Systematic Reviews of Animal Experiments Demonstrate Poor Human Clinical and Toxicological Utility

Andrew Knight

Animal Consultants International, London, UK

Summary — The assumption that animal models are reasonably predictive of human outcomes provides the basis for their widespread use in toxicity testing and in biomedical research aimed at developing cures for human diseases. To investigate the validity of this assumption, the comprehensive *Scopus* biomedical bibliographic databases were searched for published systematic reviews of the human clinical or toxicological utility of animal experiments. In 20 reviews in which clinical utility was examined, the authors concluded that animal models were either significantly useful in contributing to the development of clinical interventions, or were substantially consistent with clinical outcomes, in only two cases, one of which was contentious. These included reviews of the clinical utility of experiments expected by ethics committees to lead to medical advances, of highly-cited experiments published in major journals, and of chimpanzee experiments — those involving the species considered most likely to be predictive of human outcomes. Seven additional reviews failed to clearly demonstrate utility in predicting human toxicological outcomes, such as carcinogenicity and teratogenicity. Consequently, animal data may not generally be assumed to be substantially useful for these purposes. Possible causes include interspecies differences, the distortion of outcomes arising from experimental environments and protocols, and the poor methodological quality of many animal experiments, which was evident in at least 11 reviews. No reviews existed in which the majority of animal experiments were of good methodological quality. Whilst the effects of some of these problems might be minimised with concerted effort (given their widespread prevalence), the limitations resulting from interspecies differences are likely to be technically and theoretically impossible to overcome. Non-animal models are generally required to pass formal scientific validation prior to their regulatory acceptance. In contrast, animal models are simply assumed to be predictive of human outcomes. These results demonstrate the invalidity of such assumptions. The consistent application of formal validation studies to all test models is clearly warranted, regardless of their animal, non-animal, historical, contemporary or possible future status. Likely benefits would include, the greater selection of models truly predictive of human outcomes, increased safety of people exposed to chemicals that have passed toxicity tests, increased efficiency during the development of human pharmaceuticals and other therapeutic interventions, and decreased wastage of animal, personnel and financial resources. The poor human clinical and toxicological utility of most animal models for which data exists, in conjunction with their generally substantial animal welfare and economic costs, justify a ban on animal models lacking scientific data clearly establishing their human predictivity or utility.

Key words: *animal experiment, animal study, clinical trial, human outcome, systematic review.*

Address for correspondence: *Andrew Knight, Animal Consultants International, 91 Vanbrugh Court, Wincott Street, London SE11 4NR, UK.*
E-mail: *info@AnimalConsultants.org*

Introduction

Trends in laboratory animal use

Standards for the reporting of laboratory animal use vary internationally, with many countries failing to record or publicise statistics on animal use at all. Of those that do, most record only live animal use, and fail to record the substantial numbers of animals that may be killed prior to certain procedures, such as dissection or the collection of organs, tissues or cells. Hence, making realistic annual estimates of animal use within biomedical research and toxicity testing is difficult. Despite these limitations, it remains clear from consideration of the European Union (EU) and United States alone, that many millions of animals are used worldwide, and that certain trends are resulting in an increase in laboratory animal use.

EU

European Commission (EC) statistics on laboratory animal use in 25 EU Member States, revealed that 12,117,583 animals were used in 2005, the latest reporting period (except for France, which provided figures for 2004). The majority of these were mice (53.1%), rats (19.3%), cold-blooded animals (15.1%, consisting of fish [primarily], amphibians and reptiles), and birds (5.4%). As in previous years, France, Germany and the UK reported the greatest animal use (1).

United States

In the USA, laboratory animal use is federally regulated by the *Animal Welfare Act 1966* (amended in 1985), which excludes laboratory-bred mice and rats, as well as non-mammals, from consideration or protection (2, 3), despite the fact that mice and rats comprise the overwhelming majority of all laboratory subjects. This impedes the accurate estimation of laboratory animal use in the USA. For example, although 1,012,713 regulated animals were used in the Fiscal Year 2006 (4), the latest reporting period, Carbone (5) estimated that in excess of 100 million mice are used annually. This represents a dramatic increase from the 17–22 million vertebrates used in the mid-1980s (6).

Genetically-modified animal use

In recent years, the previous steady decreases in laboratory animal use have been reversed, in some countries, mostly as a result of dramatic increases in the use of genetically-modified (GM) animals. The production of these GM animals requires substantial breeding, which serves to further increase the numbers of animals used. Within the UK, for example, a steady and significant reduction since 1976 stabilised during the early 1990s, and then reversed. 3,012,032 procedures on living, regulated animals (vertebrates and one species of octopus, *Octopus vulgaris*) were conducted in 2006, the highest number for around 15 years (7). Greater breeding and use of GM animals have contributed to these increasing numbers (8, 9). In 1995, GM animals were used in 8% of regulated procedures. In 2004, the total was 32%, and in 2006 it was 34% (1,035,343 regulated procedures; 7). Increased GM animal use has also been recorded in Germany (10) and Switzerland (11), where total animal use is also increasing (11, 12).

Chemical testing programmes

Recently-initiated, large-scale chemical testing programmes are also important drivers of the recent and probable substantial future increases in laboratory animal use (13, 14). These programmes are intended to rectify existing knowledge gaps with regard to the toxicity of chemicals that are produced or imported into the EU and the USA in particularly high quantities (or that otherwise give rise to special concerns), and are likely to result in the use of unprecedented numbers of animals in toxicity testing. Included are three programmes initiated by the US government, and managed by the Environmental Protection Agency (EPA) since 1998: the High Production Volume (HPV) Challenge Program, the Endocrine Disruptor Screening Program, and the Voluntary Children's Chemical Evaluation Program. The 2003 EC proposal for the Registration, Evaluation and Authorisation of Chemicals (REACH), similarly aims to assess the toxicity of chemicals produced or imported in high quantities (15–20). It is reported that the HPV programme, for example, has already subjected over 150,000 animals to chemical tests (21).

Claims supporting laboratory animal use

Biomedical research using laboratory animals is highly controversial. Advocates frequently claim that such research is vital for preventing, curing or alleviating human diseases (e.g. 22, 23), that the greatest achievements of medicine have been possible only due to the use of animals (e.g. 24), and that the complexity of humans requires nothing less than the complexity of laboratory animals to serve as an effective model during biomedical investigations (e.g. 25). They even claim that medical progress would be *"severely maimed by prohibition or severe curtailing of animal experiments,"* and that *"catastrophic consequences would ensue"* (26).

However, such claims are hotly contested (e.g. 27), and the right of humans to experiment on animals has also been strongly contested philosophically (e.g. 28, 29). A growing body of empirical evidence also casts doubt upon the scientific utility of animals as experimental models of humans.

Clinical utility of animal models: case studies

Within the field of pharmaceutical development, case studies exemplifying differing human and animal outcomes — sometimes with severe adverse consequences for human patients — are sufficiently numerous to fill entire book chapters (e.g. 30, 31).

A recent notorious example was TGN1412 (also known as CD28-SuperMAB), a fully-humanised monoclonal antibody (i.e. a product developed in a non-human species and protein-engineered to possess specifically-human characteristics) that was undergoing development for the treatment of inflammatory conditions, such as leukaemia and rheumatoid arthritis (32, 33). During a Phase I clinical trial in the UK in 2006, TGN1412 caused severe adverse reactions, culminating in organ failure requiring intensive care, in all six volunteers given the drug, with one volunteer suffering permanent damage. These effects occurred despite the administration of an expected sub-clinical dose of 0.1 mg/kg — 500 times lower than the 50 mg/kg dose found not to cause adverse effects in cynomolgus monkeys. Tests on rhesus macaques, rats and mice also failed to reveal adverse effects (34, 35).

Another recent notorious example was the arthritis drug, Vioxx, which appeared to be safe, and even

beneficial to the heart, in animal studies. However, Vioxx was withdrawn from the global market in 2004, after causing as many as 140,000 heart attacks and strokes, and over 60,000 deaths, in the USA alone (36).

Since their commercial introduction in the early 1980s, non-steroidal anti-inflammatory drugs (NSAIDs) have also had a problematic clinical history. Although apparently safe in year-long studies in rhesus monkeys, benoxaprofen (Oraflex) produced thousands of serious adverse reactions in humans, including dozens of deaths, within three months of its initial marketing (37). Fenclofenac (Flenac) revealed no toxicity in ten animal species, yet produced severe liver toxicity in humans, and was subsequently withdrawn (38). Similar fates befell some other NSAIDs, including zomepirac (Zomax; 39), bromfenac (Duract; 40), and phenylbutazone (Butazolidin; 41), which produced adverse human effects undetected in animal studies.

Numerous other pharmaceuticals have also been marketed after passing limited clinical trials and more rigorous animal testing, only to subsequently be found to cause serious side-effects or death in human patients. Examples include various antibiotics (e.g. chloramphenicol, clindamycin, temafloxacin), antidepressants (e.g. nomifensine), antivirals (e.g. idoxuridine), cardiovascular medications (e.g. amrinone, cerivastatin, mibefradil, ticrynafen), and many others (e.g. 30, 42–44).

Although 92% of new drugs that pass preclinical testing, which routinely includes animal tests, fail to reach the market because of safety or efficacy failures in human clinical trials (45), adverse drug reactions detected after drugs have been approved for clinical use, nevertheless remain common. They are, indeed, sufficiently common to have been recently recorded as the 4th–6th leading cause of death in US hospitals (based on a 95% confidence interval; 46), a rate considered by investigators to be "*extremely high*".

There are also cases of safe and efficacious human pharmaceuticals that would not pass rigorous animal testing, because of severe or lethal toxicity in some laboratory animal species. Notable examples include, penicillin (e.g. 47), paracetamol (acetaminophen; e.g. 48), and aspirin (acetylsalicylic acid; e.g. 49). More rigorous animal testing may well have delayed or prevented the use of these highly beneficial drugs in human patients.

The large number of examples of apparent differences between outcomes in laboratory animals and in human patients may be the result of several factors. Flaws may occur during the pharmaceutical development and testing process, in which the design, conduct or interpretation of experiments may fail to adequately highlight the risks to human patients. Such flaws are more likely in animal studies than in human clinical trials, because the experimental quality of the former are usually significantly lower (see *Results and Discussion*). True discordance in results may also arise from interspecies differences.

Finally, the limited predictivity for wider human outcomes of human clinical trials may result from their focus on small groups of healthy young men, or from insufficient study durations. Particularly in Phases I–II, small cohorts of young men (20–300) are typically used, to minimise experimental variability and to avoid the possibility of endocrinological disruption or other risks to women of reproductive age. Although 1,000–3,000 volunteers may be used in Phase III trials, the final phase before marketing (50), it is nevertheless clear that cohort numbers, study durations or other aspects of protocol design, conduct or interpretation, are inadequate to detect the adverse side-effects of the large number of pharmaceuticals that are found to harm patients after marketing. Longer studies of more-broadly representative human populations would be more predictive, but would increase the time and cost of pharmaceutical development, and are resisted by pharmaceutical companies.

