Biostatistics and Pre-clinical Science: Design of Experiments and Data Analysis

Fundamentals of Translational Oncology Workshop

Márcio A. Diniz, PhD
Biostatistics and Bioinformatics Research Center
Cedars-Sinai Medical Center

Cedars Sinai

cedars-sinai.org

Reproducibility in Science

Published: 28 March 2012

Drug development

Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis

Nature **483**, 531–533 (2012) | Cite this article

Only 11% (6) out of 43 landmark preclinical studies were reproducible by scientists in Amgen.

Only 25% out of 67 preclinical studies were reproducible by scientists in Bayer. 70% were oncology studies.

REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012 Mar;483(7391):531-3.

Reproducibility in Science

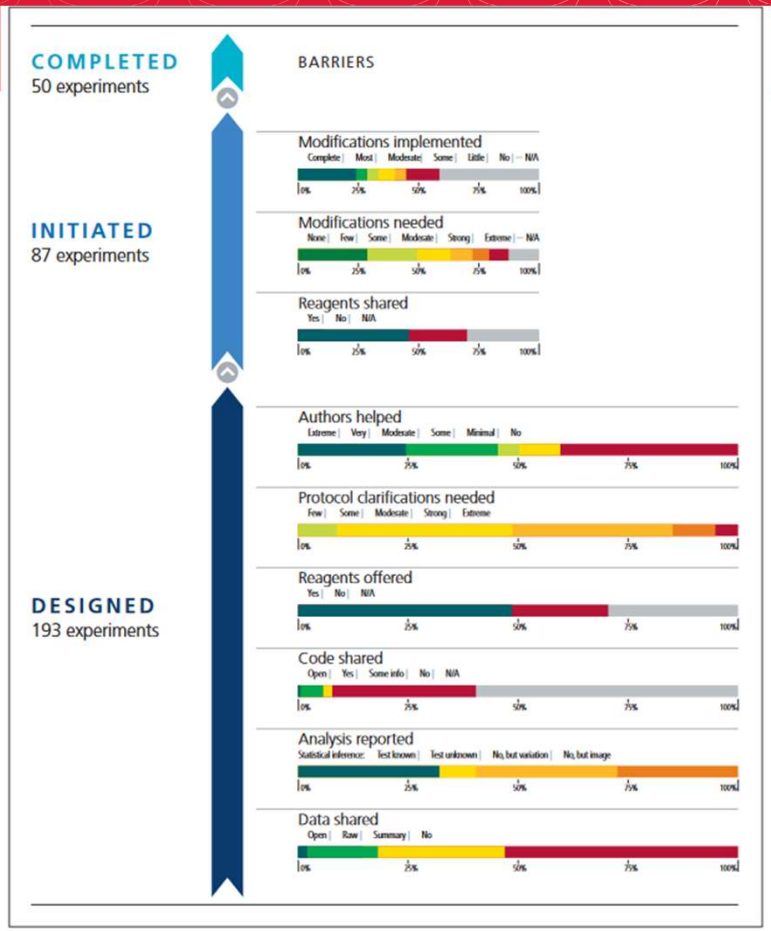
eLife

FEATURE ARTICLE

REPRODUCIBILITY IN CANCER BIOLOGY

Challenges for assessing replicability in preclinical cancer biology

TIMOTHY M ERRINGTON*, ALEXANDRIA DENIS†, NICOLE PERFITO‡, ELIZABETH IORNS AND BRIAN A NOSEK



Cedars Sinai

Errington TM, Denis A, Perfito N, Iorns E, Nosek BA. Reproducibility in cancer biology: challenges for assessing replicability in preclinical cancer biology. Elife. 2021 Dec 7;10:e67995.

Reporting

OPEN ACCESS Freely available online

PLoS one

Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals

Carol Kilkenny[1]*, Nick Parsons[2], Ed Kadyszewski[3], Michael F. W. Festing[4], Innes C. Cuthill[5], Derek Fry[6], Jane Hutton[7], Douglas G. Altman[8]

OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research

Carol Kilkenny[1]*, William J. Browne[2], Innes C. Cuthill[3], Michael Emerson[4], Douglas G. Altman[5]

Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PloS one. 2009 Nov 30;4(11):e7824.

