National Animal Ethics Advisory Committee

Occasional Paper No 12

The blind leading the blind: animal facility staff and researchers working together to reduce bias in animal research

ISSN 1173-6763 (print)
ISSN 1173-6828 (online)
ISBN 978-1-77665-079-8 (print)
ISBN 978-1-77665-078-1 (online)

October 2015

Foreword

Ensuring the rigour with which the science behind research projects using animals is planned and implemented is an important part of animal ethics committee's process. Poor scientific method will inevitably result in skewed data, meaning that animals may have been used needlessly. While lay AEC members are not expected to have a scientific background, there are a number of relatively simple concepts that are key to good science – randomisation and blinding, for example. NAEAC members feel that this paper by Dr Jack-Rivers Auty explains such concepts, as well as other pitfalls of poor scientific process, in a readily accessible manner, and therefore recommend it to AEC members as the 12th in the Occasional Paper series. We are grateful to both the author and ANZCCART for allowing republication of the paper which was presented to the 2014 ANZCCART Conference and published in the proceedings.

Virginia Williams
Chair, NAEAC

NAEAC OCCASIONAL PAPER SERIES

1 *Underreporting of the Three Rs deployment that occurs during the planning of protocols that precedes their submission to animal ethics committees,* D J Mellor, J C Schofield and V M Williams, September 2008

2 *Regulation of animal use in research, testing and teaching in New Zealand – the black, the white and the grey,* L A Carsons, April 2009

3 *Regulation of animal use in research, testing and teaching: Comparison of New Zealand and European legislation,* N Cross, L A Carsons and A C D Bayvel, October 2009

4 *Compliance monitoring: The University of Auckland approach,* J Stewart, October 2009

5 Monitoring methods for animal ethics committees, D Morgan, October 2010

6 *Planning for refinement and reduction*, D Fry, R G Das, R Preziosi and M Hudson, January 2011

7 *Avoiding duplication of research involving animals*, D Morgan, March 2011

8 *Research on Vertebrate Pesticides and Traps: Do Wild Animals Benefit?* B Warburton and C O'Connor, August 2012

9 *Ensuring regulatory compliance in the use of animals in science in New Zealand – the review process*, August 2012

10 *How to improve housing conditions of laboratory animals: The possibilities of environmental refinement*, V Baumans and P Van Loo, February 2014

11 *Use of animals in the registration of veterinary medicine products in New Zealand,* K Booth, September 2015

The blind leading the blind: animal facility staff and researchers working together to reduce bias in animal research

Dr Jack Rivers-Auty, Dept of Pathology, University of Otago, Christchurch, New Zealand (jack.rivers-auty@manchester.ac.uk)

Reprinted with the permission of the author and the Australian and New Zealand Council for the Care of Animals in Research and Teaching.

Introduction

Anyone who watches the news frequently might be aware of a very wide and deep *fissure* that exists in science today. They will be aware of this *fissure* because every week it appears that a disease is cured by a new compound. This compound might one week be found on the skin of a frog that lives deep in the Amazonian rain forest, and then the next week perhaps it was extracted from an everyday food like dark chocolate or red wine. This astounding news article would hardly cause a modest arousal to a seasoned news watcher, as they would be acutely aware that these discoveries happen every week, yet for some reason the diseases mentioned continue to flourish unabated. This phenomenon is caused by the aforementioned deep *fissure* in science that divides preclinical and clinical research. The news article would refer to new research that has shown outstanding effects in an animal model of disease or perhaps cells grown in a petri dish (preclinical research). However, these outstanding effects almost never correspond to *any* therapeutic effect when investigated in clinical research on the human condition, and with every failure of these drugs to work in the *Homo sapiens*' version of the disease the *fissure* widens. In science circles this *fissure* is referred to as a *translational failure* and the conservative estimate for the ratio of drugs that make it across the *fissure,* to those that fall into its abyss, is approximately *one* in every *three hundred* (Mak et al. 2014; Thomas et al. 2014). There are probably many factors that are causing the translation failure rate to be so exceptionally high, but fortunately there are scientists that are investigating why preclinical science is failing. And what these investigations have uncovered is that although the scientific method has come a long way, we have far from perfected it. Perhaps minor methodological improvements could build a bridge over the *fissure*, or at least reduce the gap enough to improve the translational failure rate to a more acceptable level.