The necessity of systematic reviews

The premise that laboratory animal models are generally predictive of human outcomes is the basis for their widespread use in human toxicity testing, and in the safety and efficacy testing of putative chemotherapeutic agents and other clinical interventions. However, the numerous cases of discordance between laboratory animal and human outcomes suggest that this premise may well be incorrect, and that the utility of animal experiments for these purposes may not be assured. On the other hand, only small numbers of experiments are normally reviewed in case studies, and their selection may be subject to bias. To provide more-definitive conclusions, systematic reviews of the human clinical or toxicological utility of large numbers of animal experiments are necessary. Experiments included in such reviews should be selected without bias, via randomisation, or similarly methodical and impartial means.

In support of this concept, Pound and colleagues (51) commented that clinicians and the public often consider it axiomatic that animal research has contributed to human clinical knowledge, on the basis of anecdotal evidence or unsupported claims. These constitute an inadequate form of evidence, they asserted, for such a controversial area of research, particularly given increasing competition for scarce research resources. Hence, they called for systematic reviews to examine the human clinical utility of animal experiments, and commenced by examining six existing reviews, which did not demonstrate the clinical utility expected of the experiments in question.

Soon afterwards, the UK Nuffield Council on Bioethics stated that, *It would… be desirable to*

undertake further systematic reviews and meta-analyses to evaluate more fully the predictability and transferability of animal models. They called for these to be undertaken by the UK Home Office, in collaboration with the major funders of research, industry associations and animal protection groups (52).

Since then, several such reviews and meta-analyses have been published, which collectively provide important insights into the human clinical and toxicological utility of animal models. Their identification and examination was the purpose of this review.

Methods

The *Scopus* biomedical bibliographic databases were searched for systematic reviews of the human clinical or toxicological utility of animal experiments published in the peer-reviewed biomedical literature. Among the world's most comprehensive databases, they include over 12,850 academic journals, 500 open access journals, 700 conference proceedings, and a total of 29 million abstracts (53). The *Life Sciences* database includes over 3,400 titles, and the *Health Sciences* database includes over 5,300 titles, including all of *Medline*, the leading medical and allied health profession database, which itself contains over 15 million citations from the mid-1950s onwards, sourced from more than 5,000 biomedical journals from over 80 countries (54).

All abstracts, titles and key words were searched for (*animal experiment* OR *animal model* OR *animal study* OR *animal trial*) AND (*clinical trial* OR *human outcome* OR *human relevance* OR *human result*). The results were limited to articles or reviews, but no chronological, language or other limitations were applied. Titles and, where necessary, abstracts or complete papers, were examined, in order to locate relevant papers. Additional relevant studies were obtained by examination of the reference lists of the papers retrieved, and by consultation with colleagues working in this field.

To minimise bias, reviews were included only when they had been conducted systematically, by using randomisation or similarly methodical and impartial means to select animal studies. For example, in some cases, all the animal studies within relevant subsets of toxic chemical databases were examined, without exclusion.

The examination covered only reviews which considered the human toxicological predictivity or utility of animal experiments, their contributions toward the development of prophylactic, diagnostic or therapeutic interventions with clear potential for combating human diseases or injuries, or their consistency with human clinical outcomes. Reviews which focused, for example, only on the contributions of animal experiments toward increased understanding of the aetiological, pathogenesic or other aspects of human diseases, or on the clinical utility of animal experiments in non-human species, were excluded from consideration.

Results and Discussion

Bibliographic databases are constantly updated. 2,274 articles or reviews were retrieved, by using the specified search terms, on 1 March 2007. In total, 27 systematic reviews which examined the utility of animal experiments during the development of human clinical interventions (20), or in deriving human toxicity classifications (seven), were located. Three different approaches that sought to determine the maximum human clinical utility that may be achieved by animal experiments, were of particular interest.

Clinical utility of experiments expected to lead to medical advances

Lindl and colleagues (55, 56) examined animal experiments conducted at three German universities between 1991 and 1993, that had been approved by animal ethics committees, at least partly on the basis of claims by researchers that the experiments might lead to concrete advances toward the cure of human diseases. Experiments were only included where previous studies had shown that the applications of related animal research had confirmed the hypotheses of the researchers, and where the experiments had achieved publication in biomedical journals.

For 17 experiments meeting these inclusion criteria, citations were analysed over at least 12 years. Citation frequencies and types of citing papers were recorded: whether they were reviews or animal-based, *in vitro*, or clinical studies. 1,183 citations were evident, but only 8.2% (97 citations) were in clinical publications, and only 0.3% (4 citations) demonstrated a direct correlation between the results of animal experiments and human outcomes. However, even in these four cases, the hypotheses that had been verified successfully in the animal experiment failed in every respect when applied to humans. None of these 17 experiments led to any new therapies, or had any beneficial clinical impact during the period examined.

As a result of their analysis, Lindl and colleagues called for serious, rather than cursory, evaluations of the likely benefits of animal experiments by animal ethics committees and related authorities, and for a reversal of the current paradigm, in which animal experiments are routinely approved. Instead of approving experiments because of the possibility that benefits might accrue, Lindl and colleagues suggested that where significant doubt exists, labo-

ratory animals should receive the benefit of that doubt, and that such experiments should not, in fact, be approved.

Clinical utility of highly-cited animal experiments

Hackam and Redelmeier (57) also used a citation analysis, but without geographical limitations. Based on the assumption that findings from highly-cited animal experiments would be most likely to be subsequently tested in clinical trials, they searched for experiments with more than 500 citations and published in the seven leading scientific journals, as ranked by citation impact factor.

Of 76 animal studies located, with a median citation count of 889 (range: 639–2,233), only 36.8% (28/76) were replicated in randomised human trials. 18.4% (14/76) were contradicted by randomised trials, and 44.7% (34/76) had not translated to clinical trials. Ultimately, only 10.5% (8/76) of these medical interventions were subsequently approved for use in patients, and, as stated previously, even in these cases, human benefit cannot be assumed, because adverse reactions to approved interventions are common, and a leading cause of death (46).

A low rate of translation to clinical trials of even these highly-cited animal experiments was apparent, despite 1992 being the median publication year, allowing a median of 14 years for potential translation. For studies that did translate to clinical trials, the median time for translation was seven years (range 1–15). The frequency of translation was not affected by the laboratory animal species used, the type of disease or therapy under examination, the journal, year of publication, methodological quality, and even, surprisingly, the citation rate. However, animal studies incorporating dose–response gradients were more likely to be translated to clinical trials (odds ratio [OR] = 3.3; 95% confidence interval [CI] = 1.1–10.1).

Although the rate of translation of these animal studies to clinical trials was low, as Hackam and Redelmeier stated, it is nevertheless higher than that of most published animal experiments, which are considerably less likely to be translated than these highly-cited animal studies published in leading journals. Furthermore, the selective focus on positive animal data, whilst ignoring negative results (optimism bias), was one of several factors proposed that may have increased the likelihood of translation beyond that which was scientifically merited. As Hackam (58) stated, the rigorous meta-analysis of all relevant animal experimental data would probably significantly decrease the translation rate to clinical trials.

In addition, only 48.7% (37/76) of these highly-cited animal studies were considered to be of good methodological quality. Despite their publication in leading scientific journals, few included the random allocation of animals to test groups, any adjustment for multiple hypothesis testing, or the blinded assessment of outcomes. Accordingly, Hackam and Redelmeier cautioned patients and physicians about the extrapolation of the findings of even highly-cited animal research to cases of human disease.

Clinical utility of chimpanzee experiments

Chimpanzees are the species most closely related to humans, and consequently, are considered to be the laboratory animals most likely to provide results which are predictive of human outcomes. Hence, in 2005, I conducted a citation analysis of the human clinical utility of chimpanzee experiments (59).

I searched three major biomedical bibliographic databases, and located 749 papers published between 1995 and 2004, which described experiments on captive chimpanzees or their tissues. Although published in the international scientific literature, the vast majority of these experiments were conducted within the USA (60). To obtain 95% CIs with an accuracy of at least plus or minus 10%, when estimating the proportion of chimpanzee studies subsequently cited by other published papers, a subset of at least 86 chimpanzee studies was required (61–63).

Of 95 published randomly-selected studies on chimpanzees, 49.5% (47/95) were not cited by any subsequent papers, demonstrating minimal contributions toward the advancement of biomedical knowledge. This is of particular concern, because it can be assumed that research judged to be of lesser value was not published. Hence, it appears that the majority of chimpanzee research generates data of questionable value, which make little obvious contribution toward the advancement of biomedical knowledge.

35.8% (34/95) of the 95 published chimpanzee studies were cited by 116 papers that clearly did *not* describe well-developed methods for combating human diseases. Only 14.7% (14/95) of them were cited by 27 papers that had abstracts which indicated well-developed prophylactic, diagnostic or therapeutic methods for combating human diseases. However, a detailed examination of these 27 medically-oriented papers revealed that *in vitro* studies, human clinical and epidemiological studies, molecular assays and methods, and genomic studies, contributed most to their development. 63.0% (17/27) were wide-ranging reviews of 26–300 (median 104) references, to which these cited chimpanzee studies made very small contributions. Duplication of human outcomes, inconsistency with other human or primate data, and other causes, resulted in the absence of any chimpanzee study able to demonstrate an essential contribution, or, in

most cases, a significant contribution of any kind, toward the development of the medical method described.

Despite the low utility of chimpanzee experiments in advancing human health which was indicated by these results, it remains true that chimpanzees are the species most closely related to human beings. Hence, it is highly likely other laboratory species are even less useful as experimental models of humans in biomedical research and toxicity testing.

Clinical utility of stroke and head injury models

Despite the existence of literature on the efficacy of more than 700 drugs in treating experimental models of stroke (artificially-induced focal cerebral ischaemias; 64), only recombinant tissue plasminogen activator (rt-PA) and aspirin have convincingly demonstrated efficacy in human clinical trials of treatments for acute ischaemic stroke (65–67). Hence, Macleod and colleagues (64) stated that, *This failure of putative neuroprotective drugs in clinical trials represents a major challenge to the doctrine that animals provide a scientifically-valid model for human stroke.* At least 10 published systematic reviews have described the poor human clinical utility of animal experimental models of stroke and head injuries (64, 68–76).

In some cases, clinical trials proceeded, despite equivocal evidence of efficacy in animal studies. For example, Horn and colleagues (68) systematically reviewed 20 animal studies on the efficacy of nimodipine, of which only 50% showed beneficial effects following treatment. They concluded that, *...the results of this review did not show convincing evidence to substantiate the decision to perform trials with nimodipine in large numbers of patients.* These clinical trials also demonstrated equivocal evidence of efficacy, and furthermore, proceeded concurrently with the animal studies, despite the fact that the latter are intended to be conducted prior to clinical trials, to facilitate the detection of potential human toxicity.