Reporting

RESEARCH ARTICLE

The Devil Is in the Details: Incomplete Reporting in Preclinical Animal Research

Marc T. Avey[1,2]*, David Moher[1,3], Katrina J. Sullivan[1], Dean Fergusson[1], Gilly Griffin[1], Jeremy M. Grimshaw[1,4], Brian Hutton[1,3], Manoj M. Lalu[1,7], Malcolm Macleod[5], John Marshall[6], Shirley H. J. Mei[7], Michael Rudnicki[7], Duncan J. Stewart[7,8], Alexis F. Turgeon[9,10], Lauralyn McIntyre[1,11], Canadian Critical Care Translational Biology Group[¶]

Avey MT, Moher D, Sullivan KJ, Fergusson D, Griffin G, Grimshaw JM, Hutton B, Lalu MM, Macleod M, Marshall J, Mei SH. The devil is in the details: incomplete reporting in preclinical animal research. PloS one. 2016 Nov 17;11(11):e0166733.

Reporting

PLOS BIOLOGY

COMMUNITY PAGE

Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0

Nathalie Percie du Sert[1]*, Amrita Ahluwalia[2,3], Sabina Alam[4], Marc T. Avey[5], Monya Baker[6], William J. Browne[7], Alejandra Clark[8], Innes C. Cuthill[9], Ulrich Dirnagl[10], Michael Emerson[11], Paul Garner[12], Stephen T. Holgate[13], David W. Howells[14], Viki Hurst[1], Natasha A. Karp[15], Stanley E. Lazic[16], Katie Lidster[1], Catriona J. MacCallum[17], Malcolm Macleod[18], Esther J. Pearl[1], Ole H. Petersen[19], Frances Rawle[20], Penny Reynolds[21], Kieron Rooney[22], Emily S. Sena[18], Shai D. Silberberg[23], Thomas Steckler[24], Hanno Würbel[25]

Box 2. ARRIVE Essential 10

1. Study design
2. Sample size
3. Inclusion and exclusion criteria
4. Randomisation
5. Blinding
6. Outcome measures
7. Statistical methods
8. Experimental animals
9. Experimental procedures
10. Results

Cedars Sinai

Du Sert NP, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, Clark A, Cuthill IC, Dirnagl U, Emerson M, Garner P. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLoS biology. 2020 Jul 14;18(7):e3000411.

Successful interaction between statisticians and basic scientists

Stroke

SPECIAL REPORT

The Stroke Preclinical Assessment Network: Rationale, Design, Feasibility, and Stage 1 Results

Patrick D. Lyden, MD; Francesca Bosetti, PhD; Márcio A. Diniz, PhD; André Rogatko, PhD; James I. Koenig, PhD; Jessica Lamb, BS; Karisma A. Nagarkatti, MS; Ryan P. Cabeen, PhD; David C. Hess, MD; Pradip K. Kamat, PhD; Mohammad B. Khan, PhD; Kristofer Wood, BS; Krishnan Dhandapani, PhD; Ali S. Arbab, MD, PhD; Enrique C. Leira, MD, MS; Anil K. Chauhan, PhD; Nirav Dhanesha, PhD; Rakesh B. Patel, PhD; Mariia Kumskova, MD; Daniel Thedens, PhD; Andreia Morais, PhD; Takahiko Imai, PharmD, PhD; Tao Qin; Cenk Ayata, MD, PhD; Ligia S.B. Boisserand, PhD; Alison L. Herman, BA; Hannah E. Beatty, BS; Sofia E. Velazquez, BA; Sebastian Diaz-Perez, BS; Basavaraju G. Sanganahalli, PhD; Jelena M. Mihailovic, PhD; Fahmeed Hyder, PhD; Lauren H. Sansing, MD, MS; Raymond C. Koehler, PhD; Steven Lannon, BS; Yanrong Shi, MD, MS; Senthilkumar S. Karuppagounder, PhD; Adnan Bibic, PhD; Kazi Akhter, PhD; Jaroslaw Aronowski, PhD, MD; Louise D. McCullough, MD, PhD; Anjali Chauhan, PhD; Andrew Goh, MS; on behalf of the SPAN Investigators†

Six interventions were selected assuming that dose-response studies had determined the optimal dose.

