Evidence-based Translational Medicine

David W. Howells, BSc, PhD; Malcolm R. Macleod, PhD, FRCP

Are We Abandoning the Search for Acute Stroke Therapies?

The pharmaceutical industry is in the doldrums. Food and Drug Administration drug approvals per year are falling steadily,1 and the costs of drug development are steadily increasing.2 However, although neuroscience as a whole may not be delivering the returns on investment we would like, the state of stroke medicine has never been better.

Although changing lifestyles for the better often seems intractable, pharmacotherapy has delivered effective primary prevention. Control of blood pressure, cholesterol lowering, and control of clotting in atrial fibrillation are all providing benefit. The same approaches are effective in secondary prevention, in which antiplatelet therapies already have a long-established role. We also have ever come slowly to the realization that the brain is malleable, and that therapy during the long-term rehabilitative phase of stroke is likely to help patients regain useful function.

Despite this, we are still struggling to make gains in acute stroke treatment. Thrombolysis was first introduced in the 1990s after the National Institute of Neurological Disorders and Stroke (NINDS) trial in a period that saw intensive evaluation of a range of lytic approaches. However, only a small proportion of the patients who might benefit from this powerful treatment receive it.3 Surprisingly, despite the clear benefit of tissue-type plasminogen activator, we have still to complete the analysis of novel thrombolytics, such as desmoteplase and tenecteplase.2

Moreover, we seem to be on the verge of abandoning neuroprotection, one of the most extensively studied concepts in neuroscience.4 For all intents and purposes, Big Pharma has already done so despite convincing proof of concept in humans; neuroprotection by hypothermia is in routine use to prevent the neurological consequences of cardiac arrest in adults5 and hypoxic-ischemic encephalopathy in neonates.6

The pessimism within the pharmacological industry seems reasonable. More than 100 clinical trials of neuroprotectants in stroke have failed to deliver an effective drug.3 However, despite the past expenditure, it is arguable that the neuroprotection hypothesis has not yet been refuted in ischemic stroke because it has not (yet) been adequately tested.

Although thrombolytic trials have always emphasized the importance of timing, there has rarely been the same emphasis in the clinical trials of neuroprotection. Laboratory and clinical thrombolysis experiments show similar, and unfortunately short, windows of opportunity,7,8 which have not been matched in most trials of neuroprotection.9 This disparity is illustrated by analysis of the animal and human data for tirilazad, in which the median time to successful treatment in animal experiments was 10 minutes, whereas in unsuccessful human trials, it was ≈5 hours.10 Contrast this with the data for tissue-type plasminogen activator, which works in rodents and man, and where, before the publication of the Third International Stroke Trial (IST-3), the median time to treatment in both data sets was 90 minutes.7,11

Other biases also play a large part in the over-optimistic interpretation of the preclinical data. NXY-059 was trialled robustly12 within a 6-hour time window in patients but without benefit. Yet a systematic review and meta-analysis of the preclinical data prompted by the failure of SAINT-II trial11 suggests that inadequate attention to preventing avoidable bias by randomization, allocation concealment, and blinding gave a falsely positive signal.13 Similarly, a systematic review and meta-analysis of preclinical animal studies suggests that NXY-059,14 tissue-type plasminogen activator,15 nicotinamide,16 and FK50617 are all less effective in the presence of the common human comorbidity, hypertension.

These problems are by no means restricted to the study of stroke. In emergency medicine, animal studies that fail to randomize or blind outcome assessments are more likely to report positive results,18 and for motor neuron disease, reports of therapeutic benefit in transgenic mice are most likely measurements of noise in the distribution of survival means rather than actual drug effects.19

Stroke researchers have in many ways led the way in identifying and beginning the process of correcting these deficits in our research methodology.20–22 and because of this, preclinical stroke research is now in better shape than many other branches of translational medicine. However, we believe that there is still room for substantial improvement before we can be confident that publications, funding decisions, the direction of mechanistic studies and target identification, and the prioritization of clinical trials are each based on a robust and unbiased assessment of current knowledge.

How Can We Find a Remedy?

To move forward constructively, to avoid wasteful effort and expenditure, and to provide the very highest quality evidence

