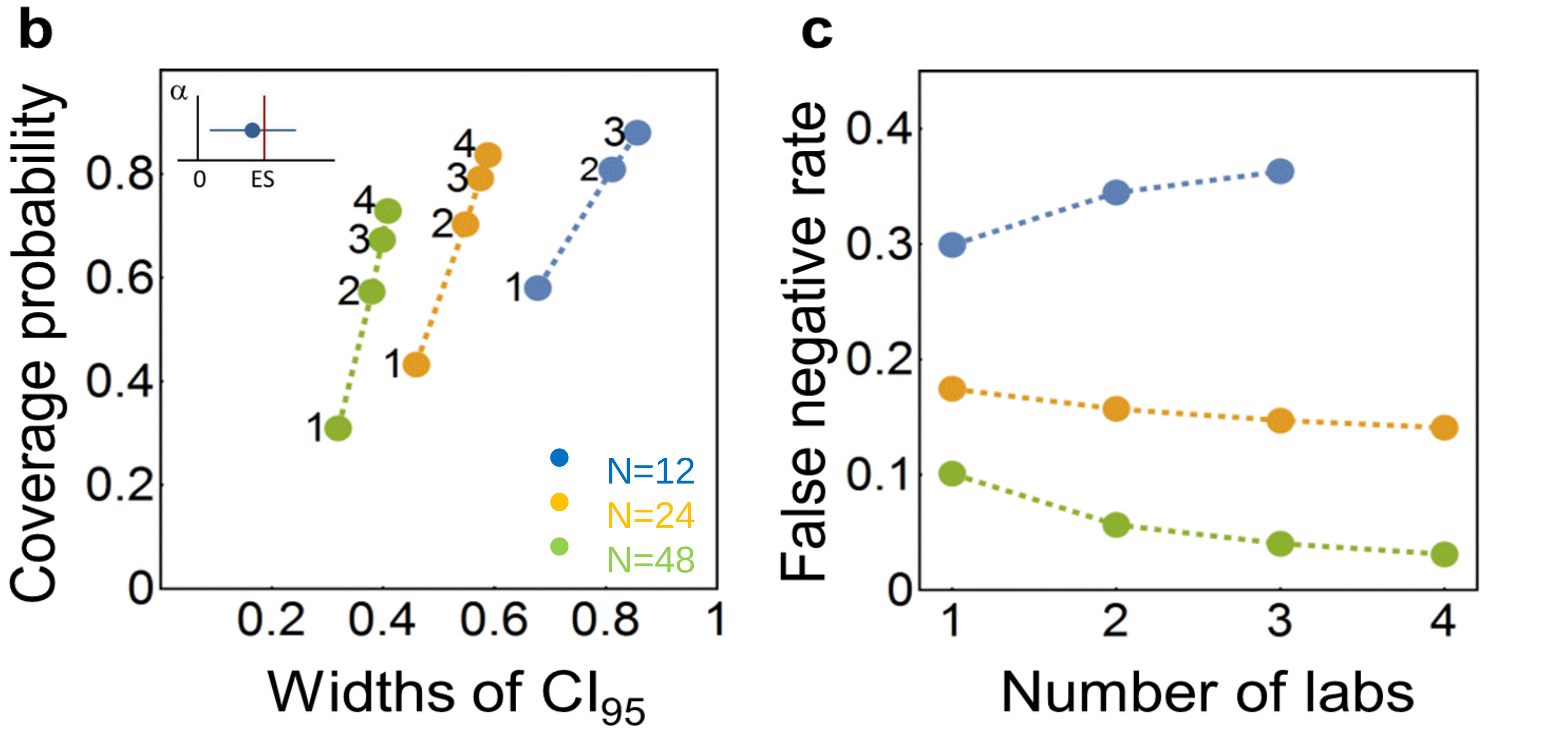
Coverage probability and false negative rate depending on sample size



