Replication Research Series-Paper 2

Empirical research must be replicated before its findings can be trusted

Journal of Clinical Epidemiology 129 (2021) 188–190

Journal of Clinical Epidemiology

REPLICATION RESEARCH SERIES

Replication Research Series-Paper 2 : Empirical research must be replicated before its findings can be trusted

Lex M. Bouter[a,b,*], Gerben ter Riet[c,d]

[a]Department of Epidemiology and Data Science, Amsterdam University Medical Centers, P.O. Box 7057, 1007 MB, Amsterdam, the Netherlands
[b]Department of Philosophy, Faculty of Humanities, Vrije Universiteit Amsterdam, De Boelelaan 1105 1081 HV, Amsterdam, the Netherlands
[c]Urban Vitality Centre of Expertise, Amsterdam University of Applied Sciences, Tafelbergweg 51, 1105 BD, Amsterdam, the Netherlands
[d]Department of Cardiology, Amsterdam University Medical Centers, Meibergdreef 9, 1105 AZ, Amsterdam, the Netherlands

Accepted 14 September 2020; Published online 25 September 2020

Keywords: Replication; Reproducibility; Responsible research practices; Research integrity; Research waste

1. Introduction

In their recent article in this journal, Vachon et al. [1] review terminology for replication research. They propose a 119-word definition which includes four continuously scaled attributes, namely 1) amount of planning of the replication study, 2) distance between the investigators, 3) similarity between the research questions or hypotheses, and 4) compliance with the methods. First, we will scrutinize their definition and the four dimensions they distinguish. We subsequently describe our own views on what replication studies are, why they are important, when irrelevant, and how open science modalities can improve replicability.

We were inspired by Vachon et al.'s [1] dimensions (see their figure 2) and explored the 16 different types of studies one gets when setting the corresponding scales at their extremes (1 and 0, so to speak) and pondered each possible combination. For example, a 1-1-0-0 study is a study that was planned to replicate an index study, by completely different investigators, asking exactly the same research question and using exactly the same methods, that is, a perfect replication study (irrespective of its findings). The opposite extreme, the 0-0-1-1 variant, should, in our view, not be called a replication study. Note that in Vachon et al.'s [1] figure 2, the opposite of 'identical question' is 'related question'. We also think that their definition can be simplified because deliberate planning is a requirement of all studies and we do not believe that unintentional replication studies are less valuable. In addition, we are convinced that the independence dimension is really about the credibility of a replication study's findings and should not be part of the definition. If we omit these two attributes, only the similarity of the research question and similarity of methods dimensions remain, leading to four possible extremes.

Moving now to our own views on how replication studies may be classified, we will for the sake of parsimony use terms derived from the word replication as much as possible, avoid other terms like reproducibility, repeatability, and duplication, and ignore various other subtleties in terminology encountered in the literature [2]. It is, however, important to make a distinction between replicability, a replication, and being replicated. Replicability means that a study is described in sufficient detail to be repeated by others. Replication refers to the act of repeating a study. In addition, being replicated is one of the possible outcomes when a study is repeated.

2. Classification of replication

Vachon et al. [1] summarize the results of their review in a classification that consists of three main categories: 1) repetition of a previous study, 2) extension of a previous study, and 3) road testing of a theory. Their first category is divided into two subcategories: 1a) intrastudy replication

Funding: No financial support was received for this work.

Conflict of interest: L.M.B. chairs the World Conferences on Research Integrity Foundation, the Netherlands Research Integrity Network, and the program committee Replication Studies of the Netherlands Organisation for Scientific Research. L.M.B. is also a member of the steering committee of the REWARD (REduce research Waste And Reward Diligence) Alliance.

* Corresponding author. Department of Epidemiology and Data Science, Amsterdam University Medical Centers, P.O. Box 7057, 1007 MB, Amsterdam, the Netherlands. Tel.: +31 20 4441285.

E-mail address: lm.bouter@vu.nl (L.M. Bouter).

https://doi.org/10.1016/j.jclinepi.2020.09.032

and 1b) interstudy replication. We would like to propose an alternative that we believe to be conceptually clearer and more tenable.

We would rather describe the two subcategories 1a and 1b as reanalysis of existing data and collection of new data with the same study protocol (direct replication), respectively. Furthermore, we are not convinced that the categories 2 and 3 of Vachon et al. [1] are different enough to warrant this distinction. In our view, both refer to what is often labeled as conceptual replication, which is in essence collection of new data using a somewhat modified study protocol. Table 1 summarizes our preferred classification proposal. Our categories 1 and 2 deal with reliability only, whereas our category 3 can additionally inform validity and generalizability. Reanalyses of the same data and direct replications will lead to the same wrong answer if the initial study was flawed. Only conceptual replications can help to detect issues of validity and generalizability. However, when a conceptual replication is unsuccessful, it can be extremely difficult to find out whether that is because the findings are not generalizable of 'just' due to a lack of precision or validity.

A closer look reveals that there are many shades of gray in replication and more specifically that the difference between a replication and the index study is not always clear. For instance, index studies can contain internal replications of various kinds, such as repeating laboratory experiments, reanalyzing the data, or performing sensitivity analyses before a study is submitted for publication. Conceptual replications can also be labeled as new index studies testing related aspects of the same theory. The third category (road testing of a theory) proposed by Vachon et al. [1] in our view broadens the concept of replication too much and seems to include practically all new studies on any aspect of the same theory including those with only a vaguely similar research question and study design. In passing, we note that many criteria can be and are in fact used for deciding what constitutes a successful replication. In Table 1, we present our simple categorization using three criteria which differ in strictness and will therefore lead to different proportions of replications labeled as being successful. A more sophisticated method may be to decide on a margin of unacceptable discordance with initial findings before embarking on a replication [3].

