

Topic	Questions
Best ranking intervention	Is there one intervention that ranks highest with respect to (a) positive predictive value, (b) consistency, (c) reproducibility (two-trial rule), (d) Edgington's p-value, (e) coverage, for all investigated data sets (asked separately for pre-clinical and clinical studies)?
	Is there one intervention that ranks highest with respect to (a) positive predictive value, (b) consistency, (c) reproducibility (two-trial rule), (d) Edgington's p-value, (e) coverage, for the absolute majority of investigated data sets (asked separately for pre-clinical and clinical studies)?
	Which intervention has the highest average rank with respect to (a) positive predictive value, (b) consistency, (c) reproducibility (two-trial rule), (d) Edgington's p-value, (e) coverage, averaged over all investigated data sets (asked separately for pre-clinical and clinical studies)?
Number of high-performing interventions	Are there interventions, that are always among the 'top three' ranks with respect to (a) positive predictive value, (b) consistency, (c) reproducibility (two-trial rule), (d) Edgington's p-value, (e) coverage, for all investigated data sets (asked separately for pre-clinical and clinical studies)?
	How many different interventions rank highest with respect to (a) positive predictive value, (b) consistency, (c) reproducibility (two-trial rule), (d) Edgington's p-value, (e) coverage, for at least one data set (asked separately for pre-clinical and clinical studies)?
	How many different interventions can be found among the top three ranks respect to (a) positive predictive value, (b) consistency, (c) reproducibility (two-trial rule), (d) Edgington's p-value, (e) coverage, at least once (asked separately for pre-clinical and clinical studies)?
Comparison of pre-clinical and clinical studies	What is the difference between pre-clinical and clinical studies in the average rank with respect to (a) positive predictive value, (b) consistency, (c) reproducibility (two-trial rule), (d) Edgington's p-value, (e) coverage, for each intervention?
Effect of power = 0.8 and 0.9	For all interventions where sample size calculations were done for a power of 0.8 and 0.9 respectively: what is the difference (in %, averaged over all data sets) in the sample size and in (a) positive predictive value, (b) consistency, (c) reproducibility (two-trial rule), (d) Edgington's p-value, (e) coverage (asked separately for pre-clinical and clinical studies)?
Effect of sample size adjustments	For the interventions, 1 and 2 (n= 10, vs. n=50), 3 and 4 (p<0.005 vs. p<0.005 with adjustment), 8 and 9a (random lab model vs. random lab model with power adjustment), 10 and 11a (multi-lab vs. multi lab with power adjustment), what is the difference (in %, averaged over all data sets) in the sample size and in (a) positive predictive value, (b) consistency, (c) reproducibility (two-trial rule), (d) Edgington's p-value, (e) coverage (asked separately for pre-clinical and clinical studies)?

Table S3: Analysis plan for the registered report “*Efficacy of Statistical and Design Interventions for Improving Reproducibility in Biomedical Research*”. The questions are all of a quantitative nature and will be answered descriptively. No hypothesis tests will be done.

	Animal data					Human data				
Question	PPV	Con	TTR	Ed P	Cov	PPV	Con	TTR	Ed P	Cov
One intervention always ranks highest	No	No	(7b)	No	No	No	No	No	No	No
One intervention ranks highest in abs majority	(4)	(7b)	(7b)	(7b)	(10)	(4)	No	(7b)	(7b)	(11b)
Which intervention has highest average rank	(4)	(7b)	(7b)	(7b)	(3)	(4)	(1)	(7b)	(7b)	(11b)
N of Interventions always in top three	0	0	1	0	0	0	0	0	0	0
N of interventions ranking highest	3	4	1	4	7	5	6	4	5	4
N of interventions among top three	7	10	6	6	11	9	13	9	9	9

Table S4: Best ranking intervention and number of high-performing interventions. PPV: positive predictive value, Con: Consistency, TTR: two-trial rule, Ed P: Edginton's p-value, Cov: Coverage. Numbers in brackets give intervention numbers according to supplementary table S3.

