Can you trust your animal study data?

Ian S. Peers, Marie C. South, Peter R. Ceuppens, Jonathan D. Bright and Elizabeth Pilling

The development of new medicines is costly and risky. As data in a recent article on attrition rates indicate (Phase II and Phase III attrition rates 2011–2012. *Nature Rev. Drug Discov.* **12**, 569 (2013))¹, the majority of Phase II clinical trials fail and most of these programmes will have used animals in the preceding discovery research phases. In the dash from bench to bedside, we need high-quality, internally valid and reproducible animal models to provide assurance that drug targets are valid, to identify safety liabilities and to avoid poor dose prediction and over-optimistic interpretation of preclinical efficacy.

So, how confident can we be in the robustness and reproducibility of preclinical animal data? Here we report the results of systematic statistical reviews of proposed animal study protocols at AstraZeneca over a recent 18-month period and identify the aspects of preclinical animal studies that are most often open to improvement. We summarize the impact of these reviews, illustrate statistical added value with an example and provide estimated costs of delivering these benefits.

Rationale and outcomes of analysis

Why do we need to improve the quality of preclinical animal studies? In the United States, it is estimated that upwards of 0.5 billion animals are used in biomedical research each year and the numbers seem to have been increasing since the mid-1980s². Despite strenuous efforts in the United Kingdom to reduce the number of animals used in research, this has not been sustained. In the past decade, the overall number of procedures has increased by approximately 1 million to 3.79 million in 2011 (REF. 3). Given the numbers of animals used in biomedical research, and the availability of a comparatively small number of validated alternatives to animal models, we are unlikely to see a drastic reduction in laboratory animal use any time soon. However, the evidence suggests that a lack of study robustness and reproducibility in preclinical research, including animal work,

persists, thus increasing the risk of poor decisions and contributing to the attrition of early-stage drug projects^{4–7}.

We therefore conclude that we have an ethical and scientific obligation to improve the quality of preclinical animal studies, ensuring that where they are used, they include as few animals as possible and are run in a way that provides reproducible and valid information. We believe that a systematic approach to reviewing animal study protocols against key statistical principles supports this goal, providing a robust challenge to study design, analysis and interpretation, thereby helping to ensure the internal validity of the studies and increasing confidence in resulting decisions.

Improvement opportunities in design and conduct of animal studies. Despite national training and support for scientists in the 3Rs (replacement, reduction and refinement in the use of animals in research) in the United Kingdom, tightly regulated animal procedures, engagement with Animal Welfare and Ethics Review Boards (AWERBs) and the best intent of scientists, statisticians in AstraZeneca have observed some studies where the statistical grounding could have been improved. These incidental examples gathered over more than 10 years from several therapeutic areas and different countries, prompted the development of a Good Statistical Practice (GSP) *in vivo* programme to assure systematic and consistent quality by design in all internal *in vivo* studies at AstraZeneca.

To substantiate our incidental observations and to identify further improvement opportunities, we aimed to complete systematic joint statistical reviews with *in vivo* scientists of protocols for planned animal studies in the 18-month period between June 2012 and November 2013. Each protocol was classified as either standard or non-standard. Standard studies are routine designs run repeatedly, usually with different compounds. These are planned for review every 3 years unless changes to their purpose or design are made. Non-standard

studies are new or novel and a review must be completed for each study. Scientists are responsible for initiating statistical input, and we set ourselves a target of completing a prior GSP review for at least 80% of studies conducted between January 2013 and November 2013. To deliver these reviews and improvements in study quality we estimated that a minimum of three statistical full-time roles were required globally in addition to resources from our scientific colleagues. The GSP programme sits alongside relevant local and national regulations and guidelines for laboratory animal work.

Outcomes of systematic statistical reviews.

Between June 2012 and November 2013, 255 protocols were jointly reviewed by statisticians and scientists against nine key principles. Options to improve the proposed design, conduct or analyses were discussed and agreed changes were documented.

TABLE 1 shows the principles, ranked according to how often improvements relating to each principle were made during reviews. For example, 57% (144) of reviews led to a change in or a documented justification for the number of animals used. Some reviews led to changes in more than one principle. More than 50% of all reviews resulted in a meaningful change to at least one of the statistical principles. These improvements are likely to be an underestimate of the broader positive impact of the GSP programme, as some studies run externally on behalf of AstraZeneca also benefited from statistical input.

From January 2013 to November 2013, 2,607 separate animal studies were conducted internally by AstraZeneca across all therapy areas and discovery-enabling functions. Of these, 82% (2,140) were covered by a completed statistical review prior to the start of the study. The number of studies covered by prior review greatly exceeds the number of reviews conducted because standard protocol reviews cover multiple studies. Across different therapy areas and enabling functions the coverage varied from 13% (20 of 160) to 98% (316 of 323). This variation is, in part, due to the type of studies conducted and the levels of previous engagement with statisticians. Areas and functions tended to have lower coverage if they had a high proportion of non-standard studies and historically less engagement with statisticians.