Voelkl et al. 2018 **PLOS Biol**

Simulation of single-lab versus multi-lab studies

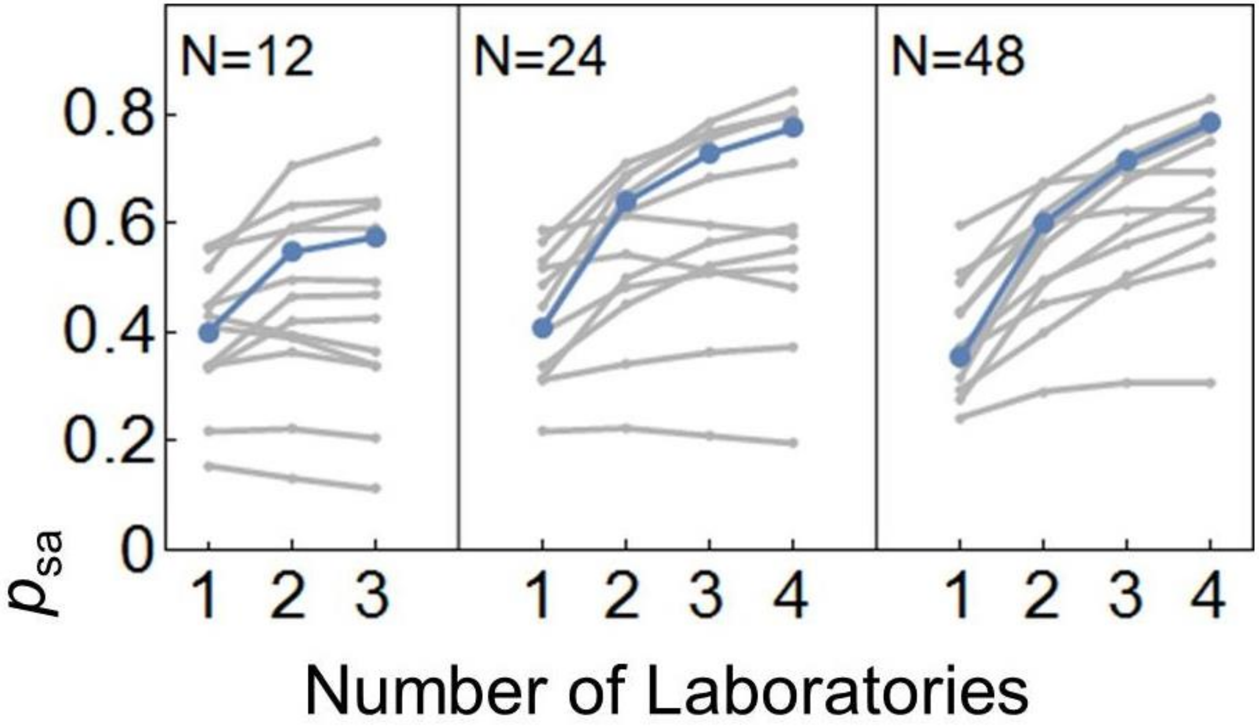
Coverage probability and false negative rate depending on sample size



Voelkl et al. 2018 **PLOS Biol**

Simulation of single-lab versus multi-lab studies

Replications for 13 different interventions in animal models of stroke, myocardial infarction, and breast cancer



Voelkl et al. 2018 **PLOS Biol**

Conclusions

- Reproducibility depends on the external validity of the results
- Standardized single-lab studies produce results of poor external validity
- Poor external validity leads to poor reproducibility

Conclusions

- Reproducibility depends on the external validity of the results
- Standardized single-lab studies produce results of poor external validity
- Poor external validity leads to poor reproducibility

Possible solutions:

► Adjusting p-values of single-lab studies by treatment x lab interaction term
(Kafkafi *et al.* 2017 *Nat Methods*)

► Heterogenization of study samples in single-lab studies
(Richter *et al.* 2010 *Nat Methods*)

► Multi-lab studies
(Voelkl *et al.* 2018 *PLOS Biol*)

Acknowldgments

My lab

- Bernhard Voelkl (senior scientist)
- Lucile Vogt (PhD student)
- Helene Richter (former PhD student)

Collaborators

- Emily Sena (Edinburgh)
- Yoav Benjamini (Tel Aviv)
- Joseph Garner (Stanford)

