Resolving the Tension Between Exploration and Confirmation in Preclinical Biomedical Research

Ulrich Dirnagl

Contents

Abstract

Confirmation through competent replication is a founding principle of modern science. However, biomedical researchers are rewarded for innovation, and not for confirmation, and confirmatory research is often stigmatized as unoriginal and as a consequence faces barriers to publication. As a result, the current biomedical literature is dominated by exploration, which to complicate matters further is often disguised as confirmation. Only recently scientists and the public have begun to realize that high-profile research results in biomedicine can often not be replicated. Consequently, confirmation has become central stage in the quest to safeguard the robustness of research findings. Research which is pushing the

U. Dirnagl (✉)
Department of Experimental Neurology and Center for Stroke Research Berlin, Charité – Universitätsmedizin Berlin, Berlin, Germany

QUEST Center for Transforming Biomedical Research, Berlin Institute of Health (BIH), Berlin, Germany
e-mail: ulrich.dirnagl@charite.de

A. Bespalov et al. (eds.), *Good Research Practice in Non-Clinical Pharmacology and Biomedicine*, Handbook of Experimental Pharmacology 257,
https://doi.org/10.1007/164_2019_278

boundaries of or challenges what is currently known must necessarily result in a plethora of false positive results. Thus, since discovery, the driving force of scientific progress, is unavoidably linked to high false positive rates and cannot support confirmatory inference, dedicated confirmatory investigation is needed for pivotal results. In this chapter I will argue that the tension between the two modes of research, exploration and confirmation, can be resolved if we conceptually and practically separate them. I will discuss the idiosyncrasies of exploratory and confirmatory studies, with a focus on the specific features of their design, analysis, and interpretation.

Keywords
False negative · False positive · Preclinical randomized controlled trial · Replication · Reproducibility · Statistics

Science, the endless frontier (V. Bush)

To boldly go where no man has gone before (G. Roddenberry)

Non-reproducible single occurrences are of no significance to science (K. Popper)

1 Introduction

Scientists' and the public's view on science is based toward innovation, novelty, and discovery. Discoveries, however, which represent single occurrences that cannot be reproduced are of no significance to science (Popper 1935). Confirmation through competent replication has therefore been a founding principle of modern science since its origins in the renaissance. Most scientists are aware of the need to confirm their own or other's results and conclusions. Nevertheless, they are rewarded for innovation, and not for confirmation. Confirmatory research is often stigmatized as unoriginal and faces barriers to publication. As a result, the current biomedical literature is dominated by exploration, which is often disguised as confirmation. In fact, many published articles claim that they have discovered a phenomenon and confirmed it with the same experiment, which is a logical impossibility: the same data cannot be used to generate and test a hypothesis. In addition, exploratory results are often garnished with far-reaching claims regarding the relevance of the observed phenomenon for the future treatment or even cure of a disease. As the results of exploration can only be tentative, such claims need to be founded on confirmed results.

2 Discrimination Between Exploration and Confirmation

Recently, scientists and the public have become concerned that research results often cannot be replicated (Begley and Ellis 2012; Prinz et al. 2011; Baker 2016) and that the translation of biomedical discovery to improved therapies for patients has a very high attrition rate (Dirnagl 2016). Undoubtedly the so-called replication crisis and the translational roadblock have complex roots, including importantly the sheer complexity of the pathophysiology of most diseases and the fact that many of the "low-hanging fruits" (e.g., antibiotics, antidiabetics, antihypertensives, treatments for multiple sclerosis, Parkinson's disease, and epilepsy, to name but a few) have already been picked. Nevertheless, results which are confounded by biases and lack statistical power must often be false (Ioannidis 2005) and hence resist confirmation. I will argue that the two modes of research need to be separated and clinical translation of preclinical evidence be based on confirmation (Kimmelman et al. 2014). Thus, exploration and confirmation are equally important for the progress of biomedical research: In exploratory investigation, researchers should aim at generating robust pathophysiological theories of disease. In confirmatory investigation, researchers should aim at demonstrating strong and reproducible treatment effects in relevant animal models. In what follows I will explore what immediate and relevant consequences this has for the design of experiments, their analysis, interpretation, and publication.