Table 1. Forms of replication and criteria for successful replication

Forms
• Reanalysis of the same data with the same data analysis plan[a]
• Direct replication: new data with the same study protocol
• Conceptual replication: new data with the modified study protocol and same research question
Criteria
• Same direction of findings[b]
• Same direction of findings and similar magnitude of effect
• Same direction of findings, similar magnitude of effect, and similar precision

[a] Or with an alternative data analysis plan to answer the same research question.

[b] If findings do not have a direction, such as incidence and prevalence estimates, this criterion does not apply.

3. Why is replication important?

It can be argued that replication is more important than innovation [4]. Without being replicated, results cannot be trusted because the likelihood of them being mere chance findings or spurious is substantial. Replication is a cornerstone of scientific and scholarly knowledge and not merely a boring way to weed out sloppy or fraudulent research [5]. In fact, research misconduct in an index study is probably only very rarely the explanation of a failed replication attempt. Arguably selective reporting and other questionable research practices such as p-hacking and hypothesizing after results are known (HARK-ing) [6] in particular are responsible for the large majority of failed replications. Selective reporting in the form of publication bias and outcome reporting bias has been elegantly documented by De Vries et al. [7] for randomized clinical trials on drugs for depression. Next to the generic determinants of questionable research practices, such as individual virtuousness, research climate, and perverse incentives, some specific actionable factors (e.g., opportunity, conflict of interest, and external pressure) may drive selective reporting and thus indirectly lower successful replication rates [8,9].

4. When is replication irrelevant?

Clearly there is no need to replicate irrelevant and methodologically unsound research. These studies should neither be performed nor funded in the first place. Unfortunately, it can be argued that this obvious principle is poorly implemented, like for instance in clinical research [10]. We believe that all relevant and methodologically sound empirical research should be successfully replicated before its results can be trusted.

There is an ongoing debate on the need for replication of qualitative research and on the relevance of replication in the humanities and some of the social sciences [11−13]. Furthermore, replication is probably not or less relevant for nonempirical scholarly work, such as modeling, logical analysis, or hermeneutic explanation in, for example, mathematics, architecture, law, and philosophy. In these instances, transparency and clarity on the line of reasoning followed should be assured by other means.

One replication attempt may not be enough, but it is difficult to say how many successful replications are needed before results can be trusted. This is often a matter of cost-effectiveness: how certain do we need to be before we can act? There are clear instances of redundant replication as well, like have been demonstrated by cumulative meta-

analyses showing that sometimes the large majority of randomized patients have only marginally improved the precision of existing evidence [14].

5. Open science modalities improve replicability

Transparency is an important condition for replicability, and open science modalities are aimed at improving transparency [15,16]. More specifically, open methods, open codes, and open data have a substantial role to play. Open methods mean that detailed study protocols are made available. When this is carried out before the data collection starts in the form of a preregistration [17] or a registered report [18], this not only makes the study replicable but it also provides the means to detect selective reporting and HARK-ing. Open codes and open data enable replication in the form of reanalysis of data, and when the data analysis plan is preregistered, they can also help in diagnosing p-hacking.

In summary, we have argued in this commentary that replication studies can be described by two dimensions and classified in three categories. Furthermore, we have explained that innovation has no real meaning without replication and that open science modalities can be instrumental in making research more replicable.

References

[1] Vachon B, Curran JA, Karunananthan S, Brehaut J, Graham ID, Moher D, et al. A concept analysis and meta-narrative review established a comprehensive theoretical definition of replication research to improve its use. J Clin Epidemiol 2020.

[2] National Academies of Sciences, Engineering and Medicine. In: Reproducibility and replicability in science. Washington: National Academies Press; 2019.

[3] ter Riet G, Storosum BWC, Zwinderman AH. What is reproducibility? [version 1; peer review: 3 approved with reservations]. F1000Res 2019;8:36.

[4] Ioannidis JPA. Why replication has more scientific value than original discovery. Behav Brain Sci 2018;41:e137.

[5] Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav 2017;1:0021.

[6] Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. Front Psychol 2016.

[7] de Vries YA, Roest AM, de Jonge P, Cuijpers P, Munafò MR, Bastiaansen JA. The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. Psychol Med 20181−3.

[8] Bouter LM. Fostering responsible research practices is a shared responsibility of multiple stakeholders. J Clin Epidemiol 2018;96:143−6.

[9] van der Steen JT, ter Riet G, van den Bogert CA, Bouter LM. Causes of reporting bias: a theoretical framework [version 2; peer review: 2 approved]. F1000Res 2019;8:280.

[10] Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009;374:86−9.

[11] Peels R, Bouter LM. The possibility and desirability of replication in the humanities. Palgrave Commun 2018;4:95.

[12] Penders B, Holbrook JB, de Rijcke S. Rinse and repeat: understanding the value of replication across different ways of knowing. Publications 2019;7(3):52.

[13] Pratt MG, Kaplan S, Whittington R. Editorial essay: the tumult over transparency: decoupling transparency from replication in establishing trustworthy qualitative research. Adm Sci Q 2020;65(1):1−19.

[14] Clarke M, Brice A, Chalmers I. Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. PLoS One 2014;9:e102670.

[15] National Academies of Sciences. Engineering, and Medicine. In: Open science by design: realizing a vision for 21st century research. Washington: National Academies Press; 2018.

[16] Allen C, Mehler DMA. Open science challenges, benefits and tips in early career and beyond. PLoS Biol 2019;17(5):e3000246.

[17] Nosek BA, Ebersole CR, DeHaven AC, Mellor D. The preregistration revolution. PNAS 2018;115:2600−6.

[18] Chambers C. What's next for registered reports. Nature 2019;573:187−9.