Received December 11, 2012; accepted December 28, 2012.
From the Florey Institute of Neuroscience and Mental Health, Heidelberg, Victoria, Australia (D.W.H.); and Department of Clinical Neurosciences, University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom (M.R.M.).
Correspondence to David W. Howells, BSc, PhD, Florey Institute of Neuroscience and Mental Health, Heidelberg, Victoria, Australia. E-mail david.howells@unimelb.edu.au
(Stroke. 2013;44:1466-1471.)
© 2013 American Heart Association, Inc.
Stroke is available at http://stroke.ahajournals.org
DOI: 10.1161/STROKEAHA.113.000469
to support future clinical trials, there are a number of simple and relatively painless measures that might be put in place. These require better education, training, and mentoring of preclinical researchers to ensure they generate the unbiased data that are most useful to the clinical community; better education, training, and mentoring of clinical researchers to ensure that they are equipped to evaluate animal modeling data; and a concerted effort by funders and publishers to ensure that experiments are well designed and accurately reported. Fundamental to this approach is the view that all scientists want to do research of the highest quality; but that perverse pressures of funding, career, and reporting systems, which reward novelty, and priority before quality mean that good ideas are not always followed up by good science. Over the years, these cumulative compromises have taken us all a long way from the scientific ideal. The challenge is to provide a defense against these perverse pressures and to create a space that allows scientists to pursue their work independent of these external considerations.

Reporting Guidelines
Clinical science was beset by the same problems until the middle of the last century, when it became clear to trialists and clinical journal editors, in particular, that the influence of bias had to be removed if clinical trials were to be successful. The Consolidated Standards of Reporting Trials (CONSORT) statement indicated that to publish clinical trials in key journals, authors should report what measures they take to avoid bias.23 This led to a demonstrable improvement in the reporting of clinical trials.24

The CONSORT statement was followed by a number of consensus statements (usefully collected together by the Equator network, www.equator.net), describing reporting standards across wider areas of research. In 2011, the UK National Center for the 3 R’s was instrumental in facilitating the Animal Research: Reporting In Vivo Experiments (ARRIVE) guidelines, a set of reporting standards for in vivo animal research. Although not derived through a formal process of consultation, these guidelines drew on the principles contained in other reporting standards and on the Stroke Good Laboratory Practice guidelines,25 and their 20 items have been adopted by >45 journals publishing in vivo research.26

Practice Guidelines
Despite the positive CONSORT experience, requiring authors to describe how randomization, group allocation, and blinding were performed (in a reporting guideline) is not the same thing as requiring improvements in the quality of the experiments themselves. In focal cerebral ischemia models, providing practice guidelines to begin the process of increasing rigor, improving quality, and enhancing the prospects for successful translation of efficacy from animal studies to human clinical trials began early. In 1999, Marc Fisher convened the first Stroke Therapy Academic Industry Roundtable meeting in Washington. This and subsequent meetings have focused on particular aspects of improving stroke translational medicine.20,27–33 From the very first Stroke Therapy Academic Industry Roundtable meeting, these meetings stressed the importance of key study design factors, such as randomization and blinding.20,33

As more evidence emerged that studies, which did not report measures to avoid bias, gave higher estimates of efficacy,7,14 it became apparent that it would be helpful to have guidelines setting out a number of critical aspects of experimental design, which should be described in all articles. Good Laboratory Practice guidelines were published in this journal in 2009, with simultaneous publication in the Journal of Cerebral Blood Flow and Metabolism and the International Journal of Stroke, setting out minimum requirements for reporting.21,25,34 These were adopted by the editorial teams of those journals, and in 2010, Stroke introduced a short checklist covering the most important of these as part of its article submission system. Although such systems have been in place for some clinical journals for some time, this was to our knowledge the first system in any research field introduced to deal specifically with preclinical animal research.

Importantly, these did not just set out the aspects of study design, which should be described, but also gave recommendations for how experiments should be done, including that they should be randomized; that the investigator should be blinded to treatment group allocation wherever possible (and specifically when determining outcome); and that wherever feasible, a priori sample size calculations should be performed and reported.

Assessing Risk of Bias at Other Stages of Research
In 2011, the US National Academy of Sciences produced a report highly critical of an Environmental Protection Agency publication about the potential environmental toxicity of formaldehyde on the grounds that the methods used lacked clarity and transparency and were not sufficient to critically evaluate data from individual studies.35 The Environmental Protection Agency, along with the National Toxicology Program (part of the National Institutes of Health), was tasked with bringing forward proposals for how they might address these shortcomings. These agencies have subsequently been involved in developing systematic, robust, and transparent methods for assessing environmental risk based on human and animal data, and these methods are based squarely on systematic review, meta-analysis, and risk-of-bias assessment. In this, they have been greatly assisted by the work of the Navigation Guide program20 at the University of California, San Francisco, and by Lisa Bero, a leading research methodologist and elected member of the Cochrane Collaboration Steering Group.

In an important development in the United Kingdom, the heads of the 2 major funders of in vivo research wrote to the deans of all UK medical schools stating their expectation that all research that they funded should, when published, comply fully with the ARRIVE reporting guidelines. However, they stopped short of saying that they would develop systems such that the risks of bias, which the ARRIVE statement is designed to illuminate, might be addressed at an earlier stage of the research cycle by making an assessment of such integral to their funding decisions.

