ORIGINAL PAPER

Effect of Initial Conditions on Reproducibility of Scientific Research

Benjamin Djulbegovic[1,2], Iztok Hozo[3]

University of South Florida, Division for Evidence-based Medicine and Health Outcomes Research, Department of Medicine, Tampa, Florida, USA[1]
H. Lee Moffitt Cancer Center & Research Institute, Departments of Hematology and Health Outcomes and Behavior, Tampa, Florida, USA[2]
Department of Mathematics and Actuarial Science, Indiana University Northwest, Gary, Indiana, USA[3]

Corresponding author: Prof. Benjamin Djulbegovic, MD, PhD. Distinguished Professor. Division of Evidence-based Medicine, Department of Internal Medicine. University of South Florida & H. Lee Moffitt Cancer Center & Research Institute. Mailing address: 3515 East Fletcher Avenue, MDT 1201, MDC 27. Tampa, FL 33612. Phone: 813-396-2349. Fax: 813-905-8909. E-mail: bdjulbeg@health.usf.edu

ABSTRACT

Background: It is estimated that about half of currently published research cannot be reproduced. Many reasons have been offered as explanations for failure to reproduce scientific research findings- from fraud to the issues related to design, conduct, analysis, or publishing scientific research. We also postulate a sensitive dependency on initial conditions by which small changes can result in the large differences in the research findings when attempted to be reproduced at later times. **Methods:** We employed a simple logistic regression equation to model the effect of covariates on the initial study findings. We then fed the input from the logistic equation into a logistic map function to model stability of the results in repeated experiments over time. We illustrate the approach by modeling effects of different factors on the choice of correct treatment. **Results:** We found that reproducibility of the study findings depended both on the initial values of all independent variables and the rate of change in the baseline conditions, the latter being more important. When the changes in the baseline conditions vary by about 3.5 to about 4 in between experiments, no research findings could be reproduced. However, when the rate of change between the experiments is ≤2.5 the results become highly predictable between the experiments. **Conclusions:** Many results cannot be reproduced because of the changes in the initial conditions between the experiments. Better control of the baseline conditions in-between the experiments may help improve reproducibility of scientific findings.
Key words: scientific research, initial conditions, reproducibility.

1. INTRODUCTION

Reproducibility of experimental results is the hallmark of science. Reproducibility has long been standard in science, and is of critical importance for policy or regulatory decisions. For the research findings to be valid, they need to be retested and reproduced (1). Scientific evidence is strengthened when important findings are replicated by multiple independent investigators using independent data, analytical methods and instruments (1). Reproducibility is a mechanism by which scientific data become viewed as less tentative and more reliable. It provides a framework for testing the findings against the empirical world, and holding up to repeated tests is generally viewed as being a necessary component of scientific process (2). However, in recent years there is increasing concern both in scientific literature and lay press that much of scientific research cannot be reproduced (3, 4, 5). For example, scientific findings were confirmed in only 11% landmark studies in preclinical cancer research (6). In clinical medicine, of 49 highly cited original clinical research studies, only 44% (20) could have been reproduced (7). Overall it is estimated that about half of published research cannot be reproduced (8).

Many reasons can explain the failure to reproduce original research findings: those that emerge from the outright fraud to those related to more subtle issues with design, conduct, analysis, or publication of scientific research (9). However, one of the reasons that is ignored in the current discussion of failure to reproduce most of contemporary research relates to neglect to take all initial conditions when the studies are performed into account. The initial conditions of experiment may vary so dramatically between the studies that in some cases it may be impossible to obtain the same or similar results. The situation was first described in the meteorology that led to establishment of the chaos theory, famously summarized as "butterfly effect" (10): a hurricane in Florida may be caused by a butterfly wing flap at the West Coast of Africa. "In chaos theory, the butterfly effect is the sensitive dependency on initial conditions in which a small change at one place in a deterministic nonlinear system can result in large differences in a later state. The name of the effect, coined by Edward Lorenz, is derived from the theoretical example of a hurricane's formation being contingent on whether or not a distant butterfly had flapped its wings several weeks earlier." (10). In this paper, we showed that similar considerations apply to biomedical research.

