Check for updates

OPINION ARTICLE

What is reproducibility? [version 1; peer review: 3 approved with reservations]

Gerben ter Riet ^{1,2}, Bram W.C. Storosum², Aeilko H. Zwinderman³

¹Dept Research and Education (O&O) - ACHIEVE, Amsterdam University of Applied Sciences, Amsterdam, North Holland, 1105BD, The Netherlands

²Department of General Practice, Amsterdam University Medical Center - University of Amsterdam, Amsterdam, North Holland, 1105AZ, The Netherlands

³Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam University Medical Center - University of Amsterdam, Amsterdam, North Holland, 1105AZ, The Netherlands

V1 First published: 09 Jan 2019, 8:36
<https://doi.org/10.12688/f1000research.17615.1>
 Latest published: 09 Jan 2019, 8:36
<https://doi.org/10.12688/f1000research.17615.1>

Abstract

The debate on reproducibility in biomedicine will gain precision only if we agree what reproducibility means. Importantly, reproducibility should be distinguished from validity ("truth"). We propose the application of an equivalence trials framework to clarify the concept of reproducibility by changing the (narrow) equivalence zone around a zero difference by a zone of reproducibility around (a) previous finding(s).

Keywords

reproducibility, replicability, repeatability, agreement, validation, truth, methodology, equivalence

This article is included in the **Research on Research, Policy & Culture** gateway.

Open Peer Review

Approval Status

	1	2	3
version 1
09 Jan 2019	view	view	view

1. **Steven N. Goodman** , Stanford University, Stanford, USA
2. **C. Glenn Begley**, BioCurate Pty. Ltd., Parkville, Australia
3. **Ksenija Bazdaric** , University of Rijeka Faculty of Medicine, Rijeka, Croatia

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Gerben ter Riet (g.terriet@amc.nl)

Author roles: **ter Riet G:** Conceptualization, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Storosum BWC:** Conceptualization, Investigation, Methodology, Writing – Review & Editing; **Zwinderman AH** : Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2019 ter Riet G *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: ter Riet G, Storosum BWC and Zwinderman AH. **What is reproducibility?** [version 1; peer review: 3 approved with reservations] F1000Research 2019, 8:36 <https://doi.org/10.12688/f1000research.17615.1>

First published: 09 Jan 2019, 8:36 <https://doi.org/10.12688/f1000research.17615.1>

Introduction

Reproducibility is said to be a core principle of scientific progress. Nevertheless, poor reproducibility has recently been shown to haunt preclinical research^{1,2}, translational research³, medicine⁴ and psychology⁵. False-positive initial results due to random chance or incorrect study design were among the reasons implicated, as well as data-dredging, publication bias and misconduct. Others called irreproducible results ‘biased’¹ and ‘unreliable’⁵.

Coming from a background of meta-analysis with its countless examples of unexplained heterogeneity and an ingrained appreciation of sampling variability, we were surprised that these outcries cited above were not accompanied by a formal definition of the concept of reproducibility. Goodman *et al.* did define three types of reproducibility (methods, results, and inferences) and stated that confusion arises when, inadvertently, people use reproducibility as a synonym for “truth”⁶. We read their paper as being about truth although its title suggests otherwise. Our paper is about reproducibility *sensu stricto* and we revisit some basic definitions of reproducibility, notice that these definitions are problematic, and argue that the concept of equivalence in randomized trials may be fruitfully applied to sharpen our understanding of what we mean by reproducibility. We propose that investigators aiming to reproduce others’ findings should pay more attention to predefining a margin of (unacceptable) discordance with existing findings.

Discussion

Box 1 shows two formal definitions of the concept of reproducibility.