In contrast, the US NINDS took a different approach, announcing in late 2011 that to ensure the quality of work, which it funded, one dimension of its funding decisions would be an assessment of the extent to which the applicants had sought to minimize bias in their experimental designs.
Optimizing the Science
In the summer of 2012, NINDS convened a 2-day meeting in Washington to discuss the importance of these issues with funders, publishers, reviewers, and scientists. Essentially, the aims were to get a sense of whether these different communities felt that too much had been done already, that progress was about right or that more needed to be done; and if more needed to be done, what this should be.

Although we cannot speak for other participants, it was our previous expectation that any consensus would be superficial and limited to a narrow range of issues rather than being broad and deep; and that, at best, this would be only the start of what might be a very protracted process.

The meeting consisted of a number of mentored discussions, all in plenary, with presentation of evidence from different research domains. What was striking was that the journey on which we have embarked for stroke has been started by others in their different areas, often using different approaches but usually with a singular conclusion: certain aspects of the animal modeling of disease are in need of improvement.

To give just 2 examples, Steve Perrin discussed the experience of the ALS Therapy Development Institute, where they sought to replicate reports of drug efficacy in the SOD1 model of Amyotrophic lateral sclerosis (Motor Neuron Disease). Remarkably, in high-quality well-conducted experiments, they found that not one of the interventions previously reported to have efficacy did so in their laboratory.20 Similarly, Khusru Asadullah recounted the experiences of Bayer in seeking to replicate published findings, which they considered to have therapeutic potential across oncology, women’s health, and cardiovascular diseases. In their high-quality intensively monitored in-house research, they found substantial inconsistencies in 66% of cases.37

There was a striking consensus among the scientists, reviewers, and funders that issues of bias were important in just about every in vivo research domain, and that there was an opportunity to substantially improve the quality of study design and therefore to increase the prospects that search findings would inform improvements in human health. There was also a clear consensus that such improvement would come about more quickly if actions were taken at a number of stages of the research cycle, including funding and publication or research and training and promotion of investigators.

Among the publishers, the consensus was perhaps less clear, recognizing a central role of peer reviewers, and funders that issues of bias were important in just about every in vivo research domain, and that there was an opportunity to substantially improve the quality of study design and therefore to increase the prospects that search findings would inform improvements in human health. There was also a clear consensus that such improvement would come about more quickly if actions were taken at a number of stages of the research cycle, including funding and publication or research and training and promotion of investigators.

An important theme was that improvement measures should have the lowest impact on researchers consistent with success (ie, they should not be overly bureaucratic or time-consuming). Although there were a large number of issues that might be included in a risk-of-bias checklist, there was consensus that it would be best to start off with an identification of a small number of critically important items (addressing which would have greatest impact), rather than seeking to change everything at once.

Again, there was remarkable consensus about what these items should be, and that they will be familiar to most readers and writers for this journal:

Randomization
Allocation of animals to different experimental groups should be done using a process based on chance alone, preferably involving computerized randomization. This included methods of, for instance, blocked design or stratified randomization.

Blinding
At every stage of the experiment where this is feasible, the experimenter should be blinded to the experimental group to which the animal belongs. This is particularly critical in the measurement of outcome but is also important at other stages of the experiment.

Reporting Exclusions
All animals, all data, and all outcomes measured should be presented in the final analysis. If animals were excluded, reasons should be given, and individual datum points should not be excluded unless for very good reasons, according to prespecified criteria and without knowledge of experimental group, and all outcomes that were measured should be reported.

Sample Size Calculations
Research applications and articles should describe how the study size was determined. This should include a formal power calculation where appropriate (and usually it is appropriate), and this should include justification of the assumptions (including the choice of effect size and the source of the variance used) underlying the power calculation.

At the close of the meeting, there was substantial enthusiasm to build on what had been achieved, including the publication of a report of the meeting, and this was carried as a Perspective article in *Nature* in October 2012.38

Future Development
It is apparent that as a community, we are somewhat ahead of the rest of the in vivo life sciences, certainly the neurosciences, in the attention we pay to these issues. The Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES) database now contains systematically identified and collected data from >4800 publications covering in vivo animal modeling of 7 different disease types; compared with other journals represented in that database by >100 publications, this journal performed highly in the reporting of 3 key measures to avoid bias (randomization, blinding, and a sample size calculation); and of note, it performed better than all journals with a higher Impact Factor (Figure 1).

This is not a manifestation of low quality just in the neurosciences; in a random sample (using PubMed id) of 1000 publications carried in PubMed, 130 publications described in vivo experiments. The prevalence of reporting of randomization, blinding, and sample size calculations was, if anything, lower than seen in the neurosciences, although there was a suggestion of improvement over time (Figure 2).

