



Universität
Zürich^{UZH}

Balancing Evidentiary Value and Sample Size of Adaptive Designs

With Application to Animal Experiments

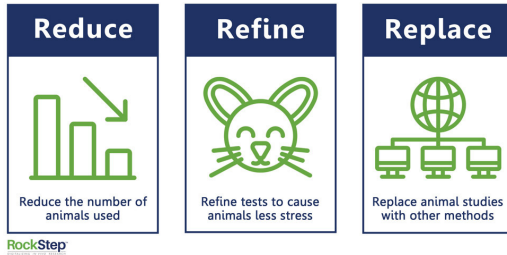
Leonhard Held

Joint work with Fadoua Balabdaoui, Saverio Fontana, and Samuel Pawel

NCS 2026 Padua, IT / 23 September 2026

Reduce the number of animals used

The 3 R's of Animal Research



*“[...] to use methods for obtaining comparable levels of information from the use of **fewer animals** in scientific procedures, or for obtaining **more information** from the same number of animals”*

swiss3rcc.org/3rs-for-the-public/

PLOS BIOLOGY

ESSAY

Is N-Hacking Ever OK? The consequences of collecting more data in pursuit of statistical significance

Pamela Reinagel *

Department of Neurobiology, School of Biological Science, University of California San Diego, La Jolla, California, United States of America

Reinagel (2023), PLOS Biology

- **Uncontrolled analysis** of results before the final number of experimental units has been reached → Type-I error (T1E) rate inflation
- **N-hacking** also known as **peeking, optional stopping**

Reinagel's “constrained sample augmentation”

- Interim analyses after $n = 8, 12, \dots, n_{\text{Max}} = 32$ animals per group

Reinagel's “constrained sample augmentation”

- Interim analyses after $n = 8, 12, \dots, n_{\text{Max}} = 32$ animals per group
- Stop for success if two-sided $p < 0.05$

Reinagel's “constrained sample augmentation”

- Interim analyses after $n = 8, 12, \dots, n_{\text{Max}} = 32$ animals per group
- Stop for success if two-sided $p < 0.05$, stop for futility if $p > 0.1$

Reinagel's “constrained sample augmentation”

- Interim analyses after $n = 8, 12, \dots, n_{\text{Max}} = 32$ animals per group
 - Stop for success if two-sided $p < 0.05$, **stop for futility if $p > 0.1$**
- Moderate T1E rate inflation to 6.25%, improved positive predictive value

Reinagel's “constrained sample augmentation”

- Interim analyses after $n = 8, 12, \dots, n_{\text{Max}} = 32$ animals per group
 - Stop for success if two-sided $p < 0.05$, **stop for futility if $p > 0.1$**
- Moderate T1E rate inflation to 6.25%, improved positive predictive value
- How do **group sequential designs** (with proper T1E control) perform?

PERSPECTIVE

Increasing efficiency of preclinical research by group sequential designs

Konrad Neumann^{1,2,*}, Ulrike Grittner^{1,2,*}, Sophie K. Piper^{1,2,3}, Andre Rex^{2,4}, Oscar Florez-Vargas⁵, George Karystianis⁶, Alice Schneider^{1,2}, Ian Wellwood^{2,7}, Bob Siegerink^{2,8}, John P. A. Ioannidis⁹, Jonathan Kimmelman¹⁰, Ulrich Dirnagl^{2,3,4,6,11,12}

Group sequential designs for in vivo studies: Minimizing animal numbers and handling uncertainty in power analysis

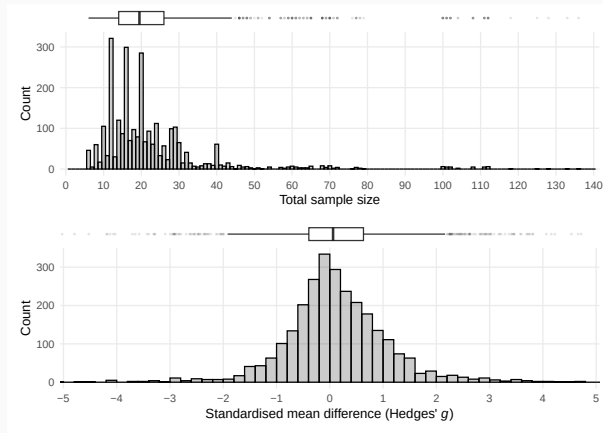
Susanne Blotwijk^{a,*}, Sophie Hernot^b, Kurt Barbé^a

