An Analysis of the Use of Animal Models in Predicting Human Toxicology and Drug Safety

Jarrod Bailey,[1] Michelle Thew[1] and Michael Balls[2]

[1]British Union for the Abolition of Vivisection (BUAV), London, UK; [2]c/o Fund for the Replacement of Animals in Medical Experiments (FRAME), Nottingham, UK

Summary — Animal use continues to be central to preclinical drug development, in spite of a lack of its demonstrable validity. The current nadir of new drug approvals and the drying-up of pipelines may be a direct consequence of this. To estimate the evidential weight given by animal data to the probability that a new drug may be toxic to humans, we have calculated Likelihood Ratios (LRs) for an extensive data set of 2,366 drugs, for which both animal and human data are available, including tissue-level effects and MedDRA Level 1–4 biomedical observations. This was done for three preclinical species (rat, mouse and rabbit), to augment our previously-published analysis of canine data. In common with our dog analysis, the resulting LRs show: a) that the absence of toxicity in the animal provides little or virtually no evidential weight that adverse drug reactions (ADRs) will also be absent in humans; and b) that, while the presence of toxicity in these species can add considerable evidential weight for human risk, the LRs are extremely inconsistent, varying by over two orders of magnitude for different classes of compounds and their effects. Therefore, our results for these additional preclinical species have important implications for their use in predicting human toxicity, and suggest that alternative methods are urgently required.

Key words: *animals, dog, drug development, mouse, preclinical testing, rabbit, rat, toxicology.*

Address for correspondence: *Jarrod Bailey, British Union for the Abolition of Vivisection (BUAV), 16a Crane Grove, London N7 8NN, UK.*
E-mail: *jarrod.bailey@mac.com*

Introduction

There exists a general assumption that animal testing helps to ensure human safety and the efficacy of new pharmaceuticals. Such preclinical testing is required by regulatory agencies worldwide (e.g. 1, 2), and involves at least two species — typically one rodent and one non-rodent species — ostensibly to aid the prediction of toxicity and pharmacokinetics. However, while there is little supportive evidence for the value or necessity of this practice (3), it continues, with apparent disregard for the dearth of new drug approvals and the drying-up of pipelines over the past decade (e.g. 4–7) that may be a consequence of the way in which the preclinical testing is currently performed.

Animals are used in significant numbers for these purposes. In the UK alone, in 2012, drug testing involved the use of more than 277,000 animals (8). The human relevance and predictive nature of these animal models have been investigated relatively rarely and then only superficially. Such a lack of evaluation (see *Discussion*) could be considered surprising, given the central role of the models in drug development (9–11), but it is chiefly due to the difficulty in accessing relevant data, most of which are unpublished and proprietary to pharmaceutical companies. Almost invariably, the human relevance of preclinical animal tests has been measured via 'concordance' metrics (e.g. 12), which have been interpreted by various authors as the true-positive rate (sensitivity) or the Positive Predictive Value (PPV). While these metrics are appropriate for assessing the reliability of a diagnostic test for a specific disorder (e.g. HIV infection), the insights they provide depend critically on the question being asked of the diagnostic test. However, they are not appropriate for assessing the salient question at issue with animal models, which is *whether or not they contribute significant weight to the evidence for or against the likely toxicity of a given compound in humans*.

Overcoming this key problem — almost entirely overlooked by previous authors — requires a precise specification of the various terms used (see *Methods*). Briefly, the appropriate metrics are Likelihood Ratios (LRs; 13): the Positive Likelihood Ratio (PLR) and the inverse Negative Likelihood Ratio (iNLR). Therefore, there is clearly a need for the kind of statistically appropriate critical analysis that we provide here. The data set we have used is unique, in that it is large and allows the conditional probabilities required for the LRs (PLR/iNLR) to be calculated.

The analysis presented here comprises data sets for the rat, mouse and rabbit. This complements our recently published analysis of dog preclinical

data (14), which contains more-complete details of the statistical methods. A précis of these methods is provided below.

Methods

Animal models are widely used to assess the likelihood that a given compound will prove toxic or non-toxic in humans. As with any diagnostic test, their reliability can only be assessed by performing tests in which the same compound is given to both animals and humans, and the presence or absence of toxicity is recorded. This leads to a 2 × 2 matrix of results, as shown in Table 1 (15).

The basis of this matrix is that the human data are correct, and the animal data are true/false, if they do/do not match them. The various cells in this matrix allow a variety of diagnostic metrics to be deduced, of which the most familiar and widely used are the true-positive rate for the test (or 'sensitivity' = a/[a + c]), and the true-negative rate (or 'specificity' = d/[d + b]). In previous research into the reliability of animal models as predictors of toxicity in humans, some authors (e.g. 12) have focused on the sensitivity, expressed as the 'true-positive concordance rate', or the so-called Positive Predictive Value (PPV), given by a/(a + b), which reflects the probability that human toxicity was correctly identified by the animal model, given that toxicity was observed in the animal model (e.g. 16). However, neither of these metrics is suitable for the role of assessing the evidential weight provided by any toxicity test. In the case of animal models, the sensitivity addresses only the ability of such models to detect toxicity that will subsequently manifest itself in humans. This is a necessary, but not sufficient, measure of evidential weight. Suppose, for example, that the animal model always indicates toxicity found in humans; it would then have a sensitivity of 100%. However, if, in addition, the model always indicated toxicity, even when such toxicity was not subsequently seen in humans, its evidential value would be no better than simply dismissing *every* compound as toxic from the outset. Thus, a useful toxicity test must also be able to give insight into when toxicity seen in the animal model is *not* observed in humans, which requires knowledge of the *specificity* of the test.

There is, of course, an obvious reason for the focus on sensitivity in animal model evaluation: if a compound is found to be positive for toxicity in an animal model, it is unlikely to go forward into human evaluation. Nevertheless, the fact remains that sensitivity alone cannot be an adequate guide to the value of animal models.

The case of the PPV is more subtle. This metric is a measure of the probability that human toxicity will be correctly identified, given that the animal model detected toxicity. As such, PPVs are conditional probabilities, the condition being the pre-existence of a positive animal test result. This makes PPVs dependent on the prevalence of toxicity in compounds, so it is an inappropriate measure of the reliability of the test with any specific compound (e.g. 13, 17).

Thus, any appropriate metric of the evidential value of animal models requires knowledge of *both* the sensitivity *and* the specificity of the model. This, in turn, implies that the appropriate metrics for the evidential weight provided by an animal model are LRs (e.g. 17). In general, these are ratios of functions of the sensitivity and specificity, which can be extracted from the 2 × 2 matrix given in Table 1. In the case of animal models, in general, two LRs are relevant. The first is the so-called PLR, which is given by:

$$\begin{aligned} \text{PLR} &= \text{sensitivity}/(1 - \text{specificity}) \\ &= (a/[a + c])/(b/[b + d]) \end{aligned}$$

It should be noted that, in a relatively small number of cases (see Table 2), this equation results in an undefined value when there are no observations in the animal (i.e. b = 0 in the 2 × 2 matrix). These cases were eliminated from consideration. This LR captures the ability of an animal model to add evidential weight to the belief that a specific compound is toxic. Any animal model that gives a PLR that is statistically significantly higher than 1.0, can be regarded as contributing evidential weight to the probability that the compound under test will be toxic in humans.

Table 1: A 2 × 2 matrix of results

	Compound toxic in humans	**Compound not toxic in humans**
Compound toxic in animal model	a: true positives (TPs)	b: false positives (FPs)
Compound not toxic in animal model	c: false negatives (FNs)	d: true negatives (TNs)

Table 2: The number of classifications of adverse effects for each species, as used in this analysis

Species	Tissue-level effects	Biomedical observations (BMOs)	Total classifications used	Classifications eliminated
Rat	62	548	610	271/881
Mouse	62	342	404	266/670
Rabbit	54	221	275	141/416
Dog	52	384	436	14/450

The numbers of classifications of effects for each species, and therefore the numbers of LRs calculated for each species, are shown. The total number of classifications used in the analysis is shown in column 3 (Total classifications used), which comprises the sum of column 1 (Tissue-level effects) and column 2 (BMOs). The number of BMO classifications for which there were no effects observed in the preclinical species of interest, and which were therefore eliminated from consideration in our analysis, are shown in column 4 (Classifications eliminated), out of the total number of classifications for which there were data. In total, 3,275 comparisons were made for each human–animal pair (human–rat, human–mouse, human–rabbit), for 2,366 compounds.

The other relevant LR is the so-called iNLR, given by:

$$\text{iNLR} = \text{specificity}/(1 - \text{sensitivity}) = (d/[b + d])/(c/[a + c])$$

This LR captures the ability of an animal model to add evidential weight to the belief that a specific compound is not toxic: any animal model that gives an iNLR that is statistically significantly higher than 1.0 can be regarded as contributing evidential weight to the probability that the compound under test will not be toxic in humans.

At this point, it is worth noting that the above definitions imply that a good animal model for detecting human toxicity is not necessarily also good for detecting an absence of toxicity. That is, a high PLR does not guarantee a high iNLR; this will emerge as a key issue in this study.

The above definitions also underscore the need for data on the human toxicity of compounds that fail initial animal tests. Again, a key feature of the current study is that this issue has been substantively overcome — at least, as much as it could ever be overcome — via data mining methods. Data were obtained from a leading pharmaceutical safety consultancy, Instem Scientific Limited (Harston, Cambridge, UK; www.instem-lss.com; 'Safety Intelligence Programme'), with funding provided by FRAME. All the information stemmed from publicly accessible sources, including: PubMed (www.ncbi.nlm.nih.gov/pubmed), the FDA Adverse Event Reporting System (FAERS), DrugBank (www.drugbank.ca), and the National Toxicology Program (ntp.niehs.nih.gov). Human and preclinical species data were available for more than 2,300 drug compounds.

Inference of the good quality of the data used in this evaluation is outlined in the *Discussion*. Compounds were selected that feature in the FAERS, FDA New Drug Applications (FDA NDAs) and DrugBank. Thus, the drugs selected for this analysis have undergone preclinical testing and are (or have been) in clinical use: human and animal data are therefore available for them. Prior to our analysis, the data provided to us by Instem had been processed — thus, a non-redundant list of parent moieties was created, for example, by normalising therapeutic products to their generic names (e.g. Lipitor to Atorvastatin). This yielded 2,366 compounds. A signature of the effects of each compound was then created, focusing on tissue-level effects (e.g. bradycardia and arrhythmic disorder would both be considered to be effects on heart tissues), as well as the individual observations, which were mapped to their MedDRA (Medical Dictionary for Regulatory Activities; www.meddramsso.com) counterparts. MedDRA observations are classified into four levels, Level 1 being the most specific, and Level 4 providing a more generic 'system organ class'. These classifications help to eliminate false positives (FPs) that may arise from species-specific observations, and help the identification of concordant observations that might otherwise have been missed by their 'rolling up' into more generic terms. A threshold of a minimum of five observations in both humans and the preclinical species had been applied, presumably to avoid the inclusion of effects considered to be 'rare'.

LRs were derived for both broad and tissue-level effects, as well as more specific biomedical observations (BMOs) mapped to MedDRA classifications (Level 1 [most specific] to Level 4 [more generic 'system organ class']). The numbers of classifications of effects for each species, and therefore the numbers of LRs calculated for each species, are shown in Table 2. Also shown are the numbers of BMO classifications not involving the species of interest, which were eliminated from further consideration. In total, 3,275 comparisons were made

for each human–animal pair (human–rat, human–mouse, human–rabbit), for 2,366 compounds. The Instem Scientific data on which our analysis was based are shown in our complementary paper (14), and the full set of data, including 95% Confidence Intervals, will be available on the *ATLA* website (www.atla.org.uk).