Further efforts to improve the reliability and validity of in vivo focal ischemia modeling must therefore start with an acknowledgment that although much has already been done, there is still room for improvement.
First, we should continue to press for higher rates of reporting of measures to avoid bias. Editorial checklists should have substantial impact, and in time, we should audit whether they have indeed had an impact. If they have not, provision of more detailed guidance for authors and indeed for reviewers may be needed. Those journals that do not currently have either their own checklist or endorse the ARRIVE guidelines should seriously consider doing so.

Second, major funders should assert (as NINDS has done) that having an experimental design that avoids bias is not a nice addition to an application but is fundamental to its likely success. Although many would assert that this is delivered by the current peer review system, the presence of biases in the publications funded in those projects would seem to suggest that this is not the case. Again, there may be an important role in the use of checklists to bring these issues to the forefront of the attention of the reviewers and funding committees.

Third (and this was discussed in the NINDS meeting) is the issue of academic promotion. At an individual and institutional level, there is still far too great a focus on the Impact Factor of the journal in which work is published rather than on the quality of the work itself, and there seems to be little consideration of whether the work was subsequently replicated or refuted. This drives a short-term ism, a focus on presenting work in a way that makes it attractive to high-impact journals. This often seems to involve gathering little bots of evidence across a range of conclusions and presenting a story, rather than providing sufficient high-quality evidence to unequivocally demonstrate a single paragraph of the story. Indeed, there is good evidence of publication bias here, with, for instance, small but positive genetic association studies being published in journals of substantially higher impact than the subsequent larger high-quality studies, which refuted those findings.

The cynical lesson for young investigators seems to be that they should do a large number of small studies, wait for the spuriously positive interesting result (which will come, in time, by chance alone), publish it, and move on with their reputation enhanced. We do not believe that this is a model that can be sustained.

Rather, performance appraisals for scientists and academic appointment boards need to engage with the more difficult task of assessing the work done not on the basis of where it was published but rather on the basis of what it reports. This

Figure 1. Prevalence of reporting of randomization (A), blinded outcome of assessment (B), and sample size calculation (C), among 1750 publications covering diverse in vivo disease models from 10 journals each contributing >100 publications to the CAMARADES database (December 8, 2012). Journals are ranked anonymously from A through J in descending order of their 2011 Journal Citation Index. (ie, Journal A has the highest Impact Factor). This journal (Stroke) is Journal D, shown in black.

Figure 2. Prevalence of reporting of randomization, blinded assessment of outcome, blinded conduct of experiment, concealment of allocation sequence, sample size calculation, and conflict of interest statement (COI) in 130 publications reporting in vivo experiments identified from 1000 randomly selected publications. From Baginskaite, CAMARADES Monograph 2012/01, available at www.camarades.info/index_files/CMs.htm.
will require some change in approach both from the assessors and from those being assessed. Stroke clinicians are well versed in the processes of appraisal, revalidation, and board certification, and although we are perhaps not yet at the stage of introducing similar systems for laboratory scientists, it is one possible approach.

Finally, we need to be more rigorous in the conduct of experiments themselves. One way of providing confidence that the hypothesis being tested has not shifted in the light of the data observed, or that the 1 positive outcome of the 10 measured was the one that was reported, or that the sample size has been inflated after a preliminary neutral analysis, or that 4 different statistical tests have been used and only 1 reported, is to have a date-stamped study protocol available to reviewers. This would give substantial confidence in the findings presented, and we suggest that researchers should certainly have such protocols and should consider making them available as part of the review process.

Experimental conduct might also be substantially improved, as has been argued previously, by the conduct of multicenter animal studies. To begin, these could sit between existing efficacy studies (phase 2 animal studies) and the decision to start clinical trials. They would provide for large, adequately powered studies to be conducted with central randomization, off-site–blinded outcome assessment, and for training, mentoring, and monitoring. We believe scientific fraud to be uncommon in stroke research, but central monitoring can identify it when it does occur. Indeed, there is good precedent here, where a National Institutes of Health–funded multicenter animal study of ischemic heart disease in dogs identified fraud at 1 center through central statistical analysis.40

We have been among the more vocal proponents of the need to improve study quality in in vivo modeling in stroke, and more generally, we have been impressed by the willingness of colleagues to engage with these issues and with the speed of progress. Stroke has played an important role in leading other fields of research and more general neuroscience journals; we now have the opportunity to consolidate that role and to continue to provide innovative and effective solutions to problems, which are in play in much of in vivo research.

Acknowledgments

Dr Howells wishes to acknowledge the support of the Victorian Government and in particular the funding from the Operational Infrastructure Support Grant.

Disclosures

Dr Howells receives funding from Australian NHMRC Program and Fellowship Grants.

References


**KEY WORDS:** guidelines n meta-analysis n neuroprotection n preclinical n reporting standards n stroke n systematic review n therapy n translation