O'Collins and colleagues (69) conducted a very large review of 1,026 experimental drugs for acute stroke that had been tested in animal models. They found that the effectiveness in animals of 114 drugs chosen for human clinical use was no greater than that of the remaining 912 drugs not chosen for clinical use, thereby demonstrating that effectiveness in animal models had no measurable effect on whether or not these drugs were selected for human clinical use. Accordingly, O'Collins and colleagues questioned whether the most efficacious drugs are, in fact, being selected for clinical trials, and called for greater rigour in the conduct, reporting, and analysis of animal experiments.

In many cases, animal models did indicate efficacy, but this did not translate to humans. In a few reviews, the authors speculated on the possible causes. For example, Jonas and colleagues (70) hypothesised that the poor clinical efficacy of neuroprotectants which had been found to be successful in animal models, was due to differences in the timing of the initiation of treatment. Curry (71) hypothesised that the human clinical failure of fourteen neuroprotective agents which were successful in animal models, was due to the antagonism of glutamate — which may be associated with neuroprotection — by drug treatment in clinically-normal individuals. He therefore proposed that clinical trials should be restricted to real stroke patients, who experience elevated plasma glutamate levels. However, such speculation has not resulted in improvements in the poor clinical record of neuroprotectants which were previously found to be successful in animal models.

The utility of the majority of these animal studies also appears to have been impeded by their poor methodological quality. Examples include: animal studies on the efficacy of melatonin (64); 20 animal studies on the efficacy of nimodipine (68); 29 animal studies on the efficacy of FK506 (72); 45 animal studies on five compounds from different classes of alleged neuroprotective agents — clomethiazole, gavestinel, lubeluzole, selfotel, and tirilazad mesylate (73); 25 animal studies on the efficacy of nitric oxide (NO) donors and L-arginine (74); and 73 animal studies of the efficacy of NO synthase inhibitors (75).

The methodological quality of animal studies was typically scored on the basis of the presence of characteristics such as: appropriate animal models (aged, diabetic or hypertensive animals are considered to more-closely model human stroke patients); power calculations of sample sizes; random allocation to treatment and control groups; use of a clinically-relevant time window for commencement of treatment; blinded drug administration; use of anaesthetics without significant intrinsic neuroprotective activity (ketamine, for example, may alter neuroprotective activity); blinded induction of ischaemia (given that the severity of induced infarcts may be subtly affected by knowledge of treatment allocation); blinded outcome assessment; assessment of both infarct volume and functional outcome; adequate monitoring of physiological parameters; assessment during both the acute (e.g. one to six days) and chronic (e.g. seven to 30 days) phases; statement of temperature control; compliance with animal welfare regulations; peer-reviewed publication; and conflict of interest statements. Typically, one point was given for the presence of each characteristic. For example, *The Stroke Therapy Academic Industry Roundtable* recommendations for standards with regard to preclinical and restorative drug development involve an eight-point scale (68, 77).

Median quality scores were: four out of 10 (13 studies; range zero to six [64]); four out of 10 (29

studies; range zero to seven [72]); three out of 10 (45 studies [73]); and three out of 8 (73 studies; range one to six [75]). Common deficiencies included lack of: sample size calculations, aged animals or those with appropriate co-morbidities, randomised treatment allocation, blinded drug administration, blinded induction of ischaemia, blinded outcome assessment, and conflict of interest statements. Some studies also used ketamine anaesthesia, and there was also substantial variation in the parameters assessed.

van der Worp and colleagues (73), for example, concluded that the collective evidence for neuroprotective efficacy which formed the basis for 21 clinical trials, was obtained in animal studies with a methodological quality that would not, in retrospect, justify such a decision. Wilmot and colleagues (74) also found considerable variations in animal experiment protocols, which concerned: animal species; physiological parameters (such as blood pressure); drug administration (timing, dosage, and route); surgical methodology; and duration of ischaemia. Statistical analysis (Egger's test) also revealed the likely existence of publication bias (an increased tendency to publish studies in which a treatment effect is apparent, or a decreased tendency to so publish, e.g. resulting from commercial pressures, particularly in the case of patented drugs under development). Macleod and colleagues (64) commented that, *These deficiencies apply to most, if not all, of the animal literature.* This is of particular concern, because Macleod and colleagues (72) reported that efficacy was apparently lower in higher quality studies, which raised concerns that the apparent efficacy may have been artificially elevated by factors such as poor methodological quality and publication bias.

A related review, not limited solely to stroke, exemplified some of these issues. Perel and colleagues (76) examined therapeutic interventions with unambiguous evidence of a treatment effect (benefit or harm), in clinical trials related to the following: corticosteroidal treatment for head injury; anti-fibrinolytics for the treatment of haemorrhage; thrombolysis, and also tirilazad, for the treatment of acute ischaemic stroke; antenatal corticosteroids in the prevention of neonatal respiratory distress syndrome; and bisphosphonates in the treatment of osteoporosis. They found that three interventions had similar outcomes in animal models, whilst three did not, suggesting that the animal studies did not reliably predict the human outcomes. Perel and colleagues reported that the animal studies varied in methodological quality and sample sizes, that randomisation and blinding were rarely reported, and that publication bias was evident.

Clinical utility of other animal experiments

Of seven systematic reviews on the utility of animal models in other clinical fields identified by this review (78–85, of which 79 and 80 described a single review), in only two cases — one of which was contentious — did the animal models appear to be clearly useful in the development of human clinical interventions, or substantially consistent with human clinical outcomes.

As in the case of stroke, some clinical trials proceeded, despite equivocal evidence of efficacy in animal studies. Upon systematically reviewing the effects of Low Level Laser Therapy (LLLT) on wound healing in 36 cell or animal studies, Lucas and colleagues (78) found that an in-depth analysis of studies with the highest methodological quality showed no significant pooled treatment effect. Despite this, the clinical trials proceeded. Furthermore, almost from the beginning of LLLT investigations, animal experiments and clinical studies occurred simultaneously, rather than sequentially. The human trials also failed to demonstrate significant benefits.

Roberts and colleagues (79), and Mapstone and colleagues (80), all systematically reviewed a group of 44 randomised, controlled animal studies on the efficacy of fluid resuscitation in bleeding animals. A previous systematic review by some of these investigators of clinical trials of fluid resuscitation had found no evidence that the practice improved outcomes, and had even identified the possibility that it might be harmful (86). In this later review (79–80), they found that fluid resuscitation reduced mortality in animal models of severe haemorrhage, but increased mortality in those with less severe haemorrhage.

After clinical trials in humans failed to provide evidence of benefit, Lee and colleagues (81) conducted a systematic review and meta-analysis of controlled trials of endothelin receptor blockade in animal models of heart failure. Meta-analysis failed to provide evidence of overall benefit, and indicated increased mortality with early administration.

In their investigation of the contributions of human clinical trial results and analogous experimental studies to asthma research — one of the most common and heavily-investigated of modern diseases — Corry and Kheradmand (82) demonstrated that failure to conduct and analyse the results of animal studies before proceeding to clinical trials is not uncommon: *Research along two fronts, involving experimental models of asthma and human clinical trials, proceeds in parallel, often with investigators unaware of their counterpart's findings.*

The clinical utility of animal models is clearly questionable in such cases, in which clinical trials proceed concurrently with, or prior to, animal studies, or continue, despite equivocal evidence of efficacy in the animals.

As in the case of stroke, the clinical utility of the majority of these animal studies also appears to have been limited by their poor methodological

quality. Examples include: 36 cell or animal studies on the effects of LLLT on wound healing (78); 44 studies on the efficacy of fluid resuscitation in bleeding animals (79–80); and studies on the efficacy of endothelin receptor blockade in animal models of heart failure (81). Common flaws included inadequate sample sizes, leaving studies underpowered, and lack of randomisation and blinding.

In some cases, obvious deficiencies within the animal models were identified. In commenting on the clinical relevance of animal models for testing the effects of LLLT on wound healing, Lucas and colleagues (78) noted that the animal models excluded common problems associated with wound healing in humans, such as ischaemia, infection, and necrotic debris.

Difficulties were also apparent, in translating animal outcomes to human clinical protocols, in at least one case. Lazzarini and colleagues (83) reviewed experimental studies on osteomyelitis, to ascertain their impacts on the systemic antibiotic treatment of human osteomyelitis. Although they found that most of the animal models reviewed were reproducible and dependable, they also found that the human predictivity of these studies was unclear, and was possibly undermined by difficulties in establishing the right dose regimen in the animals. Although they considered that the use of antibiotic combinations was associated with better outcomes in the majority of animal studies, and that these studies did provide indications of appropriate minimum treatment durations, they concluded that these studies had limited relevance to clinical practice.

In two cases, reviewers reported that animal and human outcomes were substantially consistent, although in one case this conclusion was contentious. While reviewing therapeutic approaches to streptococcal endocarditis, Scheld (84) reported good overall correlations among results obtained by *in vitro* susceptibility testing (especially killing kinetics in broth), in animal experiments, and in clinical trials on different antimicrobial regimens in humans with streptococcal endocarditis.

To investigate the efficacy of rodent models of carcinogenesis in predicting treatment outcomes in humans, Corpet and Pierre (85) conducted a systematic review and meta-analysis of colon cancer chemoprevention studies involving the use of aspirin, β-carotene, calcium, and wheat bran, in rats, mice and humans. Controlled intervention studies on the recurrence of adenomas in human volunteers were compared with chemoprevention studies of carcinogen-induced tumours in rats, and of polyps in Min (Apc[+/–]) mice. 6,714 humans, 3,911 rats and 458 mice were included in the meta-analyses. Corpet and Pierre found that comparable results were achieved in rats and humans with aspirin, calcium, β-carotene, and wheat bran. Comparable results were found in Min mice and humans with aspirin, but discordant results were obtained with calcium and wheat bran (the equivalent β-carotene results were not available). Corpet and Pierre concluded that these results suggest that the use of the rodent models can roughly predict treatment effects in humans, but that the prediction is not accurate for all agents, and that the carcinogen-induced rat model is more predictive than the Min mouse model. However, relatively few agents were tested, and two of the three agents tested in mice produced different outcomes in humans, so the conclusion that rodents are predictive of human treatment effects, albeit only roughly, is itself contentious.

Toxicological utility: carcinogenicity

Due to the limited availability of data on human exposure, the identification and regulation of exposure to potential human toxins has traditionally relied heavily on animal studies. However, systematic reviews have indicated that the utility of animal studies for these purposes is lacking in the fields of carcinogenicity (at least five reviews: 87–91) and teratology (one review: 92). No systematic review demonstrated a contrary result. The sensitivities of animal models to a range of human toxicities (i.e. the ability to identify them) highlighted by one review (93) generally appears to be accompanied by poor human specificity (i.e. the ability to correctly identify human *non*-toxins), resulting in a high incidence of false-positive results.