2. METHODS

We define reproducibility as obtaining the exact value as in a preceding experiment (within the margin of error) (e.g.„ within 95% of confidence interval of the target). We illustrate the effect of initial conditions on the convergence (reproducibility) of results in psychology, which is a field often criticized and particularly plagued with poor reproducibility of results (4). In our own field of decision-making, research to date has shown 12 minimum number of factors that may affect decisions and more than 170 measures aiming to quantify the effect of these factors (11). Table illustrates a simplifying version how these <u>minimum</u> of 12 factors may affect decisions. However, even taking only these 12 factors, we obtain 2x2x2 x2x2x2 x 4x2x2x2x2x5= 20,480 combinations, which can present as the initial condition affecting the outcome such as assessing the probability of correct treatment choice in a decision-making clinical task. This staggering number of combinations is probably an underestimate as we did not take into account a myriad of other measurable and immeasurable factors such as a time of the day of experiment, ambience, color of the walls, comfort of the chair, the manners of investigators, the current personal and social surroundings of the participant (e.g., had he/she drank alcohol a day before experiment, did not sleep well, had a marital "fight", had a long delay on the highway as he/she was driving to the venue or had interactions with other subjects; if the study included biological samples, the number of factors may further increase related to the way the sample was handled- from the moment was taken to transport to storage to thawing, etc). Nevertheless, we can use this simplifying model to illustrate the problem. To model the effects of the initial conditions on the accuracy of making a correct decision, we assume random values for the initial conditions of the 12 factors presented in Table 1. That is, we assume that each value of the factors has equal probability of being selected (e.g., probability of selecting a male = probability of selecting a female, etc). We employ a simple logistic regression equation to model the effect of factors identified in Table on the probability of correct choice of treatment:

$$\log \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots \beta_{12} * x_{12}$$

We then employ logistic map function to model (1,2) the effect of initial conditions:

$$x_{n+1} = r * x_n (1 - x_n)$$

Where x_n = outcome of interest in each separate study (n) (i.e., x = probability of correctly assigning treatment) and r is rate of change in initial conditions between the experiments. Thus, the input variables from the logistic regression represent the initial conditions for the study $n=0$ (i.e., x_0); that is, the initial conditions of the study $n = 0$ depends on the initial values of variables (i.e.,1, 2,3…12) that affect the dependent variable x. The logistic map function seems to be appropriate for the class of the problems where dependent variables is expressed as the probability (of being right vs. wrong in this case). Because each study aims to reproduce the results of the previous study, we assume dependence in the results, even if each experiment is performed independently of each other. An Excel application performing calculations according to the model described above is available from the authors upon request.

3. RESULTS

Figures 1 and 2 show typical results: the probability of reproducibility of the study findings (probability of correct treatment assignment) is a function of both the initial values of all independent variables and a parameter r, the rate of change in baseline conditions. The results are particularly dramatic if the rate of

a) Initial probability=X_0=0.0146; r=3.82   b) Initial probability=X_0=0.373; r=3.82

Figure 1. *Effect of initial conditions on reproducibility of research results.* X_0 = *initial probability of correct treatment assignment (dependent variable) as obtained by input from regression equation incorporating the factors listed in Table 1. r=rate of change in the initial conditions between the studies. The figure illustrates failure to reproduce results between the studies, largely due to effect of r parameter (when r is between 3.5 and 4, no two studies generate identical results; see text for further explanation)*

a) Initial probability=X_0=0.00439; r=2.5   b) Initial probability=X_0=0.456; r=2.5

Figure 2. *Effect of initial conditions on reproducibility of research results.* X_0 = *initial probability of correct treatment assignment (dependent variable) as obtained by input from regression equation incorporating the factors listed in Table 1. r=rate of change in the initial conditions between the studies. The figure illustrates perfect stability (reproducibility) of results between the studies, largely due to effect of r parameter (when r=2.5 the results remain virtually identical; see text for further explanation)*

A) **Decision features**
   a. Framing (e.g., gain vs. losses) (2 factors)
   b. Order of choices (e.g. A → B vs. B→A in a simple two choice decision) (2 factors)
   c. Choice justification (e.g., effect of regret, guilt etc on dissonance reduction; yes vs. no) (2 factors)

B) **Situational factors**
   a. Time pressure (e.g., yes vs no) (2 factors)
   b. Cognitive load (e.g., high vs. low) (2 factors)
   c. Social context (e.g., important vs. not important) (2 factors)

C) **Characteristics of decision-maker**
   a. Individual [e.g., age (old vs. young), gender (female vs. male) (4 factors)
   b. Group (e.g, small vs. large group) (2 factors)
   c. Cultural factors (e.g., present vs. not preset/important) (2 factors)

D) **Individual differences**
   a. Decision styles (e.g. intuitive vs. analytic) (2 factors)
   b. Cognitive ability (e.g., high vs. low) (2 factors)
   c. Personality (e.g., openness, conscientious, extraversion, agreeableness, neuroticism) ("Big 5" factors)

Table 1. *Minimum number of the factors affecting decision-making*

change in the baseline conditions varies from about 3.5 to about 4 in between experiments. The exact value in each preceding experiment varies within the large margin of error (Figure 1a and b). This can happen if, for example, we study decision-making in 15 year old children, and then repeat the study in 45 year old adults. Such a situation is often obvious to most observers; however, rate of change may not be appreciated in many other circumstances leaving us unaware of the reasons for failing to reproduce research results. Interestingly, however, when the rate of change is ≤2.5 the results are highly predictable; the effects of baseline conditions almost disappear.