Supported by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

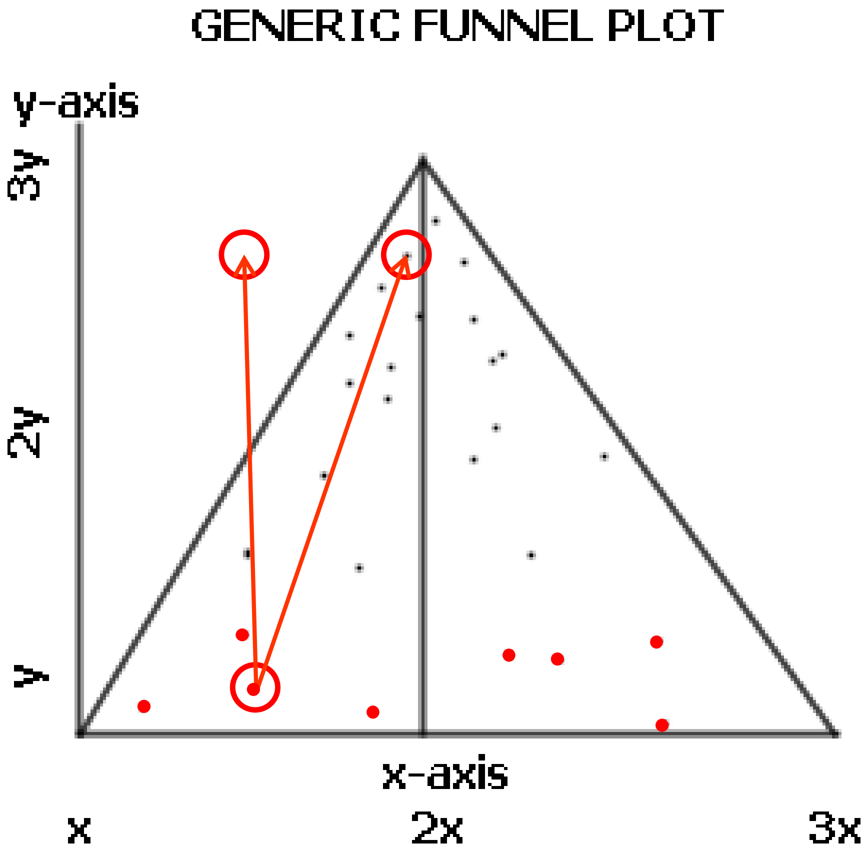
SEVENTH FRAMEWORK PROGRAMME

erc
European Research Council

Collaborative Approach to Meta Analysis and
·C·A·M·A·R·A·D·E·S·
Review of Animal Data from Experimental Studies

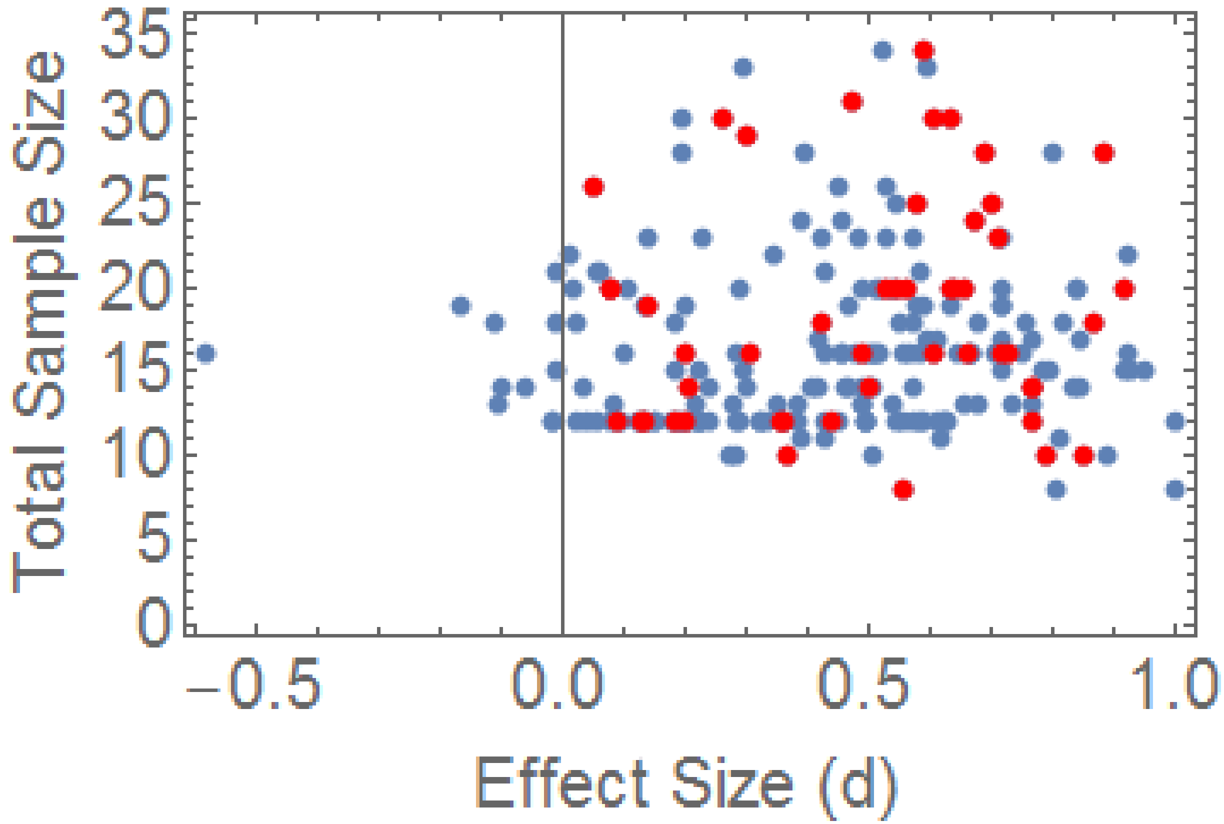
Random variation or phenotypic plasticity?