EPA survey

The regulation of human exposure to potentially carcinogenic chemicals constitutes society's most important use of animal carcinogenicity data. In 2004, to examine the utility of animal carcinogenicity data in protecting public health, I surveyed the EPA's Integrated Risk Information System (IRIS) chemicals database. This database contains the environmental contaminants of greatest concern in the USA, together with their animal, and, in a small minority of cases, human toxicity data, along with the human toxicity assessments based on this pooled data. However, of the 160 IRIS chemicals lacking even limited human exposure data, but possessing animal data, for which human toxicity assessments existed, the EPA considered the animal carcinogenicity data to be inadequate to support a classification of probable human carcinogen or non-carcinogen in the majority of cases (58.1%, 93/160; 95% CI: 50.4–65.5; 87).

Furthermore, data from the World Health Organisation's International Agency for Research on Cancer (IARC) indicated that the true utility of

animal carcinogenicity data for deriving human carcinogenicity assessments is actually substantially lower than that indicated solely by EPA assessments. Of 128 chemicals with human or animal data assessed by both the EPA and the IARC, human carcinogenicity classifications were consistent between the two agencies only for the 17 chemicals for which at least limited human data were available. For those 111 chemicals for which the classification was primarily reliant on animal data, the EPA was much more likely than the IARC to assign carcinogenicity classifications indicative of greater human risk ($p < 0.0001$; 87).

The IARC is a leading international authority on carcinogenicity assessments, and the significant differences between its human carcinogenicity classifications and those of the EPA, for identical chemicals, indicate that: i) in the absence of significant human data, the EPA is over-reliant on animal carcinogenicity data; ii) as a result, the EPA tends to over-predict carcinogenic risk; and iii) the true predictivity for human carcinogenicity of animal data is even poorer than that indicated by EPA figures alone. EPA policy erroneously assuming that tumours in animals are indicative of human carcinogenicity, was implicated as a primary cause of these errors, which have substantial US public health implications concerning the regulation of human exposures to environmental contaminants (87).

IARC Monographs *survey*

The poor human predictivity of animal carcinogenicity studies was also demonstrated in 1993 by Tomatis and Wilbourn (88), who surveyed the 780 chemical agents or exposure circumstances evaluated and listed within Volumes 1–55 of the *IARC Monographs* series (94). Of these, 502 (64.4%) had definite or limited evidence of animal carcinogenicity, and 104 (13.3%) were assessed as definite or probable human carcinogens. Virtually all of the latter group would, of course, have been members of the former; so at least 398 animal carcinogens were assessed and considered not to be definite or probable human carcinogens.

The positive predictivity of a test is the proportion of positive outcomes that are truly positive for the characteristic being tested for, while the false-positive rate refers to the proportion that are not. Hence, based on these IARC figures, the positive predictivity of the animal bioassay for definite or probable human carcinogens was, at best, only 20.7% (104/502), while the false-positive rate was at least 79.3% (398/502).

More-recent IARC classifications indicate little improvement in the positive predictivity of the animal bioassay for human carcinogens. By 1 January 2004, a decade later, only 105 additional agents had been added to the 1993 number, yielding a total of 885 agents or exposure circumstances listed in the *IARC Monographs* (95). The proportion of definite or probable human carcinogens had increased only slightly, from 13.3% in 1993 to 17.1% in 2004.

The NTP and other surveys

Surveys by other investigators have also demonstrated the poor human predictivity of animal carcinogenicity data. After examining the studies on 471 substances contained within the US National Toxicology Program (NTP) carcinogenicity database as of 1 July 1998, Haseman (89) concluded that, although 250 (53.1%) produced carcinogenic effects in at least one sex–species group, the actual proportion which posed a significant carcinogenic risk to humans was probably far lower, for reasons such as interspecies differences in mechanisms of carcinogenicity.

Similarly, around half of all chemicals tested on animals and included in the comprehensive Berkeley-based carcinogenic potency database, whether natural or synthetic, gave positive results (89). Rall (96) estimated that only around 10% of chemicals are truly carcinogenic to humans. Ashby and Purchase (97) speculated that all chemicals would eventually display some carcinogenic activity, if tested in sufficient rodent strains. Even common table salt has been classified as a tumour promoter in rats (98).

Fung and colleagues (99) estimated that, if all the 75,000 chemicals in use were tested for carcinogenicity via the standard NTP bioassay, significantly less than 50% would prove carcinogenic in animals, and less than 5–10% would warrant further investigation. They suggested that the higher positivity rate recorded is due to chemical selection based on *a priori* suspicion of carcinogenicity. However, examination of the carcinogenicity literature reveals that chemicals are selected for study for many reasons other than *a priori* suspicion, including production volumes, occupational and environmental exposure risks, and investigations of mechanisms of carcinogenesis (100). Despite this, the positivity rate of the carcinogenicity bioassay in the general literature remains around 50% (101).

Huff (90) demonstrated a significant variation in carcinogenicity test results between two major carcinogenicity testing programmes, at the NTP (Research Triangle Park, NC, USA) and the Ramazzini Foundation (RF; Bentivoglio, Italy). Both laboratories had carried out several hundred chemical carcinogenesis bioassays: around 500 at the NTP, and 200 at the RF. Of these, 21 chemicals were evaluated by both laboratories, of which published results were available for 14. The results were inconsistent for 3 of these 14 chemicals (21.4%), which had been declared carcinogenic by one labo-

ratory but not the other, questioning the reliability of these assays. Of the remaining 11 chemicals, both laboratories found nine to be carcinogenic, and two not to be carcinogenic.

Possible causes for such different toxicity results between laboratories include differences in: the test species, strain, age or gender; the quantity, duration and consistency of dosing; the route and method of administration; diet and laboratory environmental conditions; and the criteria used for the assessment of toxicity.

Ennever and Lave (91) demonstrated that neither of the two commonly-used interpretations of rodent carcinogenicity data provide valid conclusions about human carcinogenicity. If a risk avoidance interpretation is used, in which any positive result in male or female mice or rats is considered positive, then nine of the 10 known human carcinogens among the hundreds of chemicals tested by the NTP are positive (102), but so are an implausible 22% of all chemicals tested (99). If a less risk-sensitive interpretation is used, whereby only chemicals positive in both mice and rats are considered positive, then only three of the six known human carcinogens tested in both species are positive (102). The former interpretation could result in the needless denial of potentially useful chemicals to society, while the latter could result in widespread exposure to undetected human carcinogens.

Toxicological utility: teratogenicity

In 2005, my colleagues and I published an extensive survey examining the human predictivity of animal teratogenicity testing (92). We examined nearly every putative teratogen tested in more than one species, including 1,396 studies. Data for 11 groups of known human teratogens tested in 12 animal species were analysed. Discordance between species was apparent in just under 30% of these 1,396 reports. Almost a quarter of all the outcomes in the six main species used (mouse, rat, rabbit, hamster, primate and dog) were equivocal. For known human teratogens, there was high variability in positive predictivity between species, the mean of which was only 51% — hardly better than tossing a coin. Some species exhibited a high false-negative rate. Only around half of these known human teratogens were teratogenic in more than one primate species. Fewer than one in 40 of the substances designated as potential teratogens from animal studies, were conclusively linked to human birth defects.

We concluded that the poor human predictivity of animal-based teratology warrants the cessation of animal testing, and that resources should be reallocated into the further development and implementation of quicker, cheaper and more reliable, scientifically validated alternatives, such as the embryonic stem cell test.

Toxicological utility: various

Under the auspices of the International Life Sciences Institute's Health and Environmental Sciences Institute, Olsen and colleagues (93) sought to determine the extent to which various types of human toxicities evident during clinical trials could be predicted from standard toxicology studies. Based on a multi-company database of 131 pharmaceutical agents with one or more human toxicities identified during clinical trials, they reported a true-positive prediction rate of animal models for human toxicity of 69%, and also that study results from non-rodent (dog, primate) species have good potential to identify human toxicities from many therapeutic classes.

These results concur with those of the other toxicity reviews described. Animal studies are often reasonably sensitive for human toxins. However, their human predictivity and toxicological utility are limited by their poor human specificity, which results in high false-positive rates.

Causes of the poor human utility of animal models

When evaluated overall, these 27 systematic reviews clearly do not support the widely-held assumptions of animal ethics committees and the opinions of advocates of animal experimentation, that laboratory animal use is generally beneficial in the development of human therapeutic interventions and the assessment of human toxicity. On the contrary, they frequently demonstrate that animal experiments are of low utility for these purposes. This appears to result both from limitations of the animal models themselves, and also from the poor methodological quality and statistical design of many animal experiments.

Biomedical research

Chimpanzees are our closest living relatives, but despite great similarities between the structural regions of chimpanzee DNA and human DNA, important differences between the regulatory regions exert an *"avalanche"* effect on large numbers of structural genes (103). Despite nucleotide difference between chimpanzees and humans of only 1–2%, this effect results in differences of around 20%, in terms of protein expression (104), representing a marked phenotypic differences between the species. These differences manifest as: altered susceptibility to the aetiology and progression of various diseases; differences in the absorption, tissue distribution, metabolism, and excretion of chemotherapeutic agents; and differences in the toxicity and efficacy of pharmaceuticals and other agents (59, 103). Such effects appear to be responsible for the demonstrated inability of

most chimpanzee research to contribute substantially to the development of methods which are efficacious in combating human diseases (59).

Other laboratory animal species are much less similar to humans, both genetically and phenotypically, and are therefore less likely to be useful for accurately modelling the progression of human diseases or of human responses to chemicals and putative chemotherapeutic agents.

Toxicity testing

Rodents are by far the most common laboratory animal species used in toxicity studies. Several factors contribute to the demonstrated inability of rodent bioassays to reliably predict human toxicity. The stresses incurred during handling, restraint, other routine laboratory procedures, and particularly, the stressful routes of dose administration common to toxicity tests, alter immune status and disease predisposition in ways which are very difficult to accurately predict, and which distort the progression of diseases and responses to chemicals and putative chemotherapeutic agents (105, 106).

In addition, animals have a broad range of physiological defences against general toxic insults, such as epithelial shedding and inducible enzymes, which commonly prove effective at environmentally relevant doses, but which may be overwhelmed at the high doses commonly applied in routine toxicity testing (101). Carcinogenicity assays, in particular, involve chronic, high level dosing. This may result, *inter alia*, in insufficient rest intervals between doses for the effective operation of DNA and tissue repair mechanisms, which, with the unnatural elevation of cell division rates during *ad libitum* feeding, may predispose the animals to mutagenesis and carcinogenesis. Lower doses, greater intervals between exposures, shorter total periods of exposure, and intermittent feeding, which represent a more realistic approach to the environmental exposure of humans to most potential toxins, might not result in toxic changes at all (106).

Finally, differences in rates of absorption and transport mechanisms between test routes of administration and other important human routes of exposure, and the considerable variability of organ systems in response to toxic insults, between and within species, strains and genders, render profoundly difficult any attempt to accurately predict human hazard on the basis of animal toxicity data (106).