The evolution of the scientific method

For most of recorded history those who sought medical attention had a greater chance of dying than those who stayed at home (Fara 2009). Hospitals were putrid pits of disease at which the mostly well intentioned doctors administered treatments that were typically ineffective or harmful (Fara 2009). Leading medical physicians would prescribe mercury pills to induce vomiting, hysterectomies to reduce female hysteria (the word hysteria comes from the Latin word for uterus - *hystera*) or perhaps a dose of gold to cure jaundice; since jaundice causes the patient to turn gold surely gold would cure jaundice (Burgh 2009). These treatment techniques seem so unlikely to work that it is easy to think that these men were in some way unintelligent. However, for the most part this is not the case; these were brilliantly intellectual people that used their best judgment to care for the patient.

So what has changed? Well, many people would say that healthcare improved with each new discovery: aseptic techniques; penicillin; chemotherapy; statins, etc. We simply added to the pile of knowledge and this resulted in more treatments and better healthcare. However, I would

argue this misses the underlying processes of discovery and that the advancement of modern medicine should, instead, be attributed to the development and refinement of *science* itself. For example, a basic scientific approach to medicine would be collecting a group of people who are at the same stage of a particular disease and then breaking them up into smaller groups and giving each group a different treatment to assess which treatment works best. This seemingly fundamental concept of medical research became widely accepted only after it was reported by a Royal Navy surgeon named James Lind in 1747 (Singh & Ernst 2008; Burch 2009). At the time scurvy was reported to kill more naval sailors than armed conflict and as scurvy began to develop amongst the sailors of the *Salisbury* vessel, Lind collected the sufferers and split them up into groups of two which he matched for disease severity (Singh & Ernst 2008; Burch 2009). Then he gave each group a different treatment, all of which were based on his erroneous theory that acid should help the terrible condition (Singh & Ernst 2008; Burch 2009). Fortunately, one group was given citrus acid in the form of limes, which contain the only treatment for scurvy – vitamin C (Singh & Ernst 2008; Burch 2009). This group was on a miraculous path to recovery, one even returned to work, until they ran out of limes. From this study Lind concluded that a citrus syrup should be taken on future vessels (Singh & Ernst 2008; Burch 2009). Unfortunately, the process of making the syrup involved boiling which greatly reduces the levels of active vitamin C (Singh & Ernst 2008; Burch 2009). So although Lind had seemly missed the key treatment of scurvy, he was the first to describe the fundamentals of a clinical trial.

So it is wrong to say that medicine has improved as we discovered new and better therapies, because for most of recorded history the very techniques of investigation were not refined enough to make these discoveries. We simply could not have developed statins (a cholesterol-lowering drug) to treat the development of heart disease before the scientific method was refined enough to detect the unobvious impacts of the drug. Of all the refinements to the scientific method, randomisation and blinding have been particularly important to the development of new therapies for disease. Yet despite the history of these techniques, which clearly illustrates their importance to the scientific method, they are not commonly used in preclinical research (Sena et al. 2014).

One problem clinical research faces is how to divide patients into the various treatment groups. Historically, the allocation would be performed by the researcher in a *subjective* manner. Which begs the questions, did the researcher, subconsciously or consciously, place the sicker patients in one group and only give the treatment they believed should work to the patients that were likely to survive anyway? This subjective and undefined method of treatment allocation was a breeding ground for potential biases. What was needed was an *objective* method that divides the patients as opposed to a *subjective* method.

Randomisation is the best example of an objective method and could be as simple as flipping a coin to determine whether a patient receives treatment A or treatment B. One of the earliest reported clinical trials that used an *objective* method for treatment allocation was by the young Scottish surgeon Dr Hamilton in 1809 (Singh & Ernst 2008). Dr Hamilton and his colleague Dr Anderson believed that bloodletting was not an effective treatment for any aliment, while their older unnamed colleague believed, as most doctors did at the time, draining between 500 ml and 2.5 L of blood from a patient is an effective treatment for many aliments such as fever or inflammation (Singh & Ernst 2008). As patients came into the clinic they were systematically allocated to be treated by Dr Hamilton, Dr. Anderson or the unnamed doctor. It was discussed that the treatments should be standard between all doctors except Dr Hamilton and Dr Anderson would not perform bloodletting (Singh & Ernst 2008). After each doctor had seen 122 patients, the survival rates were compared. Dr Hamilton had lost four and Dr Anderson