Different predictions about effects of larger sample sizes



Simulation of single-lab versus multi-lab studies

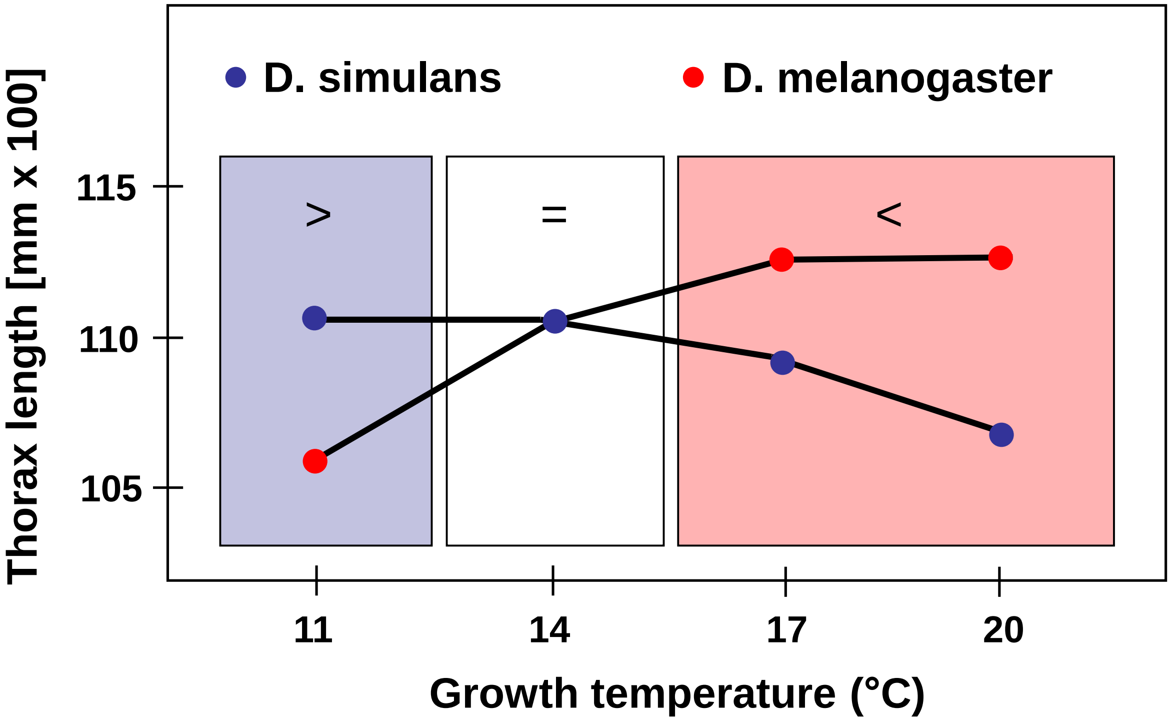
Coverage probability and false negative rate