Methodological quality

At least 11 systematic reviews (57, 64, 68, 72–76, 78–81 [of which, 79 and 80 described a single review]) demonstrated the poor methodological quality of many of the animal studies examined, and none of the reviews demonstrated good methodological quality in a majority of studies. While the omission of study details due to publication space constraints may artificially lower apparent quality, the prevalence of such deficiencies exceeds that which might reasonably be expected, and is, accordingly, grounds for considerable concern.

Common deficiencies included lack of: sample size calculations, sufficient sample sizes, appropriate animal models (e.g. aged animals or those with appropriate comorbidities), randomised treatment allocation, blinded drug administration, blinded induction of ischaemia in the case of stroke models, blinded outcome assessment, and conflict of interest statements. Some studies also used anaesthetics that may have altered the experimental outcomes, and substantial variation was evident in the parameters assessed.

These deficiencies limited the clinical utility of these studies in various significant ways. For example, it is well established that studies lacking randomisation or blinding often over-estimate the magnitude of the effects of treatments (107–109). Bebarta and colleagues (110) described the impacts of lack of randomisation or blinding on estimations of the significance of treatment effects in 389 animal studies and in 2,203 cell line studies. They found that studies lacking randomisation or blinding, but not both, were more likely to report a treatment response than studies that used these measures (OR = 3.4; 95% CI = 1.7 to 6.9, and OR = 3.2; 95% CI = 1.3 to 7.7, respectively), and that studies lacking both randomisation and blinding were even more likely to report a treatment response (OR = 5.2; 95% CI = 2.0 to 13.5).

Statistical design

Insufficient sample sizes left many studies underpowered, limiting the statistical validity of the study conclusions. Animal lives and other resources may also be wasted, if experiments subsequently require repetition as a result. As stated by the UK Medical Research Council (111), *The number of animals used… must be the minimum sufficient to create adequate statistical power to answer the question posed.*

According to Balls and colleagues (112), however, *…surveys of published papers, as well as more anecdotal information, suggest that more than half of the published papers in biomedical research have statistical mistakes, many seem to use excessive numbers of animals, and a proportion are poorly designed.* Festing (113) similarly stated that, *Surveys of published papers show that there are many errors, both in the design of the experiments and in the statistical analysis of the resulting data. This must result*

in a waste of animals and scientific resources, and it is surely unethical. De Boo and Hendriksen (114) noted the tendency to alter animal numbers based on scientifically irrelevant issues, such as availability or cost.

Factors that should be considered when calculating appropriate sample sizes include: detectability threshold (the size of the difference between treatment groups considered significant); known or expected data variation; the required significance of the test ('p' or 'α': the probability of a Type I error — assuming a difference where none exists); the acceptable probability of assuming no difference where one does exist ('β', a Type II error. The 'power' of an experiment = 1–β; 0.8 is the usual choice); and the type of statistical analysis to which the data will be subjected. Smaller thresholds, greater data variation, smaller acceptable error probabilities (greater power), and certain statistical tests for differences, all require larger samples.

No universal rule for calculating correct sample sizes exists (114). Festing (115), for example, describes two methods, the preferred 'power calculation,' and the 'resource equation.' Power calculations use formulae which are available in interactive computer programmes (e.g. 116, 117), and calculate the minimum sample sizes required to detect treatment effects with specified degrees of certainty. Mead's 'resource equation' (118) calculates sample sizes by using degrees of freedom, and incorporates statistical parameters, such as treatment effects, block effects and error degrees of freedom.

Strategies should also be considered for minimising animal numbers without unacceptably compromising statistical power. Several of these strategies aim to decrease data variability by minimising heterogeneity in experimental environments and protocols. This can be achieved by: i) the appropriate use of environmental enrichment, aimed at decreasing physiological variation resulting from barren laboratory housing and stressful procedures; ii) choosing, where possible, to measure variables with relatively low inherent variability; iii) the use of genetically homogeneous (isogenic or inbred) or specified pathogen-free animal strains; and iv) screening raw data for obvious errors or outliers (105, 114, 119–122).

Meta-analysis involves the aggregation and statistical analysis of suitable data from multiple experiments. For some purposes, treatment and control groups can be combined, permitting group numbers to be minimised. Although new information can be derived through meta-analysis, more frequently, the results allow the refinement of existing knowledge. By designing experiments and reporting protocols to maximise their utility for later meta-analyses, the benefit of individual randomised controlled experiments can be maximised (123). Strategies such as these, aimed at maximising the statistical power of small samples, are particularly appropriate when marked ethical, cost or practical constraints limit the number of animals that may be used (e.g. in experiments involving non-human primates).

Finally, the appropriate statistical analysis of the resultant data should be closely linked to the experimental design, and to the type of data produced (124). The relatively poor statistical knowledge of many animal researchers may be the cause of the high prevalence of poor sample size choices in animal studies. Solutions could include the training of researchers in statistics, and the direct input of statisticians in experimental design and data analysis (114, 125).

Raising standards: evidence-based medicine

Evidence-based medicine (EBM) bases clinical decisions on methodologically-sound, prospective, randomised, blinded, and controlled clinical trials. The gold standard for EBM is large prospective epidemiological studies, or meta-analyses of randomised and blinded, controlled clinical trials (126). The application to animal experiments of the EBM standards which are currently applied to human clinical trials, would make the results more robust and would increase their applicability (76, 127–130). However, mechanisms would be needed to ensure compliance with such standards. Compliance could, for example, be made a prerequisite for research funding, ethics committee approval, and the publication of results. These measures would require the education and cooperation of funding agencies, ethics committees and journal editors.

The UK Medical Research Council requires researchers who are planning clinical trials, to reference systematic reviews of related previous work before they are permitted to proceed (51). To facilitate the detection of toxicity and of potentially efficacious drugs, such reviews should also include all relevant animal research (76). A similar requirement to reference, or where necessary, conduct, systematic reviews of relevant animal studies, prior to the commencement of further animal studies, would encourage a more complete and impartial assessment of the existing evidence (51).

Mechanisms are also needed to encourage the reporting of negative results. The negative results of preclinical studies are much more likely to remain unpublished than are the negative results of clinical trials (131). In a systematic review of studies on the efficacy of nicotinamide in combating experimentally-induced stroke, comparisons published only in abstract form gave a significantly lower estimate of effect size than those published in full, demonstrating publication bias (132). van der Worp and colleagues (73) commented on the pressure to obtain and publish positive results: *It is*

therefore conceivable that the career of a preclinical investigator is more dependent on obtaining positive results, than that of a clinical trialist.

Fundamental constraints on the human utility of animal models

Strategies designed to increase the full and impartial examination of existing data before conducting animal studies, to improve their methodological quality, and to decrease bias during the publication of results, would minimise the consumption of animal, financial and other resources within studies of questionable merit and quality, and would increase the potential utility of animal data in addressing human situations and problems. However, the poor human clinical or toxicological utility of many animal experiments is unlikely to result solely from their poor methodological quality, or from publication bias. As stated by Perel *et al.* (76), the failure of animal models to adequately represent human disease may be another fundamental cause, which, in contrast, could be technically and theoretically impossible to correct.

The genetic modification of animal models through the addition of foreign genes (transgenic animals) or the inactivation or deletion of genes (knockout animals) is being attempted, to make them more-closely model humans. However, as well as being technically very difficult to achieve, such modification may not permit clear conclusions, due to a large number of factors, including those reflecting the intrinsic complexity of living organisms, such as the variable redundancy of some metabolic pathways between species (133). Furthermore, the animal welfare burdens incurred during the creation and use of GM animals are particularly high (134).

Implications for scientific validation of experimental models

Proposed non-animal test models are generally required to pass formal scientific validation before their use is widely or officially accepted. Pharmaceutical licensing agencies, for example, are generally unwilling to accept non-animal test data as evidence of the human safety of proposed new pharmaceuticals, until the test models used have been scientifically validated.

Scientific validation has traditionally involved the demonstration, in multiple independent laboratories, that the test in question is relevant and reliable for its specified purpose (*practical validation*; 135), such as the prediction of a certain *in vivo* outcome. It should also be preceded by an evaluation of the necessity for the test and of the adequacy of its development (136, 137). A three-stage *prevalidation* process should be utilised to improve the efficiency of the formal validation process, by ensuring satisfactory protocol refinement and transferability, and test performance (138).

However, it is not always scientifically necessary, or even logistically possible, to conduct multi-centre practical studies. Hence *weight-of-evidence validation*, also known as *validation by retrospective analysis* (139, 140), may be conducted, based on the assessment of existing data in a structured, systematic and transparent manner, provided that data of sufficient quantity and quality are available (141).

Regardless of the approach taken, the criteria required for formal validation are comprehensive (136, 141). Key objectives include: establishing the role and necessity of the test model; ensuring clarity of the defined goals; defining a prediction model, i.e. an algorithm for converting the test data into meaningful predictions of *in vivo* toxicity; examining the mechanistic relevance and credibility of the model with respect to those goals; and providing a description of the limitations of the model.

Where practical validation studies do occur, these should adhere to best practice standards, designed to ensure good methodological quality, including, for example, statistical justifications of sample sizes, randomised allocation to test groups, and blinded treatment and assessment of results. Where possible, inter-laboratory reproducibility should be demonstrated (136).

Whether validation studies are conducted by practical or weight-of-evidence approaches, experience has shown that transparency and independence from commercial, political or other interests should be maximised through the use of independent experts and the peer-reviewed publication of outcomes (136).

Scientific validation should lead to the reasoned overall assessment that sufficient evidence exists to demonstrate that a model is, or is not, relevant and reliable for the specified purpose, or that insufficient evidence exists to be reasonably certain either way. In some cases, an interim assessment can be made, until further evidence becomes available (141).

The European Centre for the Validation of Alternative Methods (ECVAM) was created by the EC in 1991, to fulfil the requirements of *Directive 86/609/EEC* on the protection of animals used for experimental and other scientific purposes. These requirements state that the EC and its Member States should actively support the development, validation and acceptance of methods which could *replace*, *refine* or *reduce* the use of laboratory animals (142). The US equivalent is the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM), which has similar goals. Despite the high standards required for successful validation, between 1998 and 2007, 21 distinct tests or categories of test methods that could *replace*, *reduce* or *refine* laboratory animal use,

had been validated and registered with ECVAM, and nine had achieved regulatory acceptance (143).

However, unlike non-animal models, animal models are generally assumed to be reasonably predictive of human outcomes in preclinical drug development, toxicity testing, and other fields of biomedical research, without the need to undergo formal validation studies. Yet the 27 systematic reviews examined in this study, demonstrate that it is insufficient to assume that animal models are reliably predictive of human outcomes, even those in use for long periods, without subjecting them to critical assessment.

Clearly, formal validation should be consistently applied to all proposed experimental models, regardless of their animal, non-animal, historical, contemporary or possible future status, and models should be chosen on the basis of critical scientific review, with appropriate consideration also given to animal welfare, ethical, legal, economic, and any other relevant factors.