had lost two, which were very good results for doctors at the time. The senior doctor's success rate explains why this unnamed doctor was unnamed, for he had lost 35 of his 122 patients (Singh & Ernst 2008). And with that, there was now robust evidence against a procedure that had been utilised by doctors since before Hippocrates stated that doctors should "first do no harm" in the 5th century BC. This demonstrates the importance of the development and refinement of science. Bloodletting, a medical practice that was performed and observed for over 2000 years, was detrimental to the patients that received the treatment, often to a lethal degree. Yet not until the scientific method had been refined could this grossly pathological practice be seen for what it was and removed from use. However, due to doctors not willing to acknowledge that their profession had been killing people, Dr Hamilton's research was largely ignored causing bloodletting to last another century until it finally faded from medical practice (Fara 2009).

Another crucial advancement in experimentation was the notion of blinding. This is where the assessor and administer of a treatment is unaware what treatment is being given. Blinding was largely devised to account for the *placebo* effect. This effect is where the patient benefits merely because they believe they are receiving an effective treatment and it is caused by both a change in the physiology of the patient and a change in the perception of the ailment. The first description of the placebo effect has a very interesting history. Any new discovery is met with theories about human health. After X-ray machines were first invented theories were proposed that exposure to X-rays invigorated the body and those who could afford it may delight in an energising daily X-ray (Sansare et al. 2011). Cell phone towers on the other hand, were initially thought to cause various diseases, including cancer and migraines, and erecting a tower near a school or kindergarten was often met with public protest (Dolan and Rowley 2009).

Nowadays, X-rays are well known to cause cancer and the only response to the erection of a new cell phone tower is gratitude for the faster Facebook updates. So in the 1780s, when Galvani Lugi used twitching frog legs to suggest that the body uses electrical fluid to activate muscle activity, it is not surprising that some entrepreneurial fellow decided that this electrical fluid must be involved in human health (Singh & Ernst 2008). The American physician Elisha Perkins proposed that noxious electrical fluid must build up in painful and inflamed areas (Singh & Ernst 2008). He developed two metal rods made from exotic materials which he named tractors; these tractors could be passed over the problematic area and would drain it of the agitating electrical evil (Singh & Ernst 2008). Due to the exotic material the rods *must* be made from, the cost of this equipment was 5 guineas, which at the time was around half the annual wage of a laborer (Singh & Ernst 2008). This business was very profitable until a skeptical and frugal British physician by the name of John Haygarth decided to investigate cheaper alternative metals that could be used to make the Perkin's tractors (Haygarth 1800; Singh & Ernst 2008). As Haygarth researched different materials he found something very odd, not only did cheap metals work as well as the exotic originals (which were actually made of the relatively cheap metals brass and steel), but non-conducting materials like wood had an equivalent therapeutic effect; in fact anything he waved over the inflicted area with convincing conviction for the required 20 minutes appeared to alleviate the patients' symptoms (Haygarth 1800). Haygarth was fascinated *"to a degree which has never been suspected, what powerful influence upon diseases is produced by mere imagination"* (Haygarth 1800; Singh & Ernst 2008). He also noted that this probably explains why treatments worked better in the hands of more famous and expensive physicians. This was a major discovery for the scientific method; the simple comparison of the patient before and after treatment had a fundamental flaw. The patients' mere imagination would corrupt the results, meaning any before and after comparison was quite likely to be completely erroneous. The problem of the patients' imagination could be solved by comparing any treatment to a

dummy treatment (placebo), just like Haygarth's wooden tractors. However, this practice did not become widely used until 150 years after Haygarth and his wooden tractors (Singh & Ernst 2008).