Voelkl et al. 2016 **unpublished results**

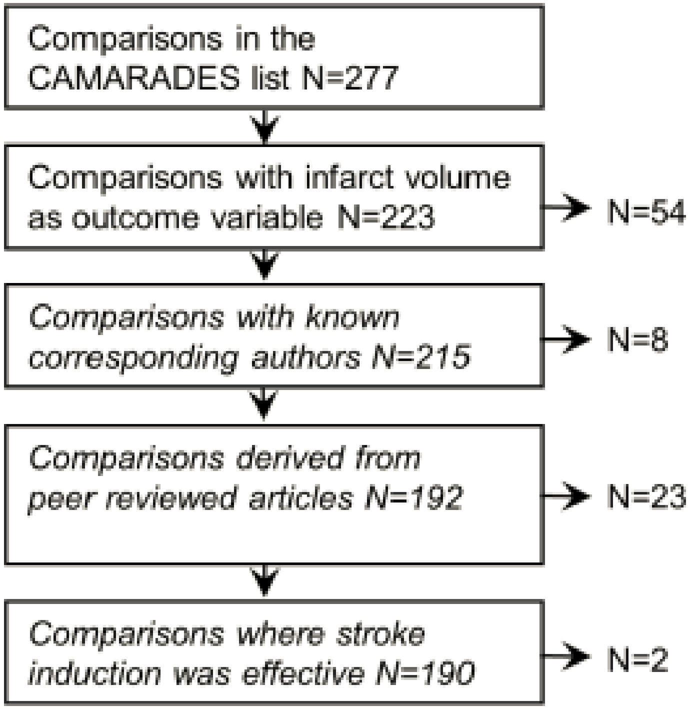
Standardization and poor reproducibility

Phenotypic plasticity can induce treatment x environment interactions



Simulation of single-lab versus multi-lab studies

Inclusion and exclusion criteria – hypothermia studies



Comparisons in the CAMARADES list N=277

Comparisons with infarct volume as outcome variable N=223

N=54

Comparisons with known corresponding authors N=215

N=8

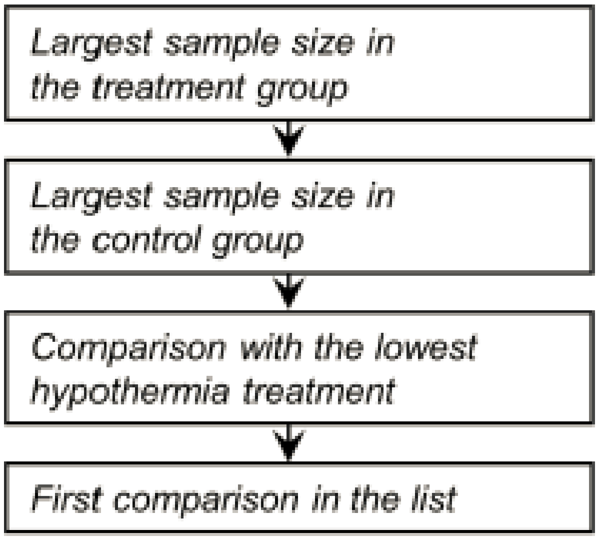
Comparisons derived from peer reviewed articles N=192

N=23

Comparisons where stroke induction was effective N=190

N=2

For selection of a single comparison per author:



Largest sample size in the treatment group

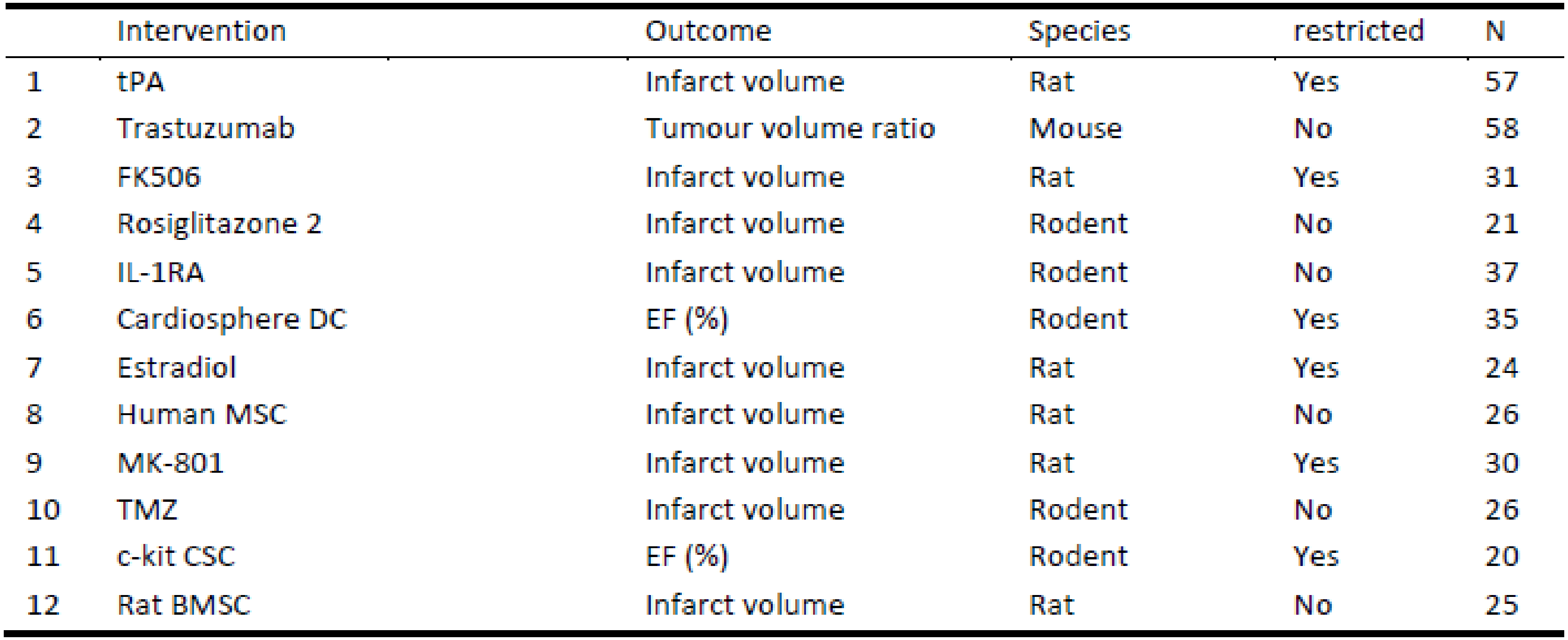
Largest sample size in the control group

Comparison with the lowest hypothermia treatment

First comparison in the list

Simulation of single-lab versus multi-lab studies

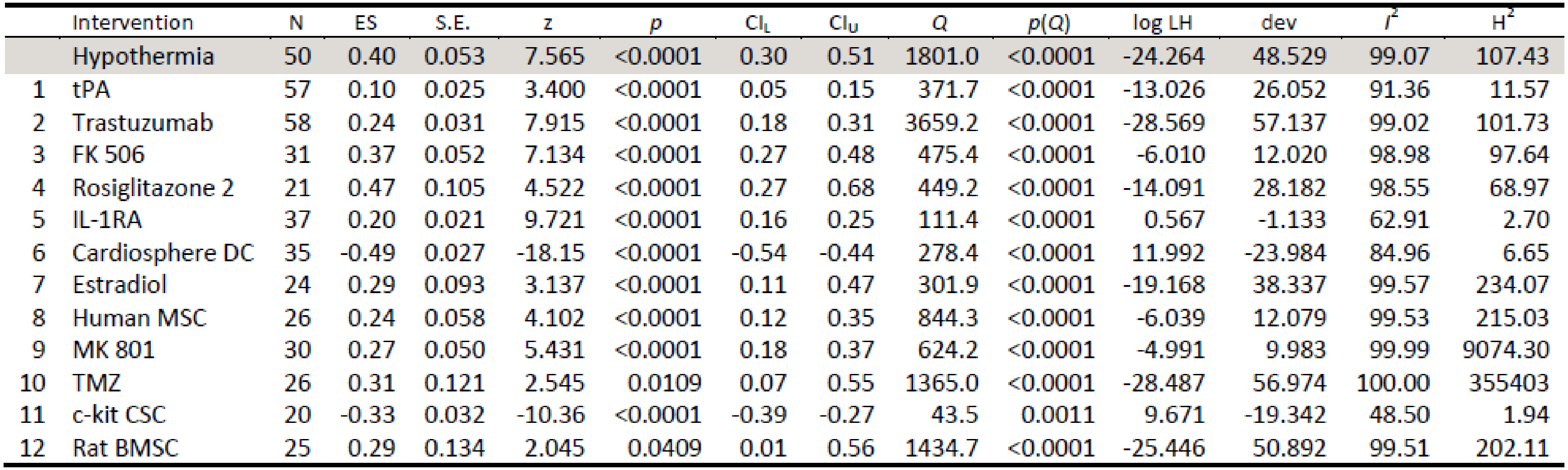
Further interventions (n=12) used to replicate hypothermia simulation