With regard to potential bias: false negatives (FNs) are more common than FPs, since there is a bias resulting from a 'precautionary principle' not to progress positives to human administration. This has been mitigated by limiting the data set to compounds reported in the FAERS database. Therefore, all the compounds are certain to have proceeded to market, and animal preclinical data are available for these compounds. Specific details of how the FPs that were identified arose were not sought, because they were not pertinent to this analysis and it was not feasible, given the nature of the data set. It must be assumed that the animal data were correlated with the human data retrospectively, and/or the human data arose from post-marketing studies, and/or clinical trials were applied for and approved, since the adverse effect(s) in animals were minor and/or mitigated by other data.

Results

Median LRs and ranges are shown in Table 3. All the PLRs were generally high: median values were 101 (rabbit), 203 (mouse) and 253 (rat), which compare favourably to the median PLR of 28 for the dog (14). In common with the canine data, these values suggest that compounds showing toxicity in these animal species are also likely to be toxic in humans. However, the range of PLRs for each of these species varies enormously. The PLR ranges were: 13–1,348 (rabbit); 23–2,361 (mouse); and 24–2,360 (rat). These ranges are considerably greater than those seen for the dog, i.e. 5–549 (14), meaning that, with no obvious pattern regarding the form of toxicity, the reliability of this aspect of animal models cannot be generalised or regarded with confidence. In contrast, the median iNLRs are substantially lower: 1.12 (rabbit); 1.39 (mouse); and 1.82 (rat). While the ranges of iNLRs, if not for the rabbit (1.01–2.33) but certainly for the mouse (1.03–50) and the rat (1.02–100), were much greater than the range for the dog iNLR values (1.01–1.92), the medians compare only slightly favourably to the median iNLR of 1.10 for the dog (14).

One major caveat regarding these data, and the associated ranges of LR values, is that they incorporate a significant number of adverse effects and events that are rarely caused and/or reported. In those specific cases, the ability to reliably estimate the sensitivity and specificity of the dog tests may be compromised. The frequency of rare events is depicted for each species in Figure 1, in which histograms show that, for the preclinical species, an average of one third (33%) of ADR classes have sample sizes of ≤ 10, and more than half (55%) have sample sizes of ≤ 20. For humans — for which the threshold for rare events was set higher on account of greater sample sizes — 64% and 78% of ADR classes have sample sizes of ≤ 100 and ≤ 200, respectively. The range of sample sizes for each species is also depicted in Figure 2, via 'box and whisker' plots. For each species, a grouped histogram representing those in Figure 1 is shown for reference, alongside the associated quantile and outlier box plots. These show clearly how, while the range of sample sizes for the ADR classifications is broad (as indicated by the maximum and minimum whiskers), the majority of the sample sizes are toward the lower end of the range, shown by the box that represents the middle 50% of the distribution.

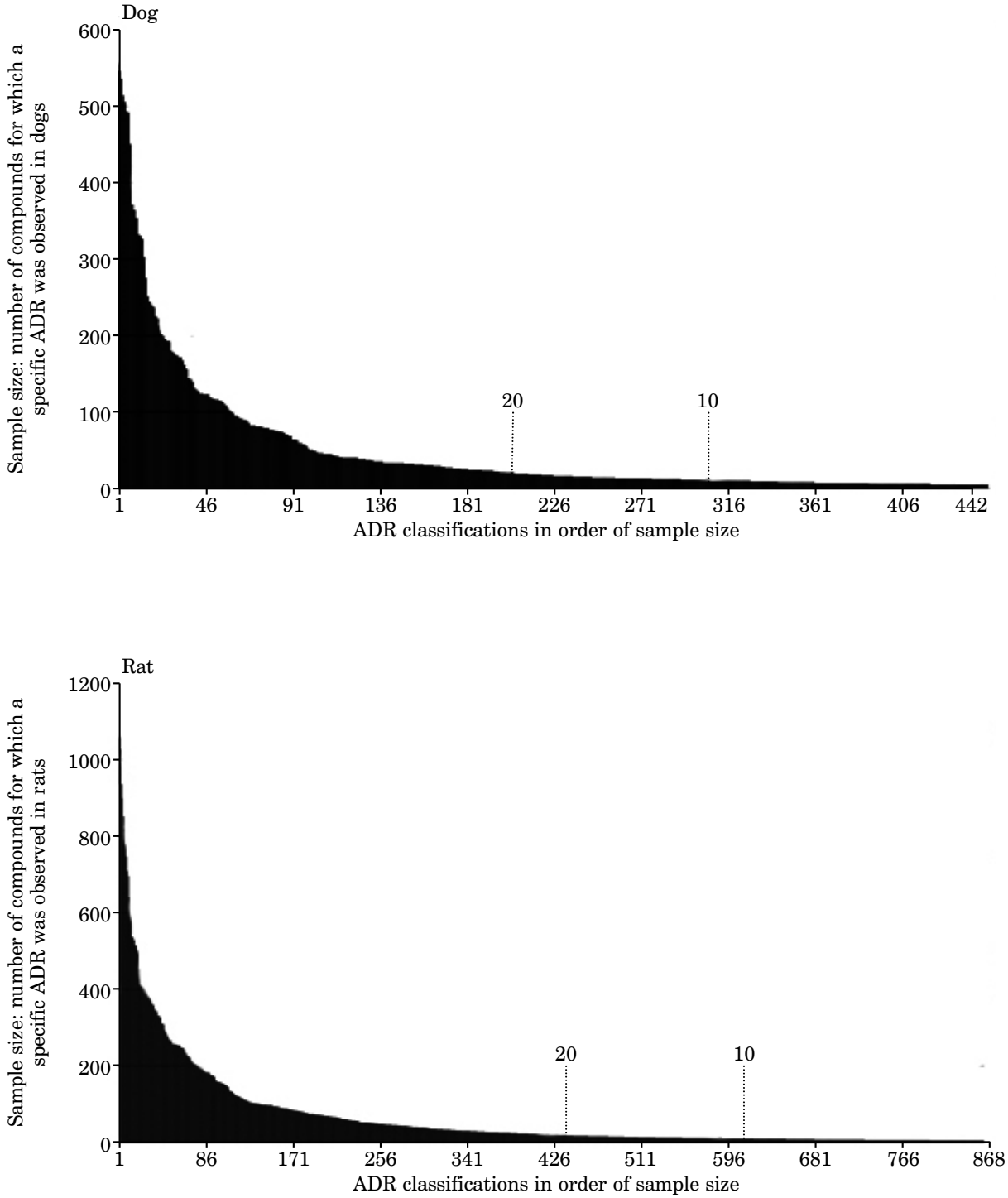
In order to assess the impact of including the rare events/low sample sizes in our analysis, we recalculated the PLRs and iNLRs for each species comparison, with the rare events in each preclinical species removed (Figures 3a and 3b), and with the rare events in both the preclinical species and humans removed (Figures 3c and 3d). Notably, removing rare events significantly reduces the PLR for each species, particularly for the rat and the mouse. With regard to iNLR values for each species, the removal

Table 3: Median LRs and ranges for the rat, mouse and rabbit

Species	PLR (median)	iNLR (median)	PLR range	iNLR range
Rat	253	1.82	24–2360	1.02–100
Mouse	203	1.39	23–2361	1.03–50
Rabbit	101	1.12	13–1348	1.01–2.33
Dog	28	1.10	5–549	1.01–1.92

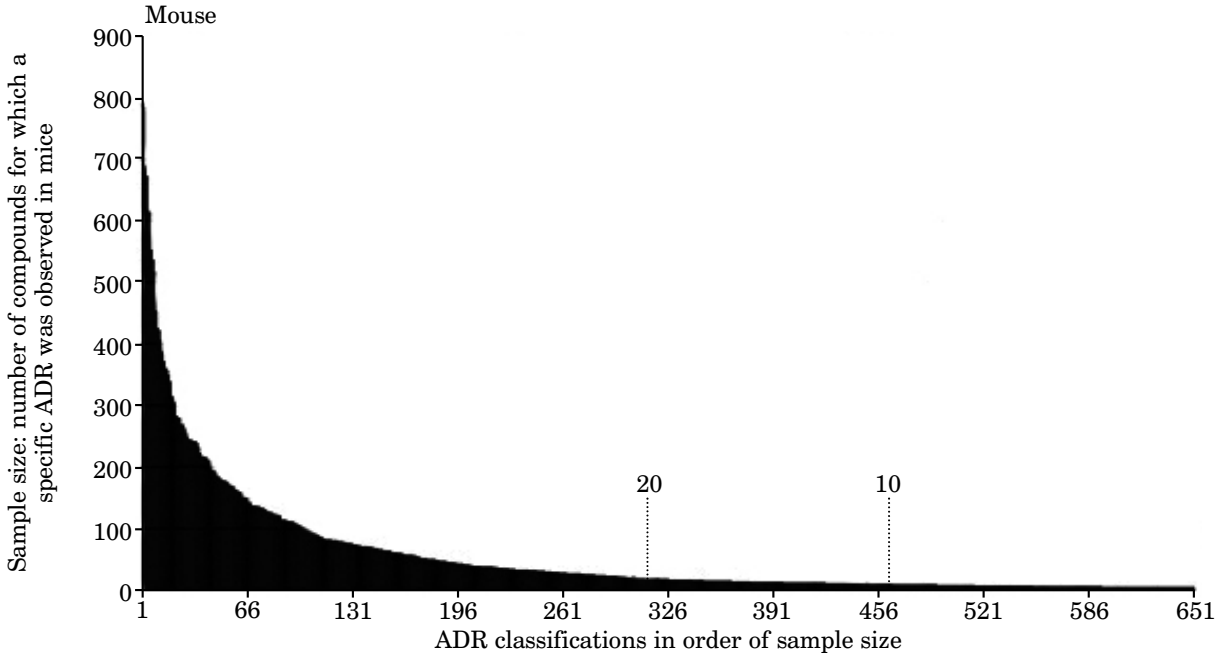
All the PLRs were generally high, and compared favourably to those for the dog, suggesting that compounds showing toxicity in those animals are also likely to be toxic in humans. However, high ranges, with no obvious pattern of toxicity, suggest the reliability of this aspect cannot be generalised or regarded with confidence. Median iNLRs were substantially lower, and compared only slightly favourably to those for the dog, supporting the view that animals provide very little or essentially no evidential weight to this aspect of toxicity testing.