Lyden PD, Bosetti F, Diniz MA, Rogatko A, Koenig JI, Lamb J, Nagarkatti KA, Cabeen RP, Hess DC, Kamat PK, Khan MB. The Stroke Preclinical Assessment Network: Rationale, Design, Feasibility, and Stage 1 Results. Stroke. 2022 Mar 31:10-161.

The Stroke Preclinical Assessment Network

Good practices:

- **Blinding;**
- **Randomization;**
- **Introduction of controlled variability;**
- **Adaptive sample sizes;**
- **Reporting followed ARRIVE 2.0.**

Opportunities of improvement:

- **Several drugs failed due lack of adequate dose-response studies.**

Randomization and Blinding

Academic Emergency Medicine
A GLOBAL JOURNAL OF EMERGENCY CARE

Free Access

Emergency Medicine Animal Research: Does Use of Randomization and Blinding Affect the Results?

Vik Bebarta MD, Dylan Luyten MD, Kennon Heard MD

First published: 28 June 2008 | https://doi.org/10.1111/j.1553-2712.2003.tb00056.x |
Citations: 123

Unconscious bias leads to more optimistic and non-reproducible results.

OPEN ACCESS Freely available online

PLOS ONE

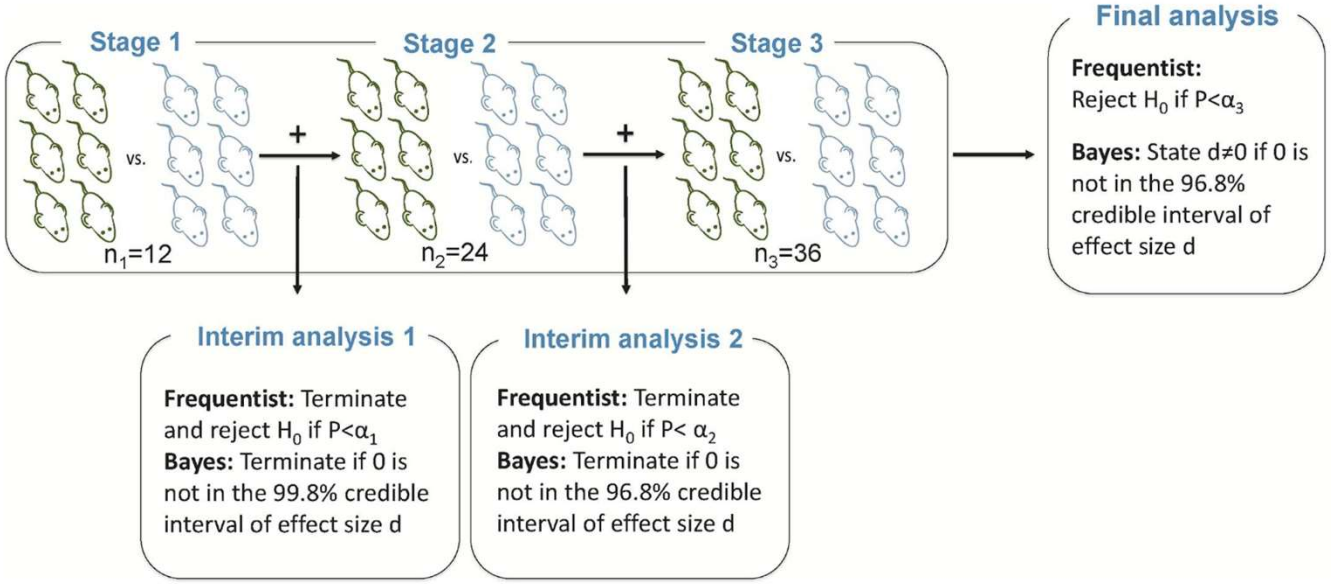
The Need for Randomization in Animal Trials: An Overview of Systematic Reviews

Jennifer A. Hirst[1*], Jeremy Howick[1*], Jeffrey K. Aronson[1], Nia Roberts[2], Rafael Perera[1], Constantinos Koshiaris, Carl Heneghan[1]

Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of randomization and blinding affect the results?. Academic Emergency Medicine. 2003 Jun;10(6):684-7.

