REVIEW ARTICLE

Journal of Pre-Clinical and Clinical Research 2020, Vol 14, No 4, 165-168
www.jpccr.eu

Irreproducibility –The deadly sin of preclinical research in drug development

Sadasivan Kalathil Pillai[1,A,C-D,F], **Katsumi Kobayashi**[2,B,E], **Mathews Michael**[3,C,E], **Meena Arumugam**[4,C,E]

[1] *International Institute of Biotechnology and Toxicology (IIBAT), Padappai, India*
[2] *Ex-Cabinet Secretary Researcher, Food Safety Commission of Japan, Akasaka, Japan*
[3] *Trichinopoly.i, Tiruchirappalli, India*
[4] *K.K. College of Pharmacy, Chennai, India*
A – Research concept and design, B – Collection and/or assembly of data, C – Data analysis and interpretation, D – Writing the article, E – Critical revision of the article, F – Final approval of article

Sadasivan Kalathil Pillai, Katsumi Kobayashi, Mathews Michael, Meena A. Irreproducibility – The deadly sin of preclinical research in drug development. J Pre-Clin Clin Res. 2020; 14(4): 165–168. doi: 10.26444/jpccr/131017

Abstract

Introduction. In recent years the irreproducibility of preclinical studies has become a serious concern in drug developmental research. The findings of preclinical studies that cannot be reproduced are a drain on public resources and slow down the drug discovery process. Among the various factors that contribute to irreproducibility in preclinical drug developmental research, poor statistical analysis and weak experimental design play a major role in the failure of drugs in clinical research.
Objective. The aim of this review is to describe key factors, such as poor statistical analysis and weak experimental design, that contribute to the irreproducibility of preclinical studies in drug development, and how such studies slow down the drug development process.
Brief description of the state of knowledge. The irreproducibility of preclinical research is a serious issue that researchers, especially those who are involved in drug discovery, are facing today. The irreproducibility of research drains public resources, time, and diminish the trust of the common man in the research community. The factors that contribute to the irreproducibility of preclinical research are related to experiment design and improper statistical analysis of the experimental data. Most of these factors can be eliminated by researchers developing a commitment to science and society.
Conclusion. Poor experimental design and lack of knowledge or limited knowledge of statistical analysis of data contribute significantly to the irreproducibility of preclinical research. A well-designed experiment with proper statistical analysis of data conducted by committed researchers rarely fails to reproduce.

Key words

data irreproducibility, preclinical studies, p-values

INTRODUCTION

Irreproducibility of research within the laboratory where it was originally conducted and/or in other laboratories is a severe crisis that the research community is facing today. An accurate and detailed scientific publication on research findings should be produced, so that others can reproduce and build on those findings. Scientists struggle to reproduce research findings from vaguely explained experimental designs, and arrive at conclusions by using poorly-described statistical analysis. Compared to other fields of research, issues of irreproducibility are more prominent in preclinical drug research [1]. A survey revealed that scientists of pharmaceutical companies failed to reproduce the conclusion of more than 75% of peer-reviewed manuscripts [2]. Another similar survey conducted by 1,576 researchers revealed that more than 70% of them failed to reproduce the experiments of other researchers, and more than half failed to reproduce their own experiments [3]. Irreproducibility of research findings has become a serious concern to funders and policy makers [4]. In the United States, approximately US$28 billion per year is spent on irreproducible preclinical research [5]. The inability to reproduce research findings diminishes public confidence in science and leads to the waste of resources [6].

Address for correspondence: Sadasivan Kalathil Pillai, International Institute of Biotechnology and Toxicology (IIBAT), Padappai, Kancheepuram Dist, India
E-mail: pillaiksp@yahoo.com

Received: 29.10.2020; accepted: 30.11.2020; first published: 21.12.2020

OBJECTIVE

This review presents the current state of knowledge regarding the key factors responsible for the irreproducibility of preclinical studies in drug development, and how such studies affect drug discovery.

