Predicting Human Drug Toxicity and Safety via Animal Tests: Can *Any* One Species Predict Drug Toxicity in Any Other, and Do Monkeys Help?

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

[1]Cruelty Free International, London, UK; [2]c/o Fund for the Replacement of Animals in Medical Experiments (FRAME), Nottingham, UK

Summary — Animals are still widely used in drug development and safety tests, despite evidence for their lack of predictive value. In this regard, we recently showed, by producing Likelihood Ratios (LRs) for an extensive data set of over 3,000 drugs with both animal and human data, that the absence of toxicity in animals provides little or virtually no evidential weight that adverse drug reactions will also be absent in humans. While our analyses suggest that the presence of toxicity in one species may sometimes add evidential weight for risk of toxicity in another, the LRs are extremely inconsistent, varying substantially for different classes of drugs. Here, we present further data from analyses of other species pairs, including non-human primates (NHPs), which support our previous conclusions, and also show in particular that test results inferring an absence of toxicity in one species provide no evidential weight with regard to toxicity in any other species, even when data from NHPs and humans are compared. Our results for species including humans, NHPs, dogs, mice, rabbits, and rats, have major implications for the value of animal tests in predicting human toxicity, and demand that human-focused alternative methods are adopted in their place as a matter of urgency.

Key words: *animal tests, dog, drug development, monkey, mouse, non-human primate, preclinical testing, rabbit, rat, toxicology.*

Address for correspondence: *Jarrod Bailey, Cruelty Free International (formerly BUAV), 16a Crane Grove, London N7 8NN, UK.*
E-mail: *jarrod.bailey@crueltyfreeinternational.org*

Introduction

Preclinical trials or 'animal tests' on new drugs are required by regulatory agencies worldwide (e.g. 1, 2), based on a presumption of human relevance and predictability, rather than robust scientific evidence (3). This requirement continues, even in the face of record levels of drug failure and drug attrition (4–7), as well as evidence that reveals the animal tests to be unfit for purpose. This evidence includes our two recent analyses, in which a data set on an unprecedented scale, of published adverse events in multiple species, induced by well over 2,000 pharmaceuticals, was used to apply the evidential weight provided by animal tests to the probability that a new drug may be toxic (or not toxic) to humans (8, 9). These studies revealed two main points: first, that toxicity in animals may be likely to also be present in humans, though this cannot be considered particularly consistent or reliable, due to considerable variability and lack of any clear pattern in types of toxic effects; and secondly, perhaps more crucially, that the absence of toxicity in animals provides essentially no insight into the likelihood of toxicity or absence of toxicity in humans.

To augment these studies, which examined the contribution of tests on dogs, rats, mice and rabbits to predicting human risk, we have now analysed further data from non-human primates (NHPs) — arguably the most likely non-human species to have significant human relevance — asking the same question. In addition, we have examined the relationships of data from other pairs of these five non-human species, to assess how reliably tests in one species can indicate the toxicological susceptibility of any other species to new drugs. Our hypothesis is that this is not possible, since major interspecies differences mean that the interspecies extrapolation of toxicity risk is inevitably poor, whatever the species, including humans.

Once again, we have used the most apposite statistical metrics in our approach — Likelihood Ratios (LRs) — which, unlike 'concordance' metrics, such as the True Positive Rate (sensitivity) or Positive Predictive Value (PPV), directly address the salient issue of the contribution of evidential weight by a test in one species for or against the toxicity of a given compound in another species.

Methods

A detailed consideration and description of the diagnostic metrics used in this analysis can be

found in our first two reports (8, 9). Briefly, the data analysed were obtained from Instem Scientific Limited (Harston, Cambridge, UK; http://www.instem-lss.com; Safety Intelligence Programme). All the data were collated by Instem from publicly available sources, including: PubMed (http://www.ncbi.nlm.nih.gov/pubmed), the FDA Adverse Event Reporting System (FAERS), DrugBank (http://www.drugbank.ca), and the National Toxicology Program (http://ntp.niehs.nih.gov), for a total of 3,028 'single active ingredients' (excluding combination drugs) in humans and preclinical species. The effects of each compound were classified by Instem according to their target tissues (e.g. bradycardia and arrhythmic disorder would both be considered to be effects on heart tissues), based on five levels of specificity classified by their MedDRA (Medical Dictionary for Regulatory Activities; http://www.meddramsso.com) counterparts, ranging from any of the (more than 70,000) very specific Low Level Terms (LLT, e.g. 'feeling queasy') up to very generic System Organ Class (SOC), such as 'gastrointestinal disorders'. These classifications help to eliminate false positives that may arise from species-specific observations, as well as in the identification of concordant observations that might otherwise have been missed, by their 'rolling up' into more-generic terms. Where it was not possible for such MedDRA mapping to be achieved, the preferred term of Instem's own Safety Intelligence Programme (SIP) knowledge base (e.g. 'musculoskeletal toxicity') was used instead.

