How well can pre-clinical stroke researchers replicate effect sizes in their experiments? A meta-analysis of "accidental replications".

Principle investigators	: Prof. Jonathan Kimmelman, Bob Siegerink PhD, Prof. Ulrich Dirnagl
Research group	: CSB | CEHRIS | STREAM
Topic / keywords	: replication crisis in pre-clinical stroke research
Date	: 6th of April 2016
Version	: 2.0
Yield of the project	: *single paper*

I. Main aims or topic of the project

To examine how well experiments in preclinical stroke research replicate by investigating their effect sizes and the predictors of effect sizes in pre-clinical stroke placebo groups.

II. Background *(short overview to show gap in knowledge)*

The term replication crisis has been coined to indicate the low number of replicated effects in contemporary research.[1–3] The field of pre-clinical stroke research is no stranger to this phenomenon, where failed translation from bed to bedside is an eminent challenge.[4,5] Various ideas exist on the origins of this phenomenon, some of which are embedded within the scientific practice, amongst them publication bias (only publishing 'positive' experiments, not the negative), lack of statistical expertise and a focus on isolated "exciting" findings.[6,7] Identification of such biasing effects is a prerequisite to remedy deleterious scientific practices. Here, meta-analytics methods have been shown to allow to identify and quantify biasing effects in research, which lead to low replicability.[8]

In this project, we will employ meta-analytic methods to analyze the extent to which effect sizes in preclinical stroke studies are reproducible and which factors might negatively influence reproducibility. The best way to perform such a study would be to identify a large number of studies employing identical protocols for testing a drug, and then compare their effect sizes in the sense of "strict replications". The availability of strict replications for stroke drugs, however, is very small owing to high study design variability, the scarcity of direct replications, and nonpublication of preclinical data. Here, a more promising approach is used by examining the reproducibility effect sizes in the control animals of the different experiments only. By testing whether researchers reproduce effect sizes in control animals when they deploy identical methods, we can observe whether scientists that use identical protocols can achieve the same effects both within labs and different labs. The approach of using control animals significantly increases the availability of studies for such an analysis: researchers testing different drugs might nevertheless treat control animals identically. We will term such overlaps "accidental replications". These offer a window into the reproducibility of laboratory techniques.

In the first step, we will analyze whether replications of experiments which are methodologically similar provide more homogeneous measures of effect sizes in control animals than methodologically different experiments. For this purpose, we will define sub-sets of studies, a) which are identical in several methodological key features, b) which are similar in those methodological key features and c) which are methodologically different. If accidental replications lead to comparable effect sizes, we would expect the most homogenous distribution of effect sizes in the subset of identical experiments, followed by the subset of similar experiments and the least homogeneous distribution in the subset of different studies. This first step will give us for the first time a quantifiable measure of how well effect sizes are replicated in pre-clinical stroke research, if identical or similar methods are employed.

In the second step, we will analyze, whether the predefined methodological key features explain all the variance of effect sizes or whether other experimental design features exert influence on effect sizes. Such other features are "experimental validity features", such as randomization and blinding. We will determine, whether these features have an influence on the effect sizes.

III. Research questions (RQ)

1. Is the spread of infarct volumes in identical or similar placebo-experiments smaller than in non-identical experiments? (=RQ1)

2. a) Can the variation in effect sizes of stroke research experiments be fully explained by methodological key features?
   b) If not, do experimental validity features exert an effect? (=RQ2)

IV. Study design and data description

Design: cross-sectional analysis

Stepwise description of the analyses:

1. Data will be collected from preclinical stroke experiment CAMARADES database. This database includes data on the methods, the result, as well as the experimental design under which these experiments were performed. In this analysis, we will focus on the data from the placebo arm in a pre-clinical stroke experiments as these are results are not confounded by active treatment.

   The following subsets (experiments pertaining a certain therapeutic that was studied) from the CAMARADES database will be included in the query (determined in collaboration with Gillian Curry, Edinburgh):

   a. Anti-inflammatory
   b. Anti-oxidant/Antioxidant
   c. Antibiotic
   d. Antidepressant
   e. cholesterol modification
   f. citicholine
   g. Environment
   h. Estrogens/oestrogens
   i. Exercise
   j. Growth factor
   k. HMG-CoA reductase antagonist
   l. Hypothermia
   m. Immunosuppressant
   n. Mixed Training
   o. Nootropic

p. NOS Donors
q. NOS Inhibitor
r. Rho GTPase inhibitor
s. Stem cells
t. Thrombolytic
u. Training
v. Vitamins

2. The studied effect sizes will be 1) infarct volume (in mm^3) and 2) the variance of the infarct volume for each experiment.