DESCRIPTION OF THE STATE OF KNOWLEDGE

Contributing factors of irreproducibility of research. There are several reasons for the irreproducibility of an experiment. The key factors that contribute to the irreproducibility are given in Table 1. Scientists involved in the preclinical development of drugs tend to submit positive and favourable results for publication in scientific journals, some of which, interestingly, are biased towards publishing flashy, positive results [10]. Several new drug molecules have failed in clinical trials because the negative results obtained in the preclinical studies conducted with these drug molecules were not disclosed [11]. For example, a new tuberculosis vaccine

Table 1. Factors contributing to the irreproducibility of research

Factors	Explanation
Pressure to publish research findings [7]	In organizations where research performance is measured by the number of publications made, there exists pressure to publish papers. In such organizations, researchers tend to publish their findings without confirming them.
Biased reporting [7]	Pressure from the investors and top management directly or indirectly put on the scientists to report scientific findings in a biased manner so as to satisfy them.
Poor experimental design [8]	In preclinical experiments, several factors affect the result of the study. A sound scientific rationale should be established while selecting the animal model and determining the sample size of each group.
Non-validation of experiments, non-qualification/ calibration of equipment, inadequate validation of reagents and biological materials [8,9]	Only validated experiment performed using qualified / calibrated equipment is reproducible. It is important to use validated reagents and biological materials in the experiment to obtain reproducible results.
Selection of appropriate statistical tool and statistical significant level [9]	The data will be wrongly interpreted if the statistical tool used to analyse the data is inappropriate. A second thought should be given for assessing significance at 5% probability level.

failed in clinical trials to show efficacy because the scientists presented only the positive results of animal studies, and did not disclose negative results [12]. The neuroprotective drug NXY-059, developed by AstraZeneca, the multinational pharmaceutical and biopharmaceutical company, was effective in experimental/animal models, but ineffective in a large clinical trial. A meta-analysis of individual animal data of 15 studies conducted with NXY-059 revealed that there were several sources of bias in the studies, e.g. publication bias, absence of sample size determination, analysis of data using limited statistical power [13], and improper randomization [14]. Randomization of animals in experimental and control groups is vital because randomization reduces bias in animal studies [15, 16], and according to Hess [17], wherever possible scientists should be blinded to treatment groups. Data from animal studies become reliable only when the animals are distributed randomly into treatment groups, and blind assessment of the outcome of the study is performed [18]. Bebarta et al. [19] stated that 'animal experiments where randomization and blind testing are not reported are five times more likely to report positive results'. Animal studies conducted with NXY-059 where randomization or blinding was not reported, showed an efficacy of 30% bigger than in studies that reported randomization or blinding [20].

In various scientific journals, sample size determination is not very well presented with due importance in animal studies [21]. The guidelines of regulatory toxicology unambiguously indicate the number of animals to be used in a group for study. In the research and development of a pharmaceutical company, where a large number of new chemical entities (NCEs) are synthesized the scientists, often carry out experiments with an 'inadequate number' of animals. Results from such studies may not be reproducible and may fail to provide the desired information on the effectiveness of the NCE [22]. Several scientists believe six animals per group is an adequate sample size on which to perform animal experiments; however, this has neither a scientific rationale nor a statistical basis [21].

Determination of sample size while designing an experiment is essential, which is illustrated in the following example: A study was conducted in rats to evaluate the hypoglycaemic potential of a herbal formulation. Twelve rats, divided into 2 groups of 6 rats each, were made hyperglycaemic by injecting streptozotocin. To Group 1 (G1), distilled water and to Group 2 (G2), herbal formulation were administered. Glucose in blood was measured in all the animals at the end of the experiment. A decrease of glucose >10% in G2 compared to G1 was considered as significant. Glucose level measured in G2 – 165±24 mg/dl, in G1 – 190 ± 24 mg/dl. Although the decrease observed was about 13% in G2, it was not statistically significant according to Student's t-test. Blood glucose in G2 would have differed significantly from G1 if the study had been conducted with a smple size of 14 rats each in G1 and G2. Several methods are available to calculate sample size [21].

A significant *P*- value is 'insignificant' in judging a significant difference. Although several decision trees are available for selecting an appropriate statistical tool for the analysis of preclinical data [23, 24], there is no congruence among scientists across different countries in the selection of statistical tool for the analysis of such data [25, 26]. In preclinical studies, the *P*-value is used to determine the significant differences in measurable items, such as functional observational battery, urinalysis, haematology, blood chemistry, organ weights, etc. [27]. The *P*-value is the probability of the observed data given that the null hypothesis is true [28]. Widely- used critical values such as $P<0.05$, $P<0.01$, and $P<0.001$ to denote specific levels of statistical significance may not always have biological relevance [29, 30]. Ronald Fisher, who introduced the *P* –value, never considered it as a definitive test to classify the data into significant and non-significant [31]. Fisher only meant the *P*-value to be used as a rough numerical guide to the strength of evidence against the null hypothesis [32].