For each compound, the presence or absence of toxicity was logged for each species of each 'species pair' being examined. These results were recorded in a standard 2 × 2 matrix of results (see Table 1), allowing various diagnostic metrics to be derived. The rationale for our use of LRs, as opposed to positive concordance rate or PPV, has been described previously (8, 9); in short, LRs — both positive (PLR) and inverse negative (iNLR) — are the only approaches that include both sensitivity and specificity, i.e. that are sufficient to measure the evidential weight provided by each test in question. Sensitivity alone is insufficient: if, for instance, an animal test always indicates toxicity present in humans, it has a *sensitivity* of 100%. However, if that test, in addition, always indicates toxicity, even when it is absent in humans, every drug tested may as well be dismissed as toxic to humans at the outset, so knowledge of that test's *specificity* is also necessary. Further, PPV is conditional and depends on the prevalence of toxicity in the compounds, and so is inappropriate (e.g. 10, 11).

With reference to the 2 × 2 results matrix (Table 1), PLR is given by: sensitivity/(1 – specificity) = (a / a + c)/(b / b + d), while iNLR = specificity/(1 – sensitivity) = (d / b + d)/(c / a + c). All the values were calculated by using the above formulae, in Microsoft Excel® worksheets containing the data. PLR and iNLR capture the ability of one animal species to add evidential weight to the belief that a specific compound is toxic/non-toxic in another, respectively. As McGee (12) has emphasised, "LRs may range from 0 to infinity. Findings with LRs greater than 1 argue for the diagnosis of interest; the bigger the number, the more convincing [the finding]. Findings whose LRs lie between 0 and 1 argue against the diagnosis of interest; the closer the LR is to 0, the less likely [the finding]. Findings whose LRs equal 1 lack diagnostic value". Note that, where McGee states, "argue [for/against] the diagnosis of interest" (as he is discussing the role of LRs in evaluating the performance of a diagnostic test clinically), we can replace this phrase with 'support the evidential value of the animal test'.

Any animal species that gives a PLR/iNLR that is statistically significantly higher than 1.0, can therefore be regarded as contributing some degree of evidential weight to the probability that the compound under test will be toxic/non-toxic in the other species. Crucially, these definitions imply that, if any particular animal species is 'good' at predicting toxicity in another, it is not necessarily also good for predicting an absence of toxicity. That is, a high PLR does not guarantee a high iNLR. The full set of Instem Scientific data on which this analysis is based, including 95% Confidence Intervals, are available on request from the author for correspondence.

A discussion of potential bias in the LR estimates is provided in the previous papers (8, 9),

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)

which should be consulted for a fuller explanation. Briefly, the data set is probably affected by selection bias/effect, given that drugs that appear highly toxic in rodent tests are unlikely to advance to tests in dogs, NHPs and humans. This means that these types of drugs will be under-represented in the data set, so the LRs will not exactly represent the population of drugs as a whole. This will be true of all data sets, however, as this precautionary principle is an inherent aspect of the testing protocol. In these circumstances, anybody conducting an analysis such as this will be faced with an identical situation, and this is acknowledged (see 8, 9). Gauging the actual contribution of animal data to human toxicology is therefore virtually impossible. All one can do is to take steps to minimise the effects of bias. We have taken such steps. The data set was limited to drugs in the FAERS database, and so all the compounds had proceeded to market, and animal and human data were available for them. The only way in which publication bias can be addressed is for the industry, as the holders of significant amounts of unpublished data, to either conduct its own investigations, and/or to facilitate such investigations by third parties. The latter could be achieved by making anonymised data available for analysis, in accordance with the promotion of transparency cited in *Directive 2010/63/EU* (13).

Results

The total number of classifications of effects for each species pair examined, and therefore the numbers of LRs calculated initially for each species pair, are shown in Table 2a. It should be noted that a large proportion of the data reflected adverse events that are rare, potentially compromising the reliability of any analysis of it. A more-detailed consideration of this point is provided in our previous paper (8). To take account of this, rare events were removed from consideration, and the LRs were recalculated. This was done for rare events at various thresholds, namely, less than 2, 5, 10 and 20 events. The number of classifications used in our analysis for each of these thresholds is shown in Table 2b.