Box 1

Definition 1:

“The value below which the absolute difference between two single test [or study, *our addition*] results may be expected to lie with a probability of 95%, when the results are obtained by the same method and equipment from identical test material in the same setting by the same operator within short intervals of time. A test or measurement [or study, *our addition*] is reproducible if the results are identical or closely similar each time it is conducted (Synonym, repeatability)”⁷

Definition 2:

“The degree of agreement among a set of observations [...] after all known sources of error are accounted for (Synonym, precision)”⁸

Note the following differences between definitions 1 and 2:

- (i) In definition 1, reproducibility is taken to be a *binary* concept: a result is either reproduced or not. Definition 2, takes reproducibility to be a *continuous* concept, like a degree of concordance.
- (ii) Related to (i), definition 1 implies the subjective choice of a difference, δ , whose value will depend on

the measurement problem at hand. Definition 2 avoids a choice of δ .

- (iii) Definition 1 chooses the value ‘95’ for the confidence interval to be used. Definition 2 avoids subjective choices of a particular confidence level, such as 95, 90, 68 etc.
- (iv) Only definition 2 emphasizes measurement that is free of bias.

Reproducibility studies may be seen as a type of equivalence trials (see [Figure 1](#)). Briefly, in classic superiority trials, we pose a statistical null hypothesis of no difference, which we then seek to reject to conclude that a difference exists. In equivalence trials, we define a (narrow) zone around a zero difference (between, say, our new drug and an existing one) and we establish equivalence if the entire confidence interval for the reproducibility study lies inside that zone. In this article, we propose to replace the difference of zero by the (pooled) value of (the) previous study or studies (vertical line in [Figure 1](#)). The width of the grey equivalence zone or “zone of reproducibility” is crucial and it seems sensible to define it pragmatically for each research situation separately. Without concrete ideas about the maximal width of this zone, judgments of when a result counts as a reproducibility can be quite subjective. For example, Begley and Ellis considered positive results as not reproduced if the replicate findings were not sufficiently robust to drive a drug-development program. Ioannidis considered the results of a therapeutic intervention as reproduced if the researcher’s final interpretation of the data in both studies was that the intervention was effective (or ineffective). [Figure 1](#), however, shows that even in situations in which one has strictly defined the width of the zone and a suitable type of confidence interval, undecided outcomes may still occur (situations 5–7, [Figure 1](#)).

Reproducibility studies imply healthy scepticism: “Can we reproduce this finding?” In contrast with the comment cited above, which states that irreproducible results are biased, we emphasize that (ir)reproducibility of results says nothing about the validity of the previous nor of the current findings. For that, we need (validity) judgments about rigor of study design and execution. Meta-analyses of many small, but concordant, studies that were subsequently negated by the result of a single mega-trial (believed by many to represent the truth) illustrate this situation⁹.

In conclusion, the concept of reproducibility (repeatability, precision) should be distinguished from validity (“truth”). Furthermore, an equivalence trials framework can be fruitfully used to clarify the concept of reproducibility if we change the (narrow) equivalence zone around a zero difference by a zone of reproducibility around (a) previous finding(s). Care should be exercised when selecting sensible margins (delta) to decide on reproducibility of results¹⁰.

Figure 1. Analogy between equivalence trials framework and reproducibility (concordance): 9 examples. Numbers in brackets refer to the 9 scenarios; horizontal lines are xx% confidence intervals (CI), where xx=95, 90, or 68 etc; short vertical lines depict point estimates; the grey area signifies the zone of reproducibility; delta (δ) refers to the maximal absolute value below which reproducibility (concordance with (an) existing finding(s)) is deemed present. Scenarios 1–4: reproducibility is present since the new point estimate and its entire 95%CI interval lie within the grey zone; scenarios 5–6: presence of reproducibility is uncertain since the point estimate lies inside the grey zone, but the xx%CI does not; scenario 7: presence of reproducibility is uncertain since the point estimate lies outside the grey zone, but part of its xx%CI lies inside; scenario 8–9: absence of reproducibility since point estimate and corresponding xx%CIs are outside the grey zone. Note, that two components are subjective: (1) the choice of δ , although preferably it should be chosen with a thorough understanding of theory or application of the research problem, and (2) the type of confidence interval since other choices than a 95%CI may be possible and defensible. Note also that, even after delta and the type of confidence limit have been chosen, uncertainty may persist if confidence limits overlap the boundaries of delta.