Hirst JA, Howick J, Aronson JK, Roberts N, Perera R, Koshiaris C, Heneghan C. The need for randomization in animal trials: an overview of systematic reviews. PloS one. 2014 Jun 6;9(6):e98856.

Experiments with adaptive sample size



Experiments with adaptive sample size allow more efficient use of animals.

Neumann K, Grittner U, Piper SK, Rex A, Florez-Vargas O, Karystianis G, Schneider A, Wellwood I, Siegerink B, Ioannidis JP, Kimmelman J. Increasing efficiency of preclinical research by group sequential designs. PLoS biology. 2017 Mar 10;15(3):e2001307.

Experiments with adaptive sample size

Adaptive experimental designs require planning and clear pre-established rules. Sample sizes cannot be increased arbitrary such as

Sample sizes. For optogenetic activation experiments, cell-type-specific ablation experiments, and in vivo recordings (optrode recordings and calcium imaging), we continuously increased the number of animals until statistical significance was reached to support our conclusions. For rabies-mediated and anterograde tracing

Weber F, Hoang Do JP, Chung S, Beier KT, Bikov M, Saffari Doost M, Dan Y. Regulation of REM and non-REM sleep by periaqueductal GABAergic neurons. Nature communications. 2018 Jan 24;9(1):1-3.

Experiments with adaptive sample size

Comment | Open Access | Published: 23 April 2019

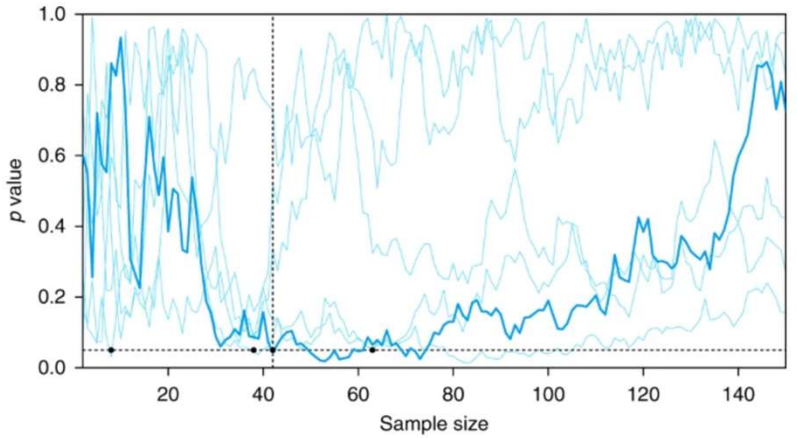
The problem with unadjusted multiple and sequential statistical testing

Casper Albers

Nature Communications **10**, Article number: 1921 (2019) | Cite this article

Without adequate statistical methods, sequential testing increases the false positive rate.

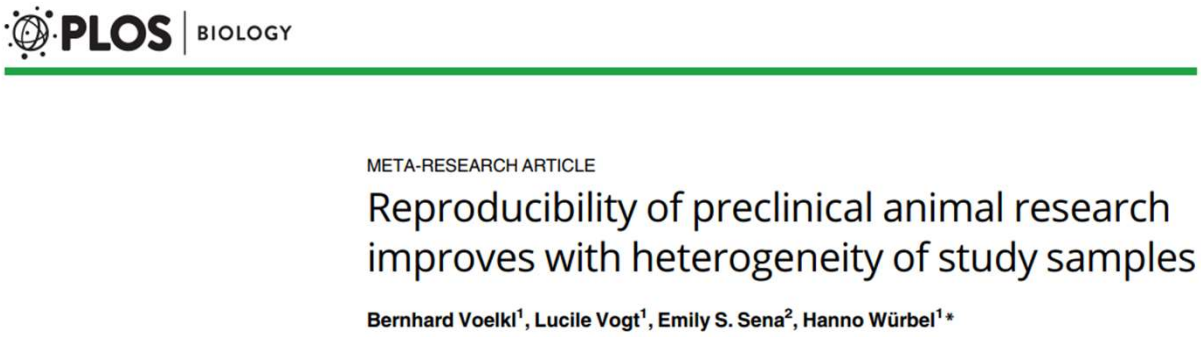
Fig. 1



A computer simulation of sequential *p*-values when there is no effect. The thick line is the instance discussed in the text; the five thin lines represent independent simulations. The black dots indicate the first instance where one of the runs falls below the 0.05 level. Two of the runs don't reach 0.05 before $n = 150$

Albers C. The problem with unadjusted multiple and sequential statistical testing. Nature Communications. 2019 Apr 23;10(1):1921.

