Bringing rigour to translational medicine

David W. Howells, Emily S. Sena and Malcolm R. Macleod

Abstract | Translational neuroscience is in the doldrums. The stroke research community was among the first to recognize that the motivations inherent in our system of research can cause investigators to take shortcuts, and can introduce bias and reduce generalizability, all of which leads ultimately to the recurrent failure of apparently useful drug candidates in clinical trials. Here, we review the evidence for these problems in stroke research, where they have been most studied, and in other translational research domains, which seem to be bedevilled by the same issues. We argue that better scientific training and simple changes to the way that we fund, assess and publish research findings could reduce wasted investment, speed drug development, and create a healthier research environment. For ‘phase III’ preclinical studies—that is, those studies that build the final justification for conducting a clinical trial—we argue for a need to apply the same attention to detail, experimental rigour and statistical power in our animal experiments as in the clinical trials themselves.


Introduction

Each year, member countries of the Organization for Economic Cooperation and Development spend, on average, 9.5% of their gross domestic product, or US$3,387 per person, on health care (2011 figures). The figure for the USA is 17.7%, or $8,811 per person.1 Because brain and mind disorders are chronic and debilitating, they account for a disproportionately large proportion of these costs. In Europe, brain diseases are estimated to account for 35% of all disease burden,2 costing €798 billion in 2010.3 Mood disorders, dementia, psychotic disorders, anxiety disorders, addiction and stroke alone accounted for a total of £16.7 billion in 2010 (£113.4 billion, €105.2 billion, €93.9 billion, €74.4 billion, €65.7 billion and €64.1 billion, respectively).3 These costs are high because many brain diseases are common, and we lack good treatments for the majority of these conditions.

Currently, we have no treatments to prevent or cure most forms of dementia.4 The symptoms of psychosis can be reduced in many cases, but not without appreciable adverse effects.5 Also, given that addiction treatments still tend to be discussed in a judicial rather than a clinical context,6 we cannot really claim to be on top of the problem. The situation is not much better for mood disorders. A 2003 US National Institutes of Mental Health report stated that “even when they do receive treatment, only slightly more than half of all patients respond well to therapy, defined as experiencing a 50% or greater reduction from baseline symptom severity. If complete symptom remission or restoration of function is the outcome, then the proportion is even lower.”7

In the case of stroke, we have seen substantial advances in risk reduction in recent years. Blood pressure control, antiplatelet and anticoagulant treatment, and cholesterol lowering have each been shown to reduce the risk of stroke. However, patient adherence to treatment is poor,8 and not all the improvements that are theoretically possible have been realized.9 Thus, despite important advances, stroke is still the second most common cause of death and disability. Approximately 15 million people have a stroke each year, with 6 million dying as a result.10 Around 30% of the 62 million stroke survivors worldwide have marked neurological impairment. We do have one very powerful acute therapy, thrombolysis with tissue plasminogen activator (tPA), but a narrow window of opportunity for therapy and risk of bleeding mean that only a small proportion of patients with stroke benefit from this potent drug.11 Moreover, despite thrombolysis, around half of patients still remain dependent or die.12

In this Review, we explore the barriers to the development and implementation of effective treatments for neurological disorders, focusing predominantly on the stroke field. We highlight the inherent deficiencies and conflicts of interest in our current systems of preclinical and clinical research, and suggest ways in which scientific rigour might be improved.

Drug development—have we lost our way?

The example of stroke indicates that we are making progress in developing new treatments, but from a modern perspective the pace of this progress is frustratingly slow. Admittedly, our expectations may be unrealistic: it took 183 years between Jenner’s demonstration that vaccination could induce immunity to smallpox and the official eradication of the disease in 1979. Surgery advanced only slowly after the introduction of anaesthesia in 1846 and Lister’s description of the virtues of cleanliness in 1867. The first blood bank was not opened until 1937, and Lister’s description of the virtues of cleanliness in 1867.
The costs of treating brain diseases are high, because many of these conditions are common, and we lack good treatments for the majority. Drug development for these diseases is hampered by a paucity of experimental and reporting rigour and a lack of statistical power, leading to overoptimistic interpretation of the literature. Recognition that we have a problem is the first step towards finding a solution.