The Heads of ECVAM and the European Chemicals Bureau, the EC agencies responsible for technical aspects of validation and for EU chemicals regulations, respectively, at that time, made a similar call in 1995, in which they urged that prevalidation and independent assessment be applied with equal force to all new or modified animal and non-animal test guidelines (144).

Conclusions

The historical and contemporary paradigm, that animal models are generally reasonably predictive of human outcomes, provides the basis for their widespread use in toxicity testing and biomedical research aimed at preventing or developing cures for human diseases. However, their use persists for historical and cultural reasons, rather than because they have been demonstrated to be scientifically valid. For example, many regulatory officials *"feel more comfortable"* with animal data (145), and some even believe that animal tests are inherently valid, simply because they are conducted in animals (146).

However, most existing systematic reviews have demonstrated that animal experiments are insufficiently predictive of human outcomes to provide substantial benefits during the development of human clinical interventions, or in deriving human toxicity assessments. In only two of 20 reviews in which clinical utility was examined, did the authors conclude that the animal models were either significantly useful in contributing to the development of clinical interventions, or were substantially consistent with clinical outcomes (84, 85), and one of these conclusions was contentious. Seven additional reviews also failed to clearly demonstrate utility in predicting human toxicological outcomes, such as carcinogenicity and teratogenicity. Consequently, animal data can be generally assumed not to be substantially useful for these purposes.

Likely causes of this inadequacy include inherent genotypic and phenotypic differences between human and non-human species, the distortion of experimental outcomes arising from experimental environments and protocols, and the poor methodological quality of many animal experiments, as was apparent in at least 11 reviews. There were no reviews in which a majority of animal experiments were of good methodological quality. Some of these problems might be minimised with concerted effort (given their widespread prevalence), but the limitations resulting from interspecies differences are likely to be technically and theoretically impossible to overcome.

Despite the fact that they have not passed and, indeed, could not pass, the formal scientific validation process required of non-animal models prior to regulatory acceptance, most animal models are incorrectly assumed to be predictive of human outcomes. The consistent application of formal validation studies to all test models is clearly warranted, regardless of their animal, non-animal, historical, contemporary or possible future status. Experimental model choices should be based on such critical scientific review, with appropriate consideration also given to animal welfare, ethical, legal, economic and other relevant factors.

Likely benefits would include greater selection of models truly predictive for human outcomes, increased safety of people exposed to chemicals that have passed toxicity tests, increased efficiency during the development of human pharmaceuticals and other therapeutic interventions, and decreased wastage of animal, personnel and financial resources.

In addition, the poor human clinical and toxicological utility of most animal models for which data exists, in conjunction with their generally substantial animal welfare and economic costs, justify a ban on the use of animal models lacking scientific data clearly establishing their human predictivity or utility.

Received 02.03.07; received in final form 10.07.07; accepted for publication 11.07.07.

References

1. Anon. (2007). *Annex to the Fifth Report on the Statistics on the Number of Animals Used for Experimental and other Scientific Purposes in the Member States of the European Union (COM(2007)675 final)*, 277pp. Brussels, Belgium: European Commission.
2. Goldberg, A.M. (2002). Use of animals in research: a science–society controversy? The American perspective: animal welfare issues. *ALTEX* **19**, 137–139.
3. Stephens, M.L., Alvino, G.M. & Branson, J.B. (2002). Animal pain and distress in vaccine testing in the

United States. *Developments in Biologicals* **111**, 213–216.
4. Anon. (2007). *FY 2006 AWA Inspections*, 11pp. Riverdale, MD, USA: United States Department of Agriculture Animal and Plant Health Inspection Service (USDA APHIS). Available at: http://www.aphis.usda.gov/animal_welfare/downloads/awreports/awreport2006.pdf (Accessed 12.12.07).
5. Carbone, L. (2004). *What Animals Want: Expertise and Advocacy in Laboratory Animal Welfare Policy*, 291pp. Oxford, UK: Oxford University Press.
6. Office of Technology Assessment, US Congress (1986). *Alternatives to Animal Use in Research, Testing and Education*, OTA-BA-273, 437pp. Washington, DC, USA: US Government Printing Office.
7. Home Office (2007). *Statistics of Scientific Procedures on Living Animals: Great Britain 2006*, 49pp. London, UK: The Stationery Office.
8. O'Shea, D. (2000). *Johns Hopkins enters suit over lab animal regulations*. Press Release, 22 September, 2000. Baltimore, MD, USA: Johns Hopkins University.
9. Fishbein, E.A. (2001). What price mice? *Journal of the American Medical Association* **235**, 939–941.
10. Sauer, U.G., Kolar, R. & Rusche, B. (2005). The use of transgenic animals in biomedical research in Germany. Part 1: Status Report 2001–2003. [Die Verwendung transgener Tiere in der biomedizinischen Forschung in Deutschland. Teil 1: Sachstandsbericht 2001–2003.] *ALTEX* **22**, 233–246.
11. Anon. (2007). Swiss animal use statistics for 2005. *Pain & Distress Report* **7**, 2. Available at: http://www.hsus.org/pain_distress_report (Accessed 12.12.07).
12. Rusche, B. (2003). The 3Rs and animal welfare — conflict or the way forward? *ALTEX* **20** Suppl. 1, 63–76.
13. Combes, R.D., Balls, M., Bansil, L., Barratt, M., Bell, D., Botham, P., Broadhead, C., Clothier, R., George, E., Fentem, J., Jackson, M., Indans, I., Loizou, G., Navaratnam, V., Pentreath, V., Phillips, B., Stemplewski, H. & Stewart, J. (2004). The Third FRAME Toxicity Committee: Working toward greater implementation of alternatives in toxicity testing. *ATLA* **32** Suppl. 1B, 635–642.
14. Green, S. & Goldberg, A.M. (2004). TestSmart and toxic ignorance. *ATLA* **32** Suppl. 1A, 359–363.
15. Fenner-Crisp, P.A., Maciorowski, A.F. & Timm, G.E. (2000). The endocrine disruptor screening program developed by the US Environmental Protection Agency. *Ecotoxicology* **9**, 85–91.
16. Green, S., Goldberg, A.M. & Zurlo, J. (2001). The TestSmart-HPV program — Development of an integrated approach for testing high production volume chemicals. *Regulatory Toxicology & Pharmacology* **33**, 105–109.
17. Armstrong, T.W., Zaleski, R.T., Konkel, W.J. & Parkerton, T.J. (2002). A tiered approach to assessing children's exposure: a review of methods and data. *Toxicology Letters* **127**, 111–119.
18. Charles, G.D. (2004). *In vitro* models in endocrine disruptor screening. *ILAR Journal* **45**, 494–501.
19. Stokes, W.S. (2004). Selecting appropriate animal models and experimental designs for endocrine disruptor research and testing studies. *ILAR Journal* **45**, 387–393.
20. Louekari, K., Sihvonen, K., Kuittinen, M. & Sømnes, V. (2006). *In vitro* tests within the REACH information strategies. *ATLA* **34**, 377–386.
21. Sandusky, C., Even, M., Stoick, K. & Sandler, J. (2006). Strategies to reduce animal testing in US EPA's HPV program. *ALTEX* **23** Special Issue, 150–152.
22. Brom, F.W. (2002). Science and society: different bioethical approaches towards animal experimentation. *ALTEX* **19**, 78–82.
23. Festing, M.F.W. (2004). Is the use of animals in biomedical research still necessary in 2002? Unfortunately, "Yes". *ATLA* **32** Suppl. 1B, 733–739.
24. Pawlik, W.W. (1998). The significance of animals in biomedical research. [Znaczenie zwierzat w badaniach biomedycznych.] *Folia Medica Cracoviensia* **39**, 175–182.
25. Kjellmer, I. (2002). Animal experiments are necessary. Coordinated control functions are difficult to study without the use of nature's most complex systems: mammals and human beings. [Djurförsök är nödvändiga. Samordnade kontrollfunktioner låter sig svårligen studeras utan tillgång till naturens mest komplexa system: däggdjur och människa.] *Lakartidningen* **99**, 1172–1173.
26. Osswald, W. (1992). Ethics of animal research and application to humans. [Etica da investigação no animal e aplicação ao homem.] *Acta Medica Portuguesa* **5**, 222–225.
27. Greek, C.R. & Greek, J.S. (2002). *4th World Congress Point/Counterpoint: Is Animal Research Necessary in 2002?*, 54pp. Los Angeles, CA, US: Americans for Medical Advancement.
28. Singer, P. (1990). *Animal Liberation: A New Ethics for our Treatment of Animals*, 2nd edn, 320pp. New York, NY, USA: New York Review/Random House.
29. La Follette, H. & Shanks, N. (1994). Animal experimentation: the legacy of Claude Bernard. *International Studies in the Philosophy of Science* **8**, 195–210.
30. Greek, C.R. & Greek, J.S. (2000). *Sacred Cows and Golden Geese*, 242pp. New York, NY, USA: Continuum.
31. Greek, C.R. & Greek, J.S. (2002). *Specious Science*, 288pp. New York, NY, USA: Continuum.
32. Anon. (2006). Statement re: TGN1412. Available at: http://www.tegenero.com/news/statement_re_tgn1412/index.php (Accessed 18.04.06).
33. Anon. (2006). Frequently asked questions regarding TGN1412. Available at: http://www.tegenero.com/news/faqs_re_tgn1412/index.php (Accessed 18.04.06).
34. Bhogal, N. & Combes, R. (2006). TGN1412: time to change the paradigm for the testing of new pharmaceuticals. *ATLA* **34**, 225–229.
35. Coghlan, A. (2006). Mystery over drug trial debacle deepens. NewScientist.com news service, 14 August, 2006. Available at: http://www.newscientist.com/article.ns?id=dn9734 (Accessed 12.12.07).
36. Graham, D.J., Campen, D., Hui, R., Spence, M., Cheetham, C., Levy, G., Shoor, S. & Ray, W.A. (2005). Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet* **365**, 475–481.
37. Dahl, S.L. & Ward, J.R. (1982). Pharmacology, clinical efficacy, and adverse effects of the nonsteroidal anti-inflammatory agent benoxaprofen. *Pharmacotherapy* **2**, 354–366.
38. Gad, S.C. (1990). Model selection in toxicology: principles and practice. *Journal of the American College of Toxicology* **9**, 291–302.
39. Ross-Degnan, D., Soumerai, S.B., Fortess, E.E. &