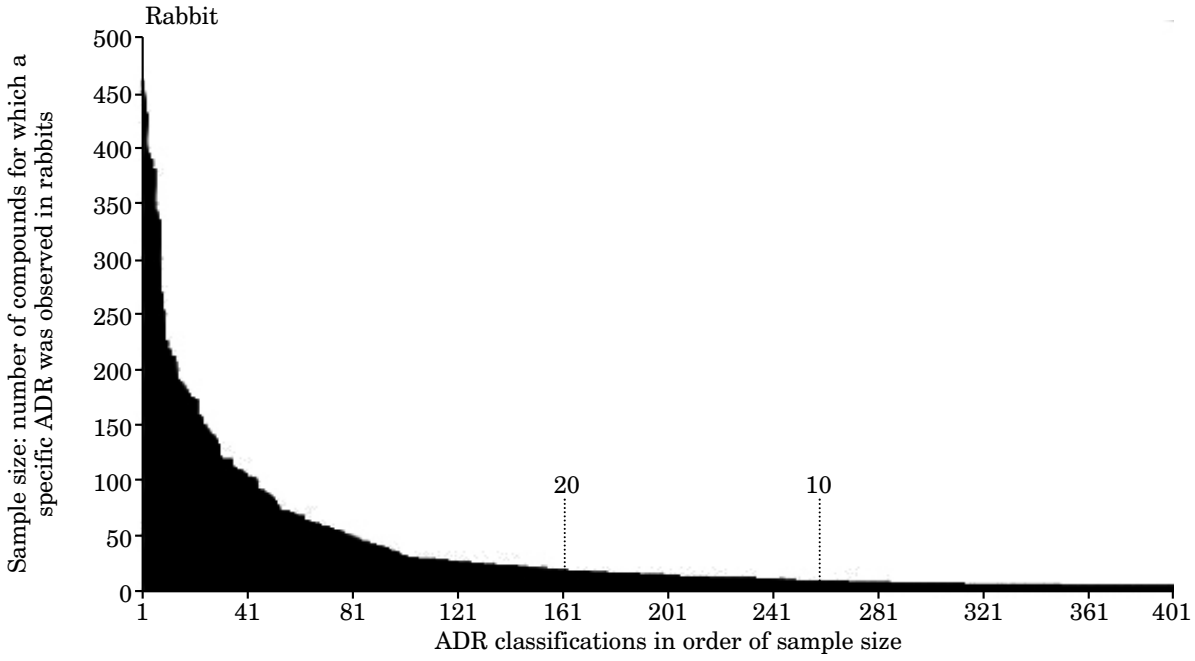
Figure 1: Sample sizes for each ADR class



For each species (including humans), the sample size (number of compounds for which a specific ADR was observed) is shown for each ADR class. Markers indicate two thresholds of rare events: for each preclinical species, sample sizes of ≤ 10 and ≤ 20, and for humans (given the much greater sample sizes), ≤ 100 and ≤ 200. The columns to the right of each marker represent ADR classes for which there are ≤ the indicated sample sizes. Many ADR classes have small sample sizes for each species. For the preclinical species, an average of one third (33%) of ADR classes have sample sizes of ≤ 10, and more than half (55%) have sample sizes of ≤ 20. For humans, 64% and 78% of ADR classes have sample sizes of ≤ 100 and ≤ 200, respectively.

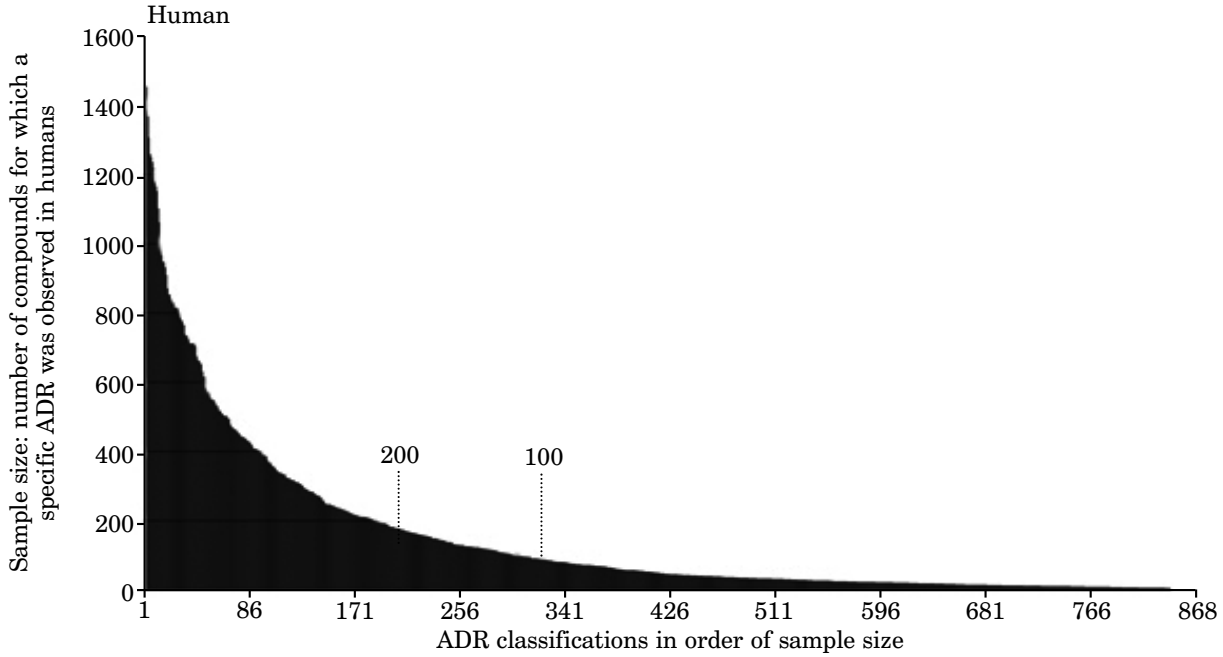
Figure 1: continued





For each species (including humans), the sample size (number of compounds for which a specific ADR was observed) is shown for each ADR class. Markers indicate two thresholds of rare events: for each preclinical species, sample sizes of ≤ 10 and ≤ 20, and for humans (given the much greater sample sizes), ≤ 100 and ≤ 200. The columns to the right of each marker represent ADR classes for which there are ≤ the indicated sample sizes. Many ADR classes have small sample sizes for each species. For the preclinical species, an average of one third (33%) of ADR classes have sample sizes of ≤ 10, and more than half (55%) have sample sizes of ≤ 20. For humans, 64% and 78% of ADR classes have sample sizes of ≤ 100 and ≤ 200, respectively.

Figure 1: continued



For each species (including humans), the sample size (number of compounds for which a specific ADR was observed) is shown for each ADR class. Markers indicate two thresholds of rare events: for each preclinical species, sample sizes of ≤ 10 and ≤ 20, and for humans (given the much greater sample sizes), ≤ 100 and ≤ 200. The columns to the right of each marker represent ADR classes for which there are ≤ the indicated sample sizes. Many ADR classes have small sample sizes for each species. For the preclinical species, an average of one third (33%) of ADR classes have sample sizes of ≤ 10, and more than half (55%) have sample sizes of ≤ 20. For humans, 64% and 78% of ADR classes have sample sizes of ≤ 100 and ≤ 200, respectively.

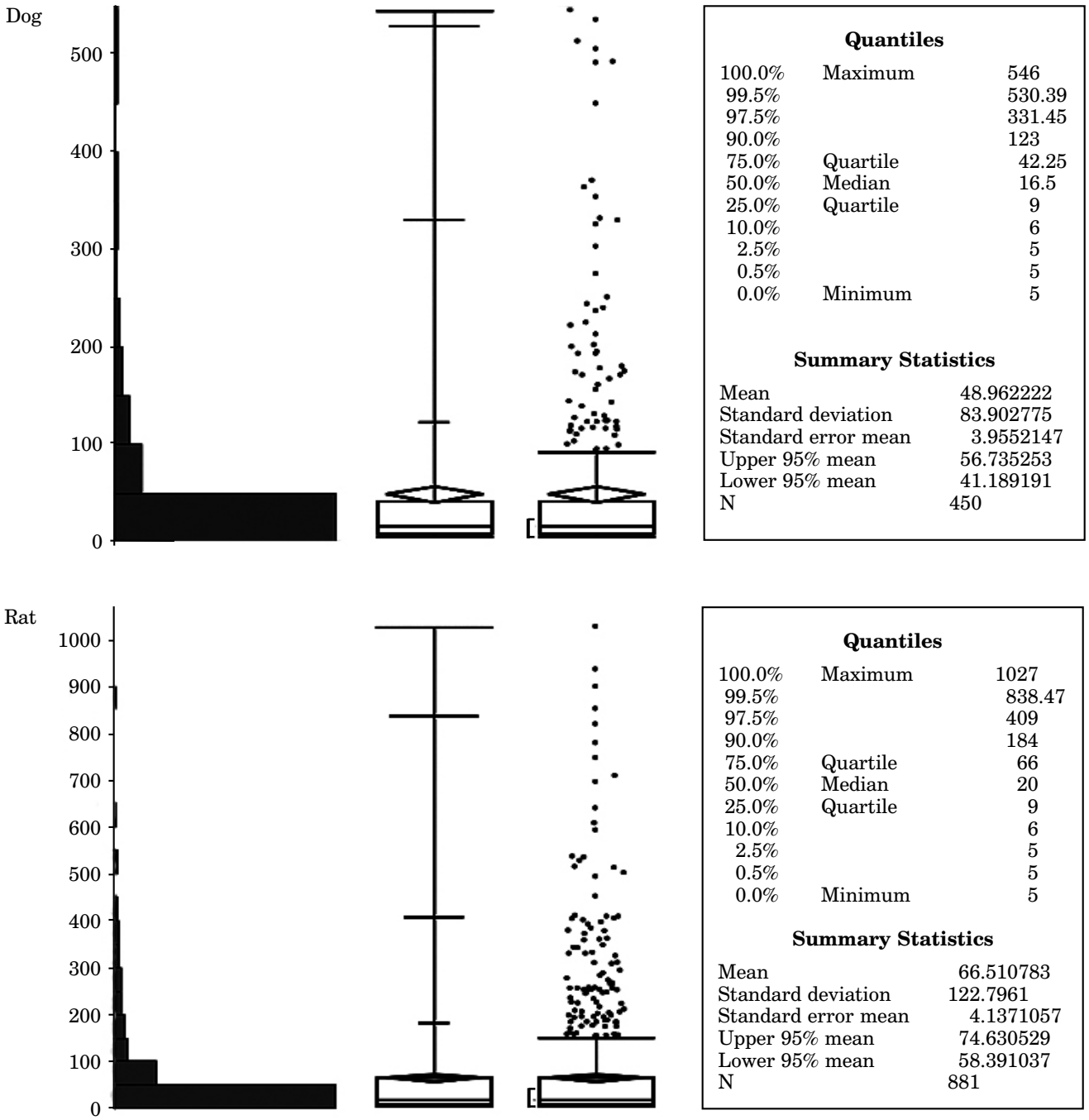
of rare events in the preclinical species results in a marginal increase in iNLR in each case, though, when rare events in humans are also removed, the iNLRs for the rat and the mouse decrease further. These observations are further supported by the scatter plots shown in Figure 4. These plots show that observed PLRs with higher values tend to have lower sample sizes, both in the preclinical animal and in humans, and that more-robust PLRs with higher sample sizes tend to have lower values. Conversely, observed iNLRs with higher values tend to have higher sample sizes, both in the preclinical animal and in humans, and more-robust iNLRs with higher sample sizes also tend to have higher values, though the increase in iNLR with sample size is slight. In summary, taking account of rare events, to illustrate their impact on our results and how making our results more statistically robust affects them, augments our conclusions. The evidential weight provided by animal tests given a positive toxicology result, is significantly decreased, most conspicuously for the rat and the mouse. The evidential weight provided by animal tests given a negative toxicology result, is also decreased for the rat and the mouse; while it is marginally increased for the rabbit and the dog, the values remain extremely low.

Therefore, in common with the canine data, our analysis of data from these other species supports the view that animals provide very little or essentially no evidential weight to this aspect of toxicity testing. Specifically, the fact that a compound shows no toxic effects in animals provides essentially no insight into whether the compound will also show no toxic effects in humans. This lack of evidential weight has important implications for the role of animals in toxicity testing, especially for the pharmaceutical industry. The critical observation for deciding whether a candidate drug can proceed to testing in humans, is the absence of toxicity in tests on animals. However, our findings show that the predictive value of the animal test in this regard is barely greater than that which would result merely by chance (see below).