We have started to provide transparency in experimental design and rigour of reporting in preclinical publications, and we now need to bring similar quantifiable measures to funding decisions. Multicentre, randomized, blinded, powered and appropriately governed preclinical trials conducted by expert consortia offer an opportunity to ensure that only the very best drug candidates reach clinical trial.

Before NASA landed men on the Moon. Even chemotherapy, which can be traced back to the 1919 discovery that mustard gas suppresses haematopoiesis,¹³ is still very much a work in progress.¹⁴ In this context, the 17 years since the use of tPA was suggested by the successful National Institute of Neurological Disorders and Stroke (NINDS) trial¹⁵ is not such a long time.

While increasing societal pressures may contribute to unrealistic expectations, many in the pharmaceutical industry argue that medicine has already plucked the low-hanging fruit.¹⁶ The pharmaceutical industry is undoubtedly in the doldrums: fewer drugs are being approved for use,¹⁷ while it is claimed that development costs are increasing steadily.¹⁸ Have we really run out of therapeutic targets, or have we just lost our way?

Some have argued eloquently for the former proposition, asserting that the law of diminishing returns explains why we have to work harder for each new drug.¹⁹ However, each new year brings further enrichment to our understanding of neurobiology. To optimists, this presents a world of therapeutic opportunity. To pessimists, the complexity of the brain is perhaps a tempting excuse for failure. However, a more plausible and testable explanation for the slowing of drug development is that our systems of scientific discovery, and of drug discovery and testing, have been compromised and we have indeed lost our way.

Deficiencies in study design

The belief that “everything works in animals but nothing works in man” came early to stroke researchers. The Stroke Therapy Academic Industry Roundtable (STAIR) first met in 1999 to discuss the translational roadblock and to proffer solutions.²⁰ From the very beginning, STAIR stressed the importance of key study design factors such as randomization and blinding, and over the years these important meetings have drawn attention to deficiencies in clinical trial design, including inappropriate extrapolation of animal data to the human setting, and a narrow approach to animal modelling that might limit generalizability in the clinic.²¹–²⁵ By 2006, however, it was clear that despite these efforts, the situation was not improving rapidly. With our colleagues, we demonstrated that the agents selected for clinical trials were not necessarily those that were most effective in animals, and that within the animal data, the more a drug was tested, the less effective it seemed to be.²⁶ Systematic review and meta-analysis of the published data for individual drugs selected because they were considered good candidates for clinical trial revealed a persistent pattern of failure to report a priori sample size calculations, or to take measures to reduce the risk of bias. Importantly, where comparisons have been made, studies that do consider these issues usually report less benefit.²⁷–³⁵ Analyses in the wake of the failure of the SAINT II randomized controlled trial of the free radical trapping agent NXY-059 highlighted the potential magnitude of the impact of bias on reported effect size.³²

Bias and lack of statistical power

To our knowledge, only one study has used a randomized, blinded and appropriately powered approach to re-examine efficacy in animals of drugs that seemed to have therapeutic potential for stroke, and the results failed to confirm the promise of the published literature.³⁶ These problems are clearly not unique to the study of stroke. The NINDS Facilities of Research Excellence in Spinal Cord Injury (FORE-SCI) programme, which sponsored independent replication studies of novel strategies to treat spinal cord injury, confirmed beneficial effects for only half of the strategies studied.³⁷ In studies of transgenic models of motor neuron disease, it now seems likely that apparent improvements in survival were not mediated by drug effects, but were a manifestation of bias and lack of statistical power.³⁸

Most experimental stroke studies use fewer than 10 experimental animals in each cohort (Figure 1). Statisticians are uneasy about the retrospective analysis of statistical power because, in testing their hypotheses, investigators have usually established de facto whether the study was adequately powered. However, it is reasonable to ask whether, across a field of research, studies are adequately powered to detect the effects that they report. If not, it implies, first, that selective reporting (an excess of significance) could have occurred; second, that the positive predictive value of those studies that report statistically significant effects is reduced; and, last, that a number of studies falsely conclude that an effect is not present (false-negative studies).