Most researchers believe that the ultimate objective of a study is to calculate '$P < 0.05$' resulting from null hypothesis significance tests [33]. This belief is based on the endorsement of the *P*-value by Fisher on certifying experimental result significant or non-significant: 'A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this [$P < 0.05$] level of significance' [34]. A false belief prevails among researchers that $P < 0.05$ does not apply to 'noise' (random data that obscure deterministic data of interest), hence, replicability is assured [35]. However, statistical significance ($P < 0.05$) can easily be obtained from mere noise [36]; hence, most of the studies with a statistically significant difference cannot be reproduced. Distinctly classifying the results of a study into statistically significant and non-significant, is itself considered as a statistical error [37]. Hence, the misuse of *P*-value is a major reason for the irreproducibility of the research findings [38]. According to Trafimow and Marks [39], the significance level of *P* at 0.05 is too easy to pass, and sometimes serves as an excuse for lower quality research.

Relying more on *P*- values for judging a significant difference may encourage unethical research practices [40]; for example, a *P*-value of 0.05 does not mean that the probability data that arose by chance alone is 1 in 20 [41].

Is the *P*-value misinterpreted? Some statisticians opined that *P*-values are commonly misunderstood and misinterpreted [42]. Several authors stated that the hypothesis set for *P*-values is itself irrelevant; hence, *P*-values overstate the evidence against those hypotheses [43, 44]. A common belief among scientists is that $P < 0.05$ indicates a 95% chance that a given hypothesis is correct. In fact, $P < 0.05$ only signifies whether the null hypothesis is true, all other assumptions made are valid, and there is a 5% chance of obtaining a result at least as extreme as the one observed [38]. A small *P*-value can be obtained by increasing the number of observations. In experiments where the number of observations is large, the *P*-value usually becomes statistically significant. Such significant *P*- values do not provide evidence for the quantitative significance of the effect (magnitude of effect) in the study, therefore, they lead to erroneous conclusions and interpretations [43, 45, 46]. One suggestion to overcome this situation is to change the *P*-value threshold for statistical significance from 0.05 to 0.005 [47]. Several statisticians have suggested replacing the *P*-value with Bayes' rule, which explains probability as the plausibility of an outcome, rather than as the potential frequency of that outcome [31]. Another suggestion is to use the *D*-value, which connects effect size and discrimination error along with *P*-value. *D*-value, unlike the *P*-value, is not affected by sample size [48]. It has also been suggested to judge the significance of *P*-value together with confidence intervals. This will give a better understanding of whether the observed difference represents a true difference in the entire population from which the sample has been drawn. The data on which a significant difference is assessed using *P*-value, along with confidence intervals, is reproducible [49]. Confidence intervals combine the concepts of both biological and statistical significance [50]. In this context, it should be mentioned that the editors of *Basic and Applied Social Psychology* decided in 2015 not to publish papers containing *P*-values reported without confidence intervals [39], and the journal *Osteoarthritis and Cartilage* prefer confidence intervals to *P*- values [49].

It is essential that the scientists are aware of the underlying principle of the statistical tool used for the analysis of the data. Statistical analysis should not override the experience of the researcher in interpreting the results of experiments [22].

CONCLUSION

Research findings that are not reproducible misguide the research community, and irreproducible preclinical research has hampered the pace of drug development. Reproducibility of preclinical research can be achieved by the robust design of an experiment, and by understanding the underlying principle of the statistical tool intended for analyzing the data gathered from the experiment. It would be ideal to identify the elements that might affect the reproducibility of the results of the research right at the stage of design of the experiment, and implement a plan to mitigate them.