3. including all placebo data from the CAMARADES database would provide the largest statistical power. However, it would also entail a substantial amount of noise, caused by the high variability of experimental settings. By creating sets of experiments with identical methodological key features, this problem can be circumvented. These sets will be called "identity sets", as they are identical regarding these pre-defined methodological key features. As we can expect that these identity sets will be small, the validity of the results is limited by the small number of experiments. To solve this problem, we will additionally create a set of experiments which are not identical, but similar regarding the predefined methodological key features. This will allow to create bigger sets called "similarity sets". These approach is likely to results into an optimal tradeoff between statistical noise and power.

4. To identify similarity sets, we will use categorical principal component analysis (CATPCA), where each set will be assigned values for a predefined number of components. In the next step, k-means clustering will identify experiments, which are similar in these component values. Each cluster will then be proven to be a similar set. The number of clusters will be identified based on the maximum possible similar-groups yielded by each clusters-number according to the following criteria: $0.1<CoV<1$ to both avoid identity groups and limit dispersion of probability distribution. From the spread in effect sizes in a) identity sets, b) similarity sets and c) the whole set of unique experiments we can answer research question 1. See below for an overview of the way we will identify identity sets and similarity sets in a detailed and stepwise fashion:

    a. Find **identity sets** of experiments identical in methodological key features. The methodological key features that need to be identical to form an identity set are predefined as:

        1. Species ID
        2. Sex
        3. Anasthetic ID
        4. Ventiliation
        5. Type of Injury
        6. Induction of Injury
        7. Use of Comorbid Animals
        8. Sample size in control group (dichotomized ≥10 vs. 1-9)
        9. Time of assessment
        10. Duration of Ischemia

b. Find **<u>similarity sets</u> of experiments** similar in methodological key features through categorical **principle component analysis** (CATPCA). Likely to yield more and larger sets than the identity sets, but this method also allows more lenience for non-similarity possibly increasing the noise in further analyses. PCA analysis will be executed on the following predefined list of variables (identical to 3a):

    1. Species ID
    2. Sex
    3. Anasthetic ID
    4. Ventiliation
    5. Type of Injury
    6. Induction of Injury
    7. Use of Comorbid Animals
    8. Sample size in control group (dichotomized ≥10 vs. 1-9)
    9. Time of assessment
    10. Duration of Ischemia

c. It is predefined that if a parameter in a/b shows more than 50% values labeled as "unknown" or "missing", it will not be included in the analysis.

d. We will compare the variation in the infarct volume and variation in the variance of infarct volumes in

    i. identical sets compared to
    ii. similar sets compared to
    iii. whole of unique experiments

e. The spread for each set will be given as the coefficient of variation and the interquartile range for each effect size parameter in the set.

f. It is known from that database that all studies exhibit a value for the effect size "final infarct volume", but not all studies for the effect size "variation of infarct volume", i.e. standard deviation. Thus, the analysis of the effect size "variation of infarct volumes" will be performed in the subset of studies, where this parameter is available.

5. Within each set (identical, similar and whole), what circumstances explain the variation in the effect size? Can these be explained by the methodological key features of the experiments? Or do experimental validity features also exert an effect? This analysis will answer RQ2.

    a. First, the methodological key features will be included as predictors of the predefined effect sized (infarct volume, variation of infarct volume) in a linear regression model. The percentage of explained variation will be inferred from the R2 value of the model.

    b. If the variation in effect size parameters cannot be explained by methodological key features ($R2 \leq 80\%$), we will analyze whether experimental validity features exert an

influence on the effect sizes. These features are available in the CAMARADES database. All parameters will be converted to dummy variables. To account for the fact that a regression model should have at least 10 cases per predictor, we will predefine a ranking of the available experimental validity features. For each bin of 10 experiments in a set an additional predictor will be included in the prediction models.

1. Explanation of Animal Exclusions
2. Allocation concealment
3. Random allocation to group
4. Blinded Assessment of Infarct Volume
5. Year of Publication (divided a priori in 5 year bins, with a lenience of +/- 2 at the upper and lower borders)
6. Sample size calculation
7. Monitoring of Physiological Variables
8. Control of Temperature
9. Compliance with animal welfare regulations
10. Statement of Potential Conflicts of Interest
11. Peer reviewed publication

VII. Statistical analyses

All statistical analyses will be performed in SPSS Version 23 with all code saved in syntax. In the following a step by step description of the different steps is shown.

1. **Cleaning of Database I**

    a. All variables will be changed to numerical variables.

2. **Studies, which are *not* Rat or Mouse will be deleted**

3. **Cleaning of Database II**

    The predefined methodological variables (see 3a) will be tested for frequency distribution. If more than 50% of the entries are “missing” or “unknown”, that variable will not be used in the analysis.