In studies modelling stroke, the efficacy of tPA, the only drug that currently works in both animals and humans, is around 25%.³⁹ This is, therefore, a clinically relevant threshold to guide power calculations in experiments testing other drugs. Along with the average reported SD of 30% for experiments using Sprague Dawley rats (the strain most commonly used in preclinical stroke studies),³⁹ this suggests that for 80% power, the required cohort size is more than 20 animals. If the threshold effect size is inflated by publication bias,⁴⁰ the corrected target effect size of 16% requires over 50 rats, and if it is further inflated by bias arising from methodological issues (for example, randomization and blinding),⁴¹ then the cohort sizes will need to be yet larger. Variance may be reduced by using rat strains that produce more-consistent infarct sizes, such as the spontaneously hypertensive (SHR) rat (average SD

Thereafter, we have started to provide transparency in experimental design and rigour of reporting in preclinical publications, and we now need to bring similar quantifiable measures to funding decisions. Multicentre, randomized, blinded, powered and appropriately governed preclinical trials conducted by expert consortia offer an opportunity to ensure that only the very best drug candidates reach clinical trial.
Figure 1 | Numbers of animals used in experimental stroke studies. Graph shows number of Sprague Dawley rats in each control and treatment cohort in 446 neuroprotection experiments extracted from a database published by O’Collins et al. (2006).26

per experiment = 20%; Figure 2, Table 1),29 or by using models (for example, Tamura or photothrombosis) that provide more-consistent vascular lesions. However, if consistency comes at the expense of generalizability (for instance, through use of models with a very small ischaemic penumbra),41,42 these strategies will prove to be false economies.

Evidence of publication bias and methodological bias also exists for other disease models, such as experimental autoimmune encephalomyelitis, and should be taken into account in power calculations.43

Lack of generalizability
The heterogeneity of drug response in humans is compounded by well-known comorbidities and the fragility that accompanies ageing, and we know that these same problems worsen outcomes in our animal models.44 However, only rarely do we choose to study candidate drugs in experimental settings that are important for generalizability in the clinic; instead, the majority of animal experiments are performed in the young healthy male animals most likely to provide a positive signal.45

For stroke, therefore, an additional source of bias is found in the experiments that the stroke community chooses to perform.

The strengths and weaknesses of the models available to stroke researchers and the impact of comorbidities on outcome in experimental animals and humans have been reviewed elsewhere.24,44-46 No single experiment—or, indeed, single experimental model—is sufficient to explore all the variables that might limit a drug’s utility. It is logical to start the process of drug evaluation in a setting that is cost-effective, and in which a signal is easy to detect (for example, immediate therapy after thread occlusion in young, male, normotensive Wistar Kyoto rats). To reduce the risks of failure before proceeding to clinical trial, however, we need to know whether the drug works in a clinically useful time window in aged male and female animals. We would also like to know whether the drug works if these same animals have hypertension and/or diabetes, or if they have metabolic syndrome or are exposed to cigarette smoke. In addition, does the drug retain its activity when given with other drugs such as anti-clotting agents and tPA, as would be expected in patients with stroke?

We know that we can perform such experiments; for example, we can induce stroke in aged SHR and diabetic (streptozotocin-treated) rats,39 and can combine therapies with tPA.50 Moreover, detection of broadly the same limits to the efficacy of tPA in experimental animals and humans suggests that although each of the models used by the stroke community has different strengths and weaknesses, they are predictive of outcome in humans.33 Moreover, a drug need not satisfy all of the safety and efficacy criteria; for example, tPA is contraindicated in hypertensive patients but still has an important place in the clinic.32

