Is Translational Stroke Research Broken, and if So, How Can We Fix It?

Ulrich Dirnagl, MD

Abstract—Based on research, mainly in rodents, tremendous progress has been made in our basic understanding of the pathophysiology of stroke. After many failures, however, few scientists today deny that bench-to-bedside translation in stroke has a disappointing track record. I here summarize many measures to improve the predictiveness of preclinical stroke research, some of which are currently in various stages of implementation: We must reduce preventable (detrimental) attrition. Key measures for this revolve around improving preclinical study design. Internal validity must be improved by reducing bias; external validity will improve by including aged, comorbid rodents of both sexes in our modeling. False-positives and inflated effect sizes can be reduced by increasing statistical power, which necessitates increasing group sizes. Compliance to reporting guidelines and checklists needs to be enforced by journals and funders. Customizing study designs to exploratory and confirmatory studies will leverage the complementary strengths of both modes of investigation. All studies should publish their full data sets. On the other hand, we should embrace inevitable NULL results. This entails planning experiments in such a way that they produce high-quality evidence when NULL results are obtained and making these available to the community. A collaborative effort is needed to implement some of these recommendations. Just as in clinical medicine, multicenter approaches help to obtain sufficient group sizes and robust results. Translational stroke research is not broken, but its engine needs an overhauling to render more predictive results. (Stroke. 2016;47:2148-2153. DOI: 10.1161/STROKEAHA.116.013244.)

Key Words: biostatistics • neuroprotection • research design • stroke • translational medical research

Over the past few decades, preclinical research on stroke has led to tremendous progress in our basic understanding of the pathophysiological events that follow focal cerebral ischemia.12 Numerous cellular and molecular targets have been identified for brain protective and restorative therapies, and many of these are highly effective in rodents. The incidence and morbidity and mortality of stroke have decreased. Stroke units and recanalization via intravenous tissue-type plasminogen activator or thrombectomy are impressive clinical success stories benefitting many patients. Intriguingly, however, practically none of these clinical breakthroughs are the result of bench-to-bedside translation. On the contrary, almost all therapies that were preclinically successful have failed in patients with actual stroke.7 This exceedingly high rate of attrition in translational stroke research has already been the subject of many articles. There are no simple explanations, and stroke research is certainly not the only biomedical field struggling with a translational roadblock. In the following, I would like to emphasize factors for which quantitative meta-analytic evidence exists that suggests that they contribute to attrition. My selection is biased toward items in the preclinical realm that pose straightforward opportunities for improvement. Inspired by work from bioethics8,9 and from meta-research,10 I would like to propose, counterintuitively there are instances where we need to embrace attrition. Collectively, I argue for an update of our intellectual framework for translational research.

To a large extent, bench-to-bedside translation is a black box. Innumerable factors affect whether and to what extent preclinical evidence is transferable to clinical evidence. Attrition lurks on all levels: preclinically, when moving to first in man, when trying to obtain safety or initial signs of efficacy, and in large clinical trials aiming at regulatory approval.

Can Mice Mimic Human Stroke Pathophysiology?

The most basic and dramatic threat to the validity of bench-to-bedside translation concerns the question of whether preclinical models can predict human pathophysiology and therapeutic outcomes. This relates to the concept and construct validity of how we model the different types of strokes, and in a more general sense to whether nonhuman (in particular...
rodent) physiology and pathobiology are sufficiently similar to that of humans. There are indeed few best practice cases that unequivocally demonstrate translational success for a treatment. Unfortunately, tissue-type plasminogen activator, which was effective in a rabbit model of embolic stroke before its clinical efficacy, had been established in the seminal NINDS (National Institute of Neurological Disorders and Stroke) trial, has proved to be the exception rather than the rule. We therefore have to rely on indirect evidence, such as similar phenotypes of pathophysiologic phenomena in experimental and human stroke. Examples include immunodepression after stroke or spreading depolarization. More examples, and a more elaborated argument for why modeling of stroke in rodents can indeed be predictive for human pathobiology and treatments, can be found in the study by Dirnagl and Endres.