Interestingly, the placebo effect can work not only on the patient but also on the physician or the scientist as well. One famous example of this was the discovery that homeopathy works at the cellular level. Now to those who are unaware of the details of homeopathy this claim might seem quite normal, but to most scientists homeopathy is an amusing way to teach undergraduate students about dilutions. The principles of homeopathy state that a substance which causes symptoms will cure those symptoms in *extreme* dilutions; so caffeine keeps you awake and therefore will cure insomnia when diluted suitably (Singh & Ernst 2008). Now a normal dilution in homeopathy would be in homeopathic jargon a 30C dilution. C is the Roman numeral for 100 and so 30C indicates the solution is diluted 1 in 100, thirty times. On top of this extreme dilution, the normal dose given to a patient is one hundredth of a millilitre, which is less than one drop. To really explain what this means, imagine taking half a teaspoon of pure caffeine, dissolving it in a ball of water the size of our solar systems (using Pluto's orbit), then taking less than one drop from the ball of water and administering it for the treatment of insomnia (Singh & Ernst 2008). The patient would be more likely to win lotto three times in a row than receive a single molecule of caffeine in their treatment.

If not for the fact that in 2013 the homeopathic industry's estimated worth was $6.4 billion in the United States alone, homeopathy would be quite a humorous subject to most scientists (Singh & Ernst 2008). However, in 1988 a paper was published by Dr Benveniste and his laboratory group in the prestigious journal *Nature* which provided inexplicable evidence for homeopathy (Dayenas et al. 1988). The subjects of the paper were not patients reporting on subjective feelings or overall well-being, but cells under a microscope. The diluted solution used in Dr Benveniste's study was effective at eliciting a response in the cells at a 120C dilution, which is more than 1 billion billion billion billion billion billion times more diluted than the solar system caffeine example given above (Dayenas et al. 1988). *Nature* published the paper; however, as no phenomenon in science could explain the reported results, scientific observers selected by the *Nature* journal went to the laboratory that produced the paper to observe the experiment for themselves (Dayenas et al. 1988; Singh & Ernst 2008). The experiment involved applying a stimulant solution to the cells and observing the cells to see if they "degranulate", which is a process where the cells eject their signaling molecules reservoirs. Dr Benveniste's laboratory group repeated the experiment for the observers and they got the same result; however, the person looking down the microscope at the cells and counting the number of degranulated cells was aware of what solution had been applied (Singh & Ernst 2008). The *Nature* observers asked them to repeat the experiment whilst blinded to cells which had been given normal saline and cells which had been given the extreme dilutions of the stimulant (which, given the dilution, was by all probability also normal saline) (Singh & Ernst 2008). Once blinded the effect disappeared, showing that it was the scientist observing the cells that was affected by the treatment and not the cells themselves. This isn't to say that Dr Benveniste and his laboratory were actively skewing the results; perhaps just minor differences in their interpretation of the cells degranulated state caused an unintentional and detectable level of bias. After the *Nature* journal reviewers published their observations, all scientists could then take a sigh of relief as this inexplicable result was now explicable. But, should they breathe a sigh of relief given that a vast majority of all preclinical scientific experiments at that time were performed in an unblinded manner? What other scientific "discoveries" were actually just results of a subconscious change in interpretation or implementation of a method due to the all

too human desire to produce interesting results which confirm the proposed hypothesis?

After Dr Hamilton's research clearly showing bloodletting was killing patients the technique took around 100 years to disappear from common practice. Similarly, the use of a placebo control only became common 150 years after the work of Dr Haygarth on the Perkin's tractors, as now again research techniques are in a period of delay. Dr Benveniste's work clearly shows that preclinical scientists must be blinded to the treatment groups and Dr Hamilton's research shows the unarguable robustness of randomised research and yet the practices of blinding and randomisation have not become common practice in preclinical research. In 2014, an analysis of preclinical animal research in the field of treating diseases of the brain found that only 20 percent of the research reported blinding of the assessor of the brain damage and 15 percent reported randomly allocating the animals into treatment groups (Sena et al. 2014). The percentage of research that is done in what many would describe as a rigorous manner appears dismal; however, medicine is improving; new drugs, new therapies, new diagnostic techniques and new surgeries are being developed every year. This begs the question, are these biases actually affecting research output? Do we need to change? Isn't preclinical animal research working anyway?

Does preclinical animal research work?

Animal research is incredibly useful and certainly does work. Surgical techniques are developed and practised on animals, which are then used in the clinic with huge translational success. Toxicology research in animals is very good at predicting which compounds are toxic to humans; generally compounds which are safe at high doses in several mammalian families are very likely to be safe for humans. But the topic of this article is drug development, so does preclinical animal research work for drug development? To answer this question the definition of "work" must be established and to do that we must look at the aim of animal models of disease. Preclinical research investigating novel treatments of disease induces pathologies in animals using a range of techniques. As the human disease cannot be replicated exactly, the aim is to produce a model that replicates the real condition as close as possible with the overall goal of *accurately predicting if a new therapy will work in the human disease.* Therefore, one way of answering the question "does preclinical animal research work?" is to answer the question "does preclinical animal research accurately predict if a new therapy will work in the human disease?" To this question we have some answers.