Gurwitz, J.H. (1993). Examining product risk in context. Market withdrawal of zomepirac as a case study. *Journal of the American Medical Association* **270**, 1937–1942.
40. Peters, T.S. (2005). Do preclinical testing strategies help predict human hepatotoxic potentials? *Toxicologic Pathology* **33**, 146–154.
41. Venning, G.R. (1983). Identification of adverse reactions to new drugs. I: What have been the important adverse reactions since thalidomide? *British Medical Journal* **286**, 199–202.
42. Wallenstein, L. & Snyder, J. (1952). Neurotoxic reaction to chloromycetin. *Annals of Internal Medicine* **36**, 1526–1528.
43. Blum, M.D., Graham, D.J. & McCloskey, C.A. (1994). Temafloxacin syndrome: review of 95 cases. *Clinical Infectious Diseases* **18**, 946–950.
44. Mulder, P., Richard, V. & Thuillez, C. (1998). Different effects of calcium antagonists in a rat model of heart failure. *Cardiology* **89** Suppl. 1, 33–37.
45. Food and Drug Administration, US Department of Health and Human Services (2004). *Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products*, 31pp. Available at: http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.pdf (Accessed 12.12.07).
46. Lazarou, J. & Pomeranz, B. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Journal of the American Medical Association* **279**, 1200–1205.
47. Koppanyi, T. & Avery, M.A. (1966). Species differences and the clinical trial of new drugs: a review. *Clinical Pharmacology & Therapeutics* **7**, 250–270.
48. Villar, D., Buck, W.B. & Gonzalez, J.M. (1998). Ibuprofen, aspirin and acetaminophen toxicosis and treatment in dogs and cats. *Veterinary & Human Toxicology* **40**, 156–162.
49. Wilson, J.G., Ritter, E.J., Scott, W.J. & Fradkin, R. (1977). Comparative distribution and embryotoxicity of acetylsalicylic acid in pregnant rats and rhesus monkeys. *Toxicology & Applied Pharmacology* **41**, 67–78.
50. National Institutes of Health (2006). *Information on Clinical Trials and Human Research Studies*. Available at: http://clinicaltrials.gov/ct/info/whatis;jsessionid=B9D601AD55432DBDD59314931CA8385C#phases (Accessed 17.04.07).
51. Pound, P., Ebrahim, S., Sandercock, P., Bracken, M. & Roberts, I. (2004). Where is the evidence that animal research benefits humans? *British Medical Journal* **328**, 514–517.
52. Nuffield Council on Bioethics (2005). *The Ethics of Research Involving Animals*, 376pp. London, UK: Nuffield Council on Bioethics.
53. Anon. (2006). *Scopus* in detail: what does it cover? Available at: http://www.info.scopus.com/detail/what/ (Accessed 01.03.07).
54. National Center for Biotechnology Information (2006). *PubMed* overview. Available at: http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html (Accessed 14.04.07).
55. Lindl, T., Völkel, M. & Kolar, R. (2005). [Animal experiments in biomedical research. An evaluation of the clinical relevance of approved animal experimental projects.] [German.] *ALTEX* **22**, 143–151.
56. Lindl, T., Völkel, M. & Kolar, R. (2006). Animal experiments in biomedical research. An evaluation of the clinical relevance of approved animal experimental projects: No evident implementation in human medicine within more than 10 years. [Lecture abstract.] *ALTEX* **23**, 111.
57. Hackam, D.G. & Redelmeier, D.A. (2006). Translation of research evidence from animals to humans. *Journal of the American Medical Association* **296**, 1731–1732.
58. Hackam, D.G. (2007). Translating animal research into clinical benefit: poor methodological standards in animal studies mean that positive results may not translate to the clinical domain. *British Medical Journal* **334**, 163–164.
59. Knight, A. (2007). The poor contribution of chimpanzee experiments to biomedical progress. *Journal of Applied Animal Welfare Science* **10**, 281–308.
60. Conlee, K.M., Hoffeld, E.H. & Stephens, M.L. (2004). A demographic analysis of primate research in the United States. *ATLA* **32** Suppl. 1A, 315–322.
61. Morris, E. (Undated). *Sampling from Small Populations*. Available at: http://uregina.ca/~morrisev/Sociology/Sampling%20from%20small%20populations.htm (Accessed 12.12.07).
62. Guenther, W.C. (1973). A sample size formula for the hypergeometric. *Journal of Quality Technology* **5**, 167–170.
63. Green, J. (1982). Asymptotic sample size for given confidence interval length. *Applied Statistics* **31**, 298–300.
64. Macleod, M.R., O'Collins, T., Horky, L.L., Howells, D.W. & Donnan, G.A. (2005). Systematic review and meta-analysis of the efficacy of melatonin in experimental stroke. *Journal of Pineal Research* **38**, 35–41.
65. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group (1995). Tissue plasminogen activator for acute ischemic stroke. *New England Journal of Medicine* **333**, 1581–1588.
66. Chinese Acute Stroke Trial (CAST) Collaborative Group (1997). Randomised placebo-controlled trial of early aspirin use in 20,000 patients with acute ischaemic stroke. *Lancet* **349**, 1641–1649.
67. International Stroke Trial Collaborative Group (1997). The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, or both, or neither, among 19,435 patients with acute ischaemic stroke. *Lancet* **349**, 1569–1581.
68. Horn, J., de Haan, R.J., Vermeulen, M., Luiten, P.G.M. & Limburg, M. (2001). Nimodipine in animal model experiments of focal cerebral ischemia: a systematic review. *Stroke* **32**, 2433–2438.
69. O'Collins, V.E., Macleod, M.R., Donnan, G.A., Horky, L.L., van der Worp, B.H. & Howells, D.W. (2006). 1026 experimental treatments in acute stroke. *Annals of Neurology* **59**, 467–477.
70. Jonas, S., Aiyagari, V., Vieira, D. & Figueroa, M. (2001). The failure of neuronal protective agents versus the success of thrombolysis in the treatment of ischemic stroke: the predictive value of animal models. *Annals of the New York Academy of Sciences* **939**, 257–267.
71. Curry, S.H. (2003). Why have so many drugs with stellar results in laboratory stroke models failed in clinical trials? A theory based on allometric relationships. *Annals of the New York Academy of Sciences* **993**, 69–74.
72. Macleod, M.R., O'Collins, T., Horky, L.L., Howells, D.W. & Donnan, G.A. (2005). Systematic review and meta-analysis of the efficacy of FK506 in experimental stroke. *Journal of Cerebral Blood Flow & Metabolism* **25**, 1–9.

73. van der Worp, H.B., de Haan, P., Morrema, E. & Kalkman, C.J. (2005). Methodological quality of animal studies on neuroprotection in focal cerebral ischaemia. *Journal of Neurology* **252**, 1108–1114.
74. Willmot, M., Gray, L., Gibson, C., Murphy, S. & Bath, P.M. (2005). A systematic review of nitric oxide donors and L-arginine in experimental stroke; effects on infarct size and cerebral blood flow. *Nitric Oxide* **12**, 141–149.
75. Willmot, M., Gibson, C., Gray, L., Murphy, S. & Bath, P. (2005). Nitric oxide synthase inhibitors in experimental ischemic stroke and their effects on infarct size and cerebral blood flow: a systematic review. *Free Radical Biology & Medicine* **39**, 412–425.
76. Perel, P., Roberts, I., Sena, E., Wheble, P., Briscoe, C., Sandercock, P., Macleod, M., Mignini, L.E., Jayaram, P. & Khan, K.S. (2007). Comparison of treatment effects between animal experiments and clinical trials: systematic review. *British Medical Journal* **334**, 197–200.
77. Stroke Therapy Academic Industry Roundtable (1999). Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* **30**, 2752–2758.
78. Lucas, C., Criens-Poublon, L.J., Cockrell, C.T. & De Haan, R.J. (2002). Wound healing in cell studies and animal model experiments by Low Level Laser Therapy; were clinical studies justified? A systematic review. *Lasers in Medical Science* **17**, 110–134.
79. Roberts, I., Kwan, I., Evans, P. & Haig, S. (2002). Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. *British Medical Journal* **324**, 474–476.
80. Mapstone, J., Roberts, I. & Evans, P. (2003). Fluid resuscitation strategies: a systematic review of animal trials. *Journal of Trauma* **55**, 571–589.
81. Lee, D.S., Nguyen, Q.T., Lapointe, N., Austin, P.C., Ohlsson, A., Tu, J.V., Stewart, D.J. & Rouleau, J.L. (2003). Meta-analysis of the effects of endothelin receptor blockade on survival in experimental heart failure. *Journal of Cardiac Failure* **9**, 368–374.
82. Corry, D.B. & Kheradmand, F. (2005). The future of asthma therapy: integrating clinical and experimental studies. *Immunologic Research* **33**, 35–51.
83. Lazzarini, L., Overgaard, K.A., Conti, E. & Shirtliff, M.E. (2006). Experimental osteomyelitis: What have we learned from animal studies about the systemic treatment of osteomyelitis? *Journal of Chemotherapy* **18**, 451–460.
84. Scheld, W.M. (1987). Therapy of streptococcal endocarditis: correlation of animal model and clinical studies. *Journal of Antimicrobial Chemotherapy* **20** Suppl. A, 71–85.
85. Corpet, D.E. & Pierre, F. (2005). How good are rodent models of carcinogenesis in predicting efficacy in humans? A systematic review and meta-analysis of colon chemoprevention in rats, mice and men. *European Journal of Cancer* **41**, 1911–1922.
86. Roberts, I., Evans, A., Bunn, F., Kwan, I. & Crowhurst, E. (2001). Normalising the blood pressure in bleeding trauma patients may be harmful. *Lancet* **357**, 385–387.
87. Knight, A., Bailey, J. & Balcombe, J. (2006). Animal carcinogenicity studies: 1. Poor human predictivity. *ATLA* **34**, 19–27.
88. Tomatis, L. & Wilbourn, J. (1993). Evaluation of carcinogenic risk to humans: the experience of IARC. In *New Frontiers in Cancer Causation* (ed. O. Iversen), pp. 371–387. Washington, DC, USA: Taylor and Francis.
89. Haseman, K. (2000). Using the NTP database to assess the value of rodent carcinogenicity studies for determining human cancer risk. *Drug Metabolism Reviews* **32**, 169–186.
90. Huff, J. (2002). Chemicals studied and evaluated in long-term carcinogenesis bioassays by both the Ramazzini Foundation and the National Toxicology Program. *Annals of the New York Academy of Sciences* **982**, 208–230.
91. Ennever, F.K. & Lave, L.B. (2003). Implications of the lack of accuracy of the lifetime rodent bioassay for predicting human carcinogenicity. *Regulatory Toxicology & Pharmacology* **38**, 52–57.
92. Bailey, J., Knight, A. & Balcombe, J. (2005). The future of teratology research is *in vitro*. *Biogenic Amines* **19**, 97–145.
93. Olson, H., Betton, G., Stritar, J. & Robinson, D. (1998). The predictivity of the toxicity of pharmaceuticals in humans from animal data — an interim assessment. *Toxicology Letters* **102–103**, 535–538.
94. International Agency for Research on Cancer (IARC) (1972–1992). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans,* Volumes 1–55. Lyon, France: IARC.
95. International Agency for Research on Cancer (IARC) (undated). *IARC Monographs Programme on the Evaluation of Carcinogenic Risks to Humans*. Available at: http://monographs.iarc.fr (Accessed 01.01.04).
96. Rall, D.P. (2000). Laboratory animal tests and human cancer. *Drug Metabolism Reviews* **2**, 119–128.
97. Ashby, J. & Purchase, I.F.H. (1993). Will all chemicals be carcinogenic to rodents when adequately evaluated? *Carcinogenesis* **8**, 489–495.
98. Shirai, T., Fukushima, S., Ohshima, M. & Ito, N. (1984). Effects of butylated hydroxyanisole, butylated hydroxytoluene, and NaCl on gastric carcinogenesis initiated with *N*-methyl-*N*-nitro-*N*-nitrosoguanidine in F344 rats. *Journal of the National Cancer Institute* **72**, 1189–1198.
99. Fung, V., Barrett, J. & Huff, J. (1995). The carcinogenesis bioassay in perspective: application in identifying human hazards. *Environmental Health Perspectives* **103**, 680–683.
100. Gold, L.S., Bernstein, L., Magaw, R. & Slone, T.H. (1989). Interspecies extrapolation in carcinogenesis: prediction between rats and mice. *Environmental Health Perspectives* **81**, 211–219.
101. Gold, L.S., Slone, T.H. & Ames, B.N. (1998). What do animal cancer tests tell us about human cancer risk? Overview of analyses of the carcinogenic potency database. *Drug Metabolism Reviews* **30**, 359–404.
102. Johnson, F.M. (2001). Response to Tennant *et al.*: Attempts to replace the NTP rodent bioassay with transgenic alternatives are unlikely to succeed. *Environmental Molecular Mutagenesis* **37**, 89–92.
103. Bailey, J. (2005). Non-human primates in medical research and drug development: a critical review. *Biogenic Amines* **19**, 235–255.
104. Glazko, G., Veeramachaneni, V., Nei, M. & Makalowski, W. (2005). Eighty percent of proteins are different between humans and chimpanzees. *Gene* **346**, 215–219.
105. Balcombe, J., Barnard, N. & Sandusky, C. (2004). Laboratory routines cause animal stress. *Contemp-*