Discussion

The analysis presented here complements and augments our recently published similar analysis of canine toxicity data (14). These analyses are

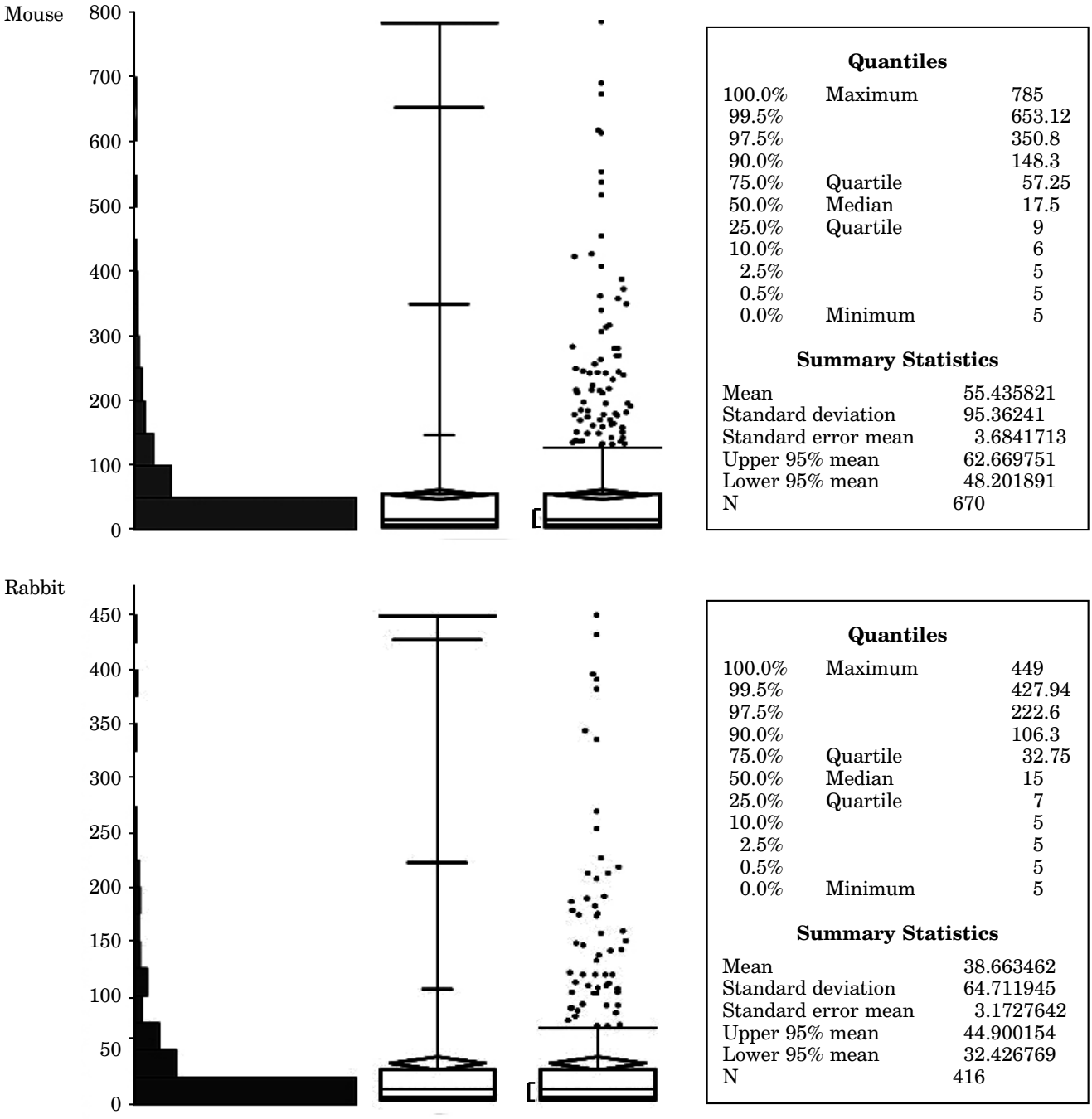
Figure 2: Box plots to illustrate sample-size distribution



For each species (including humans), the distribution of sample sizes (number of compounds for which a specific ADR was observed) is illustrated via 'box and whisker' plots. In each case, column 1 depicts a condensed 'grouped' sample size histogram, using the same data that generated the histograms in Figure 1, for ease of comparison. Column 2 shows a standard quantile box plot, and column 3 an outlier box plot. The 'box' is bounded by the lower (25%) and upper (75%) quartiles, and therefore represents the interquartile range, i.e. the 50% of sample sizes that lie either side of the median value, which itself is shown by the horizontal line bisecting the box. The mean value is indicated by the diamond. The maximum and minimum sample sizes are shown by the whiskers at the top and bottom of the plots, respectively. For the outlier plot in column 3, outlier values are shown as dots, which are greater than 1.5× the interquartile range above the upper (75%) quartile.

The precise values for each marker are provided in the accompanying box (Quantiles and Summary Statistics) beside the plots. These box plots demonstrate that many ADR classes have small sample sizes, for each species. For the preclinical species, an average of one third (33%) of ADR classes have sample sizes of ≤ 10, and more than half (55%) have sample sizes of ≤ 20. For humans, 64% and 78% of ADR classes have sample sizes of ≤ 100 and ≤ 200, respectively.

Figure 2: continued

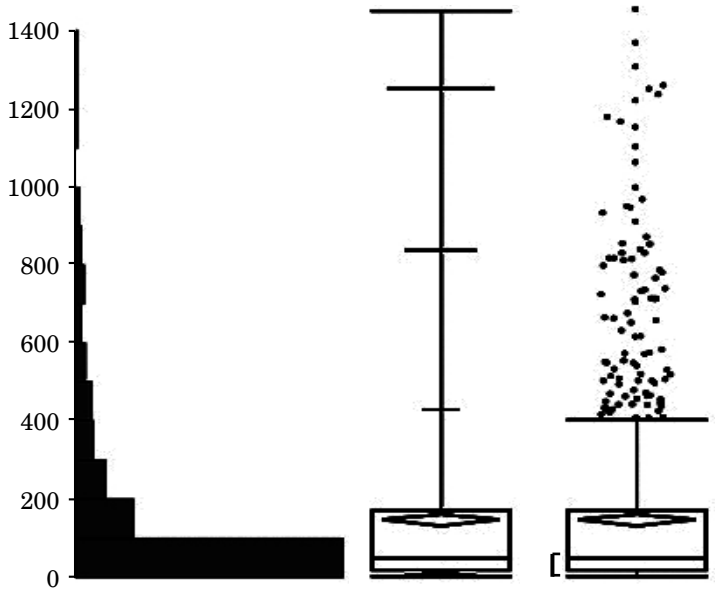


For each species (including humans), the distribution of sample sizes (number of compounds for which a specific ADR was observed) is illustrated via 'box and whisker' plots. In each case, column 1 depicts a condensed 'grouped' sample size histogram, using the same data that generated the histograms in Figure 1, for ease of comparison. Column 2 shows a standard quantile box plot, and column 3 an outlier box plot. The 'box' is bounded by the lower (25%) and upper (75%) quartiles, and therefore represents the interquartile range, i.e. the 50% of sample sizes that lie either side of the median value, which itself is shown by the horizontal line bisecting the box. The mean value is indicated by the diamond. The maximum and minimum sample sizes are shown by the whiskers at the top and bottom of the plots, respectively. For the outlier plot in column 3, outlier values are shown as dots, which are greater than 1.5× the interquartile range above the upper (75%) quartile.

The precise values for each marker are provided in the accompanying box (Quantiles and Summary Statistics) beside the plots. These box plots demonstrate that many ADR classes have small sample sizes, for each species. For the preclinical species, an average of one third (33%) of ADR classes have sample sizes of ≤ 10, and more than half (55%) have sample sizes of ≤ 20. For humans, 64% and 78% of ADR classes have sample sizes of ≤ 100 and ≤ 200, respectively.

Figure 2: continued

Human



Quantiles

100.0%	Maximum	1454
99.5%		1255.31
97.5%		838.5
90.0%		429.8
75.0%	Quartile	174
50.0%	Median	50
25.0%	Quartile	20
10.0%		11
2.5%		6
0.5%		5
0.0%	Minimum	5

Summary Statistics

Mean	148.76731
Standard deviation	226.61868
Standard error mean	7.6349773
Upper 95% mean	163.7522
Lower 95% mean	133.78242
N	881

For each species (including humans), the distribution of sample sizes (number of compounds for which a specific ADR was observed) is illustrated via 'box and whisker' plots. In each case, column 1 depicts a condensed 'grouped' sample size histogram, using the same data that generated the histograms in Figure 1, for ease of comparison. Column 2 shows a standard quantile box plot, and column 3 an outlier box plot. The 'box' is bounded by the lower (25%) and upper (75%) quartiles, and therefore represents the interquartile range, i.e. the 50% of sample sizes that lie either side of the median value, which itself is shown by the horizontal line bisecting the box. The mean value is indicated by the diamond. The maximum and minimum sample sizes are shown by the whiskers at the top and bottom of the plots, respectively. For the outlier plot in column 3, outlier values are shown as dots, which are greater than 1.5× the interquartile range above the upper (75%) quartile.

The precise values for each marker are provided in the accompanying box (Quantiles and Summary Statistics) beside the plots. These box plots demonstrate that many ADR classes have small sample sizes, for each species. For the preclinical species, an average of one third (33%) of ADR classes have sample sizes of ≤ 10, and more than half (55%) have sample sizes of ≤ 20. For humans, 64% and 78% of ADR classes have sample sizes of ≤ 100 and ≤ 200, respectively.

urgently required, to support informed debate about the value of animal models in preclinical testing. It is acknowledged among some stakeholders (if not universally among all stakeholders) that assessment of the scientific value of animal data in drug development is necessary, has been scarce, and has been thwarted for decades by the lack of availability of relevant data for analysis (e.g. 18–20). Nevertheless, primarily due to concerns over privacy and commercial interests, data sharing and making data available continue to be resisted, in spite of assurances to the contrary from the industry (18).