Internal Validity: Keeping Cognitive Biases in Check

To provide a solid basis for clinical development, evidence at the bench must be robust and reliable. Lack of these attributes can lead to attrition and wasted resources, unethical use of animals in research, and can potentially put patients at risk. Robustness and reliability of research are threatened by many biases and consequently low internal validity. Selection bias is controlled by randomization, which safeguards that experimental groups are similar except for the experimental manipulation. Concealing the allocation to experimental groups, a form of blinding prevents performance bias. Finally, detection bias is kept in check by blinded assessment of outcomes. The conceptual framework of these measures, which are intended to keep all other things equal (save the intervention), was well developed decades ago for clinical trials. In this highly regulated area of biomedical research, internal validity is a central consideration when planning and reporting a study and a key criterion of review boards and regulators. Surprisingly, in experimental biomedicine, internal validity seems to be much less of a concern. Indeed, in the wake of what has been termed a reproducibility crisis, the internal validity of preclinical research has recently been put into question. Metaresearch has provided ample evidence that despite an international discussion and the introduction of guidelines, improvements in experimental design, conduct, analysis, and reporting are overdue and very slow in coming. We have recently studied the effect of attrition bias, which to date has received relatively little attention although it seems highly prevalent and may substantially skew experimental evidence. We reviewed 100 randomly selected reports published between 2000 and 2013 describing 522 experiments that used rodents to test cancer or stroke treatments and compared the number of animals reported in the articles’ Methods and Results sections. In close to two thirds of the experiments, it was impossible to trace the flow of animals through the study, and thus to decide whether any animals had been dropped from their final analysis. Of those that did report numbers, ≈30% reported that they had dropped rodents from their study analysis, but <25% of those explained why. Using simulated data, we demonstrated that this can lead to a major distortion of the results, especially when group sizes are small.

Power Failure: False-Positives and Inflated Effect Sizes

Group sizes in preclinical medicine are exceedingly small. An analysis of >2000 experimental stroke studies performed over the past few decades reveals a mean group size of 8. The Collaborative Approach to Meta-Analysis and Review of Animal Data From Experimental Studies (CAMARADES) database also contains the normalized effect sizes of all these studies, which can be used to calculate the average statistical power. With a mere 45%, the statistical power of the preclinical stroke literature is slightly lower than that in effect when we toss a coin (50%). Yet, perplexingly, this is still superior to the 23% median power calculated for >700 primary neuroscience studies. A power of 45% means not only that an effect, if indeed present, can only be detected in 45% of those cases (high false-negative rate) but also that in these cases effect sizes will be overestimated by >40% (Winner’s curse). In addition, and most worryingly, given reasonable previous probabilities for the effectiveness of the tested compounds (or hypotheses), a power of only 45% will lead to false-positive rates of ≈50%. Ioannidis concluded in his 2005 landmark article that because the synergistic effects of insufficiently controlled bias (eg, by nonblinding, see above) and low statistical power, most published research findings are false.

Publication Bias: Show Me the Evidence

In clinical medicine, where lawmakers, regulators, and journal editors mandate preregistration of trials, it is estimated that the results of only 50% of all studies are eventually published. In preclinical medicine, where preregistration is virtually nonexistent and studies often have no clearly defined beginning or end, we cannot know what and how much high-quality evidence is produced but never reported. However, judging from the almost complete dearth of any studies in the preclinical literature that reject the NULL hypothesis, we can only speculate that there is an exceedingly strong publication bias toward effective drugs or confirmed hypotheses. Sena et al systematically reviewed 525 original publications in the preclinical stroke literature. Only 10 publications (2%) reported no significant effects on infarct volume, and only 6 (1.2%) did not report even 1 significant finding. If this correctly reflects the success rate of hypothesizing and drug treatments in the experimental stroke field, I argue that it is wasteful and potentially unethical to conduct experiments at all: Experiments must necessarily demonstrate that what has been hypothesized must be true (or that drugs work)! Oddly, at the risk of producing a false-negative! Meta-analytic evidence has unequivocally demonstrated the prevalence of publication bias and the detrimental effects it can have. In the experimental stroke literature, publication bias accounts for at least 30% of the published effect sizes.

Low External Validity: Stroke Treatments for Healthy Young Male Rodents

With few exceptions, stroke research is conducted in healthy, very young, male inbred rodent strains raised under specific pathogen-free conditions and fed a diet optimized for high fertility and overall health. The equivalent human cohort would be healthy pubertal twins raised in 6 m² isolator tents on an...
enriched granola diet (Figure), but patients with stroke are elderly, of both sexes, comorbid, on multiple medications, and have had exposure to numerous pathogens and antigens throughout their life. Studies on outbreds, different strains, comorbidities, aging, sex, and diet, as well as housing conditions, have demonstrated the strong impact of these factors on stroke outcome in experimental animals and on treatment efficacy. As a general rule, the closer stroke models have mimicked patients with stroke, for example, after aged animals and comorbidities had been studied, the more substantial has been the loss in efficacy of the treatment effects.27–29 Counterintuitively, in view of the lack of efficacy of the same drugs in clinical trials, this may serve as a further indicator for the good prediction of rodent models of stroke!