3 Exploration Must Lead to a High Rate of False Positives

Research which is pushing the boundaries of or challenges what is currently known must result in a plethora of false positive results. In fact, the more original initial findings are, the less likely it is that they can subsequently be confirmed. This is the simple and straightforward result of the fact that cutting edge or even paradigm shifting (Kuhn 1962) research must operate with low base rates, that is, low prior probabilities that the tested hypotheses are actually true. Or in other words, the more mainstream and less novel research is, the likelier it is that it will find its hypotheses to be true. This can easily be framed statistically, either frequentist or Bayesian. For example, if research operates with a probability that 10% of its hypotheses are true, which is a conservative estimate, and accepts type I errors at 5% (alpha-level) and type II errors at 20% (i.e., 80% power), almost 40% of the times, it rejects the NULL hypothesis (i.e., finds a statistically significant result), while the NULL hypothesis is actually true (false positive) (Colquhoun 2014). In other words, under those conditions, which are actually unrealistic in preclinical medicine in which power is often 50% and less (Button et al. 2013; Dirnagl 2006), the positive predictive value (PPV) of results is much worse than the type I error level. This has grave consequences, in particular as many researchers confuse PPV and significance level, nurturing their delusion that they are operating at a satisfactory level of wrongly accepting their hypothesis in only 5% of the cases. As a corollary, under

those conditions true effect sizes will be overestimated by almost 50%. For a detailed discussion and R-code to simulate different scenarios, see Colquhoun (2014).

4 The Garden of Forking Paths

The "Garden of Forking Paths" (1944) is a metaphor which the statisticians Gelman and Loken (2013) have borrowed from Jorge Luis Borges' (1899–1986) novel for their criticism of experimental design, analysis, and interpretation of psychological and biomedical research: In exploratory investigation, researchers trail on branched-out paths through a garden of knowledge. And as poetic as this wandering might seem, it withholds certain dangers. On his way through the labyrinth, the researcher proceeds in an inductively deterministic manner. He (or she) does not at all notice the many levels of freedom available to him. These will arise, for example, through an alternative analysis (or interpretations) of the experiment, fluke false positive or false negative results, the choice for an alternative antibody or mouse strain, or from the choice of another article as the basis for further experiments and interpretations. The labyrinth is of infinite size, and there is not only one way through it, but many, and there are many exits. And since our researcher is proceeding exploratively, he has set up no advance rules according to which he should carry out his analyses or plan further experiments. So the many other possible results escape his notice, since he is following a trail that he himself laid. In consequence, he overestimates the strength of evidence that he generates. In particular, he overestimates what a significant *p*-value means regarding his explorative wanderings. He should compare his results with all the alternative analyses and interpretations that he could have carried out, which is obviously an absurd suggestion. Indeed, in the Garden of Forking Paths, the classic definition of statistical significance (e.g., $p < 0.05$) does not apply, for it states that in the absence of an effect, the probability of bumping coincidentally into a similarly extreme or even more extreme result is lower than 5%. You would have to factor in all data and analyses that would be possible in the garden. Each of these other paths could also have led to statistically significant results. Such a comparison however is impossible in explorative research. If you nonetheless do generate *p*-values, you will, according to Gelman and Loken, get a "machine for the production and publication of random patterns." A consequence of all this is that our knowledge derived from exploration is less robust than the chain of statistically significant results might have us believe and that the use of test statistics in exploration is of little help or may even be superfluous if not even misleading.