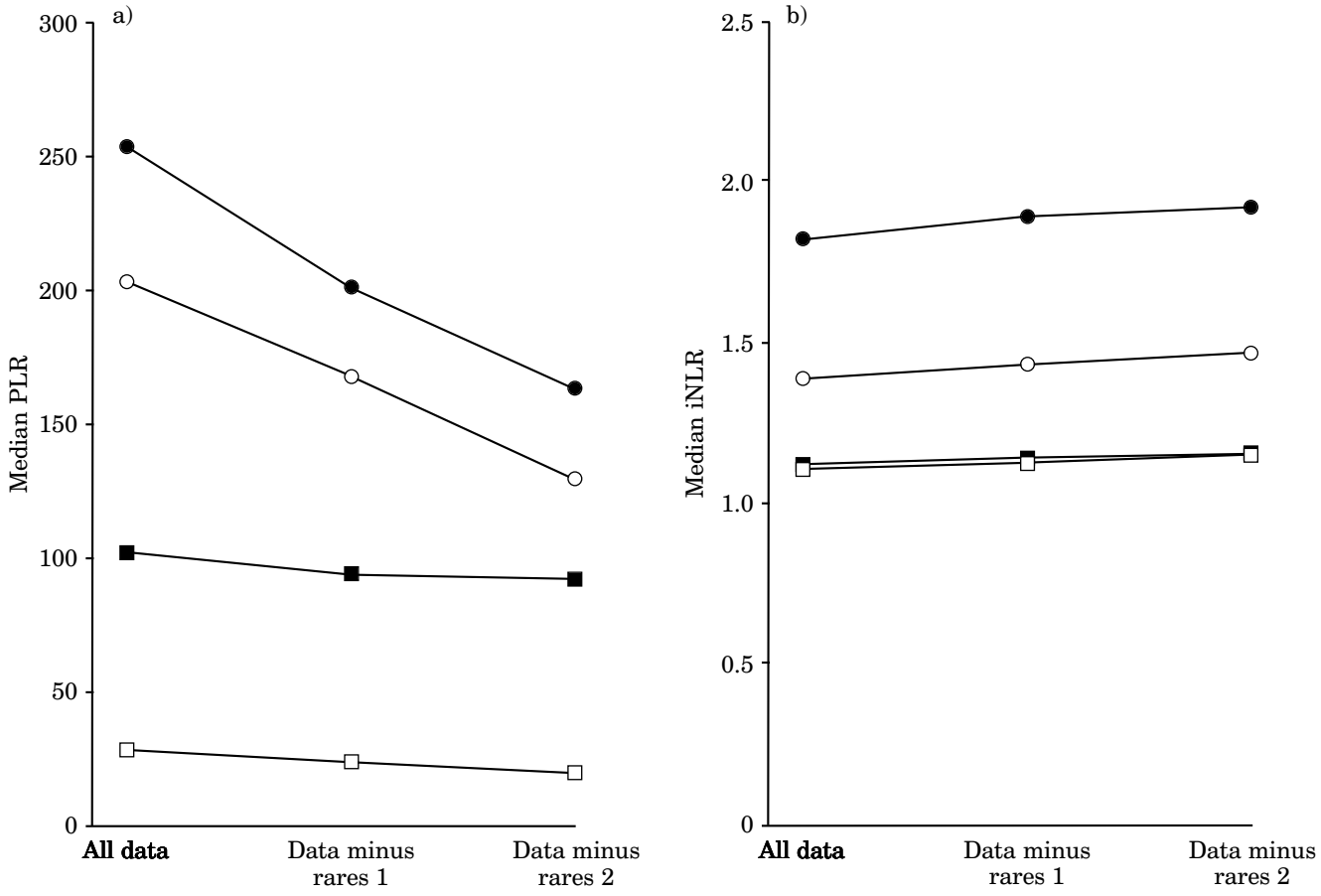
Those few analyses that have been done, tend to reflect unfavourably on animal models. In 2012, a study that expressly set out to minimise bias, showed that 63% of serious ADRs had no counterparts in animals, and less than 20% of serious ADRs had an actual positive corollary in animal studies (21). Other similar examples exist for testing generally (e.g. 22–24), and more-specifically, for example, in teratology (e.g. 25, 26) and drug-induced liver injury (e.g. 3, 27). One notable study claimed a good concordance between animal and human toxicology (12), though neither the predictive nature of the animal data for humans, nor the evidential weight provided by those data, were addressed (28).

We have, for the first time, addressed the salient question of the *contribution of evidential weight for or against the toxicity of a given compound in humans* by data from animal tests, by using the appropriate metrics of LRs. Furthermore, we have applied the apposite LRs to a data set of an unprecedented scale, to critically question the value of the use of the main preclinical animal species in the testing of new human pharmaceuticals.

Substantiation of data quality is evidenced by: a) the methods used to source the data and the assured quality of the supplying databases (listed earlier); b) the ways in which the data had been used recently as a basis for scientific publications and presentations (e.g. 29–32); and c) the international corporate and academic clients that have used the consultancy and its data (e.g. AstraZeneca; see 29–32). In addition, the impact of 'missing data' (i.e. unpublished data held by pharmaceutical companies) was mitigated by strictly limiting the data set to drugs "with the greatest chance of having been evaluated in all the species included in the study". In other words, "…lack of evidence for an association between a compound and a specific BMO demonstrates a real absence of effect, and is not due to missing data" (direct quotes from the Instem Scientific Ltd Analysis Report, unpublished).

Naturally, there must be caveats. Our analysis was limited to data that are published and publicly available. It is widely acknowledged that many animal experimental results/preclinical data remain unpublished and/or proprietary, for a variety of reasons (e.g. 21, 33–36). Such publication bias is a major problem (e.g. 37–40), and, compounded by

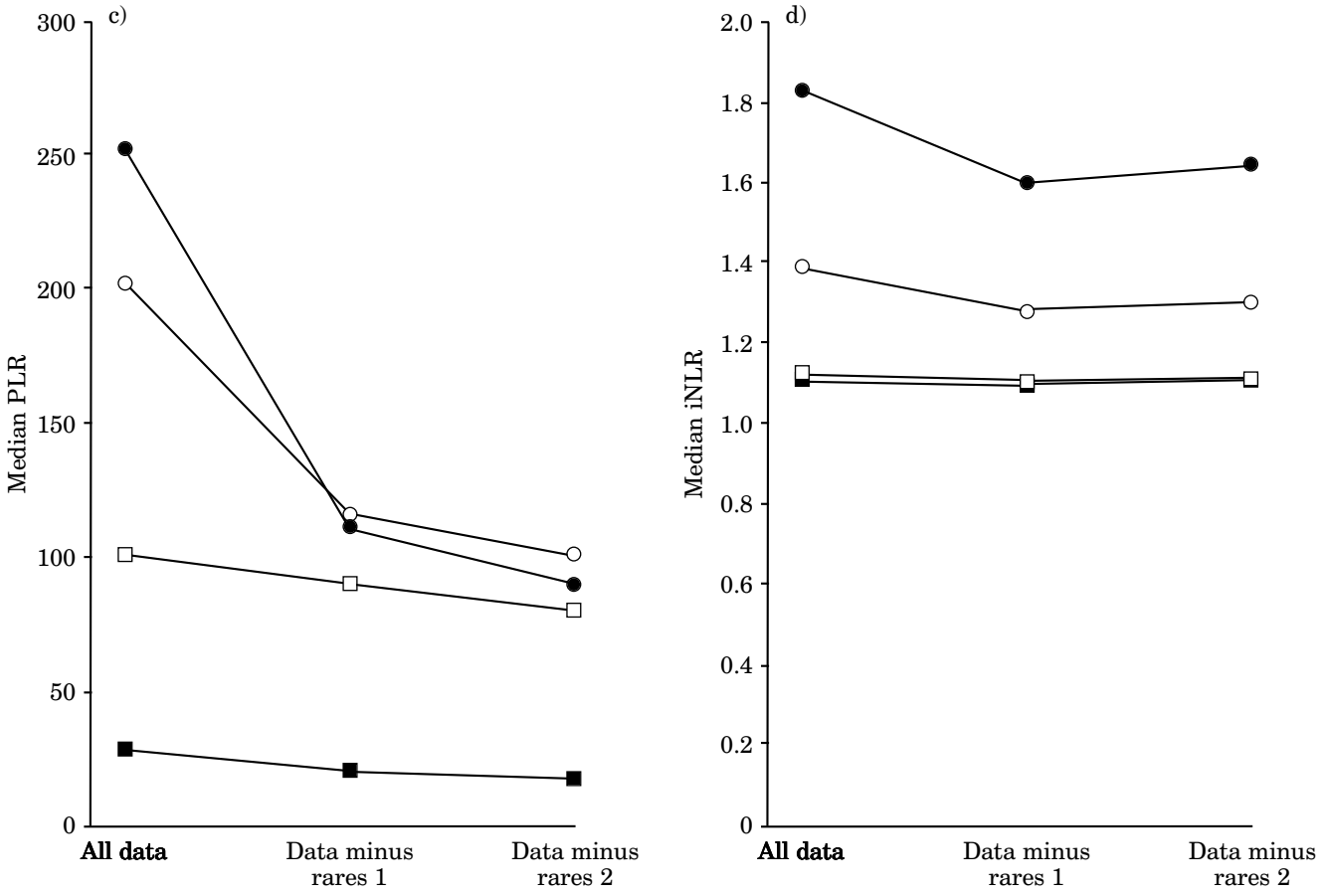
Figure 3: Impact of rare ADRs on LRs



● = *Rat;* ○ = *mouse;* ■ = *rabbit;* □ = *dog.*

Median PLRs and iNLRs were recalculated for each preclinical species–human data set, with rare events removed for illustrative purposes. Graphs a and b show the median PLR and iNLR values, respectively, with rare events removed from the animal data only; graphs c and d show the median PLR and iNLR values, respectively, with rare events removed from both the animal data set and the human data set. For each animal species–human pair, the first point on each line shows the median PLR or iNLR for the entire data set; the second point (data minus rares 1) shows the median PLR or iNLR for the data set with rare events below the first threshold removed (see the Results *section); and the third point (data minus rares 2) shows the median PLR or iNLR for the data set with rare events below the second threshold removed (see the* Results *section). Removing rare events significantly reduces the PLR for each species, particularly for the rat and the mouse. It results in a marginal increase in iNLR in each case, although, when rare events in humans are also removed, the iNLRs for the rat and the mouse decrease further.*

Figure 3: continued



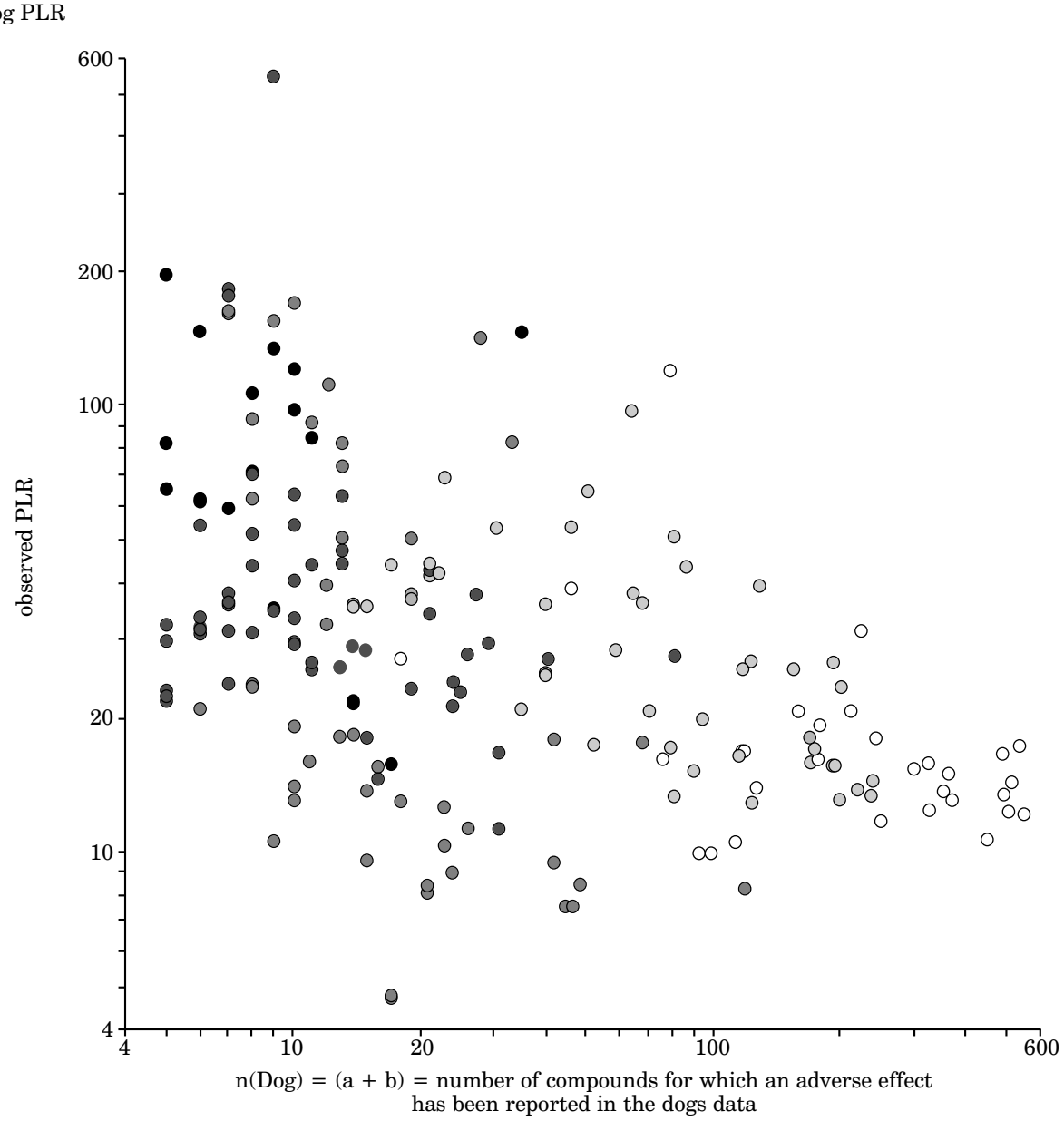
● = *Rat and human;* ○ = *mouse and human;* ■ = *dog and human;* □ = *rabbit and human.*