4. **Removing duplicates to create a placebo-only database.**
    In the original database each entry is defined by a unique effect size. However, studies with variations of drug related variables (drug itself, doses, application etc.) will result in multiple entries, but with the same placebo parameters. Thus, only one entry should be kept.
    A duplicate for this purpose is defined as identical regarding the “infarct volume” plus

"Animal" plus "number in control group" and all other predefined method parameters except of the dichotomized number of animals. By including the effect size, which is defined by a value of two decimals, it is prevented that identity sets are deleted, which are required for the analysis. By including "Animal", a cross-delete by chance is prevented.

5. Creation of identity sets

Identity sets will be created using the parameters in 3a. The threshold for the minimal number of experiments to enter the following statistical analysis is set to 5.

6. **Creation of similarity sets**

Similarity sets will be identified by categorical principal component analysis implemented in SPSS and following k-means clustering.

a. The method variables are set as defined in 3a.
b. Weighting of all variables is 1
c. Both continuous variables (duration of ischemia; time relative to reperfusion) are set to numeric. All other variables are set to nominal.
d. Discretization of data. All originally categorical data is set to "ranking". Duration of ischemia and time relative to reperfusion are set to grouping (data is put into bins).
e. At least 90% of data variation must be explained by the new dimensions.
f. The derived object scores are used to find similar sets by cluster analysis.
g. K-means clustering
   As the new object scores for each dimension are numerical values, a standard k-means clustering can be performed. The number of optimal clusters for the clustering step will be determined by the highest number of usable clusters for each cluster n. A usable cluster is defined by the coefficient of variation within a cluster being $0.1<CoV<1$ for all cluster variables.

7. Analyzing the spread of effect sizes (RQ1)

The spread of effect sizes will be compared by the **coefficient of variation** and the **IQR** for each identity set, each similarity set and the whole set of unique experiments.

	Infarct volume CoV	Infarct volume IQR
Identity set N		
Similarity Set N		
Whole dataset		

While CoV is a dimensionless number and therefore can be used to compare different groups of measurements with completely different properties, the IQR is an actual range which directly refers to the measurements dimensions. Therefore, while the CoV provides an abstract measure of the data spread in each group, the IQR gives a more intuitive measure with a direct relation to the measured subject.

8. **Analyzing the predictors of effect sizes (RQ2)**

a. **Do methodological key features predict effect sizes?**

The pre-defined methodological key features (see 3a) will be used in the identity sets, in the similarity sets and the whole set as predictors of the effect sizes in a linear regression analysis. We will infer from the R2 of the models, whether all variance of effect sizes is explained by methodological key features (R2>80%).

b. **Do experimental validity features predict effect sizes?**

In case the variation cannot be explained by methodological key features alone in a certain set (R2≤80%), the experimental validity features pre-defined (see 4c) for the regression analysis will be used as predictors of the effect sizes. We will build the model based on all variables if sample size of the different sets allows us to do so (see 5b). If not, a forward stepwise regression model will be used allow for less predictors to be tested at once.

	Model	R2	predictors in the model*
Identity set N	standard model (all predictors)		This will give a list of all abbrevaietd variables that we the full model
	forward model		Here we show only those which remain

Similarity set N	standard model (all predictors)		
	forward model		

all unique datasets	standard model (all predictors)		
	forward model		

*The strongest predictor(s) will be mentioned in the text separately

VIII sample size / power considerations

The data that we will be using is already collecting within the CAMARADES dataset. Therefore, the sample size is fixed at the final number of individual experiments, which will be left after steps VII 1-4.

IX post publication data integrity measures

data + code will be made available whenever possible (if in line with CAMARADES guidelines).

X Research team

Vince I. Madai (project researcher)
Jonathan Kimmelman (main project supervisor)
Bob Siegerink (main statistical supervisor)
Ulrich Dirnagl (conceptual support)
Gillian Curry (database support)
Emily Sena (conceptual support)
Michelle LIvne (statistical support with CATPCA and cluster analysis)
X

XI description of possible conflicts of interest

»» None

XII Cost

No additional costs for material and measurements

XIII References

1. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10,** 712–712 (2011).
2. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).
3. Collaboration, O. S. An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspect. Psychol. Sci.* **7,** 657–660 (2012).
4. Perel, P. *et al.* Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ* **334,** 197 (2007).
5. Sena, E., van der Worp, H. B., Howells, D. & Macleod, M. How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci.* **30,** 433–439 (2007).
6. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med* **2,** e124 (2005).
7. Steckler, T. Editorial: preclinical data reproducibility for R&D-the challenge for neuroscience. *SpringerPlus* **4,** 1 (2015).
8. Sena, E. S., Currie, G. L., McCann, S. K., Macleod, M. R. & Howells, D. W. Systematic Reviews and Meta-Analysis of Preclinical Studies: Why Perform Them and How to Appraise Them Critically. *J. Cereb. Blood Flow Metab.* **34,** 737–742 (2014).