Let us look at the disease with the most preclinical research – cancer. Animal research in cancer normally involves growing cancer cells in a petri dish and then inserting the cells under the skin of a mouse and observing the growth rates of the cells. Then a treatment that had been shown to kill cancer cells in a petri dish experiment would now be given to the mice at different doses to assess if the drug can get to the cancer cells and kill them in the new setting of an animal. Now imagine a laboratory that is very successful and discovers 500 new compounds that appear to be effective in treating a range of cancers in these mouse models. Other laboratories will read about the compounds and perform similar experiments and report their results. The reported evidence will build until a pharmaceutical company (or public entity) deems the evidence to be worthy of investing. The company will fund a phase one clinical trial. Of the 500 new compounds originally discovered only 150 will make it to this stage of research (Mak et al. 2014; Thomas et al. 2014). A phase one clinical trial will assess what the drug does in healthy males: Where does it go? What does it do? What does it turn into? How fast is it excreted? Based on these results the company may decide to go to a phase two trial, which is a moderate sized trial on sufferers of the disease. Less than 13 of the original 500 compounds will make it on to this stage (Mak et al. 2014; Thomas et al. 2014). The drugs that appear to have some therapeutic effects will go on to

a large-scale phase three trial of disease sufferers. Of the 13 drugs in phase two trials, 5 will be tested in phase three trials and of these maybe 1 or 2 will receive food and drug administration (FDA) approval and be used in the clinic (Mak et al. 2014; Thomas et al. 2014) So from 500 original drugs that were therapeutic in preclinical animal models, 1 or 2 are found to be therapeutic in the clinic (Mak et al. 2014; Thomas et al. 2014). From this it seems that animal models fail to accurately predict what occurs in the clinic. Keep in mind the fact that this failure rate from this example is what you get when you boil the effects of the drugs down to binary data, *therapeutic* or *not*; when the effect sizes are compared the predictive power of the animal model drops even further. Even if a drug is found to be successful in the clinic setting, it is often far less effective than the results reported in the animal studies.

From this it appears the preclinical animals models are doing a poor job of predicting what would happen in the human condition but is this because of blinding and randomisation? What about other obvious shortcomings of animal research such as using mice that are not human? Mice are smaller and have different metabolisms to *Homo sapiens*. Is it not possible that the 300 failures for every 1 success is the expected failure rate given the biological differences between the species they are tested on? One answer to this is that in mice: caffeine increases activity; methamphetamine is addictive, marijuana effects memory and appetite; Prosac™ reduces anxiety; antibiotics selectively kill bacteria and not the cells of the mouse; cocaine is addictive; nicotine is addictive; aspirin reduces pain and swelling; sunscreen prevents UV damage; agent orange causes deformities; and the list goes on. The number of drugs that have similar effects in humans and mice is much larger than those which act substantially different. Another answer to what is causing the 300 to 1 failure rate lies in the history of an antioxidant drug named NXY059.

Case study: NXY059

According to the current paradigm of human physiology, our bodies are constantly producing free radicals (oxidants) which are highly reactive and potentially damaging to the cell. To counteract this, our cells are also producing antioxidants which safely react with free radicals diffusing their damaging properties. In healthy tissues, antioxidants and free radicals are in balance and both play an integral role in normal physiology. However, in many disease states these become imbalanced and free radicals can reach high concentrations within the tissues. Free radicals are seen as dangerous molecules as they can react with components of the cell and cause them to malfunction; DNA can mutate, proteins can change shape and membranes can become leaky. The health and nutraceutical industry has caught on to this idea and that is why you will see "rich in antioxidants" everywhere as you peruse through the supermarket or pharmacy. Following stroke, there is a huge rise in the concentration of free radicals within the brain and this was/is believed to contribute to the brain damage caused following a stroke. Therefore, preclinical research went into developing antioxidants that were potent and could penetrate into the brain to the site of injury; this led to the development of the very promising antioxidant NXY059 (Shuaib et al. 2007). Following the first neuroprotective animal study in 1999, animal research piled up with a vast majority of studies reporting dramatic therapeutic effects (Macleod et al. 2008). Some studies reported almost complete protection from brain damage in animal models of stroke (Mak et al. 2014; Thomas et al. 2014). A private company decided it was worth investing and designed a moderately sized phase 2 trial (Shuaib et al. 2007). This trial reported promising yet inconclusive results and so the company decided to pursue it further and organised another phase 2 trial with more patients (Shuaib et al. 2007). This time over 3000 subjects were enlisted in the study which became the largest clinical stroke trial in history (Shuaib et al. 2007). The company never released the true cost of the studies but