REFERENCES

1. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature. 2014; 505: 612–613. https://doi.org/10.1038/505612a
2. Begley C, Ellis L. Raise standards for preclinical cancer research. Nature. 2012; 483: 531–533. https://doi.org/10.1038/483531a
3. Baker M. 1,500 Scientists lift the lid on reproducibility. Nature. 2016; 533: 452–454. https://doi:10.1038/533452a
4. Goodman SN, Fanelli D, Loannidis JPA. What does research reproducibility mean? Sci Transl Med. 2016; 8: 1–6. https://doi:10.1126/scitranslmed.aaf5027
5. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol. 2015; 13: e1002165. https://doi.org/10.1371/journal.pbio.1002165
6. Johnson VE. Revised standards for statistical evidence. Proc Nat Acad Sci. 2013; 110 (48): 19313–19317. https://doi.org/10.1073/pnas.1313476110
7. Boulbes DR, Costello TJ, Baggerly KA, Fan F, Wang R, Bhattacharya R, et al. A survey on data reproducibility and the effect of publication process on the ethical reporting of laboratory research. Clin Cancer Res. 2018; 24: 3447–3455. https://doi.org/10.1158/1078-0432.CCR-18-0227
8. Daniel C. Poorly designed animal experiments in the spotlight. Nature. 2015; doi: 10.1038/nature.2015.18559
9. Freedman LP, Venugopalan G, Wisman R. Reproducibility 2020: Progress and priorities. F1000Res. 2017; 6: 604. https://doi.org/10.12688/f1000research.11334.1
10. Loannidis JP. Why most published research findings are false. PLoS Med. 2005; 8: e124. https://doi.org/10.1371/journal.pmed.0020124
11. Weaver J. Animal studies paint misleading picture. Nature. 2010; doi.org/10.1038/news.2010.158
12. Cohen D. Oxford TB vaccine study calls into question selective use of animal data. BMJ. 2018; 360: j5845. https://doi.org/10.1136/bmj.j5845
13. Bath PM, Gray LJ, Bath AJ, Buchan A, Miyata T, Green AR. Effects of NXY-059 in experimental stroke: an individual animal meta-analysis. Br J Pharmacol. 2009; 157: 1157–1171. https://doi.org/10.1111/j.1476-5381.2009.00196.x
14. Savitz SI. A critical appraisal of the NXY-059 neuroprotection studies for acute stroke: A need for more rigorous testing of neuroprotective agents in animal models of stroke. Exper Neurol. 2007; 205: 201–205. https://doi.org/10.1016/j.expneurol.2007.03.003
15. Festing MFW, Altman DG. Guidelines for the Design and Statistical Analysis of Experiments Using Laboratory Animals. ILAR J. 2002; 43: 244–458. https://doi.org/10.1093/ilar.43.4.244
16. Bespalov A, WickeK, Castagné V. Blinding and Randomization. In: Bespalov A, Michel M, StecklerT, editors. Good research practice in non-clinical pharmacology and biomedicine. Handbook of Experimental Pharmacology. Springer, Cham, 2019. Vol. 257.
17. Hess KR. Statistical design considerations in animal studies. Cancer Res. 2011; 71(2): 625. https://doi.org/10.1158/0008-5472.CAN-10-3296
18. Macleod M. Why animal research needs to improve. Nature. 2011; 477: 511. https://doi.org/10.1038/477511a
19. Bebarta V, Luyte D, Heard K. Emergency medicine research: Does use of randomization and blinding affect the results? Acad Emerg Med. 2003; 10: 684–687. https://doi.org/ 10.1111/j.1553-2712.2003.tb00056.x
20. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke. 2008; 39: 2824–2829. https://doi.org/10.1161/STROKEAHA.108.515957
21. Charan J, Kantharia ND. How to Calculate Sample Size in Animal Studies? J Pharmacol Pharmacother. 2013; 4: 303–306. https://doi.org/10.4103/0976-500X.119726
22. Kobayashi K, Pillai KS. A handbook of applied statistics in pharmacology. New York: CRC Press; 2013.
23. Pillai KS. Statistical analysis in non-clinical GLP studies. In: Mohanan PV, editor. Good laboratory practice and regulatory issue. Bombay: Education Book Centre; 2006.
24. Pillai KS. Statistical methods in regulatory toxicology. In: Sengupta R. editor. Regulatory toxicology – essentially practical aspects. Delhi: Narosa Publishing House Pvt Ltd; 2016.
25. Kobayashi K, Pillai KS, Sakuratani Y, SuzukiM,Jie W. Do we need to examine the quantitative data obtained from toxicity studies for both normality and homogeneity of variance? J Environ Biol. 2008; 29: 47–52.
26. Kobayashi K, Pillai KS, Guhatakurta S, Cherian KM. Statistical tools for analysing the data obtained from repeated dose toxicity studies with rodents: A comparison of the statistical tools used in Japan with that of used in other countries. J Environ Biol. 2011; 32: 11–16.
27. Kobayashi K, Pillai KS, Michael M, Cherian KM, Ohnishi M. Determining NOEL/NOAEL in repeat-dose toxicity studies, when the