orary Topics in Laboratory Animal Science **43**, 42–51.
106. Knight, A., Bailey, J. & Balcombe, J. (2006). Animal carcinogenicity studies: 2. Obstacles to extrapolation of data to humans. *ATLA* **34**, 29–38.
107. Poignet, H., Nowicki, J.P. & Scatton, B. (1992). Lack of neuroprotective effect of some sigma ligands in a model of focal cerebral ischemia in the mouse. *Brain Research* **596**, 320–324.
108. Aronowski, J., Strong, R. & Grotta, J.C. (1996). Treatment of experimental focal ischemia in rats with lubeluzole. *Neuropharmacology* **35**, 689–693.
109. Marshall, J.W., Cross, A.J., Jackson, D.M., Green, A.R., Baker, H.F. & Ridley, R.M. (2000). Clomethiazole protects against hemineglect in a primate model of stroke. *Brain Research Bulletin* **52**, 21–29.
110. Bebarta, V., Luyten, D. & Heard, K. (2003). Emergency medicine animal research: does use of randomisation and blinding affect the results? *Academic Emergency Medicine* **10**, 684–687.
111. Medical Research Council (MRC) (1993). *Responsibility in the Use of Animals in Medical Research*, 12pp. London, UK: MRC.
112. Balls, M., Festing, M.F.W. & Vaughan, S. (eds) (2004). Reducing the use of experimental animals where no replacement is yet available. *ATLA* **32** Suppl. 2, 1–104.
113. Festing, M.F.W. (2004). Good experimental design and statistics can save animals, but how can it be promoted? *ATLA* **32** Suppl. 1A, 133–135.
114. De Boo, J. & Hendriksen, C. (2005). Reduction strategies in animal research: a review of scientific approaches at the intra-experimental, supra-experimental and extra-experimental levels. *ATLA* **33**, 369–377.
115. Festing, M.F.W. (1997). Experimental design and husbandry. *Experimental Gerontology* **32**, 39–47.
116. van Wilgenburg, H., van Schaick Zillesen, P.G. & Krulichova, I. (2003). Sample power and ExpDesign: tools for improving design of animal experiments. *Laboratory Animals* **32**, 39–43.
117. van Wilgenburg, H., van Schaick Zillesen, P.G. & Krulichova, I. (2004). Experimental design: computer simulation for improving the precision of an experiment. *ATLA* **32** Suppl. 1B, 607–611.
118. Mead, R. (1988). *The Design of Experiments*, 634pp. New York, NY, USA: Cambridge University Press.
119. Balcombe, J. (2006). Laboratory environments and rodents' behavioural needs: a review. *Laboratory Animals* **40**, 217–235.
120. Eskola, S., Lauhikari, M., Voipio, H., Laitinen, M. & Nevalainen, T. (1999). Environmental enrichment may alter the number of rats needed to achieve statistical significance. *Scandinavian Journal of Laboratory Animal Science* **26**, 134–144.
121. Schauber, E.M. & Edge, W.D. (1999). Statistical power to detect main and interactive effects on the attributes of small-mammal populations. *Canadian Journal of Zoology* **77**, 68–73.
122. Festing, M.F.W. & Altman, D.G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR Journal* **43**, 244–257.
123. Phillips, C.J.C. (2005). Meta-analysis — A systematic and quantitative review of animal experiments to maximise the information derived. *Animal Welfare* **14**, 333–338.
124. Festing, M.F.W., Baumans, V., Combes, R.D., Halder, M., Hendriksen, C.F.M., Howard, B.R., Lovell, D.P., Moore, G.J., Overend, P. & Wilson, M.S. (1998). Reducing the use of laboratory animals in biomedical research: problems and possible solutions. *ATLA* **26**, 283–301.
125. Balls, M., Goldberg, A.M., Fentem, J.H., Broadhead, C.L., Burch, R.L., Festing, M.F.W., Frazier, J.M., Hendriksen, C.F., Jennings, M., van der Kamp, M.D., Morton, D.B., Rowan, A.N., Russell, C., Russell, W.M.S., Spielmann, H., Stephens, M.L., Stokes, W.S., Straughan, D.W., Yager, J.D., Zurlo, J. & Van Zutphen, B.F. (1995). The Three Rs: the way forward: The report and recommendations of ECVAM Workshop 11. *ATLA* **23**, 838–866.
126. Evidence-Based Medicine Working Group (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *Journal of the American Medical Association* **286**, 2420–2425.
127. Watters, M.P.R. & Goodman, N.W. (1999). Comparison of basic methods in clinical studies and *in vitro* tissue and cell culture studies in three anaesthesia journals. *British Journal of Anaesthesia* **82**, 295–298.
128. Moher, D., Schulz, K.F. & Altman, D.G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* **357**, 1191–1194.
129. Arlt, S. & Heuwieser, W. (2005). [Evidence based veterinary medicine.] [German.] *Deutsche Tierärztliche Wochenschrift* **112**, 146–148.
130. Schulz, K.F. (2005). Assessing allocation concealment and blinding in randomised controlled trials: why bother? *Equine Veterinary Journal* **37**, 394–395.
131. Brown, C.M., Calder, C., Linton, C., Small, C., Kenny, B.A., Spedding, M. & Patmore, L. (1995). Neuroprotective properties of lifarizine compared with those of other agents in a mouse model of focal cerebral ischaemia. *British Journal of Pharmacology* **115**, 1425–1432.
132. Oktem, I.S., Menku, A., Akdemir, H., Kontas, O., Kurtsoy, A. & Koc, R.K. (2000). Therapeutic effect of tirilazad mesylate (U-74006F), mannitol, and their combination, on experimental ischemia. *Research in Experimental Medicine* **199**, 231–242.
133. Houdebine, L.M. (2007). Transgenic animal models in biomedical research. *Methods in Molecular Biology* **360**, 163–202.
134. Sauer, U.G., Kolar, R. & Rusche, B. (2006). [The use of transgenic animals in biomedical research in Germany. Part 2: Ethical evaluation of the use of transgenic animals in biomedical research and perspectives for the changeover in research to research animal-free methods.] [German.] *ALTEX* **23**, 3–16.
135. Balls, M., Blaauboer, B.J., Fentem, J.H., Bruner, L., Combes, R.D., Ekwall, B., Fielder, R.J., Guillouzo, A., Lewis, R.W., Lovell, D.P., Reinhardt, C.A., Repetto, G., Sladowski, D., Spielmann, H. & Zucco, F. (1995). Practical aspects of the validation of toxicity test procedures: The report and recommendations of ECVAM Workshop 5. *ATLA* **23**, 129–147.
136. Balls, M. & Combes, R. (2005). The need for a formal *invalidation* process for animal and non-animal tests. *ATLA* **33**, 299–308.
137. Hoffmann, S. & Hartung, T. (2006). Toward an evidence-based toxicology. *Human & Experimental Toxicology* **25**, 497–513.
138. Curren, R.D., Southee, J.A., Spielmann, H., Liebsch, M., Fentem, J.H. & Balls, M. (1995). The role of prevalidation in the development, validation and acceptance of alternative methods. *ATLA* **23**, 211–217.

139. US Interagency Coordinating Committee on the Validation of Alternative Methods in National Institutes of Health (1997). *Validation and Regulatory Acceptance of Toxicological Test Methods. A Report of the* ad hoc *Interagency Coordinating Committee on the Validation of Alternative Methods*, 123pp. Research Triangle Park, NC, USA: National Institute of Environmental Health Sciences.
140. Organisation for Economic Cooperation and Development (OECD) (2003). *OECD Series on Testing and Assessment: No. 34: Guidance Document on the Validation and International Acceptance of New and Updated Test Methods for Hazard Assessment, Environment Directorate*, 96pp. Paris, France: OECD.
141. Balls, M. & Combes, R. (2006). Validation via weight-of-evidence approaches. *ALTEX* **23**, 332–335.
142. European Centre for the Validation of Alternative Methods (ECVAM), Joint Research Centre, European Commission Directorate General (Undated). About ECVAM. Available at: http://ecvam.jrc.cec.eu.int/index.htm (Accessed 12.12.07).
143. European Centre for the Validation of Alternative Methods, Joint Research Centre, European Commission Directorate General (Undated). Validated methods. Available at: http://ecvam.jrc.cec.eu.int/index.htm (Accessed 12.12.07).
144. Balls, M. & Karcher, W. (1995). The validation of alternative test methods. *ATLA* **23**, 884–886.
145. O'Connor, A.M. (1997). Barriers to regulatory acceptance. In *Animal Alternatives, Welfare and Ethics* (ed. L.F.M. van Zutphen & M. Balls), pp. 1173–1176. Amsterdam, The Netherlands: Elsevier Science B.V.
146. Balls, M. (2004). Are animal tests inherently valid? *ATLA* **32** Suppl. 1B, 755–758.