Implausible analyses
Some of the publications describing the in vitro and early in vivo experiments that lay the groundwork for translational research present implausible analyses. Often, such publications report a sequence of five or six experiments using independent samples, each testing a different stage or prediction of a single mechanistic hypothesis. If that hypothesis is correct, each experiment may either support this hypothesis (true positive) or appear to refute it (false negative). Over a series of publications, the proportion of true positive studies should reflect the statistical power of those experiments. We might then calculate the probability that a sequence of experiments will all be true positives if the mechanistic hypothesis is true. Assuming that individual studies are powered at 60%, then two experiments would each be true positives in 36% of publications, and four experiments in 13%—or one-eighth—of publications. In our experience, the vast majority of publications report experiments that all reach statistical significance.
Occlusion
12
Spontaneously
27–35
1.5
19
Figure 2
| Infarct volume variability in different strains of rat after MCAo. Graph shows infarct volumes after transient or permanent MCAo for 1.5 h or 2 h in three rat strains: Sprague Dawley, Wistar Kyoto and spontaneously hypertensive. Abbreviations: MCAo, middle cerebral artery occlusion; p, permanent; t, transient.

Table 1 | Cohort sizes required for preclinical drug testing in animal models of stroke

<table>
<thead>
<tr>
<th>Rat strain</th>
<th>Occlusion time</th>
<th>50% effect size</th>
<th>40% effect size</th>
<th>30% effect size</th>
<th>20% effect size</th>
<th>10% effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sprague Dawley</td>
<td>1.5h (young)</td>
<td>43</td>
<td>67</td>
<td>119</td>
<td>268</td>
<td>1,062</td>
</tr>
<tr>
<td></td>
<td>2h (young)</td>
<td>56</td>
<td>87</td>
<td>154</td>
<td>345</td>
<td>1,389</td>
</tr>
<tr>
<td></td>
<td>24h (young)</td>
<td>29</td>
<td>45</td>
<td>80</td>
<td>179</td>
<td>715</td>
</tr>
<tr>
<td>Wistar Kyoto</td>
<td>2h (aged)</td>
<td>12</td>
<td>19</td>
<td>33</td>
<td>74</td>
<td>295</td>
</tr>
<tr>
<td>Spontaneously</td>
<td>hypertensive</td>
<td>2h (young)</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>2h (aged)</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>20</td>
<td>80</td>
</tr>
</tbody>
</table>

*Calculated to provide a power of 80% and α = 0.05 under different experimental circumstances. †Effect size of drug under investigation.

For 80% of publications describing a sequence of four experiments, to find them all truly positive would require that individual experiments were powered at 95%, which is simply not plausible. Alternatively, more experiments might have been conducted but not reported; again, with power of individual experiments of 60%, seven independent experiments testing aspects of that hypothesis would provide at least four true positives suitable for publication 70% of the time.

Selective analysis and reporting might also explain the excess of significant studies seen in the in vivo modelling of neurological disease. In the same way that it is not reasonable to conduct multiple statistical tests on data from the same cohort of animals without correcting for type I errors, it is also unreasonable to penalize a series of independent studies testing the same hypothesis in the same publication; journals should encourage the use of alternative statistical approaches that might address this issue, and be more tolerant of manuscripts in which some experiments do not give a statistically significant result. Indeed, journals should be suspicious of manuscripts in which a large number of positive experiments are presented. Even if the hypothesis is correct, such an eventuality is highly unlikely to have occurred without some help. For authors, it may be helpful, before a series of experiments, to specify which experiment is to be considered as the primary test of the hypothesis, or to use tests of significance that consider the totality of evidence presented.

Implications
Taken together with evidence of failure to avoid experimental bias and widespread publication bias, it seems likely that at best the true effect of many candidate drugs for stroke and other indications is significantly less than suggested by the literature, and that some might be completely inert. These biases in the animal literature may result in biologically inert or even harmful substances being taken forward to clinical trials, thus exposing patients to unnecessary risk and wasting scarce research funds. Chalmers and colleagues have estimated that we might waste as much as 85% of our research effort.

For another estimate of the costs of research failure, we can assess the impact on the market capitalization of pharmaceutical companies when an apparently promising drug fails in clinical trials. The day before publication of the neutral results of the SAINT II trial, AstraZeneca shares were trading at $66.49, but within a few days had fallen to $58.68. Given the number of shares in circulation, we can estimate that the market’s valuation of a positive trial result was at least $12 billion. Overall, the evidence suggests that we have indeed lost our way. If this is the case, how do we find our way back to the true path?