Data availability

No data is associated with this article.

Grant information

The author(s) declared that no grants were involved in supporting this work.

References

- Begley CG, Ellis LM: **Drug development: Raise standards for preclinical cancer research.** *Nature*. 2012; **483**(7391): 531–533.
[PubMed Abstract](#) | [Publisher Full Text](#)
- 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*. 2011; **10**(9): 712.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Hackam DG, Redelmeier DA: **Translation of research evidence from animals to humans.** *JAMA*. 2006; **296**(14): 1731–1732.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ioannidis JP: **Contradicted and initially stronger effects in highly cited clinical research.** *JAMA*. 2005; **294**(2): 218–228.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Open Science Collaboration: **PSYCHOLOGY. Estimating the reproducibility of psychological science.** *Science*. 2015; **349**(6251): aac4716.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Goodman SN, Fanelli D, Ioannidis JP: **What does research reproducibility mean?** *Sci Transl Med*. 2016; **8**(341): 341ps312.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Porta M: **A Dictionary of Epidemiology**, VI ed. Oxford: Oxford University Press, 2014.
[Publisher Full Text](#)
- Miettinen OS: **Epidemiological Research: Terms and Concepts**. Dordrecht: Springer, 2011.
[Publisher Full Text](#)
- Cappelleri JC, Ioannidis JP, Schmid CH, *et al.*: **Large trials vs meta-analysis of smaller trials: how do their results compare?** *JAMA*. 1996; **276**(16): 1332–1338.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Dekkers OM, Cevallos M, Buhner J, *et al.*: **Comparison of noninferiority margins reported in protocols and publications showed incomplete and inconsistent reporting.** *J Clin Epidemiol*. 2015; **68**(5): 510–517.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 25 February 2019

<https://doi.org/10.5256/f1000research.19261.r44140>

© 2019 Bazdaric K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ksenija Bazdaric

Department of Medical Informatics, University of Rijeka Faculty of Medicine, Rijeka, Croatia

Thank you for giving me the opportunity to read this manuscript. It was very interesting. As opinion pieces are not supposed to be very long I understand that not all concepts/constructs could have been explained in detail. I think the article is about the definition of reproducibility and should be understood as such. I would advise acceptance with minor changes.

Comments:

Introduction

The aim of the article is clear but the title is not. I would advise adding a change to the title to 'What is reproducibility? - a definition proposal'.

I would advise repeating at least one of the most known definitions in order to ease the reading to general audience. Readers must understand the flaws of existing definitions in order to embrace the new one(s). Perhaps this one: NSF report as "replicability," which refers to "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."¹ or some other.

Discussion

When you state "we were surprised that these outcries cited above were not accompanied by a formal definition of the concept of reproducibility". I wonder what do you mean by formal, a statistical definition or a more narrow definition, or a more exact definition? Please make your statement more clear if possible.

I really like Box 1 and the 2 definitions proposed. The first model might work and be valuable for life sciences and biomedicine, while the second can be more used in psychology and other social sciences.

Figure 1. is very clear with clear examples. I especially like example 5 and the explanation in the discussion.

Conclusion

Of course, the concept of reproducibility should be distinguished from validity. If a measurement is not valid there is no need for replication at all. But if you think the general audience is

misunderstanding the terms and that they have to be distinguished please give a short definition of validation in brief, because it is a widely used term in psychology but not in other disciplines.

References

1. Goodman S, Fanelli D, Ioannidis J: What does research reproducibility mean?. *Science Translational Medicine*. 2016; **8** (341). [Publisher Full Text](#)

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Partly

Are arguments sufficiently supported by evidence from the published literature?

Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: research integrity, open science, methodology, plagiarism, publishing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 15 February 2019

<https://doi.org/10.5256/f1000research.19261.r44139>

© 2019 Begley C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

C. Glenn Begley

BioCurate Pty. Ltd., Parkville, VIC, Australia

The issue of data reproducibility is central to science and is worthy of ongoing discussion. Although the authors state at the outset that "Reproducibility is said to be a core principle of scientific progress", to me it IS a core principle.

Data that cannot be reproduced does not serve as a foundation upon which others can build. The authors propose that a formal definition of reproducibility be pre-defined, pre-agreed when

investigators attempt to reproduce others' findings. Pre-defining criteria for failure or success is always valuable. It removes the natural bias to interpret results to suit one's prejudice post-hoc, and may also be useful in this context.

However, the paper that is cited on which I am the first author (Begley and Ellis, 2012)¹ does not really support the authors' argument. The authors state that we considered results as not reproduced if findings were not sufficiently robust to drive a drug development program. That is correct: we were focused on developing new drugs, and could not justify moving forward if the results were not reproduced. But what was truly shocking, was that in the majority of cases it was the original authors themselves who were unable to reproduce their own findings. Our 'standard operating procedure' when unable to reproduce key findings was to go to the original laboratory and watch them repeat their experiments (which required a confidentiality agreement and precludes disclosure of those laboratories). Their failure to reproduce their findings certainly negated that "research" as being sufficiently robust to drive a drug development program.

Using the criteria outlined in Figure 1 of this paper, the published experiments are illustrated in Scenario 8, while the repeated (and unpublished) experiments are illustrated in Scenario 9. In our experience therefore, the pre-definition of confidence intervals appeared unnecessary. It was this experience that led us to conclude that the fundamental problem was not really one of "reproducibility", nor a problem of definition, it was rather a problem of cherry-picking, p-hacking, HARKING, lack of controls, lack of repeats, lack of blinding. This poor experimental methodology was employed so as to generate an initial data set that was sufficiently exciting to justify publication.

Therefore, I do not think the issue regarding lack of reproducibility is simply one of a lack of clear definition, rather, in my view, it is systematic and driven by the perverse incentives that govern our current system. Thus focusing solely on agreeing on a definition, does not lead us toward finding a solution to a problem that is deeply embedded in our system, and in fact has been used by some to distract and argue that there isn't really an issue of irreproducibility - its simply about a definition.

From my perspective, it would be valuable for these Authors to acknowledge these wide-spread scientific practices as central to the issue of "reproducibility".

References

1. Begley CG, Ellis LM: Drug development: Raise standards for preclinical cancer research. *Nature*. 2012; **483** (7391): 531-3 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the topic of the opinion article discussed accurately in the context of the current literature?

Partly

Are all factual statements correct and adequately supported by citations?

Partly

Are arguments sufficiently supported by evidence from the published literature?

Partly

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Particular interest in the area of scientific rigour and research methodology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 13 February 2019

<https://doi.org/10.5256/f1000research.19261.r42838>

© 2019 Goodman S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Steven N. Goodman

Departments of Health Research and Policy (Epidemiology) and Medicine, Stanford University, Stanford, CA, USA

This is a thoughtful piece that attempts to offer a construct that will help define research reproducibility. They say that their purpose is to offer an operational definition of reproducibility that they claim a previous paper entitled "What does research reproducibility mean?" (with myself as first author) did not address:

"Goodman et al. did define three types of reproducibility (methods, results, and inferences) and stated that confusion arises when, inadvertently, people use reproducibility as a synonym for "truth"⁶. We read their paper as being about truth although its title suggests otherwise."

The claim that the 2016 paper ¹ was about truth and not reproducibility is not quite right. Let us see exactly what the prior paper said:

Results reproducibility (previously described as replicability) refers to obtaining the same results from the conduct of an independent study whose procedures are as closely matched to the original experiment as possible. ... this might be clear in principle but is operationally elusive. The problem arises in settings where there is substantial random error in any result, making unclear the criteria for considering results to be "the same." The intuition and logic of results reproducibility are derived from systems that are deterministic or for which the signal-to error ratio is exceedingly high. But, when the same intuition and logic are applied to studies with substantive stochastic components, the paradigm of accumulating evidence might be more appropriate than any binary criteria for successful or unsuccessful replication.