Blotwijk et al. (2022), Research in Veterinary Science

Neumann et al. (2017), PLOS Biology

Application

Data on 2738 experiments with mice and rats from Neuroscience and Metabolism



Group-sequential designs with one interim analysis

One interim analysis at information fraction 2/3:

Method	α_1	α_2	T1E rate (in %)
Pocock	0.032	0.032	5.00
O'Brien-Fleming	0.014	0.045	5.00
Haybittle-Peto	0.010	0.050	5.20
N-hacking	0.050	0.050	7.42

α_1, α_2 : Nominal significance thresholds at analysis 1, 2

How many animals could have been saved?

Simulate **interim z-value** z_1 (at sample size $n_1 \approx 2/3 \cdot n_2$), given the **final z-value** z_2 :

$$Z_1 \mid Z_2 = z_2 \sim N\left(z_2 \sqrt{n_1/n_2}, 1 - n_1/n_2\right)$$

Method	Mean n	H_0 rejections (%)	Interim efficacy stops (%)	Interim futility stops (%)	Animals saved
No interim analysis	22.5	27.8	0	0	0
Haybittle-Peto	21.5 (21.4, 21.5)	28 (27.9, 28.3)	14.3 (13.6, 14.9)	0 (0, 0)	2'945 (2'772, 3'121)
O'Brien-Fleming	21.4 (21.3, 21.4)	27.4 (27.1, 27.7)	15.7 (15, 16.4)	0 (0, 0)	3'232 (3'051, 3'417)
Pocock	21.1 (21, 21.1)	26.3 (25.8, 26.9)	20.1 (19.3, 21)	0 (0, 0)	4'095 (3'892, 4'304)
Haybittle-Peto + futility	18.4 (18.3, 18.6)	27.5 (27.2, 27.8)	14.3 (13.6, 14.9)	47 (45.7, 48.3)	11'238 (10'886, 11'590)
O'Brien-Fleming + futility	18.2 (18.1, 18.4)	26.9 (26.5, 27.3)	15.7 (15, 16.4)	48.5 (47.2, 49.7)	11'792 (11'430, 12'139)
Pocock + futility	17.6 (17.5, 17.7)	25.9 (25.3, 26.5)	20.1 (19.3, 21)	53 (51.8, 54.3)	13'506 (13'157, 13'850)
Reinagel	16.2 (16.1, 16.3)	26.6 (25.7, 27.4)	23.5 (22.6, 24.4)	69.5 (68.5, 70.5)	17'403 (17'178, 17'610)

Additional **futility stopping** of group sequential designs if **predictive power** $\leq 10\%$

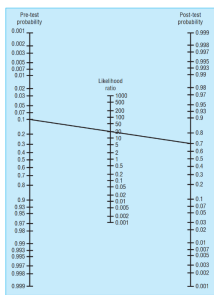
Which method should be used?

1. Adopt **diagnostic likelihood ratios** and **diagnostic odds ratio** to hypotheses testing
2. Incorporate sample size to define the **Experimental Unit Information Index (EUII)**, the **average information from one animal**
3. Extend formulation to **adaptive designs**
4. Compare different adaptive designs in **simulation study**

Likelihood ratios

- Likelihood ratios (LRs) quantify the change in odds after having observed a **positive (significant, +)** or **negative (not significant, -)** test result:

$$\text{Posterior odds}(H_1) = \left\{ \begin{array}{l} \text{LR}^+ \\ \text{LR}^- \end{array} \right\} \cdot \text{Prior odds}(H_1)$$



