Show me some discipline

Discipline of Physiology, Anderson Stuart Building, University of Sydney

This year's AuPS meeting in Sydney will feature a symposium on '***Rigor and reproducibility in physiological research***' (organized by Severine Lamon and me)**.** We will hear from three Australians who have thought long and hard about this troubling issue: Professors Simon Gandevia, Miranda Grounds and David Vaux. The aim is to focus the collective intelligence of AuPS members upon what we, as a discipline, can do to develop more consistent practices and reporting of physiological research.

The need for improved rigor has been highlighted by fraternal scientific societies and by prominent journals. Many drugs that appeared very promising when tested in animal models of disease failed to provide any benefit in subsequent clinical trials. Alarmingly, in many cases attempts to simply replicate the original animal findings also failed. Together these failures have cast doubt on the value of animal models more generally (Prinz *et al.*, 2011; Steward & Balice-Gordon, 2014). Who will be willing to invest millions of dollars in clinical trials to test a new drug if there are doubts about whether the animal-based evidence can be replicated?

Many issues could account for failure of an attempt to replicate animal findings. Differing results might be due to details in the experimental intervention, procedural differences or differences in the way the outcomes are measured and statistically analysed (web talks: http://neuronline.sfn.org/collections/promoting-awareness-and-knowledge-to-enhance-scientific-rigor-in-neuroscience). The sample size may have been too small, or the statistics inappropriate. It is also possible that subconscious bias affected the assignment of animals to treatment groups or the grading of outcome measures. The challenge now for biomedical

researchers is to rebuild trust and confidence in our findings through improved transparency and more consistent discipline standards.

Back to Basics

Much of physiological theory rests upon empirical evidence. We develop a hypothesis then test it with experiments. Confidence in the results of any experiment depends upon our willingness to assume that the findings were **not** the result of:

1/ inadequate sampling (observations that were not representative) OR

2/ a subconscious bias by the experimenters that may have distorted their results.

As scientists, reviewers and journal editors we have been collectively a little slack about both these assumptions (Landis *et al.*, 2012; Perrin, 2014). Given biological variability, single observations or small sample sizes can be misleading. This is what led our scientific forebears to insist on replication, and to develop statistical tests for the null hypothesis. David Vaux has previously identified many examples of serious errors in the use and interpretation of statistics in the peer reviewed literature, including some of the most prestigious of journals. As he points out, the right statistics depend upon the kind of experiments one does. It really requires development of field-specific expectations (Vaux, 2012).

Spiral staircase, Anderson Stuart Building, University of Sydney

Then there is the insidious problem of confirmation bias. Two thousand years ago the Roman lawyer, Pliny, observed that: *"…everyone is prejudiced in favour of his own powers of discernment, and will always find an argument most convincing if it leads to the conclusion he*

had reached for himself…"[1] We humans are all prone to confirmation bias, whether we are aware of it or not. A scientific hypothesis may develop in a lab over many years. The more work we put into developing our hypothesis the greater will be our emotional attachment to it. Subconscious confirmation bias might distort every step in the design and execution of an experiment and it's analysis, if we don't work to guard against it.

Salesman versus scientist

Science is competitive. Success, or even feeding the family, may depend upon numbers of published papers and publishing in 'big name' journals. Competition to get published in 'good' journals, combined with competition among journals (for journal impact factors) has resulted in an ever-greater emphasis upon perceived significance. We must now all focus on 'selling' our manuscripts to reviewers and editors. This imperative may affect the way we report, what we include (and don't), and the degree to which we are openly self-critical about our results. Selling the story can come at the expense of some of the nitty-gritty of scientific quality control. In recent years some journals have stipulated better reporting standards, but progress in adoption has a long way to go (Kilkenny *et al.*, 2010; Landis *et al.*, 2012).

Plenary speaker

Reviewer fatigue is another problem. Quality control in science assumes that experts are willing to commit the time and effort to carefully read and critically evaluate the work of others. With the explosion of (for-profit) online journals, journal editors can have difficulty recruiting expert reviewers with the right scientific and technical backgrounds who will review a manuscript fairly and thoroughly. This is particularly a problem with multidisciplinary papers where two or three reviewers may struggle to critically evaluate all the various technical aspects of the paper. While we are all busy advancing our own individual careers the 'scientific commons' grows weeds.

[1] *The letters of the younger Pliny,* translated by Betty Radice (Penguin Classics).

Developing consistency of protocols and practice

There is much to be gained if researchers within a field share common experimental protocols and assessment criteria rather using than lab-specific or individual-specific protocols. However, many fields of biomedical science resemble manufacturing before the standardization of screw threads. In 2014 I took part in a meeting sponsored by the Myasthenia gravis Foundation of America and NIH to develop guidelines for preclinical myasthenia gravis research. What became very evident to me were the many subtle protocol differences in how 'standard' animal models were implemented by different researchers, despite it being a quite mature field. Weakness grades are usually the primary outcome measure yet the criteria for assigning these grades often differed, or were vague and open to differing interpretations. What one lab described as 'grade 2 weakness' in an experimental mouse might have been scored as 'grade 3' or even 'grade 1' by another lab. Up until recently blinded grading has been rarely mentioned in published papers in the field. Multiple independent studies all came to the conclusion that autoantibodies against Muscle specific kinase (MuSK) cause myasthenia gravis but, as a field, we are still far from adopting standard practices and assessment criteria that would allow us to pool data (Phillips *et al.*, 2015).