Average rank: (animal, human) difference

Intervention number	Intervention name	PPV	Con	TTR	Ed P	Cov
0	t-test	(19.2,16.6) -	(14.1,11.6) -	(13.2,12.8) -	(13.2,12.8) -	(3.9,8.9) +
1	Rule of 10	(17.4,20.2) +	(12.1,6.2) -	(11.3,18.6) +	(11.3,18.6) +	(7.2,7.2) =
2	N=50	(12.1,14.7) +	(7.4,12.6) +	(5.9,11.3) +	(5.9,11.3) +	(15.4,10.3) -
3	p<0.005	(2.3,3.1) +	(14.9,6.7) -	(18.2,18.7) +	(18.2,18.9) +	(3.8,11.8) +
4	p<0.005 adj	(1.3,1.6) +	(15.1,9.5) -	(14.8,15.3) +	(15.0,15.4) +	(9.2,13.6) +
5a	Power one study 0.8	(19.3,13.9) -	(17.1,17.8) +	(14.4,10.2) -	(14.4,10.3) -	(5.0,14.1) +
5b	Power one study 0.9	(18.2,12.4) -	(15.3,16.5) +	(12.3,8.5) -	(12.3,8.6) -	(7.1,17.3) +
6a	Power meta 0.8	(15.4,10.9) -	(10.7,13.5) +	(8.9,6.8) -	(8.9,6.8) -	(10.2,18.5) +
6b	Power meta 0.9	(14.1,9.4) -	(8.9,10.6) +	(7.1,4.4) -	(7.1,4.4) -	(13.1,20.2) +
7a	MID small 0.8	(8.8,7.9) -	(2.7,8.8) +	(2.0,3.0) +	(2.1,3.0) +	(19.9,10.7) -
7b	MID small 0.9	(8.3,7.1) -	(1.6,7.4) +	(1.0,1.8) +	(1.1,1.8) +	(20.9,12.6) -
7c	MID medium 0.8	(11.0,13.4) +	(6.2,13.3) +	(4.9,9.8) +	(4.9,9.8) +	(16.5,4.3) -
7d	MID medium 0.9	(10.8,11.7) +	(5.0,13.6) +	(3.7,7.9) +	(3.7,7.9) +	(17.9,6.3) -
7e	MID large 0.8	(15.0,18.4) +	(10.3,8.4) -	(9.0,15.6) +	(9.0,15.6) +	(11.4,7.4) -
7f	MID large 0.9	(14.0,16.8) +	(9.0,10.1) +	(7.5,13.6) +	(7.5,13.5) +	(13.3,13.9) +
8	RLM	(13.5,16.7) +	(12.7,6.8) -	(18.5,17.0) -	(18.4,16.9) -	(4.2,17.2) +
9a	RLM power 0.8	(6.5,7.4) +	(13.6,12.8) -	(15.8,12.9) -	(15.7,12.8) -	(10.1,4.0) -
9b	RLM power 0.9	(5.7,6.7) +	(13.3,12.3) -	(14.3,11.0) -	(14.2,11.0) -	(13.3,10.6) -
10	Multi-Lab (ML)	(8.5,12.6) +	(10.7,6.5) -	(20.1,16.7) -	(20.1,16.6) -	(4.8,14.0) +
11a	ML power 0.8	(5.4,5.4) =	(15.7,14.0) -	(15.1,9.0) -	(15.0,9.0) -	(10.4,5.9) -
11b	ML power 0.9	(4.2,4.1) -	(14.5,12.2) -	(12.9,6.1) -	(12.7,6.1) -	(13.3,2.3) -

Table S5: Comparison of pre-clinical and clinical studies. PPV: positive predictive value, Con: Consistency, TTR: two-trial rule, Ed P: Edginton's p-value, Cov: Coverage.

Percent increase (median over all studies) from unadjusted to power 08

	Animal data						Human data					
	N	PPV	Con	TTR	Ed P	Cov	N	PPV	Con	TTR	Ed P	Cov
1 vs 2	400	0.6	19	45	40	-51	400	4.8	-20	304	286	-12
3 vs 4	75	0.1	3	78	74	-21	71	0.3	-10	104	100	-6
8 vs 9a	558	2.0	-1	33	30	-22	3057	5.6	-13	133	142	110
10 vs 11a	558	0.9	-12	211	208	-31	3057	2.6	-22	460	443	98

Table S6: PPV: positive predictive value, Con: Consistency, TTR: two-trial rule, Ed P: Edginton's p-value, Cov: Coverage. Increase in % between interventions 1 vs 2 (Rule of 10 and N=50), 3 vs 4 ($p<0.005$ and $p<0.005$ adj.), 8 vs 9a (Random Lab Model and Random Lab Model with power=0.8) 10 vs 11a (Multi-Lab Study and Multi-Lab Study with power=0.8).

Percent increase (median over all studies) from 0.8 to 0.9

	Animal data						Human data					
	N	PPV	Con	TTR	Ed P	Cov	N	PPV	Con	TTR	Ed P	Cov
5a vs 5b	14	0.2	2.1	9	9	-6	15	0.2	1.5	12	11	-13
6a vs 6b	17	0.1	3.6	6	6	-9	16	0.2	3.3	10	9	-13
9a vs 9b	17	0.1	0.1	5	4	-9	16	0.0	0.0	7	6	-38
11a vs 11b	17	0.1	2.5	22	21	-11	16	0.1	3.5	17	16	7

Table S7: Effect of power = 0.8 and 0.9. PPV: positive predictive value, Con: Consistency, TTR: two-trial rule, Ed P: Edginton's p-value, Cov: Coverage.