	Intervention	Outcome	Species	restricted	N
1	tPA	Infarct volume	Rat	Yes	57
2	Trastuzumab	Tumour volume ratio	Mouse	No	58
3	FK506	Infarct volume	Rat	Yes	31
4	Rosiglitazone 2	Infarct volume	Rodent	No	21
5	IL-1RA	Infarct volume	Rodent	No	37
6	Cardiosphere DC	EF (%)	Rodent	Yes	35
7	Estradiol	Infarct volume	Rat	Yes	24
8	Human MSC	Infarct volume	Rat	No	26
9	MK-801	Infarct volume	Rat	Yes	30
10	TMZ	Infarct volume	Rodent	No	26
11	c-kit CSC	EF (%)	Rodent	Yes	20
12	Rat BMSC	Infarct volume	Rat	No	25

Simulation of single-lab versus multi-lab studies

Results of random effects meta-analyses in R (*metafor 1.9-9*)



	Intervention	N	ES	S.E.	z	*p*	CI_L	CI_U	*Q*	*p*(*Q*)	log LH	dev	I^2	H^2
	Hypothermia	50	0.40	0.053	7.565	<0.0001	0.30	0.51	1801.0	<0.0001	-24.264	48.529	99.07	107.43
1	tPA	57	0.10	0.025	3.400	<0.0001	0.05	0.15	371.7	<0.0001	-13.026	26.052	91.36	11.57
2	Trastuzumab	58	0.24	0.031	7.915	<0.0001	0.18	0.31	3659.2	<0.0001	-28.569	57.137	99.02	101.73
3	FK 506	31	0.37	0.052	7.134	<0.0001	0.27	0.48	475.4	<0.0001	-6.010	12.020	98.98	97.64
4	Rosiglitazone 2	21	0.47	0.105	4.522	<0.0001	0.27	0.68	449.2	<0.0001	-14.091	28.182	98.55	68.97
5	IL-1RA	37	0.20	0.021	9.721	<0.0001	0.16	0.25	111.4	<0.0001	0.567	-1.133	62.91	2.70
6	Cardiosphere DC	35	-0.49	0.027	-18.15	<0.0001	-0.54	-0.44	278.4	<0.0001	11.992	-23.984	84.96	6.65
7	Estradiol	24	0.29	0.093	3.137	<0.0001	0.11	0.47	301.9	<0.0001	-19.168	38.337	99.57	234.07
8	Human MSC	26	0.24	0.058	4.102	<0.0001	0.12	0.35	844.3	<0.0001	-6.039	12.079	99.53	215.03
9	MK 801	30	0.27	0.050	5.431	<0.0001	0.18	0.37	624.2	<0.0001	-4.991	9.983	99.99	9074.30
10	TMZ	26	0.31	0.121	2.545	0.0109	0.07	0.55	1365.0	<0.0001	-28.487	56.974	100.00	355403
11	c-kit CSC	20	-0.33	0.032	-10.36	<0.0001	-0.39	-0.27	43.5	0.0011	9.671	-19.342	48.50	1.94
12	Rat BMSC	25	0.29	0.134	2.045	0.0409	0.01	0.56	1434.7	<0.0001	-25.446	50.892	99.51	202.11