Introduction of controlled variability



Multi-laboratory studies are more reproducible than single-laboratory studies.

Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS biology. 2018 Feb 22;16(2):e2003693.

Introduction of controlled variability

Systematic heterogenization for better reproducibility in animal experimentation

S Helene Richter

Lab Animal **46**, 343–349 (2017) | Cite this article

Variability can be introduced in single laboratory studies with more than one mouse model/strain and mini-batches of experiments.

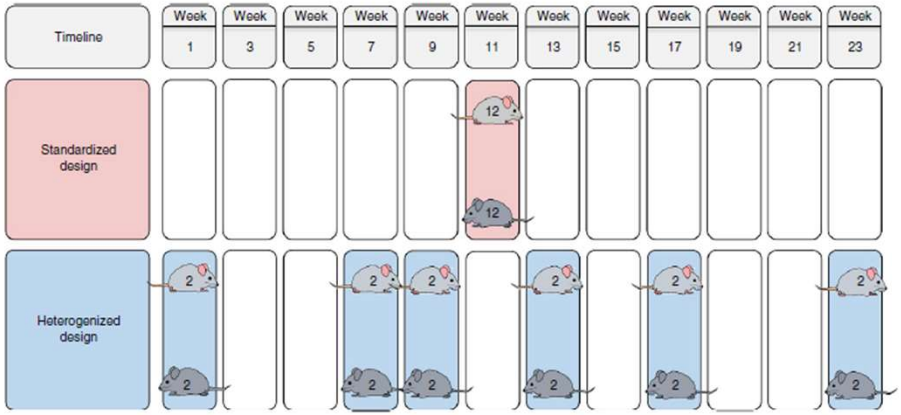
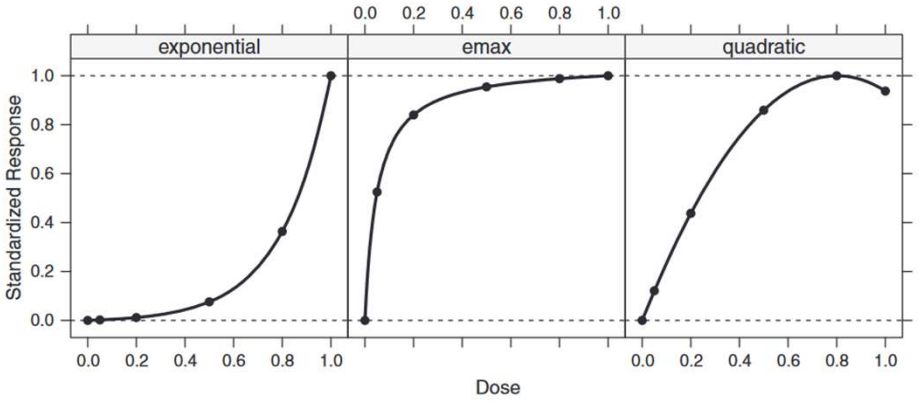


FIGURE 2 | Systematic heterogenization over time ("batch heterogenization"). Batch heterogenization aims to split experiments into small batches of animals that are tested some time apart (heterogenized design) instead of testing them at once in just one large batch (standardized design). Combining these "mini-experiments" in one big experiment is then assumed to increase representativeness of the whole study population, resulting in findings that are more reproducible between experiments and laboratories.

Richter SH. Systematic heterogenization for better reproducibility in animal experimentation. Lab animal. 2017 Sep;46(9):343-9..

Dose-Response studies



MCP-Mod approach is the current recommended statistical method by FDA to perform dose-response studies in humans.