Improving research conduct
Removing incentives that compromise conduct
The problems that we face are multifaceted, and lasting solutions will require the cooperation of educators, scientists, funders, publishers and the community at large. In many regards, we reward ‘quick and dirty science’, favouring for promotion and funding those who report ‘exciting’ findings in high-impact journals, with less attention paid to the veracity or long-term value of the findings. NINDS-sponsored studies have failed to replicate many exciting studies in spinal cord injury, and while exciting genetic association studies reporting disease associations have been published in high-impact journals, their subsequent refutation with better-quality data has not received such favourable treatment. Conversely, as the biomedical sciences industry grows, we are confronted with a plethora of new journals, which ensure that almost all articles can eventually find a home. One part of the solution would be to empower editors and reviewers to make publishing decisions on the basis of quantifiable criteria. We should judge novelty according to evidence from systematic search strategies. We should judge the veracity of the data on the basis of demonstrated experimental rigour and evidence of institutional oversight. We should judge the completeness of
the story, and only then—if at all—consider the purely subjective potential for ‘impact’ in the future. Under such a schema, journal impact and citations would have true value, and a scientist’s publication record would facilitate informed decisions on the risks involved in funding and employment. Expensive data-capture schemes to help governments rank institutions and scientists are perhaps missing the point: that it may be more cost-effective to fix the way in which data enter the literature than to manipulate existing but compromised data. In a world where electronic rather than paper publishing is becoming the norm, peer review needs to add more value if scientists are to continue to expend their time and effort on the process.

Better publication policies are on their own unlikely to solve the problem. Researchers inhabit a complex world of rewards and challenges relating to granting, publishing and promotion systems. The challenge is to provide incentives that reward the best science and scientists. The quantitative criteria outlined to assess publication quality might contribute to this process if they were also used to assess the quality of prior work in considering funding applications; those scientists whose work had been read and cited most widely and had been replicated by others would have greater access to funding and, thus, opportunity for continued success. Applications from groups with established systems for institutional oversight and monitoring, audit of research quality, and provision for the continuing professional development and appraisal of the research workforce might similarly be favoured.

Practical steps

The beginnings of the broad structural changes outlined above are in place and are gaining momentum. Perhaps most important is the recognition that we have a large problem, which requires a cooperative and multinational solution. In the USA, NINDS has sponsored replication studies and symposia to discuss these issues. Major stroke journals on either side of the Atlantic have carried editorials making the case for bringing the rigour of randomized, controlled, multicentre clinical trialling to the preclinical setting. Both clinical and basic science focused meetings have made the same plea, and the European Union has provided seed funding to develop MultiPART, the first randomized controlled trial network for preclinical stroke research.

Multicentre animal studies are not at present a routine part of drug development. In the early 1980s, the US National Heart, Lung and Blood Institute was concerned about inconsistencies in findings from different laboratories that were testing the same hypotheses using very similar experimental designs. To address this issue, they funded a study, Animal Models for the Protection of Ischemic Myocardium, which involved five groups testing the efficacy of ibuprofen and verapamil in dogs (four centres) and rats (one centre). As well as demonstrating the feasibility of this approach, concerns about differences in the data from different centres led to detailed statistical analysis that revealed that an investigator in one centre had been fabricating data. Since multicentre animal studies were proposed for focal ischaemia, other disease communities have advocated a similar approach. Operation Brain Trauma Therapy describes a network of laboratories that would systematically collect evidence to support a decision to move to clinical trial. However, while this consortium advocates central coordination of their research effort, they do not propose that the same experiment should be conducted across a number of sites, or that opportunities for central randomization, data monitoring and statistical analysis should be leveraged.