it is estimated to have cost well over $100 million USD.

When the results came in from the 362 centres from 31 different countries, there was no detectable effect of NXY059 on stroke (Shuaib et al. 2007). It was a resounding failure and the search for why it had failed began. Although there was probably more than one cause of this failure, some very clever preclinical researchers produced a meta-analysis (a study of studies), that clearly implicated failures of preclinical experimental technique resulting in NXY059 being found to be far more therapeutic than it may actually be (Macleod et al. 2008). McLeod and his colleagues' meta-analysis compared the reported effectiveness of NXY059 in multiple animal studies with what steps the studies used to reduce bias. What they found was very convincing; studies which reported the use of randomisation found the NXY059 was 2-fold *less* effective than studies which did *not* report randomisation (Macleod et al. 2008). Furthermore, producing a stroke in an animal requires surgery and what McLeod's study found was that if the surgeon was aware of which treatment the animal was going to receive, NXY059 was ~2-fold more effective than the studies where the surgeons were blinded to the treatment group (Macleod et al. 2008). Somehow the severity of the stroke was subconsciously manipulated by the surgeon!

This study went further and found that perhaps it is not that mice and rats are inappropriate animals to predict human conditions but that we are using the wrong mice and rats. A vast majority of rodent stroke studies use healthy young animals, when strokes normally occur in old and hypertensive humans (Ford 2008). The meta-analysis found that studies which used hypertensive animals reported NXY059 to be 2.2-fold *less* effective compared with studies that used young healthy animals (Macleod et al. 2008). Collectively, this meta-analysis showed that if animal studies had used randomisation, blinding the assessor of brain damage, blinding the surgeon of the treatment groups and old hypertensive rodents, NXY059 would have been found to be substantially less effective than what was previously reported (Macleod et al. 2008). This would have been considerably less appealing to the pharmaceutical company and perhaps the human trial would never have been done. Despite the failure of NXY059 and despite the beautiful work by McLeod and his colleagues, a recent meta-analysis found that there has been no increase in the use of randomisation, blinding or hypertensive animals since the NXY059 failure (Philip et al. 2009). Sadly, it seems that the lag between discovery and acceptances, which we saw with Dr Hamilton and Dr Haygarth's work, seems to be unavoidable, even in this information age we live in today.

Blinding – more complicated than you think

Most pharmacological animal research requires two steps which must be blinded. First, the induction of the disease must be blinded to what treatment the animals will receive and then after the treatment the scientist must be blinded again while assessing the severity of the disease. The first blinding can occur through appropriate randomisation. First the animals are randomly assigned to either have the disease induced or not. Then simply randomly allocating the diseased animals to the treatment groups after the induction of the disease will blind the scientist performing the procedure to what treatment the animal is about to receive. This could be as easy as flipping a coin to see if they receive the drug or the placebo control. The next stage in blinding should be done after all the surgeries and drug treatments have been completed. This can be done by a colleague entering the animal room and replacing the animal information cards with a card with just a letter on it and recording which letter corresponds to which animal.

There are examples where things are not that simple and each situation must be worked through to develop an appropriate method. One of the most difficult situations is the use of genetically engineered (GE) mice which are visibly different to the normal control mice, such as the

hairless GE mouse that is lacking the vitamin D receptor. How can you perform a surgery or an assessment blinded if merely looking at the mice unblinds you? Well this has a less simple solution and requires systems set up in university departments and private organisations to provide a solution. This could be that there are qualified staff in the animal care facilities who are capable of inducing the disease and giving the drugs without being aware of the proposed hypothesis. Bias can be reduced by acting independently and in a systematic fashion. Given the 2-fold increase in the effectiveness of NXY059 in animal studies that did not blind, it seems that it is not viable to simply say sometimes blinding is too hard.