....

Statistical significance by itself tells very little about whether one study has "replicated" the results

of another. For example, two studies that show identical 10% survival differences between the treatment and control arms would have very different degrees of statistical significance if their sample sizes were substantially different. If one was highly significant and the other far from significance, the two studies might be reported individually as supporting opposite conclusions, in spite of the fact that they are mutually corroborative. An interpretive error complementary to the one described above involves the assumption that multiple studies that fail to demonstrate statistical significance necessarily confirm the absence of an effect.

.....

It is easier to statistically define non-replication than replication, through statistical tests of heterogeneity, which can evaluate whether the difference between two or more experimental results might be due to the play of chance. Two or more studies are judged to be statistically heterogeneous when the between-study variance in reported effects is substantially greater than what is expected from sampling error. Such tests, however, are greatly underpowered and therefore unreliable when comparing several studies, particularly when they are small or imprecise (17). Conversely, when there are many large studies, tests for heterogeneity might demonstrate statistical heterogeneity (and, therefore, lack of results reproducibility) even if the effect sizes of different studies are close (17) and regarded as scientifically equivalent. Therefore, a preferred way to assess the evidential meaning of two or more results with substantive stochastic variability is to evaluate the cumulative evidence they provide vis-à-vis a hypothesis of interest and not whether one contradicts or discredits the other through the lens of statistical significance.

So it should be apparent that the 2016 paper does indeed address exactly what this paper addresses, including the notion that results can differ in statistical significance yet be regarded as "scientifically equivalent". That is essentially what this paper goes on to try to define, with a "zone of equivalence" that defines "scientifically equivalence". But as these authors acknowledge in the legend of Figure 1, *"... even after delta and the type of confidence limit have been chosen, uncertainty may persist if confidence limits overlap the boundaries of delta."* The problem is that the confidence intervals will quite often cross the boundaries of delta, and so we are left with the same conundrum that the original paper said was inescapable.

The point of the original paper was that trying to define "reproducibility" was in the end not very constructive, and that if we turned our attention instead to the cumulative evidence represented by several studies, instead of whether they "reproduced" or not, we could avoid these distinctions, which ultimately serve little purpose. The authors here are right that I believe that the goal of science, and of scientific studies, is to move us closer to the truth. I contend that debates about which and how many studies reproduced do not, particularly when that definition is elusive. The prior paper did indeed tell us that convergence on the truth should be our lodestar, not an arbitrarily defined reproducibility criterion, which even with the improved version offered here does not provide a clear verdict in the vast majority of cases.

I agree with the authors that it is helpful to have some notion of differences that make a difference, and thereby scientific equivalence. But the degree of imprecision in most health studies precludes an unambiguous conception of reproducibility even if one introduces that interval. I also agree with the authors' conclusion that *"...the concept of reproducibility (repeatability, precision) should be distinguished from validity ("truth")"*, but disagree that the purpose of assessing reproducibility is anything other than getting at the truth, and still believe that the cumulative evidence model and not the reproducibility model - which cannot be clearly defined - is what gets us there.

References

1. Goodman S, Fanelli D, Ioannidis J: What does research reproducibility mean?. *Science Translational Medicine*. 2016; **8** (341). [Publisher Full Text](#)

Is the topic of the opinion article discussed accurately in the context of the current literature?

Partly

Are all factual statements correct and adequately supported by citations?

Partly

Are arguments sufficiently supported by evidence from the published literature?

Partly

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Partly

Competing Interests: I was a lead author on the 2016 article which is being discussed here.

Reviewer Expertise: Statistical inference, research reproducibility, epidemiology, clinical research.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research