Use of Fagan's nomogram for calculating post-test probabilities

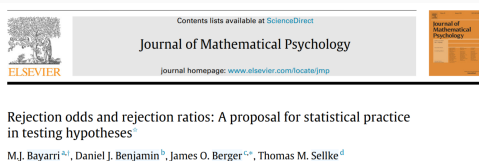
Statistics Notes

Diagnostic tests 4: likelihood ratios

Jonathan J Deeks, Douglas G Altman

Deeks and Altman (2004), BMJ

Likelihood ratios for hypotheses testing



Received: 30 March 2022 | Revised: 11 November 2022 | Accepted: 15 November 2022
DOI: 10.1002/simj.202201012

RESEARCH ARTICLE

Biometrical Journal →

Relative likelihood ratios for neutral comparisons of statistical tests in simulation studies^a

Qiuxi Huang¹ | Ludovic Trinquart^{1,2,3}

positive LR (rejection ratio): $LR^+ = \frac{\text{Power}}{\text{T1E rate}} \left(\text{e.g. } \frac{80\%}{5\%} = 16 \right)$

Likelihood ratios for hypotheses testing



Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses^o

M.J. Bayarri^{a,1}, Daniel J. Benjamin^b, James O. Berger^{c,*}, Thomas M. Sellke^d

Received: 30 March 2022 | Revised: 11 November 2022 | Accepted: 15 November 2022
DOI: 10.1002/bimj.202200102

RESEARCH ARTICLE

Biometrical Journal →

Relative likelihood ratios for neutral comparisons of statistical tests in simulation studies 

Qiuxi Huang¹  | Ludovic Trinquart^{1,2,3} 

positive LR (rejection ratio): $LR^{+} = \frac{\text{Power}}{\text{T1E rate}} \text{ (e.g. } \frac{80\%}{5\%} = 16 \text{)}$

negative LR (nonrejection ratio): $LR^{-} = \frac{1 - \text{Power}}{1 - \text{T1E rate}} \text{ (e.g. } \frac{20\%}{95\%} = 4/19 \text{)}$

Likelihood ratios for hypotheses testing



Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses^o

M.J. Bayarri^{a,1}, Daniel J. Benjamin^b, James O. Berger^{c,*}, Thomas M. Sellke^d

Received: 30 March 2022 | Revised: 11 November 2022 | Accepted: 15 November 2022
DOI: 10.1002/bimj.202200102

RESEARCH ARTICLE

Biometrical Journal →

Relative likelihood ratios for neutral comparisons of statistical tests in simulation studies 

Qiuxi Huang¹  | Ludovic Trinquart^{1,2,3} 

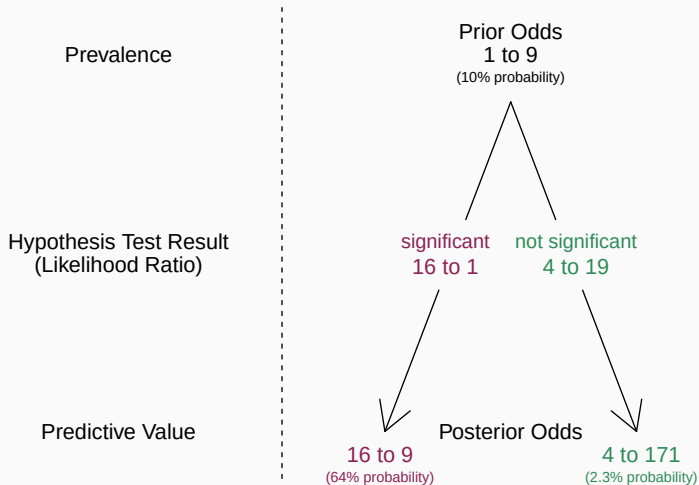
positive LR (rejection ratio): $LR^+ = \frac{\text{Power}}{\text{T1E rate}} \text{ (e.g. } \frac{80\%}{5\%} = 16 \text{)}$

negative LR (nonrejection ratio): $LR^- = \frac{1 - \text{Power}}{1 - \text{T1E rate}} \text{ (e.g. } \frac{20\%}{95\%} = 4/19 \text{)}$

Diagnostic odds ratio: $\text{DOR} = LR^+ / LR^- \text{ (e.g. } 16 / (4/19) = 76 \text{)}$

Glas et al. (2003), J Clin Epi

Bayesian Updating with Likelihood Ratios



Diagnostic odds ratios for hypotheses testing

Important reformulations:

1. Frequentist:

$$\text{DOR} = \frac{\text{Power Odds}}{\text{T1E Odds}}$$

Diagnostic odds ratios for hypotheses testing

Important reformulations:

1. Frequentist:

$$\text{DOR} = \frac{\text{Power Odds}}{\text{T1E Odds}}$$

2. Bayesian:

$$\text{DOR} = \frac{\text{Odds}(H_1 \mid \text{significant})}{\text{Odds}(H_1 \mid \text{not significant})}$$

Experimental unit-information index (EUII)

- Motivation: Likelihood ratios act **multiplicatively** on the odds of H_1 , so the average contribution from one experimental unit is $LR^{1/n}$:

$$LR = \underbrace{LR^{1/n} \times LR^{1/n} \times \dots \times LR^{1/n}}_{n \text{ times}}$$

→ $DOR^{1/n}$ represents the **evidentiary value of one experimental unit**:

$$EUII = \frac{(LR^+)^{1/n}}{(LR^-)^{1/n}} = DOR^{1/n}$$

One-sample one-sided z - and t -test

Power (in %)	Type-I error rate (in %)	DOR	z -test		t -test	
			n	EUII	n	EUII
80	5.0	76.0				
80	2.5	156.0				
90	5.0	171.0				
90	2.5	351.0				

One-sample one-sided z - and t -test

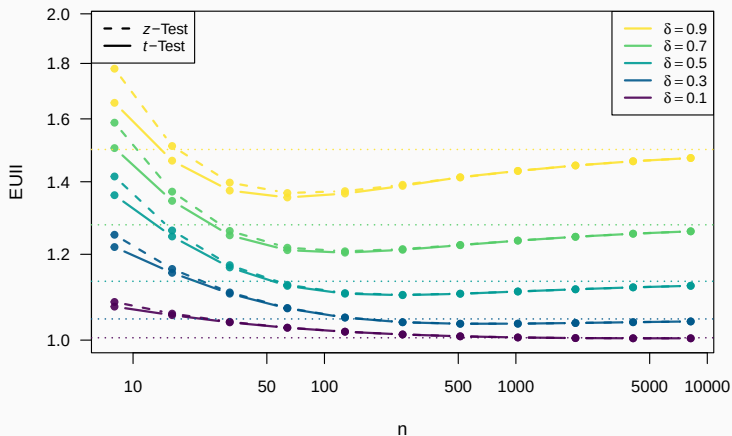
Power (in %)	Type-I error rate (in %)	DOR	z-test		t-test	
			n	EUII	n	EUII
80	5.0	76.0	12.6	1.41	14.1	1.36
80	2.5	156.0	16.0	1.37	18.0	1.32
90	5.0	171.0	17.5	1.34	18.9	1.31
90	2.5	351.0	21.4	1.31	23.4	1.28

↑

↑

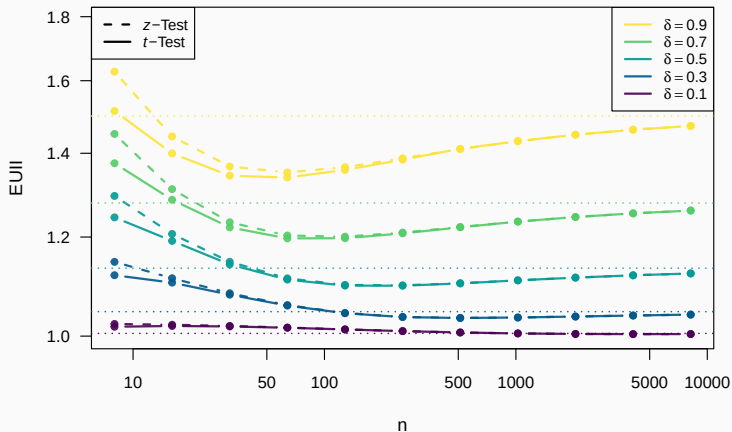
Required sample size to detect a
standardised effect $\delta = 0.7$

EUII for increasing n and fixed $\alpha = 5\%$ (one-sample, one-sided)