Median PLRs and iNLRs were recalculated for each preclinical species–human data set, with rare events removed for illustrative purposes. Graphs a and b show the median PLR and iNLR values, respectively, with rare events removed from the animal data only; graphs c and d show the median PLR and iNLR values, respectively, with rare events removed from both the animal data set and the human data set. For each animal species–human pair, the first point on each line shows the median PLR or iNLR for the entire data set; the second point (data minus rares 1) shows the median PLR or iNLR for the data set with rare events below the first threshold removed (see the Results *section); and the third point (data minus rares 2) shows the median PLR or iNLR for the data set with rare events below the second threshold removed (see the* Results *section). Removing rare events significantly reduces the PLR for each species, particularly for the rat and the mouse. It results in a marginal increase in iNLR in each case, although, when rare events in humans are also removed, the iNLRs for the rat and the mouse decrease further.*

other factors, such as size and quality of the animal studies, variability in the requirements for reporting animal studies, 'optimism bias', and lack of randomisation and blinding (34, 41), it means that gauging the true contribution of animal data to human toxicology is virtually impossible — at least for third parties without access to pharmaceutical company files. It would be an interesting exercise to speculate on how such biases affect analyses such as ours. Such speculation is acknowledged to be difficult, however, due to a lack of empirical studies of toxicological bias, and the absence of knowledge of its prevalence and impact (33).

All data sets are imperfect to varying degrees. However, it is only possible to use data which are available, and to ensure, as far as is feasible, that those data are of good quality and as free from bias as possible, and that their analysis and derived conclusions are as objective as possible. It must be made abundantly clear that we, the authors of this report, did not make decisions regarding the toxicity/non-toxicity of the drugs, or decide upon or apply any criteria to such decisions. The mining of the data, and the decisions on toxicity of the drugs, were independent of the authors of this paper, and were made by one or more of the authors of the

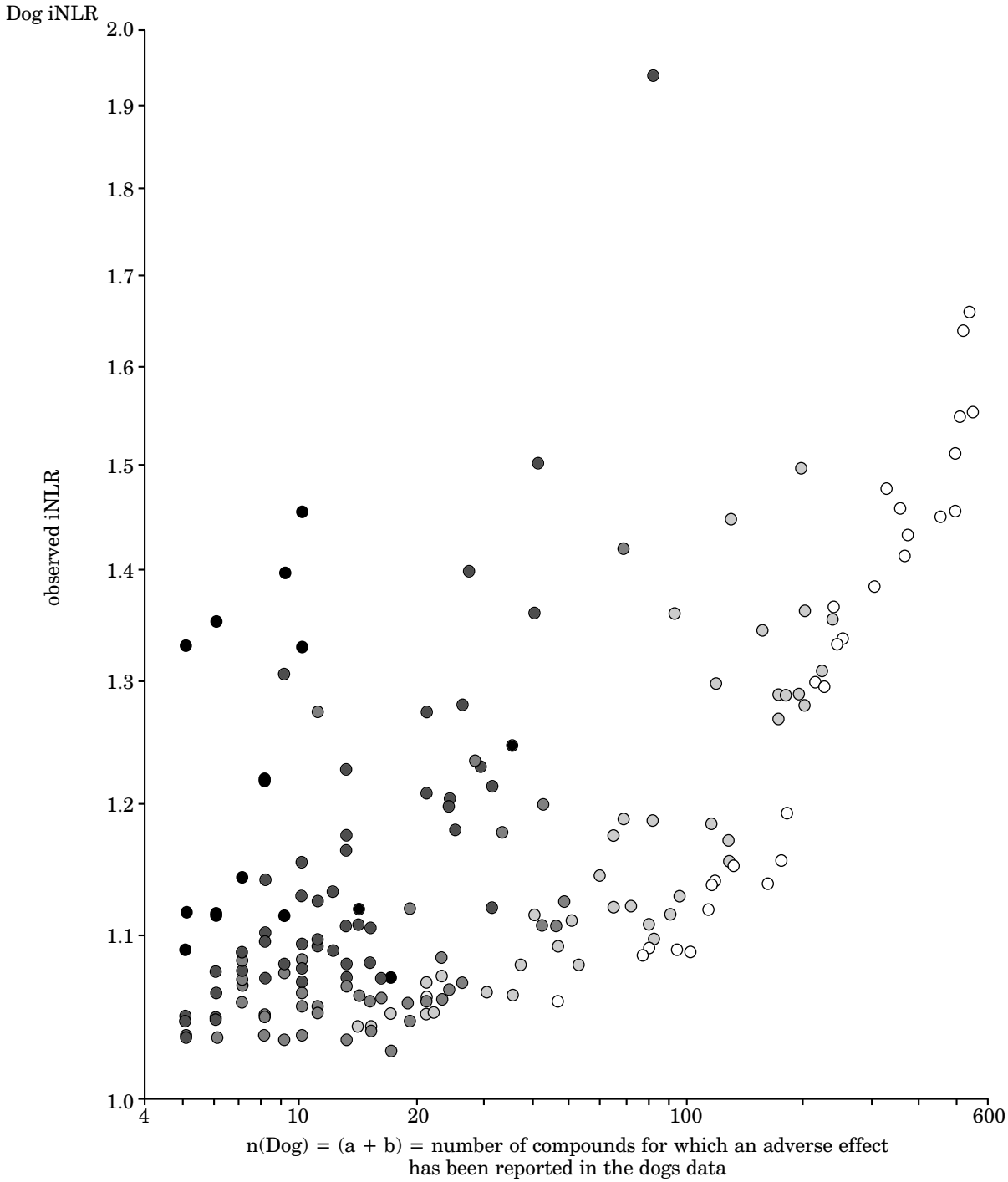
Figure 4: Scatter plots to illustrate the dependence of PLR and iNLR values on sample size



● = *0.9–1.4 log10(nHuman);* ● = *1.4–1.9 log10(nHuman);* ● = *1.9–2.4 log10(nHuman);* ○ = *2.4–2.9 log10(nHuman);* ○ = *2.9–3.4 log10(nHuman).*

Each plot shows a sample of 200 LR values for the dog, for reasons of clarity, and shows PLR (a) and iNLR (b) against animal (x axis) and human (dot intensity) sample sizes. Observed PLRs with higher values tend to have lower sample sizes, both in the preclinical animal and in humans, and more robust PLRs (with higher sample sizes) tend to have lower values. Conversely, observed iNLRs with higher values tend to have higher sample sizes, both in the preclinical animal and in humans, and more robust iNLRs (with higher sample sizes) also tend to have higher values, though the increase in iNLR with sample size is slight. The evidential weight provided by animal tests given a positive toxicology result is significantly decreased by the omission of rare events from the data set, most conspicuously for the rat and the mouse. The evidential weight provided by animal tests given a negative toxicology result is also decreased for the rat and the mouse, when rare events are excluded; while it is marginally increased for the rabbit and dog, the values remain extremely low.

Figure 4: continued



● = *0.9–1.4 log10(nHuman);* ● = *1.4–1.9 log10(nHuman);* ● = *1.9–2.4 log10(nHuman);* ○ = *2.4–2.9 log10(nHuman);* ○ = *2.9–3.4 log10(nHuman).*

Each plot shows a sample of 200 LR values for the dog, for reasons of clarity, and shows PLR (a) and iNLR (b) against animal (x axis) and human (dot intensity) sample sizes. Observed PLRs with higher values tend to have lower sample sizes, both in the preclinical animal and in humans, and more robust PLRs (with higher sample sizes) tend to have lower values. Conversely, observed iNLRs with higher values tend to have higher sample sizes, both in the preclinical animal and in humans, and more robust iNLRs (with higher sample sizes) also tend to have higher values, though the increase in iNLR with sample size is slight. The evidential weight provided by animal tests given a positive toxicology result is significantly decreased by the omission of rare events from the data set, most conspicuously for the rat and the mouse. The evidential weight provided by animal tests given a negative toxicology result is also decreased for the rat and the mouse, when rare events are excluded; while it is marginally increased for the rabbit and dog, the values remain extremely low.

drug/toxicity papers and/or database submissions used, and the data-mining consultancy/curators of the Safety Intelligence Programme, Instem Scientific Limited. Therefore, if any pharmaceutical industry stakeholder disagrees with our conclusions, it is incumbent on them, as the holders of significant amounts of unpublished data, to either conduct an investigation into the worth of the animal models they use routinely, or to facilitate such an investigation by a third party. The latter course of action could be facilitated by making their anonymised data available for analysis, in accordance with the promotion of transparency cited in EU *Directive 2010/63/EU* (42).

Our findings have practical implications for the use of animal models for toxicity testing, especially in the pharmaceutical industry. Reliance on flawed models of toxicity testing leads to two types of failure. If the models have poor PLRs, then there is a risk that many potentially useful compounds will be wrongly discarded, because of FPs produced by the toxicity model. On the other hand, if the models have poor iNLRs, then many toxic compounds will wrongly find their way into human tests, and will fail in clinical trials. The relatively high PLRs found in this study show that animal models may not be leading to the loss of many potentially valuable candidate drugs through the generation of FPs. However, our results do imply that many toxic drugs are not being detected by animal models, leading to the risk of unnecessary harm to humans. Notably, the removal of rarely caused/reported adverse events from our data, which to some degree dominate the data set, in order to make our analysis more statistically robust, further substantiated and validated our conclusions.

In this regard, our findings are entirely consistent with the acknowledged failure of animal models in general to provide guidance on likely toxicity ahead of the entry of compounds into human trials. Drug attrition has increased significantly over the past two decades (e.g. 4, 5, 43–48): 92–94% of all the drugs that pass preclinical tests fail in clinical trials, mostly due to unforeseen toxicities (49–51), and half of those that succeed may be subsequently withdrawn or re-labelled due to ADRs not detected in the animal tests (52). ADRs are a major cause of premature death in developed countries (53). A major contributing factor is the inadequacy of preclinical animal tests: one recent study showed that 63% of ADRs had no counterpart in animals, and less than 20% had a positive corollary in animal studies (21).