Timing of Treatment: Same Tissue Clock in Mice and Men?
Clearly, the exceedingly high attrition rate of clinical stroke trials cannot be blamed on low internal or external validity of preclinical stroke research alone. A multitude of reasons may have led to false-negative clinical results, for example, insufficient sample sizes paired with overoptimistic expectations on effect size. Another problem frequently quoted may have been study designs in which time to treatment exceeded the biological time window within which brain tissue is not yet fully committed to cell death. In a recent review,10 we looked at the time window of major recent neuroprotection trials and found a median targeted inclusion time window of 16 hours. In animal models of stroke robust protection is usually observed only with the first 1 to 2 hours after occlusion of the artery. The fact that the time window for efficacy of thrombolysis is almost identical in rodents30 and man is indirect evidence that the tissue clock of brain tissue after focal cerebral ischemia is indeed similar in rodent and man. Except for the FASTMAG (Field Administration of Stroke Therapy - Magnesium) trial,32 none of the plethora of stroke trials were aimed at treatment within the golden hour. Although this may be understandable from a practical standpoint, it may be speculated that many potent neuroprotectants have been wasted because treatment was too late. Because of the neutral study results, the drugs will probably never be tested again. Fortunately, innovations and improvements in hyper acute stroke care and trial methodology now allow us to put neuroprotection to the ultimate test in the golden hour. Several ongoing trials use randomization in the field and ultra-acute treatment either by paramedics (FRONTIER [Field Randomization of NA-1 Therapy in Early Responders] trial, ClinicalTrials.gov Identifier: NCT0231544333) or by dedicated mobile stroke units with specialist teams deployed to the home of the patient.34 These approaches enable a 1-hour...
time window for treatment and thus offer the ultimate test for neuroprotection.

**Two Study Modes: Discovery and Confirmation**

Most preclinical stroke researchers aim at finding new pathophysiological mechanisms or drugs (exploration and discovery). To emphasize the clinical relevance of their findings, they often at the same time use the results of these exploratory studies to support inferences confirming the utility of their discovery for treatment in humans. This may increase the chances of publishing this work in prestigious journals. But these claims are often not backed up by the sort of robustness and external validity required to advocate translation to humans. Measured against their promise, all attempts to translate preclinically effective treatments into guideline-based stroke therapy have failed. I propose that confounding exploration with confirmation can be a major contributor to the translational roadblock. In a recent article, we have posited that distinguishing between exploratory and confirmatory preclinical research will improve translation. In exploratory investigation, researchers should aim at generating robust pathophysiological theories of disease. In confirmatory investigation, researchers should collect strong and reproducible treatment effects in relevant animal models. We should disentangle these 2 modes and customize design and reporting guidelines for each mode. The Table gives a tentative overview on how discriminating between exploratory and confirmatory preclinical study designs might entail different study designs. Such a policy can leverage the complementary strengths of both modes and 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 this approach would also reveal that most preclinical stroke research date is heavily biased toward exploration, and that confirmation is often missing.

**Can Failure Foster Progress?**

The standard model of bench-to-bedside translation is linear. It leads from novel mechanism or compound to a new and effective therapy in a straight line, and as quickly as possible. Disruptions of this process at any stage we call attrition. The metrics for translational success are time spent from discovery to licensing and the ratio of licensed drugs to failed attempts (attrition rate). In this model, attrition equals failure. Provocatively, London and Kimmelman and Ioannidis argue that failures in the translational process may not only be necessary but also may often be more important than success. Although the quest for novel effective treatments remains the prime mover, they argue that the ultimate goal should be to increase useful information, which includes high-quality negative results. Any negative result in well-designed studies that provide good-quality evidence results in information which can, among other benefits, correct mechanistic concepts, define dosing and timing of treatments, and free up resources for other avenues of investigation. Failures can rule out dead ends, identify research lines that should be modified and inappropriate methods. In particular in the clinical stages of translation, unsuccessful translation trajectories can be critical for maximizing efficient healthcare. Drugs are only clinically useful if we know dosages, treatment schedules, and timing at the bedside. We collect this knowledge by probing the windows beyond which a drug is no longer clinically useful. Kimmelman and London remind us that in basic and preclinical biomedical research identifying promising interventions is akin to exploring a vast, multidimensional landscape of agents, doses, disease indications and treatment schedules. The methods used to explore this landscape often rely on small sample sizes and/or surrogate end points. This allows large areas of the landscape to be explored quickly and at relatively low cost. However, economy and speed come at a cost, since small and less rigorous studies tend to produce more false positives (ie, studies that show spurious clinical promise due to bias or random variation). In particular, base rates for discovering truly effective interventions are likely to be low in areas where our knowledge of disease process, mechanism and pharmacology is underdeveloped. As is well known in diagnosis, when base rates are low, false-positive tests due to random variation are frequent, even if tests are sensitive.