Limiting value $\lim_{n \rightarrow \infty} \text{EUII} = \exp(\delta^2/2)$

EUII for increasing n and fixed $\alpha = 5\%$ (one-sample, two-sided)



Limiting value $\lim_{n \rightarrow \infty} \text{EUII} = \exp(\delta^2/2)$

EUII for adaptive designs

- Sample size N is now a random variable
- Distinguish sample size for **significant** (N^+) and **not significant** (N^-) findings:

$$\text{EUII} = \frac{(\text{LR}^+)^{E(1/N^+)}}{(\text{LR}^-)^{E(1/N^-)}}$$

Computation of $E(1/N^+)$ and $E(1/N^-)$ requires specification of prior probability $\Pr(H_1)$, where we use $\Pr(H_1) \in \{0.01, 0.1, 0.5\}$.

Second-order Taylor approximation

Second-order Taylor approximation

$$E(1/N^{\bullet}) \approx \frac{1 + CV(N^{\bullet})^2}{E(N^{\bullet})}$$

uses mean sample size $E(N^{\bullet})$ and **coefficient of variation**

$$CV(N^{\bullet}) = \frac{\sqrt{\text{Var}(N^{\bullet})}}{E(N^{\bullet})} \quad (\text{where } \bullet \in \{+, -\})$$

$$\rightarrow EU_{II} \approx \frac{(LR^+)^{(1+CV(N^+)^2)/E(N^+)}}{(LR^-)^{(1+CV(N^-)^2)/E(N^-)}}$$

→ Mean and variance of sample size are taken into account

Simulation study

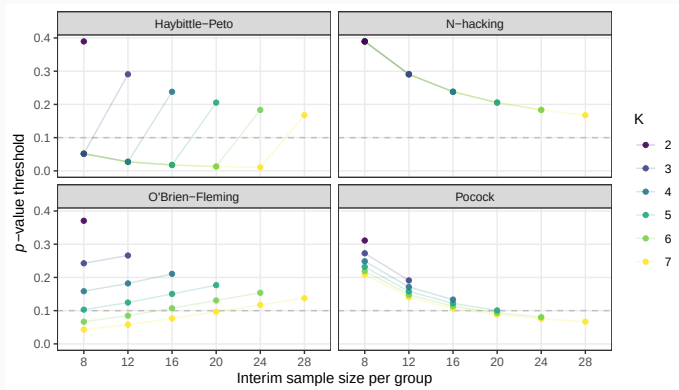
- Compare Reinagel's method with four group sequential designs:
 - Haybittle-Peto
 - N -hacking
 - O'Brien-Fleming
 - Pocock

for different values of

- $n_{\text{Max}} \in \{12, 16, \dots, 32\}$
- δ (standardised mean difference, SMD) $\in \{0.3, 0.5, 0.8, 1.0\}$
- $\Pr(H_1) \in \{0.01, 0.1, 0.5\}$

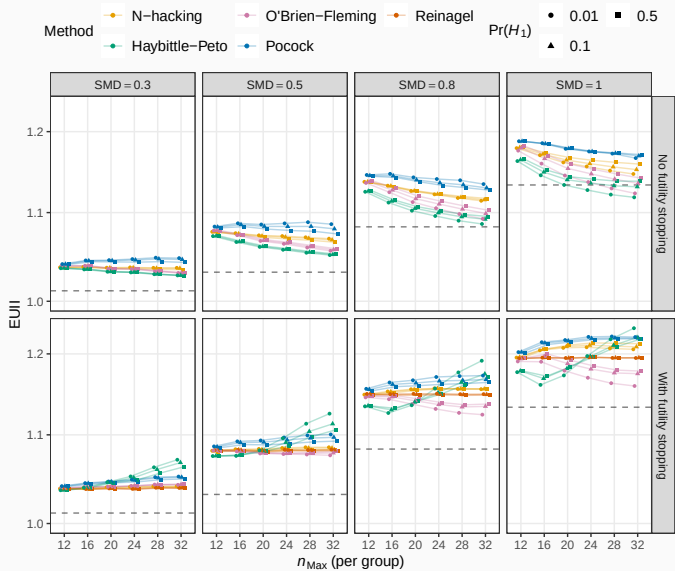
Futility Thresholds for P-Values

Futility stopping if **predictive power** < 10% at the **next interim analysis**:

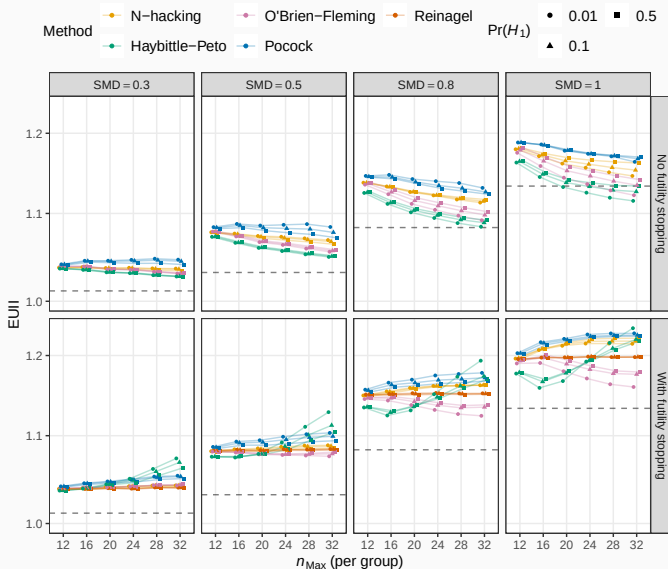


(The futility threshold of Reinagel's method is always 0.1)

Comparison of EUI



Comparison of EUII (Second-Order Approximation)



A new measure of statistical information that takes into account Type-I error rate, power, and sample size

- Future applications:
 - Group-sequential designs based on nonparametric tests
 - Bayesian adaptive designs
 - Dual-criterion
 - More complex experimental (block) designs
- Average the power over plausible values of the effect size
- Monte Carlo standard error of (log) EUII

Balancing Evidentiary Value and Sample Size of Adaptive Designs with Application to Animal Experiments

Leonhard Held^{1*} , Fadoua Balabdaoui² , Saverio Fontana¹ , Samuel
Pawel¹ 



References

- Blotwijk, S., Hernot, S., and Barbé, K. (2022). Group sequential designs for in vivo studies: Minimizing animal numbers and handling uncertainty in power analysis. *Research in Veterinary Science*, 145:248–254. doi:10.1016/j.rvsc.2022.03.003.
- Bonapersona, V., Hoijsink, H., Sarabdjitsingh, R. A., and Joëls, M. (2021). Increasing the statistical power of animal experiments with historical control data. *Nature Neuroscience*, 24(4):470–477. doi:10.1038/s41593-020-00792-3.
- Deeks, J. J. and Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *BMJ*, 329:168–169.
- Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., and Bossuyt, P. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*, 56(11):1129–1135. doi:10.1016/S0895-4356(03)00177-X.
- Neumann, K., Grittner, U., Piper, S. K., Rex, A., Florez-Vargas, O., Karystianis, G., Schneider, A., Wellwood, I., Siegerink, B., Ioannidis, J. P. A., Kimmelman, J., and Dirnagl, U. (2017). Increasing efficiency of preclinical research by group sequential designs. *PLOS Biology*, 15(3):e2001307. doi:10.1371/journal.pbio.2001307.
- Reinagel, P. (2023). Is N-Hacking Ever OK? The consequences of collecting more data in pursuit of statistical significance. *PLOS Biology*, 21(11):e3002345. doi:10.1371/journal.pbio.3002345.

Backup: Characteristics of group-sequential designs

One-sample, 4 equally spaced analyses with nominal significance thresholds α_i :

Method	α_1	α_2	α_3	α_4	T1E rate (in %)	n_{Max}
Pocock	0.00911	0.0091	0.0091	0.0091	2.50	59.2
O'Brien-Fleming	0.00003	0.0021	0.0097	0.0215	2.50	51.1
Haybittle-Peto	0.00050	0.0005	0.0005	0.0250	2.54	50.0

Comparison with fixed design ($n = 50$)

Comparison of group-sequential designs