4. DISCUSSION

History of science is replete with examples of contradiction or disconfirmation of initial results (13), and the only possible way to confirm or contradict the initial findings can be achieved by reproducing the existing research. Only ideas that stand the test of times survive and become part of a body of scientific knowledge. Indeed, progress in science depends on the reproducibility of research. As stated in the Introduction, the vast proportion of the contemporary scientific research cannot be reproduced (4, 6, 7). Some reasons for the failure to reproduce research findings are rare and others are probably more common. The most serious one, which is probably the rarest one is fraud, which occurs in about 1-2% of cases (9, 14). The most pervasive and perfidious reason is probably selective reporting of "positive" results that occur in 70-90% publications across all sciences (4). Such a cherry-picking and accentuating the "positive" results prevent the effort to reproduce results.

To this list, we add another unappreciated cause of failure to reproduce research: a role of initial conditions. Our intent was to present a conceptual framework and not necessarily an empirically accurate model. Further research is needed to better characterize the exact mathematical form of the model, which can then be submitted for empirical verification. Nev-

ertheless, our results indicate that the change in initial conditions and the rate in changes of these conditions may dramatically affect the findings between experiments (Figures 1 and 2). The rate of change (parameter r) seems to be exerting more effect, but interestingly within relatively narrow range of values (from about 3.5 to 4). We postulate that convergence (reproducibility) in results is a function of a type of science. In "hard sciences" (e.g., physics, chemistry, etc) where one can control experiments much better, r is expected to be small (<2.5), which makes the convergence easier to achieve. However, "soft" sciences (e.g. social sciences, psychology, clinical medicine, etc) are characterized by larger r (>3.5) making reproducibility of the results much more difficult to achieve.

What are the implications of our conceptual model for research reproducibility? Theoretically, randomization can equalize all baseline (initial) factors. However, when the number of potential combinations that can affect result becomes large (20,480 in our example assessing the probability of correct treatment choice in a decision-making clinical task), the sample size may become prohibitively large to effectively deal with each disbalance in baseline conditions. Nevertheless, such a disbalance may often be assumed to occur due to chance alone that theoretically can be dealt in the analytical phase of the research experiment (15). Randomization , however, cannot control for the rate of change in the baseline conditions (parameter r) between experiments. Parameter r can only be controlled by attempts to replicate the findings under as identical conditions as it is possible. However, replicability should be distinguished from reproducibility (16). Reproducibility requires changes in the experimental conditions to reproduce the research findings of interest ; on other hand, replicability avoids the changes, which is a reason that some authors argued that replicability is an "impoverished version of reproducibility and is one not worth having" (16). Indeed, replicability is often impossible on practical grounds (16). Experimental techniques, instrumentation, the way we collect data, etc too often change between the treatment study periods. However, to the extent the rate of change between initial conditions is suspected as a reason for poor reproducibility, replicating the results is the only way to control for it. In some cases, this may even prove to be impossible, as when one attempts to replicate the results that are conducted with long time delay. Because experimental technology constantly evolves it may be extremely difficult to conduct experiments under the same conditions when they are undertaken after many months or years of the original study. This means that important research findings should be replicated by scientific community as soon as possible; waiting to reproduce the results in the future may leave the important scientific results unheeded.

CONFLICT OF INTEREST: NONE DECLARED.

REFERENCES

1. Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible Research: Moving toward Research the Public Can Really Trust. Ann Intern Med. 2007;146: 450-453.
2. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. Am J Epidemiol. 2006;163: 783-789.
3. Collins FS, Tabak LA. NIH plans to enhance reproducibility. Nature. 2014; 505: 612-613.
4. Yong E. Bad copy. Nature. 2012; 485: 298-300.
5. McNutt M. Reproducibility. Science. 2014; 343: 229.
6. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012; 483: 531-533.
7. Ioannidis JPA. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. JAMA. 2005; 294: 218-228.
8. Annonymous. Problems with scientific research. How science goes wrong. Economist. 2013. http://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong/print. Accessed April 27, 2014.
9. Steneck N. Introduction to the responsible conduct of research. Washington, DC: Health and Human Services Dept., Office of Research Integrity. 2004.
10. Anonymous. Butterfly effect. 2014. http://en.wikipedia.org/wiki/Butterfly_effect. Accessed April 27, 2014.
11. Appelt KC, Milch KF, Handgraaf MJJ, Weber EU. The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgment and decision-making research. Judgment and Decision Making. 2011; 6: 252-262.
12. May RM. Simple mathematical models with very complicated dynamics. Nature. 1976; 261: 459-467.
13. Freedman DH. Wrong. Why experts keep failing us-and how to know when not to trust them. New York: Little Brown & Co., 2010.
14. Martinson BC, Anderson MS, de Vries R. Scientists behaving badly. Nature. 2005; 435: 737-738.
15. Senn S. Seven myths of randomisation in clinical trials. Stat Med. 2013; 32: 1439-1450.
16. Drummond C. Replicability is not Reproducibility:Nor is it Good Science. Proc. of the Evaluation Methods for Machine Learning Workshop at the 26th ICML. Montreal, Canada, 2009.