At this point we must summarize that exploratory research, even if of the highest quality and without selective use of data, p-hacking (collecting, selecting, or analyzing data until statistically nonsignificant results become significant), or HARKING ("hypothesizing after the results are known"), leads to at best tentative results which provide the basis for further inquiry and confirmation. In this context it is sobering that Ronald Fisher, a founding father of modern frequentist statistics, considered results which are significant at the 5% level only "worth a second look" (Nuzzo 2014). Today clinical trials may be based on such preclinical findings.

5 Confirmation Must Weed Out the False Positives of Exploration

Since discovery is unavoidably linked to high false positive rates and cannot support confirmatory inference, dedicated confirmatory investigation is needed for pivotal results. Results are pivotal and must be confirmed if they are the basis for further investigation and thus drain resources, if they directly or indirectly might impact on human health (e.g., by informing the design of future clinical development, including trials in humans (Yarborough et al. 2018)), or if they are challenging accepted evidence in a field. Exploration must be very sensitive, since it must be able to faithfully capture rare but critical results (e.g., a cure for Alzheimer disease or cancer); confirmation on the other hand must be highly specific, since further research on and development of a false positive or non-robust finding is wasteful and unethical (Al-Shahi Salman et al. 2014).

6 Exact Replication Does Not Equal Confirmation

Many experimentalists routinely replicate their own findings and only continue a line of inquiry when the initial finding was replicated. This is laudable, but a few caveats apply. For one, a replication study which exactly mimics the original study in terms of its design may not be as informative as most researchers suspect. Counterintuitively, in an exact replication, with new samples but same sample size and treatment groups, and assuming that the effect found in the original experiment with $p < 0.05$ (but $p \approx 0.05$) equals the true population effect, the probability of replication (being the probability of getting again a significant result of the same or larger effect) is only 50% (Goodman 1992). In other words, the predictive value of an exact replication of a true finding that was significant close to the 5% level is that of a coin toss! Incidentally, this is the main reason why phase III clinical trials (which aim at confirmation) have much larger sample sizes than phase II trials (which aim at exploration). For further details on sample size calculation of replication studies, see Simonsohn (2015).

A second problem of exact replications, in particular if performed by the same group which has made the initial observation, is the problem of systematic errors. One of the most embarrassing and illustrative examples for a botched replication in the recent history of science relates to the discovery of neutrinos that travel faster than the speed of light. The results of a large international experiment conducted by physicists of high repute convulsed not only the field of physics; it shook the whole world. Neutrinos had been produced by the particle accelerator at CERN in Geneva and sent on a 730 km long trip. Their arrival was registered by a detector blasted through thousands of meters of rock in the Dolomites. Unexpectedly, the neutrinos arrived faster than would the photons travelling the same route. The experiment was replicated several times, and the results remained significant with a p-value of less than 3×10^{-7}. In the weeks following the publication (the OPERA Collaboration 2011) and media excitement, the physicists found that the GPS used to measure distances was not correctly synchronized and a cable was loose.