There is a scientific basis for the inadequacy of preclinical animal tests. Foremost, are differences in the main enzymes responsible for the metabolism of drugs — the cytochrome P450 (CYP) enzymes (54), which are believed to be involved in the metabolism of more than 90% of drugs (55). While members of the CYP superfamily of enzymes (consisting of 18 families and 43 subfamilies) are highly conserved (typically 75–80% amino-acid sequence identity), it is noted that minor changes in their amino-acid sequences — even one single, conservative substitution — may result in significant differences in activity and/or substrate specificity (56, 57). With regard to differences between species, there is an acknowledged paucity of available comparable data (54). However, reports of important species differences do exist, and suggest that this is a widespread phenomenon that may have significant consequences for the extrapolation of animal data to humans (e.g. 58–62). Important differences with consequences for human extrapolation exist, not just in rodents, but also in the non-rodent species used in drug testing, such as monkeys (58) and dogs (55, 63).

These differences comprise not only differences in amino-acid sequence and catalytic activity, but also in the cellular levels of specific P450 enzymes, which, as in the case of P450 1A2, can show a 25-fold difference between certain species, or be entirely absent in other species (57, 64). Indeed, there is significant variation in the complement of P450 isoforms between humans and other species: for example, humans actually have fewer functional P450 genes than mice, which exhibit substrate specificities and regulatory patterns that can differ markedly from the P450s in other species (65). For example, CYP2B is well conserved between rodents and rabbits, though it is poorly expressed in human liver; CYP3A is the major component of human hepatic P450, but is generally at a low level in other animals, notably in experimental species (64). Four subfamilies of enzymes have been lost in humans compared to rodents, while some genes present in humans are absent in mice: species differences in the seven main CYP gene clusters "pose serious problems in interpretation, when extrapolating from the mouse to human" (66). The 2A subfamily differ markedly in catalytic specificity, despite their sequence similarity (57), and levels of several subfamily forms are relatively low in humans, meaning that comparisons of their activities across species "may be problematic" (67). In the 2C subfamily, "…similarity of one catalytic activity among animal species may have little predictive value for the other reactions catalyzed by the enzymes", and several genetic polymorphisms have been noted in humans, but not in other animals (57). Extrapolation from animal P450 activities to humans must be done with "some caution" for subfamilies 1A1, 1A2, 17A, 1B1 and 4A, "more caution" for subfamilies 2D and 3A, and "major problems" are noted for subfamilies 2A, 2B and 2C (57).

There are marked species differences in the nuclear receptors involved in the activation of CYP pathways, which "make the prediction of cyto-

chrome P450 (CYP) induction in humans from data derived from animal models problematic" (68, 69). Non-genotoxic inducers of CYPs 2B and 4A cause liver tumours in rats and mice, for example, but not in humans (69). There are also notable differences, not only between rats and mice, but also between strains of the same species, which may be further confounded by differences resulting from dietary and other environmental factors (64). It is believed that human polymorphic drug metabolism, underpinned by intra-species variability in P450 enzymes, is a leading cause of adverse drug reactions (70).

Conclusions

This analysis of the most comprehensive quantitative database of publicly-available animal toxicity studies yet compiled, suggests that results from tests on animals (specifically rat, mouse and rabbit models) are highly inconsistent predictors of toxic responses in humans, and are little better than what would result merely by chance — or tossing a coin — in their most important role of providing a basis for deciding whether a compound should proceed to testing in humans. In other words: "…for any putative source of evidential weight to be deemed useful, its specificity and sensitivity must be such that LR_+ *[i.e. PLR]* > 1. Tossing a coin contributes no evidential weight to a given hypothesis, as the sensitivity and specificity are the same — 50% — and thus the LR_+ *[i.e. PLR]* is equal to 1" (28).

This analysis complements, and to a large degree is in accordance with, our recent findings for predictions from dog studies (14), as follows: PLRs were generally high, showing that a drug that is toxic in these species is likely to be toxic in humans. The rat was the species with the highest median PLR, followed by the mouse, then the rabbit. All three were higher than the one for the dog. However, the PLRs for each species were extremely variable — even more so than for the dog — and with no obvious pattern, suggesting that this aspect of animal tests cannot be considered particularly reliable or helpful for any specific new drug. Notably, removing rare events from the data set in order to strengthen the analysis resulted in a marked decrease in PLR values, particularly for the rat and the mouse. More importantly, while iNLRs were much more consistent than PLRs for each species, the range of values was relatively high for the rat and mouse, in contrast to the ones for the rabbit and dog. While the removal of rare events from the data set resulted in a slight increase in iNLR values in two out of eight cases, the values remained very low, indicating the provision of little evidential weight. The median values were better than the median for the dog and, though the order of species from 'best' to 'worst' was the same, all medians were very low, showing that these species provide very little, or essentially no, evidential weight concerning this aspect of toxicity testing. Specifically, if a compound shows no toxic effects in rats, mice, rabbits or dogs, this provides essentially no insight into whether the compound will also show no toxic effects in humans. This is crucial: a critical observation for deciding whether a candidate drug can proceed to testing in humans is the absence of toxicity in tests on animals, and our findings show that the predictive value of the animal tests in this regard is barely greater than that by chance.

This can be illustrated quantitatively. Suppose researchers wish to investigate a candidate compound belonging to a family which prior experience indicates has a 70% probability of absence of ADRs in humans. Before conducting tests in humans, the drug is tested in animals. By using the median iNLR figures found by our study, if the compound shows no sign of toxicity in the rabbit, the probability that the compound will also show no toxic effects in humans will have been increased by the animal testing from 70% to 72%: this is identical to the increase from the performance of tests in dogs (14). These results suggest that testing in the dog or the rabbit contribute essentially no additional confidence in the outcome, but at considerable extra cost, both in monetary terms and in terms of animal welfare. This also has obvious practical relevance to the issue of high attrition rates in clinical trials on new drug candidates. The mouse and rat studies performed marginally better in this respect: lack of toxicity in the mouse increased the probability of no toxic effects in humans from 70% to 76%, and lack of toxicity in the rat increased the probability from 70% to 81%.

A mean increase in the probability that a new drug will not be toxic to humans of just 5% in these four species, suggests that these tests are not fit for purpose (i.e. to validate the progression of testing from animals to humans), and that they are not worth the cost in terms of monetary expense, man-hours, and animal suffering and lives. It is argued that a comprehensive suite of more reliable alternative methods is available (18, 71, 72), and it is difficult to conceive that they would add less evidential weight than the animal tests. Combined with considerable public concern over the use of animals in science (73), and the high ethical costs of doing so, we conclude that preclinical testing of pharmaceuticals in animals cannot currently be justified on scientific or ethical grounds.

Acknowledgements

The authors are grateful to the British Union for the Abolition of Vivisection (BUAV), the Fund for the

Replacement of Animals in Medical Experiments (FRAME), and The Kennel Club (via FRAME), for funding. They thank Robert Matthews for advice on inferential issues, Bob Coleman for his help and encouragement during the inception of this undertaking, and Instem Scientific (previously BioWisdom; Harston, Cambridge, UK) for scientific consultancy and for data analysis on integrated data relating to adverse events in model animal species. The research described in this article is based on the analysis and conclusions of the authors: it has not been subjected to each agency's peer review and policy review; therefore, it does not necessarily reflect the views of the organisations, and no official endorsement should be inferred.

Received 30.01.14; received in final form 15.05.14; accepted for publication 23.05.14.

References

1. Anon. (2004). *Directive 2004/27/EEC* of the European Parliament and the Council of 31 March 2004, amending *Directive 2001/83/EC* on the Community code relating to medicinal products for human use. *Official Journal of the European Union* **L136**, 30.04.2004, 34–57.
2. Anon. (1938). *Federal Food, Drug and Cosmetics Act*. Silver Spring, MD, USA: US Food and Drug Administration. Available at: http://www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/default.htm (Accessed 28.05.14).
3. Aithal, G.P. (2010). Mind the gap. *ATLA* **38**, Suppl. 1, 1–4.
4. Duyk, G. (2003). Attrition and translation. *Science, New York* **302**, 603–605.
5. Kola, I. & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery* **3**, 711–715.
6. Wehling, M. (2011). Drug development in the light of translational science: Shine or shade? *Drug Discovery Today* **16**, 1076–1083.
7. DiMasi, J.A. (2014). Pharmaceutical R&D performance by firm size: Approval success rates and economic returns. *American Journal of Therapeutics* **21**, 26–34.
8. UK Home Office (2013). *Statistics of Scientific Procedures on Living Animals — Great Britain 2012*. HC 549, 60pp. London, UK: The Stationery Office.
9. Greaves, P., Williams, A. & Eve, M. (2004). First dose of potential new medicines to humans: How animals help. *Nature Reviews Drug Discovery* **3**, 226–236.
10. Heywood, R. (1981). Target organ toxicity. *Toxicology Letters* **8**, 349–358.
11. Schein, P.S., Davis, R.D., Carter, S., Newman, J., Schein, D.R. & Rall, D.P. (1970). The evaluation of anticancer drugs in dogs and monkeys for the prediction of qualitative toxicities in man. *Clinical Pharmacology & Therapeutics* **11**, 3–40.
12. Olson, H., Betton, G., Robinson, D., Thomas, K., Monro, A., Kolaja, G., Lilly, P., Sanders, J., Sipes, G., Bracken, W., Dorato, M., Van Deun, K., Smith, P., Berger, B. & Heller, A. (2000). Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regulatory Toxicology & Pharmacology* **32**, 56–67.
13. Altman, D.G. & Bland, J.M. (1994). Diagnostic tests 2: Predictive values. *BMJ* **309**, 102.
14. Bailey, J., Thew, M. & Balls, M. (2013). An analysis of the use of dogs in predicting human toxicology and drug safety. *ATLA* **41**, 335–350.
15. Anon. (2012). *Likelihood Ratios*. Oxford, UK: Centre for Evidence Based Medicine (CEBM). Available at: http://www.cebm.net/index.aspx?o=1043 (Accessed 28.05.13).
16. Greek, R. & Menache, A. (2013). Systematic reviews of animal models: Methodology *versus* epistemology. *International Journal of Medical Sciences* **10**, 206–221.
17. Grimes, D.A. & Schulz, K.F. (2005). Refining clinical diagnosis with likelihood ratios. *Lancet* **365**, 1500–1505.
18. Hasiwa, N., Bailey, J., Clausing, P., Daneshian, M., Eileraas, M., Farkas, S., Gyertyan, I., Hubrecht, R., Kobel, W., Krummenacher, G., Leist, M., Lohi, H., Miklosi, A., Ohl, F., Olejniczak, K., Schmitt, G., Sinnett-Smith, P., Smith, D., Wagner, K., Yager, J.D., Zurlo, J. & Hartung, T. (2011). Critical evaluation of the use of dogs in biomedical research and testing in Europe. *ALTEX* **28**, 326–340.
19. CAAT–Europe (2011). Critical evaluation of the use of dogs in biomedical research and testing in Europe. *Centre for Alternatives to Animal Testing–Europe Workshop, 21–23 June 2011*. Konstanz, Germany: *Centre for Alternatives to Animal Testing–*Europe.
20. CAAT (2011). Critical evaluation of the use of dogs in biomedical research and testing. *Centre for Alternatives to Animal Testing Workshop, 21–23 January 2011*. Baltimore, MA, USA: Johns Hopkins Bloomberg School of Public Health.
21. van Meer, P.J., Kooijman, M., Gispen-de Wied, C.C., Moors, E.H. & Schellekens, H. (2012). The ability of animal studies to detect serious post marketing adverse events is limited. *Regulatory Toxicology & Pharmacology* **64**, 345–349.
22. Igarashi, T., Nakane, S. & Kitagawa, T. (1995). Predictability of clinical adverse reactions of drugs by general pharmacology studies. *Journal of Toxicological Sciences* **20**, 77–92.
23. Broadhead, C.L., Jennings, M. & Combes, R. (1999). *A Critical Evaluation of the Use of Dogs in the Regulatory Toxicity Testing of Pharmaceuticals,* 106pp. Nottingham, UK: Fund for the Replacement of Animals in Medical Experiments (FRAME).
24. Litchfield, J.T.J. (1962). Symposium on clinical drug evaluation and human pharmacology. XVI. Evaluation of the safety of new drugs by means of tests in animals. *Clinical Pharmacology & Therapeutics* **3**, 665–672.
25. Bailey, J. (2008). Developmental toxicity testing: Protecting future generations? *ATLA* **36**, 718–721.
26. Schardein, J. (2000). *Chemically Induced Birth Defects*, 3rd edn, 1019pp. Boca Raton, FL, USA: CRC Press.
27. Spanhaak, S., Cook, D., Barnes, J. & Reynolds, J. (2008). *Species Concordance for Liver Injury*, 6pp. Cambridge, UK: Biowisdom Ltd. Available at: http://www.biowisdom.com/files/SIP_Board_Species_Concordance.pdf (Accessed 28.05.14).