As discussed above, biomedical research is heavily biased against NULL results. As a consequence, and if Kimmelman and London are correct, this must be highly wasteful because available information remains un- or underutilized. Embracing attrition has many important implications for experimental design and reporting: We should plan experiments so that they lead to useful information even when the NULL is rejected;

---

**Table. Suggested Differences Between Exploratory and Confirmatory Preclinical Study Designs**

<table>
<thead>
<tr>
<th></th>
<th>Exploratory (Discovery)</th>
<th>Confirmatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td>Establish pathophysiology</td>
<td>+++</td>
<td>+</td>
</tr>
<tr>
<td>Sequence and details of experiments established at onset</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td>Defined primary end point</td>
<td>–</td>
<td>++</td>
</tr>
<tr>
<td>Sample size calculation</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td>Blinding</td>
<td>+++</td>
<td>+</td>
</tr>
<tr>
<td>Randomization</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>External validity (aging, comorbidities, etc)</td>
<td>–</td>
<td>++</td>
</tr>
<tr>
<td>Predefined inclusion/exclusion criteria</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Test statistics</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td>Preregistration</td>
<td>–</td>
<td>++</td>
</tr>
<tr>
<td>High sensitivity (high type I error rate, low type II error rate): find what might work</td>
<td>+++</td>
<td>+</td>
</tr>
<tr>
<td>High specificity (low type I error rate, high type II error rate): weed out false-positives</td>
<td>+</td>
<td>+++</td>
</tr>
</tbody>
</table>
we should not stop our experiments as soon as first signs of a potential NULL results appear, and of course, we should publish NULL results.

**Overcoming the Roadblock**

Few scientists today deny that bench-to-bedside translation in stroke has a disappointing track record. Analyzing the strengths, weaknesses, opportunities, and threats of this complex process can help to improve its efficiency. Based on the evidence provided by recent metaresearch and several best practice examples, I propose the following measures, many of which are currently in various stages of implementation:

We should try to reduce preventable (detrimental) attrition. Key measures to achieve this revolve around improving preclinical study design. Internal validity should be improved by reduction of bias, for example, by randomization, blinded treatment allocation and outcome assessment, and predefined inclusion/exclusion criteria, among others. External validity can be improved by including aged, comorbid rodents of both sexes in our modeling. False-positives and inflated effect sizes can be reduced by increasing statistical power, which essentially means increasing group sizes. Adherence to reporting guidelines and checklists (such as Animals in Research: Reporting In Vivo Experiments [ARRIVE]) is currently low and needs to be enforced by journals and funders. When planning, conducting, and reporting preclinical studies, we should discriminate between exploration and confirmation. Customizing study designs can leverage the complementary strengths of both modes of investigation. In particular, for confirmatory studies, we should consider publication of study protocols. All studies should publish their full data sets (Open science).

We should embrace inevitable NULL results. Failures are a necessary element of the translational research enterprise and may actually promote our understanding of pathophysiology and successful translation in the long run. This implies planning experiments in such a way that they produce high-quality evidence if the NULL results were obtained, and of course, then making these available to the community.

Some of these recommendations are hard or even impossible to implement for individual laboratories. A collaborative effort is needed to overcome these bottlenecks. Just as in clinical medicine, multicenter approaches help to obtain large enough group sizes and robustness of results. Multicentre Preclinical Animal Research Team, an European Union–funded project with participation of National Institutes of Health (NIH)/NINDS will provide a scalable framework for such efforts. First multicenter preclinical stroke studies have recently been published.

Translational stroke research is not broken, but in need of an overhaul. Its engine must be made more efficient; its results must be made more predictive. I have focused on a few potential remedies informed by recent meta research and limited my analysis to preclinical research. Researchers, funders, journals, and professional societies must work together to develop desperately needed novel and effective therapies for a disease that puts a tremendous burden on patients, their families, and health systems and economies.

**Acknowledgments**

I thank John Ioannidis, Malcolm Macleod, and Jonathan Kimmelman for inspiration and guidance.

**Disclosures**

None.

**References**


**Disclosures**

None.


Thomas Willis Lecture: Is Translational Stroke Research Broken, and if So, How Can We Fix It?
Ulrich Dirnagl

Stroke. 2016;47:2148-2153; originally published online June 28, 2016;
doi: 10.1161/STROKEAHA.116.013244

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/47/8/2148

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org/subscriptions/