7 Design, Analysis, and Interpretation of Exploratory vs Confirmatory Studies

Thus, exploratory and confirmatory investigation necessitates different study designs. Confirmation is not the simple replication of an exploratory experiment. Exploratory and confirmatory investigation differs in many aspects. While exploration may start without any hypothesis ("unbiased"), a proper hypothesis is the obligatory starting point of any confirmation. Exploration investigates physiological or pathophysiological mechanisms or aims at drug discovery. The tentative findings of exploration, if relevant at all, need to be confirmed. Confirmation of the hypothesis is the default primary endpoint of the confirmatory investigation, while secondary endpoints may be explored. Both modes need to be of high internal validity, which means that they need to effectively control biases (selection, detection, attrition, etc.) through randomization, blinding, and prespecification of inclusion and exclusion criteria. Of note, control of bias is as important in exploration as in confirmation. To establish an experimental design and analysis plan before the onset of the study may be useful in exploration but is a must in confirmation. Generalizability is of greater importance in confirmation than in exploration, which therefore needs to be of high external validity. Depending on the disease under study, this may include the use of aged or comorbid animals. Statistical power is important in any type of experimental study, as low power results not only in a high false negative rate but also increases the number of false positive findings and leads to an overestimation of effect sizes (Button et al. 2013). However, as exploration aims at finding what might work or is "true," type II error should be minimized and therefore statistical power high. Conversely, as confirmation aims at weeding out the false positive, type I error becomes a major concern. To make sure that statistical power is sufficient to detect the targeted effect sizes, a priori sample size calculation is recommended in exploratory mode but obligatory in confirmation where achievable effect sizes and variance can be estimated from previous exploratory evidence. Due to manifold constraints, which include the fact that exploration (1) often collects many endpoints and hence multiple group comparisons are made, (2) is usually underpowered, and (3) comes with an almost unlimited degree of freedom of the researcher with respect to selection of animal strains, biologicals, and selection of outcome parameters and their analysis ("Garden of Forking Paths," see above and (Gelman and Loken 2013)), statistical significance tests (like t-test, ANOVA, etc.) are of little use; the focus in exploration should rather be on proper descriptive statistics, including measures of variance and confidence intervals. Conversely, in confirmatory mode, the prespecified analysis plan needs to describe the planned statistical significance test. To prevent outcome switching and publication bias, in confirmatory studies the hypothesis, experimental, and analysis plan should be preregistered (Nosek et al. 2018). Preregistration can be embargoed until the results of the study are published and are therefore not detrimental to intellectual property claims. Table 1 gives a tentative overview of some of the idiosyncrasies of exploratory and confirmatory investigation.

Table 1 Suggested differences between exploratory and confirmatory preclinical study designs

	Exploratory	Confirmatory
Establish pathophysiology, discover drugs, etc.	+++	(+)
Hypothesis	(+)	+++
Blinding	+++	+++
Randomization	+++	+++
External validity (aging, comorbidities, etc.)	−	++
Experimental and analysis plan established before study onset	+	+++
Primary endpoint	−	++
Inclusion/exclusion criteria (prespecified)	++	+++
Preregistration	(−)	+++
Sample size calculation	(+)	+++
Test statistics	+	+++
Sensitivity (type II error): find what might work	++	+
Specificity (type I error): weed out false positives	+	+++

Modified with permission (Dirnagl 2016), for details see text

8 No Publication Without Confirmation?

Vis a vis the current masquerading of exploratory preclinical investigations as confirmation of new mechanisms of disease or potential therapeutic breakthroughs, Mogil and Macleod went as far as proposing a new form of publication for animal studies of disease therapies or preventions, the "preclinical trial." In it, researchers besides presenting a novel mechanism of disease or therapeutic approach incorporate an independent, statistically rigorous confirmation of the central hypothesis. Preclinical trials would be more formal and rigorous than the typical preclinical testing conducted in academic labs and would adopt many practices of a clinical trial (Mogil and Macleod 2017).

It is uncertain whether scientists or journals will pick up this sensible proposal in the near future. Meanwhile, another novel type of publication, at least in the preclinical realm, is gaining traction: preregistration. When preregistering a study, the researcher commits in advance to the hypothesis that will be tested, the study design, as well as the analysis plan. This provides full transparency and prevents HARKING, p-hacking, and many other potential barriers to the interpretability and credibility of research findings. Preregistration is not limited to but ideally suited for confirmatory studies of high quality. In fact, it may be argued that journals should mandate preregistration when processing and publishing confirmatory studies.