More progress has been made over the past decade in developing standard protocols for the mdx mouse model of muscular dystrophy, through the work of Miranda Grounds and her international and national colleagues (Willmann *et al.*, 2012). Developing such protocols requires a deep understanding of the biology, together with a sustained commitment to work with others and proselytize the cause. These efforts will only be successful if the protocols and guidelines become widely adopted, and this requires persistence.

Lasting value of experimental data

Just as important is the need to report the experiments in sufficient detail to permit meta-analysis and replication (Landis *et al.*, 2012). Detailed reporting might include showing data points on graphs (rather than just mean and SEM) and making spread-sheets of raw data freely available as supplementary files. Providing such details will help build trust and confidence. Standardizing animal models and detailed reporting will allow results of multiple studies from multiple research groups to be combined, thereby building statistical power and confidence about reproducibility (when present). However, detailed reporting will only happen if authors feel they can trust that their openness will be treated with respect, and not be abused by those who are not yet in the tent. There is a role for journal reviewers and

editors and for societies like AuPS, in encouraging openness and in building trust. The papers we publish really consist of two parts, the experimental results themselves and the interpretations we build upon them. In my original field, the theory of how neuromuscular synapses are formed during development has changed several times over the past three decades due to persuasive new findings. Nevertheless, experimental results in earlier papers (when well documented), could then be reinterpreted in the light of the new evidence. We should archive our annotated raw data in an accessible way because it is the evidence base for the paper, but also because the data may well be useful to colleagues in the future (who, no doubt, will then cite your paper).

Poster session

Preclinical versus exploratory studies

It has become increasingly obvious that authors, reviewers and journals will, in future, need to distinguish between exploratory studies and formal preclinical testing of a drug. Many of us work at the exploratory end of biomedical research. We investigate physiological and pathophysiological mechanisms. A single paper might contain half a dozen ad hoc experiments, each helping to develop a hypothesis.

Exploratory studies➔Hypothesis➔ Preclinical trials?➔Clinical trial?➔ Better health?

By contrast, a preclinical trial in an animal model has more in common with a human clinical trial. It starts with an established formal hypothesis (that drug X will reduce a specific primary measure of disease severity in a particular animal model of a disease). A preclinical study should then be fully planned in advance, be adequately powered and address all the current procedural and reporting expectations (Kilkenny *et al.*, 2010; Landis *et al.*, 2012).

Conference coffee queue

Improved rigor also needed in exploratory studies

Those of us undertaking exploratory research should also heed the call for greater rigor. A formal preclinical trial requires considerable time, effort and cost, and offers very little joy to the researchers if the results turn out negative. If we hope our exploratory studies might one day lead to a new therapeutic intervention that improves health outcomes then our experiments, and the way they are reported, must be rigorous to justify the risk associated with a preclinical trial. We live in an era when the public expresses much less trust in the opinions of experts that hitherto. The public and politicians are increasingly aware of the failings of scientists (real or imagined). If we are to be viewed as something more than just another special interest group, we need to work individually and collectively to 'get our act together' scientifically. We need to represent intellectual discipline rather than just marketing.

At the symposium

Simon Gandevia is a world expert in human motor control studies, who has published on the problem of questionable reproducibility in his, technically-demanding, field of research. Simon will spell out why we need to take seriously the issue of rigor and reproducibility.

David Vaux is a renowned cellular immunologist and champion of the need for improved reporting standards. He will highlight what to look for in weak science.

Miranda Grounds is a world expert in animal models of neuromuscular disease with a long experience in working with international colleagues to improve reproducibility and reporting standards. She will explain how we might move forward.

Deficiencies of rigor and reproducibility in research won't be fixed with a single symposium or by the efforts of a few champions. We all need to become intellectually engaged this multi-faceted challenge. For this reason we have left time at the end of the symposium for a panel Q&A session. We hope this will be just the beginning.

William (Bill) Phillips

References

Kilkenny C, Browne WJ, Cuthill IC, Emerson M & Altman DG. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* **8,** e1000412.

Landis SC, Amara SG, Asadullah K, Austin CP, et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490,** 187-191.

Perrin S. (2014). Make mouse studies work. *Nature* **507,** 423-425.

Phillips WD, Christadoss P, Losen M, Punga A, Shigemoto K, Verschuuren J & Vincent A. (2015). Guidelines for pre-clinical animal and cellular models of MuSK-myasthenia gravis. *Experimental Neurology* **270,** 29-40.

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.

Steward O & Balice-Gordon RJ. (2014). Rigor or Mortis: Best Practices for Preclinical Research in Neuroscience. *Neuron* **84,** 572-581.

Vaux D. (2012). Research methods: Know when your numbers are significant. *Nature* **492,** 180–181.

Willmann R, De Luca A, Benatar M, Grounds M, Dubach J, Raymackers JM, Nagaraju K & Network. R-NN. (2012). Enhancing translation: guidelines for standard pre-clinical experiments in mdx mice. *Neuromuscul Disord* **22,** 43-49. .