FIGURE 1 provides an example of the benefits of engaging with statisticians when planning and conducting *in vivo* experiments. It shows a design modification to a

Table 1 | Improvements implemented based on a review against nine statistical principles*

Characteristic	Question	Number of protocols amended (%)
Animal numbers	Is the number of animals per group appropriate and justified?	144 (56.5%)
Analysis	Are planned statistical analyses clearly defined?	70 (27.5%)
Randomization	Are allocation procedures unbiased (of animals to cages, cages to groups and so on)?	51 (20.0%)
Experimental procedures	Are in-life and sample processing procedures unbiased?	48 (18.8%)
Design	Is the proposed experimental design efficient and effective?	46 (18.0%)
Blocking	Is the experiment split into appropriate subunits if it cannot be processed as a single unit?	32 (12.5%)
Monitoring (standard studies only)	Is the model being appropriately monitored for ongoing consistent performance?	18 (7.1%)
Controls	Are all appropriate control groups included?	16 (6.3%)
Blinding	Is blinding used where needed?	11 (4.3%)

*A total of 255 animal study protocols in the 18-month period between June 2012 and November 2013 were jointly reviewed against the nine statistical principles above by scientists and statisticians, together with the number and proportion of protocols with improvements implemented relating to each principle.

pig coagulopathy study that resulted in a 36% reduction in study duration and a 36% reduction in the number of animals used.

Conclusions

Animal experiments have contributed much to our understanding of mechanisms of disease and selection of drug candidates, but their value in both cases is diminished if studies are not internally valid and robust. In our analysis, we identified improvements in the design, conduct and/or analysis in the majority of cases we reviewed. This supports our own previous observations, and those in the literature, that it is both possible and desirable to improve the quality of *in vivo* studies through the rigorous application of statistical principles.

The systematic nature of the review, documenting action against nine statistical principles for each study, allows us to have greater confidence that good practice is underpinning our project decisions more consistently globally and that the *in vivo* data generated are more robust. It also had unexpected benefits, including the involvement of the statistician in decision-making during the study when unplanned events had consequences related to the GSP principles.

The nine principles between them provide confidence that a study has a suitable design and size, with precautions to control bias and variation, and an appropriate means of analysing and interpreting the results to support effective decision-making. We suggest that there are likely to be opportunities for improvements in all the above areas

wherever biomedical researchers conduct *in vivo* research with limited access to statistical expertise. We particularly suggest that AWERBs systematically engage expert statistical input.

We have developed a repository of *in vivo* statistical reviews, a valuable resource for retrospective evaluation that will enable anyone to see how a study should run, why it was designed that way, and how it should be analysed alongside the study outcomes.

Although a review of individual studies does not directly address the question of translation of animal models, we suggest that translation can only be properly evaluated once *in vivo* models are internally valid and run reproducibly.

Ian S. Peers, Marie C. South, Peter R. Ceuppens, Jonathan D. Bright and Elizabeth Pilling are at AstraZeneca, Alderley Park, Macclesfield SK10 4TG, UK. Correspondence to I.S.P. e-mail: ian.peers@astrazeneca.com

doi: 10.1038/nrd4090-c1

1. Arrowsmith, J. & Miller, P. Phase II and Phase III attrition rates 2011–2012. *Nature Rev. Drug Discov.* **12**, 569 (2013).
2. Rice, M. J. The institutional review board is an impediment to human research: the result is more animal-based research. *Philos. Eth. Humanit. Med.* **6**, 12 (2011).
3. Home Office. *Statistics of Scientific Procedures on Living Animals, Great Britain 2011*. London Stationery Office (2012).
4. Landis, S. C. *et al.* A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490**, 187–191 (2012).
5. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
6. Peers, I. S. *et al.* In search of preclinical robustness. *Nature Rev. Drug Discov.* **11**, 733–734 (2012).
7. Prinz, F. *et al.* Believe it or not: how much can we rely on published data on potential drug targets? *Nature Rev. Drug Discov.* **10**, 712 (2011).

Acknowledgements
The authors would like to thank S. Robinson, K. Pritchard, K. Nelander and K. Hansson, together with all the other AstraZeneca scientists and statisticians who made valuable contributions to the *in vivo* Good Statistical Practice (GSP) programme.

Competing interests statement
The authors are employees of AstraZeneca.

Figure 1 | Example of the impact of statistical input: adaptive design of a pig coagulopathy study to identify the minimal effective dose. Compared with an originally proposed non-adaptive design using 56 animals and taking 28 weeks, this revised approach enabled the identification of a 2 mg per kg dose as the minimal effective dose (within the dose range of interest) using 36 animals (36% reduction), and taking 18 weeks (36% reduction). Animal totals below each arm included concurrent controls required for interpretation: eight positive, eight vehicle and four negative.