More recently, the International League Against Epilepsy/American Epilepsy Society (ILAE/AES) Working Groups joint meeting to optimize preclinical epilepsy therapy discovery suggested that translational efforts in epilepsy might be enhanced by adopting what they describe as a phase II multicentre trial paradigm employing an approach similar to human multicentre clinical trials. They argue that this approach “would reduce biases related to individual laboratory practices and conditions, implement rigorous blinding and statistical design, and incorporate independent monitoring of data collection and analysis.” In many ways, this sentiment echoes the efforts of the stroke community in establishing Multi-PART. However, the ILAE/AES group claims that there may be circumstances where sufficient evidence is available from single-centre studies to justify a decision to proceed to clinical trial without a confirmatory multicentre animal study. Their use of the term ‘double-blinded’ is unclear, implying that they would take steps to ensure that the animal subjects were not aware of treatment group allocation. Knowledge of treatment group allocation on the part of the investigator might affect how the animal is handled, thereby possibly affecting corticosterone levels and, thus, animal behaviour, but it would be more precise to describe this as blinding of the investigator to treatment group allocation rather than implying that the animals were in some way blinded.

Conclusions and future prospects

Preclinical and translational journals have embraced the principles of rigour and transparency by adopting publication guidelines similar to those which allowed clinical journals to improve the rigour of clinical trials after publication of the CONSORT statement in 1996. The benefits of this development, in terms of the quality of published research, should start to become apparent in the near future.

The presence of a substantial publication bias must be addressed. Less-exciting findings might be left on the investigator’s hard drive, or are published in journals that are not readily available to the community, or in languages other than English. To know which research has been done, repositories of work that has been funded or received ethical approval should be established, and deposition in these repositories should be a precondition of funding, ethical approval and publication.

Investigators might measure a number of outcomes and report only the most interesting or those that
support their hypothesis; and statistical analysis plans might be adjusted after the event in light of the data. Both of these practices will introduce bias, and could be avoided by prior publication (or deposition) of the study protocol. At first, this information might only be available to those reviewing manuscripts for publication, with a sunrise clause that after a given period, or on publication of study results, it becomes more widely available. Publication of all outcomes will inevitable mean that those manuscripts contain some findings that do not unequivocally support the study hypothesis; journals and reviewers should be tolerant of these blemishes, and consider the merits and interpretation of a body of work as a whole.

The adoption of meaningful sample size calculations, and the avoidance of models with low variance but restricted generalizability, will mean that more animals need to be used in each experiment. Funders and ethics committees must understand that it is better to fund a large study that gives useful and reliable results than to fund any number of smaller studies that might appear to provide a reduction in animal numbers but in fact provide data of limited use. We believe that multicentre animal studies might provide the capacity required for such studies.

The experience of attempted replication suggests that a systematic approach to confirming findings is as important as the initial research. There should be funding and recognition for those carrying out such studies. In areas where translational failure is apparent or where clinical trials are based on findings from animal studies, replication, in adequately powered studies at low risk of bias, should be a precondition for further clinical development.

The Nature and Science journals have taken important steps to improve the rigour of the work that they publish,\(^9,10\) Risk of bias might be reduced further if all journals publishing in vivo research were to take action to reach the same standards. The next challenge will be to ensure that the measures taken to reduce the risk of bias within the laboratory were appropriate. For instance, coin-toss randomization is much less rigorous than the use of computerized randomization systems.

In most professions, there is an expectation of continuing development, such that individuals do not come to a role fully formed. Scientists are by nature ‘improvers’, but there is a dearth of organized activities to aid their career progression. Such activities might include opportunities for developing new skills in relevant areas, time set aside for audit and quality control activities, and opportunities for mentoring and reflection at all levels. Other professions have instituted formal programmes for continuing professional development; while some research establishments are developing similar programmes for their workforce, the focus is still largely on grants in and papers out.

Many of the concerns highlighted above have arisen through a combination of translational research and empirical research that has explored the epistemology of translational medicine: how we know what we know. There is an ongoing need to enrich our knowledge of what makes for effective translation, including the risk of bias and how to address it, and to study the impact of the changes suggested here and elsewhere. This effort will require research, an appropriately trained research workforce, and appropriate funding.

**Review criteria**

We selected articles for inclusion in this Review on the basis of our existing knowledge. We did not employ a formal search strategy but, rather, selected articles that illustrate the points we thought relevant to the subject matter being discussed. While we have attempted to provide a fair summary of the relevant literature, there is likely to be an unconscious bias in favour of publications that support our point of view.

---