We must at every stage try and reduce bias in research as biased results are simply unethical. If the results of the work cannot be applied clinically due to bias, which was at least a contributing factor in NXY059 case, how can it be justified ethically? Biased research causes unnecessary animal suffering, unnecessary expenditures by governments and private organisations on preclinical and clinical research, unnecessary human experimentation and it undermines the integrity of science. With continued failures of animal research, there is a risk it will lose public support and, with that, governmental support both financially and in the policies made. So what can be done?

What can we do about bias?

Scientific rigour has always been monitored by the trusted and respected "peer review system". This is where a study is written up and sent to a journal, then the journal editor sends this on to several experts in the field who voluntarily review the research. Documents are drawn up by the reviewers and are most often full of suggested new experiments and revisions, or the reviewers could outright reject the research due to poor design. Once the study is written up in a way that is acceptable to the reviewers it is published by the journal. This system is full of problems that would take at least a whole new article to address, so this article will only address one: money. The journals have to make money and they normally do this through selling subscriptions to read the research published in the journal. Therefore, universities have to pay huge subscription fees to ensure all good research that is published is available to its students and researchers. This also means that the general public will find it very hard to access journal articles, yet they will find it easy to access blogs and "information sites" on the internet (and we wonder why unfounded cures, fad diets, homeopathy, iridology, etc, persist in the age of modern science). This model essentially means that private journals appear to own knowledge and can sell it on the free market. But this also means that the journal is looking to have a quality *product* to ensure universities will purchase subscriptions. Because of this, the journal would rather reject low-quality research in order to maintain the standards of the journal.

Nevertheless, the fact that these journals essentially own and sell knowledge is seen by some as a crime against freedom of information. This perspective resulted in the proliferation of a new model – the open access journal. These journals still needed money to operate and so would charge the scientists for publishing their work in the journal (often around $3000). The journal would then allow the public to view the research for free. But the problem with this model is that there is less motive to reject low-quality research. As they are not selling subscriptions the *product* quality is not as important as *quantity* for the business to be profitable. There is no governing body controlling creation or operation of these journals and, as the journalist and biologist John Bohannon found, this combination of monetary motive and science does not have good results. Bohannon organised a sting operation where he used a computer program to generate more than 250 papers which were completely fabricated, with made up authors from fictional universities (Bohannon 2013). The papers were designed to have fundamental flaws in them including a dose-dependent effect that was non-existent, no relevant controls and blatant

failures in basic experimental design (Bohannon 2013). He submitted these papers to open access journals and was accepted and ready to publish in 157 of them and rejected by 98 journals (Bohannon 2013). Bohannon found that the open access journal model had generated a system where scientists could essentially pay to have their work published regardless of the quality of the research. It is clear from Bohannon's sting operation that scientific rigour cannot be left to the ungoverned "peer review" publication systems, so where can the standards be set?

There are two points before publication in which the experimental design can be assessed. The first point is the funding body. Most of science relies on research grants to fund their work. Applications are drawn up and submitted where they are reviewed by respected scientists and a few are selected to be funded. However, these applications are not full of experimental detail; this is because a scientist from one field would find it hard to review the methods of hundreds of applications from other fields of research when they are not familiar with the methods of that field. As a result minor details like blinding and randomisation are left out along with the technical details of the methodology. Research applications are about selling an idea, so they are filled with how important, novel and publishable the research would be, not the finicky details of methodology. So while new systems could be introduced to assess experimental rigour at this stage, it currently seems like something which would be seen as too hard.

The second review process all animal research must go through is the ethical approval process. Now first it should be stated that poor quality research *has* profound implication in the ethical use of animals. What justification can there be for the use of the animals that suffered in researching the effect of NXY059 on stroke? At last count there have been over 1300 compounds that have been found to be effective in animal models of stroke and yet only one of these compounds is currently used in the clinic. The quality and reliability of animal research is integral to it being ethical. Now unlike the research funding applications, the animal ethics applications are long and probing of the technical details of the methods. The Animal Ethic Committees (AEC) have to know every interaction with the animal to ensure it is ethical. Therefore, I would argue that the animal ethical review process is the best point in which the methods of blinding and randomisation can be assessed. It should be at this point that universities and research facilities can ensure that the research performed is of a high standard and therefore has the greatest applicability to human disease and consequently is the most ethical use of animals in research. Currently in New Zealand a human ethics application must be filed to perform a clinical study and this is reviewed by a committee. The human ethics committees require information on randomisation and blinding. Therefore, there is a working precedent of evaluating experimental rigour which could be followed in the animal ethics application and review process. I would argue that we owe it to the animals to consider this minor change in the review process.