low dose group shows significant difference in quantitative data. Lab Anim Res. 2010; 26: 133–137. https://doi.org/10.5625/lar.2010.26.2.133
28. Altman DG. Practical statistics for medical research. 1st ed. London: Chapman and Hall; 1991.
29. Kobayashi K, Pillai KS. Applied statistics in toxicology and pharmacology, Science Publishers Inc., USA; 2003.
30. OECD Guidance document 116 on conduct and design of chronic toxicity and carcinogenicity studies, Supporting test guidelines 451, 452 and 453- 2nd ed. Paris; 2012. p. 114–143.
31. Nuzzo R. Scientific Method: Statistical Errors. Nature. 2014; 506: 150–152. https://doi.org/ 10.1038/506150a
32. Goodman S. A dirty dozen: Twelve p-value misconceptions. Semin Hematol. 2008; 45: 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003
33. Hubbard R, Haig BD, Parsa RA. The limited role of formal statistical inference in scientific inference. Am Stat. 2019; 73: 91–98. https://doi.org/10.1080/00031305.2018.1464947
34. Fisher RA. The arrangement of field experiments. J Ministry Agri Great Britain. 1926; 33: 503–513.
35. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. Am Stat. 2019; 73: 235–245. https://doi.org/10.1080/00031305.2018.1527253
36. Bem DJ. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. J PersSoc Psychol. 2011; 100: 407–425. https://doi.org/10.1037/a0021524
37. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. Am Stat. 2006; 60: 328–331. https://doi.org/10.1198/000313006X152649
38. Baker M. Statisticians issue warning over misuse of p values. Nature. 2016; 531: 151. https://doi.org/10.1038/nature.2016.19503
39. Trafimow D, Marks M. Editorial. Basic Appl Social Psych. 2015; 37: 1, 1–2. https://doi.org/10.1080/01973533.2015.1012991
40. Head ML, Holman L, Lanfear R, KahnAT, Jennions MD. The extent and consequences of p-hacking in Science. PLoS Biol. 2015; 13, e1002106. https://doi.org/10.1371/journal.pbio.1002106
41. Price R, Bethune R, Massey L. Problem with p values: Why p values do not tell you if your treatment is likely to work. Postgrad Med J. 2020; 96: 1–3. doi: 10.1136/postgradmedj-2019-137079
42. Sterne JA, Smith DG. Sifting the evidence-What's wrong with significance tests? BMJ. 2001; 322: 226–231. https://doi.org/10.1136/bmj.322.7280.226
43. WasserestinRL, Lazar NA. The ASA's statement on p-values: Context, process, and purpose. Am Stat. 2016; 70: 129–133. https://doi.org/10.1080/00031305.2016.1154108
44. Greenland S. Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with S-values. Am Stat. 2019; 73: 106–114. https://doi.org/10.1080/00031305.2018.1529625
45. Kruschke JK. What to believe: Bayesian methods for data analysis. Trends Cogn Sci. 2010; 14: 293–300. https://doi.org/ 10.1016/j.tics.2010.05.001
46. Chander NG. Beyond statistical significance. J Indian Prosthodont Soc. 2019; 19: 201–202.
47. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers, EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018; 2: 6–10. https://doi.org/10.1038/s41562-017-0189-z
48. Demidenko E. The p-value you can't buy. Am Stat. 2016; 70: 33–38. https://doi.org/10.1080/00031305.2015.1069760
49. Ranstam J. Why the p-value culture is bad and confidence intervals a better alternative. Osteoarthr Cartil. 2012; 20: 805–808. https://doi.org/10.1016/j.joca.2012.04.001
50. Redmond AC, Keenan A. Understanding statistics – Putting p-values into perspective. J Am Podiatric Med Assoc. 2002; 92: 297–305. https://doi.org/10.7547/87507315-92-5-297