9 Team Science and Preclinical Multicenter Trials

Confirmation lacks the allure of discovery and is usually more resource intense. It requires higher sample sizes and benefits from multilab approaches which come with considerable organizational overhead. Permission from regulatory authorities may

be hard to obtain, as the repetition of animal experiments combined with upscaling of sample sizes may antagonize the goal to reduce animal experimentation. Publication of confirmatory results faces bigger hurdles than those of novel results. Failure to confirm a result may lead to tensions between the researchers who published the initial finding and those who performed the unsuccessful confirmation. Potential hidden moderators and context sensitivity of the finding dominate the resulting discussion, as does blaming those who failed confirmation of incompetence. In short, confirmatory research at present is not very attractive. This dire situation can only be overcome if a dedicated funding stream supports team science and confirmatory studies, and researchers are rewarded (and not stigmatized) for this fundamentally important scientific activity. It is equally important to educate the scientific community that results from exploratory research even of the highest quality are inherently tentative and that failure of competent replication of such results does not disqualify their work but is rather an element of the normal progression of science.

It is promising that several international consortia have teamed up in efforts to develop guidelines for international collaborative research in preclinical biomedicine (MULTIPART 2016) or to demonstrate that confirmatory preclinical trials can be conducted and published (Llovera et al. 2015) with a reasonable budget.

10 Resolving the Tension Between Exploration and Confirmation

Science advances by exploration and confirmation (or refutation). The role of exploration is currently overemphasized, which may be one reason of the current "replication crisis" and the translational roadblock. In addition, generation of postdictions is often mistaken with the testing of predictions (Nosek et al. 2018); in other words exploration is confounded with exploration. We need to leverage the complementary strengths of both modes of investigation. This will help to improve the refinement of pathophysiological theories, as well as the generation of reliable evidence in disease models for the efficacy of treatments in humans. Adopting a two-pronged approach of exploration-confirmation requires that we shift the balance which is currently biased toward exploration back to confirmation. Researchers need to be trained in how to competently engage in high-quality exploration and confirmation. Funders and institutions need to establish mechanisms to fund and reward confirmatory investigation.

References

Al-Shahi Salman R et al (2014) Increasing value and reducing waste in biomedical research regulation and management. Lancet 383:176–185

Baker M (2016) 1,500 scientists lift the lid on reproducibility. Nature 533:452–454

Begley CG, Ellis LM (2012) Drug development: raise standards for preclinical cancer research. Nature 483:531–533

Button KS et al (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376
Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of P values. R Soc Open Sci 1:1–15. https://doi.org/10.1098/rsos.140216
Dirnagl U (2006) Bench to bedside: the quest for quality in experimental stroke research. J Cereb Blood Flow Metab 26:1465–1478
Dirnagl U (2016) Thomas Willis lecture: is translational stroke research broken, and if so, how can we fix it? Stroke 47:2148–2153
Gelman A, Loken E (2013) The garden of forking paths: why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf. Accessed 23 Aug 2018
Goodman SN (1992) A comment on replication, p-values and evidence. Stat Med 11:875–879
Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2:e124
Kimmelman J, Mogil JS, Dirnagl U (2014) Distinguishing between exploratory and confirmatory preclinical research will improve translation. PLoS Biol 12:e1001863
Kuhn TS (1962) The structure of scientific revolutions. The University of Chicago Press, Chicago
Llovera G et al (2015) Results of a preclinical randomized controlled multicenter trial (pRCT): anti-CD49d treatment for acute brain ischemia. Sci Transl Med 7:299ra121
Mogil JS, Macleod MR (2017) No publication without confirmation. Nature 542:409–411
MULTIPART. Multicentre preclinical animal research team. http://www.dcn.ed.ac.uk/multipart/. Accessed 23 May 2016
Nosek BA, Ebersole CR, DeHaven AC, Mellor DT (2018) The preregistration revolution. Proc Natl Acad Sci U S A 115:2600–2606
Nuzzo R (2014) Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 506:150–152
Popper K (1935) Logik der Forschung. Springer, Berlin
Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 10:712
Simonsohn U (2015) Small telescopes. Psychol Sci 26:559–569
The OPERA Collaboration et al (2011) Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. https://doi.org/10.1007/JHEP10(2012)093
Yarborough M et al (2018) The bench is closer to the bedside than we think: uncovering the ethical ties between preclinical researchers in translational neuroscience and patients in clinical trials. PLoS Biol 16:e2006343