ELSEVIER

Reproductive Toxicology

Volume 72, September 2017, Pages 97-105

Combined exposure to low doses of pesticides causes decreased birth weights in rats

Ulla Hass [a] ✉, Sofie Christiansen [a], Marta Axelstad [a], Martin Scholze [b], Julie Boberg [a]

Dose-biomarker-response modeling of the anticancer effect of ethaselen in a human non-small cell lung cancer xenograft mouse model

Suo-fu Ye, Jian Li, Shuang-min Ji, Hui-hui Zeng ✉ & Wei Lu ✉

Acta Pharmacologica Sinica **38**, 223–232 (2017) | Cite this article

Cedars Sinai

Hass U, Christiansen S, Axelstad M, Scholze M, Boberg J. Combined exposure to low doses of pesticides causes decreased birth weights in rats. Reproductive Toxicology. 2017 Sep 1;72:97-105.

Ye, Sf., Li, J., Ji, Sm. *et al.* Dose-biomarker-response modeling of the anticancer effect of ethaselen in a human non-small cell lung cancer xenograft mouse model. *Acta Pharmacol Sin* **38,** 223–232 (2017). https://doi.org/10.1038/aps.2016.114

Do all comparisons matter?

- We often consider all pairwise comparisons using Tukey test or Dunn test.

- However, some comparisons are less important.
- Hierarchical test procedures can helpful.
  - Fixed sequence procedure: H1 is more important than H2, and H2 is more than important than H3;
    - In this case, H1 can be tested at 5% significance level; Otherwise, we stop.
    - If we reject H1, then H2 can be tested at 5% significance level; Otherwise, we stop
    - If we reject H3, then we stop.

Do all comparisons matter?

- Hierarchical test procedures can helpful.
  - Fallback procedure: H1 is more important than H2, and H2 is more than important than H3;
    - H1 is tested at 1.66%.
    - H2 is tested at 3.22% if H1 is rejected; Otherwise, H2 is tested at 1.66%
    - H3 is tested at 5%, if H1 and H2 is rejected; Otherwise, H3 is tested at 1.66%

How should we handle missing data?

- Should we ignore animals that die before the collection of an endpoint of interest?
- Example: Animals with large stroke receiving ineffective drugs die before Day 30 assessment.
- Is the missing data informative?

JAMA Guide to Statistics and Methods

September 28, 2020

Worst-Rank Score Methods—A Nonparametric Approach to Informatively Missing Data

John M. Lachin, ScD[1]

» Author Affiliations

JAMA. 2020;324(16):1670-1671. doi:10.1001/jama.2020.7709

Lachin JM. Worst-Rank Score Methods—A Nonparametric Approach to Informatively Missing Data. JAMA. 2020 Oct 27;324(16):1670-1.

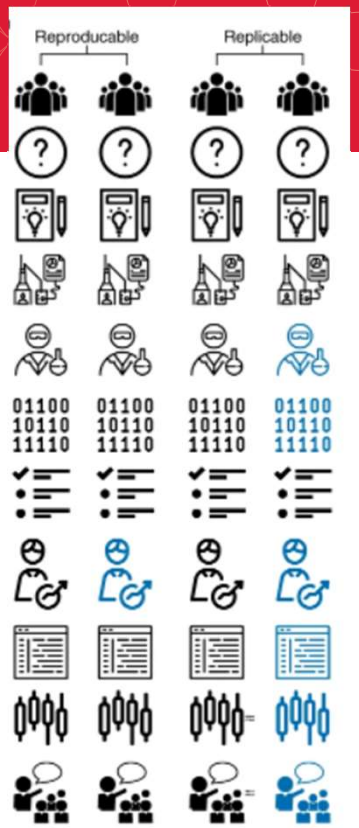
Opportunities of Interaction

- **Design of experiments**

- **Strategies for randomization and blinding;**
- **Strategies to introduce controlled variability in experiments;**
- **Power Considerations including experiments adaptive sample sizes.**

- **Data analyses for in-vitro and in-vivo studies**

- **Statistical rigor to conduct test of pre-established hypotheses;**
- **Code for analysis available to share with publication;**
- **Reporting according to ARRIVE guidelines.**



Cedars Sinai

Patil, P., Peng, R.D. & Leek, J.T. Publisher Correction: A visual tool for defining reproducibility and replicability. Nat Hum Behav 3, 886 (2019).

Opportunities of Interaction

Questions?

Marcio.diniz@cshs.org