28. Matthews, R.A. (2008). Medical progress depends on animal models — doesn't it? *Journal of the Royal Society of Medicine* **101**, 95–98.
29. Barnes, J.C., Matis, S., Kenna, G., Swinton, J., Bradley, P.M., Day, N.C., Reed, J.Z., Reynolds, J. & Cook, D. (2008). *The Safety Intelligence Program: An Intelligence Network for Drug-induced Liver Injury*, 2pp. Cambridge, UK: Biowisdom Ltd. Available at: http://bioblog.instem.com/downloads/Carboxylic_acids_A4.pdf (Accessed 28.05.14).
30. Sidaway, J.R.M., Roberts, S., Huby, R., Nicholson, A., Pemberton, J., South, M., Noeske, T., Engkvist, O., Bradley, P. & Reed, J. (2012). *Drug Toxicities Associated With Pharmacological Activity: Using Harmonised Data to Make the 'Known' Visible*, 2pp. Macclesfield, UK: Safety Assessment, AstraZeneca.
31. Fourches, D., Barnes, J.C., Day, N.C., Bradley, P., Reed, J.Z. & Tropsha, A. (2010). Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. *Chemical Research in Toxicology* **23**, 171–183.
32. Greco, I., Day, N., Riddoch-Contreras, J., Reed, J., Soininen, H., Kloszewska, I., Tsolaki, M., Vellas, B., Spenger, C., Mecocci, P., Wahlund, L.O., Simmons, A., Barnes, J. & Lovestone, S. (2012). Alzheimer's disease biomarker discovery using *in silico* literature mining and clinical validation. *Journal of Translational Medicine* **10**, 217.
33. Wandall, B., Hansson, S.O. & Rudén, C. (2007). Bias in toxicology. *Archives of Toxicology* **81**, 605–617.
34. Hackam, D.G. (2007). Translating animal research into clinical benefit. *BMJ* **334**, 163–164.
35. ter Riet, G., Korevaar, D.A., Leenaars, M., Sterk, P.J., Van Noorden, C.J., Bouter, L.M., Lutter, R., Elferink, R.P. & Hooft, L. (2012). Publication bias in laboratory animal research: A survey on magnitude, drivers, consequences and potential solutions. *PLoS One* **7**, e43404.
36. Briel, M., Muller, K.F., Meerpohl, J.J., von Elm, E., Lang, B., Motschall, E., Gloy, V., Lamontagne, F., Schwarzer, G. & Bassler, D. (2013). Publication bias in animal research: A systematic review protocol. *Systematic Reviews* **2**, 23.
37. van der Worp, H.B., Howells, D.W., Sena, E.S., Porritt, M.J., Rewell, S., O'Collins, V. & Macleod, M.R. (2010). Can animal models of disease reliably inform human studies? *PLoS Medicine* **7**, e1000245.
38. Sena, E.S., van der Worp, H.B., Bath, P.M., Howells, D.W. & Macleod, M.R. (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biology* **8**, e1000344.
39. 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. *BMJ* **334**, 197.
40. Schott, G., Pachl, H., Limbach, U., Gundert-Remy, U., Ludwig, W.D. & Lieb, K. (2010). The financing of drug trials by pharmaceutical companies and its consequences. Part 1: A qualitative, systematic review of the literature on possible influences on the findings, protocols, and quality of drug trials. *Deutsches Arzteblatt International* **107**, 279–285.
41. Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F., Cuthill, I.C., Fry, D., Hutton, J. & Altman, D.G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* **4**, e7824.
42. Anon. (2010). *Directive 2010/63/EU* of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. *Official Journal of the European Union* **L276**, 20.10.2010, 33–79.
43. US FDA (2004). *Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products*, 31pp. Silver Spring, MD, USA: US Department of Health and Human Services, Food and Drug Administration.
44. Issa, A.M., Phillips, K.A., Van Bebber, S., Nidamarthy, H.G., Lasser, K.E., Haas, J.S., Alldredge, B.K., Wachter, R.M. & Bates, D.W. (2007). Drug withdrawals in the United States: A systematic review of the evidence and analysis of trends. *Current Drug Safety* **2**, 177–185.
45. Bennani, Y.L. (2011). Drug discovery in the next decade: Innovation needed ASAP. *Drug Discovery Today* **16**, 779–792.
46. Eichler, H.G., Aronsson, B., Abadie, E. & Salmonson, T. (2010). New drug approval success rate in Europe in 2009. *Nature Reviews Drug Discovery* **9**, 355–356.
47. Hughes, B. (2008). 2007 FDA drug approvals: A year of flux. *Nature Reviews Drug Discovery* **7**, 107–109.
48. Hartung, T. (2009). Toxicology for the twenty-first century. *Nature, London* **460**, 208–212.
49. Harding, A. (2004). *More compounds failing phase I.* [*The Scientist*, 06.08.04]. Available at: www.the-scientist.com/?articles.view/articleNo/23003/title/More-compounds-failing-Phase-I/ (Accessed 28.05.14).
50. Okie, S. (2006). Access before approval — a right to take experimental drugs? *New England Journal of Medicine* **355**, 437–440.
51. Aurup, P. (2012). *Er Danmark et Attraktivt Land for Klinisk Forskning?* (*Is Denmark an Attractive Country for Clinical Research?*), 23pp. Ballerup, Denmark: MSD Laboratories. Available at: http://di.dk/SiteCollectionDocuments/Opinion/Sundhed/Høring/Præsentation%20-%20Peter%20Aurup,%20Merck.pdf (Accessed 28.05.14).
52. Anon. (1990). *FDA Drug Review: Post Approval Risks 1976–1985.* GAO/PEMD-90-15, 132pp. Washington, DC, USA: US General Accounting Office.
53. Lazarou, J., Pomeranz, B.H. & Corey, P.N. (1998). Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *JAMA* **279**, 1200–1205.
54. Nishimuta, H., Nakagawa, T., Nomura, N. & Yabuki, M. (2013). Species differences in hepatic and intestinal metabolic activities for 43 human cytochrome P450 substrates between humans and rats or dogs. *Xenobiotica* **43**, 948–955.
55. Martinez, M.N., Antonovic, L., Court, M., Dacasto, M., Fink-Gremmels, J., Kukanich, B., Locuson, C., Mealey, K., Myers, M.J. & Trepanier, L. (2013). Challenges in exploring the cytochrome P450 system as a source of variation in canine drug pharmacokinetics. *Drug Metabolism Reviews* **45**, 218–230.
56. Zhou, S.F., Liu, J.P. & Chowbay, B. (2009). Polymorphism of human cytochrome P450 enzymes and its clinical impact. *Drug Metabolism Reviews* **41**, 89–295.
57. Guengerich, F.P. (1997). Comparisons of catalytic

selectivity of cytochrome P450 subfamily enzymes from different species. *Chemico-Biological Interactions* **106**, 161–182.
58. Nishimuta, H., Sato, K., Mizuki, Y., Yabuki, M. & Komuro, S. (2011). Species differences in intestinal metabolic activities of cytochrome P450 isoforms between cynomolgus monkeys and humans. *Drug Metabolism & Pharmacokinetics* **26**, 300–306.
59. Komura, H. & Iwaki, M. (2011). *In vitro* and *in vivo* small intestinal metabolism of CYP3A and UGT substrates in preclinical animals species and humans: Species differences. *Drug Metabolism Reviews* **43**, 476–498.
60. Shimada, T., Mimura, M., Inoue, K., Nakamura, S., Oda, H., Ohmori, S. & Yamazaki, H. (1997). Cytochrome P450-dependent drug oxidation activities in liver microsomes of various animal species including rats, guinea pigs, dogs, monkeys, and humans. *Archives of Toxicology* **71**, 401–408.
61. Turpeinen, M., Ghiciuc, C., Opritoui, M., Tursas, L., Pelkonen, O. & Pasanen, M. (2007). Predictive value of animal models for human cytochrome P450 (CYP)-mediated metabolism: A comparative study *in vitro*. *Xenobiotica* **37**, 1367–1377.
62. Antonovic, L. & Martinez, M. (2011). Role of the cytochrome P450 enzyme system in veterinary pharmacokinetics: Where are we now? Where are we going? *Future Medicinal Chemistry* **3**, 855–879.
63. Gad, S.C. (2006). *Animal Models in Toxicology*, 952pp. Boca Raton, FL, USA: CRC Press.
64. Lewis, D.F., Ioannides, C. & Parke, D.V. (1998). Cytochromes P450 and species differences in xenobiotic metabolism and activation of carcinogen. *Environmental Health Perspectives* **106**, 633–641.
65. Gonzalez, F.J. (2004). Cytochrome P450 humanised mice. *Human Genomics* **1**, 300–306.
66. Nelson, D.R., Zeldin, D.C., Hoffman, S.M., Maltais, L.J., Wain, H.M. & Nebert, D.W. (2004). Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics* **14**, 1–18.
67. Honkakoski, P. & Negishi, M. (1997). The structure, function, and regulation of cytochrome P450 2A enzymes. *Drug Metabolism Reviews* **29**, 977–996.
68. Hewitt, N.J., Lecluyse, E.L. & Ferguson, S.S. (2007). Induction of hepatic cytochrome P450 enzymes: Methods, mechanisms, recommendations, and *in vitro–in vivo* correlations. *Xenobiotica* **37**, 1196–1224.
69. Graham, M.J. & Lake, B.G. (2008). Induction of drug metabolism: Species differences and toxicological relevance. *Toxicology* **254**, 184–191.
70. Zhang, W., Roederer, M.W., Chen, W.Q., Fan, L. & Zhou, H.H. (2012). Pharmacogenetics of drugs withdrawn from the market. *Pharmacogenomics* **13**, 223–231.
71. Spielmann, H., Kral, V., Schafer-Korting, M., Seidle, T., McIvor, E., Rowan, A. & Schoeters, G. (2011). *The AXLR8 Consortium. Alternative Testing Strategies, Progress Report 2011*, 364pp. Berlin, Germany: Institute of Pharmacy, Free University of Berlin. Available at: http://scrtox.eu/~scrtox/images/stories/AXLR8-2011.pdf (Accessed 28.05.14).
72. Committee on Toxicity Testing and Assessment of Environmental Agents, National Research Council (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy,* 216pp. Washington, DC, USA: National Academies Press.
73. YouGov plc. (2009). *Public Opinion*. London, UK: European Coalition to End Animal Experiments. Available at: http://www.eceae.org/en (Accessed 28.05.14).