Conclusion

The number of compounds which cross the *translational fissure* between preclinical and clinical success is dismal. Excellent research has illustrated that this translational failure is, at least in part, caused by a lack of experimental rigour in animal research. The basic steps of blinding and randomisation are not commonly performed despite historical and contemporary research that clearly demonstrates their importance. Part of the problem is the lack of regulation in the performance and reporting of scientific research. There is no governing body that is ensuring that the standards of preclinical research are such that the results are unlikely to be affected by biases. One point of regulation that has the ability and the desire to perform this regulatory role is the AEC. These committees must approve animal research and currently assess the importance of the research relative to the suffering the animals will experience during the study. Here I argue

that scientific rigour is an ethical issue and should also be assessed by animal ethics committees. After all, how can animal suffering be justified if the results are of little relevance to the human condition?

References

1. Bohannon, J. 2013: Who's afraid of peer review? *Science 342*: 60-65.
2. Burch, D. 2009: Taking the medicine. Great Britain: Chatto & Windus.
3. Dayenas, E.; Beauvais, F.; Amara, J.; Oberbaum, M.; Robinzon, B.; Miadonna, A.; Tedeschit, A.; Pomeranz, B.; Fortner, P.; Belon, P.; Sainte-Laudy, J.; Poitevin, B.; Beneveniste, J. 1988: Human basophil degranulation triggered by very dilute antiserum against IgE. *Nature 333*: 816-818.
4. Dolan, M., and Rowley, J. 2009. The Precautionary Principle in the Context of Mobile Phone and Base Station Radiofrequency Exposures. *Environmental Health Perspectives 117:* 1329-1332.
5. Fara, P. 2009: Science: A four thousand year history. Great Clarendon Street, Oxford, England: Oxford University Press.
6. Ford, G. A. 2008: Clinical pharmacological issues in the development of acute stroke therapies. *British Journal of Pharmacology 153*: S112-S119.
7. Haygarth, J. 1800: Of the imagination, as a cause and as a cure of disorders of the body: exemplified by fictitious tractors, and epidemical convulsions. Bath: R. Crutwell.
8. Macleod, M. R.; van der Worp, H. B.; Sena, E. S.; Howells, D. W.; Dirnagl, U.; Donnan, G. A. 2008: Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke 39*: 2824-2829.
9. Mak, I. W.; Evaniew, N.; Ghert, M. 2014: Lost in translation: animal models and clinical trials in cancer treatment. *American Journal of Translational Research 6*: 114-118.
10. Philip, M.; Benatar, M.; Fisher, M.; Savitz, S. I. 2009: Methodological quality of animal studies of neuroprotective agents currently in phase II/III acute ischemic stroke trials. *Stroke 40*: 577-581.
11. Sansare, K.; Khanna, V.; Karjodkar, F. 2011: Early victims of X-rays: a tribute and current percetpion. *Dentomaxilofacial Radiology 40*: 123-125.
12. Sena, E. S.; Currie, G. L.; McCann, S. K.; Macleod, M. R.; Howells, D. W. 2014: Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *Journal of Cerebral Blood Flow & Metabolism (2014) 34,* 737–742.
13. Singh, S.; Ernst, E. 2008: Trick or Treatment. 61-63 Uxbridge Road, London, England: Bantam Press.
14. Shuaib, A.; Lees, K. R.; Lyden, P.; Grotta, J.; Davalos, A, Davis SM, Diener H, Ashwood T, Wasiewski WW, Emeribe U, Investigators SIT (2007). NXY-059 for the treatment of acute ischemic stroke. *New England Journal of Medicine 357*: 562-571.
15. Thomas, D. W.; Craighead, J. L.; Economides, C.; Rosenthal, J. 2014: Clinical development success rates for investigational drugs. *Nature Biotechnology 32*. 40–51

National Animal Ethics
Advisory Committee

New Zealand Government