Median PLRs and iNLRs, involving the entire data set for each species pair (i.e. with no classifications containing rare observations removed), in each direction (i.e. 'mouse for rat' and 'rat for mouse'), are shown in Table 3a. Almost all the values show the evidential weight provided by an animal test in one species for toxicity/lack of toxicity in another species to be zero (the exceptions being the rat–mouse pairs). These conclusions, however, must be affected by the predominance of rare observations in the data set, as mentioned above. For example, 97% of the classifications were discounted for the NHP–human species pair, when accounting for rare observations (< 2; see Table 2), due to the extremely low numbers of observations in one or more species. The recalculated LRs, at four different thresholds with classifications containing rare events removed, are depicted graphically in Figure 1. The PLR and iNLR values for the '< 5' threshold, at which any classifications in the data set that contained observations of toxicity in either species that numbered fewer than five were discounted, are listed in Table 3b. This threshold was chosen, as it has served as a basis for removing the potential bias caused by the inclusion of rare observations in previous analyses, and is the cut-off point used by the data provider, Instem. The range of LRs (both PLRs and iNLRs) for each species pair, also with rare LRs (< 5 observations) removed from the data set, is shown in Figure 2.

All the PLRs were generally quite high (though not dramatically so, bearing in mind that LRs may have values up to infinity), particularly for the rat–mouse pairs, suggesting that compounds showing toxicity in one species are also likely to be toxic in the other. This includes the NHP–human species pair. Notably, however, these are significantly lower than median PLRs for the dog, mouse, rabbit and rat with respect to humans, reported in our previous analyses (8, 9). However, in common with these previous reports, high ranges, with no obvious pattern of toxicity, suggest the reliability of this aspect cannot be generalised or regarded with confidence. Crucially, median iNLRs were substantially lower than PLRs, and were barely greater than unity, supporting the view that, when toxicity appears to be absent in one species, this result provides essentially no evidential weight to the likelihood of lack of toxicity in any other species. As for PLRs, this includes the NHP–humans species pair.

It has been put to us that it may be difficult to appreciate the significance of our results, given that LRs can extend to infinity. In other words, what, for instance, do PLRs of approximately 10, 20 or 50 actually mean? We have tried to provide a rough, though relevant, 'yardstick' by taking data from the comparison of the two most-commonly used preclinical species (rat and dog), and comparing the rat with itself, and the dog with itself. Clearly, the LRs from these 'control' comparisons were always going to have infinite values, as a data set must show absolute identity with itself. We therefore factored in a small percentage of FP and FN values (0.1% of the total sample size for each observation, in the dog–rat data set with rare events [< 5] removed, to derive a value for a hypothetical inter-species 'self' comparison that is not perfect, but which shows a high level of identity. The PLR for the hypothetical rat–rat pair was 898, and for the dog–dog 845. The iNLRs were 17.3 and 6.6, respectively. While these values are artificial,

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

a) The total number of classifications

	PLR				iNLR				
Species pair	**Tissue-level effects**	**Biomedical observations (BMOs)**	**Total classifications used**	**Classifications eliminated**	**Tissue-level effects**	**Biomedical observations (BMOs)**	**Total classifications used**	**Classifications eliminated**	**Total**
NHP–human	44	483	527	16,230	86	16,563	16,649	194	16,757
NHP–dog	56	394	450	1791	63	1814	1877	364	2241
Dog–NHP	58	405	463	1778	58	724	782	1459	2241
NHP–rat	55	495	550	7965	86	8428	8514	1	8515
Rat–NHP	50	626	676	7839	51	619	670	7845	8515
NHP–mouse	57	493	550	4022	82	4250	4332	240	4572
Mouse–NHP	57	530	587	3985	57	722	779	3793	4572
Dog–rat	59	1119	1178	7772	86	8263	8349	601	8950
Rat–dog	65	1238	1303	7647	59	1675	1734	7216	8950
Dog–mouse	64	829	893	4402	83	4227	4310	985	5295
Mouse–dog	65	854	919	4376	64	1781	1845	3450	5295
Rat–mouse	83	2779	2862	6842	83	3875	3958	5746	9704
Mouse–rat	83	2565	2648	7056	86	8137	8223	1481	9704

The number of classifications of effects for each species pair, and therefore the numbers of LRs calculated for each species pair, are shown: the total number of classifications the data provided (column 9), as well as those data able to be used to calculate both PLR (columns 1–4), and iNLR (columns 5–8). Note that each pair is shown bi-directionally (e.g. dog–rat and rat–dog), for reasons described in the text; briefly, the test hypothesis was: 'Does species 2 predict/add value to species 1?', so each pair had to be done in both directions. The exception was the 'NHP–human' pair, as we are not interested in to what degree humans predict/add value to NHP toxicity (Row 1 of results).

The total number of classifications used in the analysis is shown in columns 3 and 7 (Total classifications used; PLR and iNLR respectively), which comprise the sum of column 1 (PLR) or 5 (iNLR) (tissue-level effects) and column 2 (PLR) or 6 (iNLR) (BMOs). The number of BMO classifications for which there were no effects observed in the 'test' species of interest, and which were therefore eliminated from consideration in our analysis, are shown in columns 4 (PLR) and 8 (iNLR) (Classifications eliminated), out of the total number of classifications for which there were data. In total, data were available for 3,028 'active pharmaceutical ingredients'.

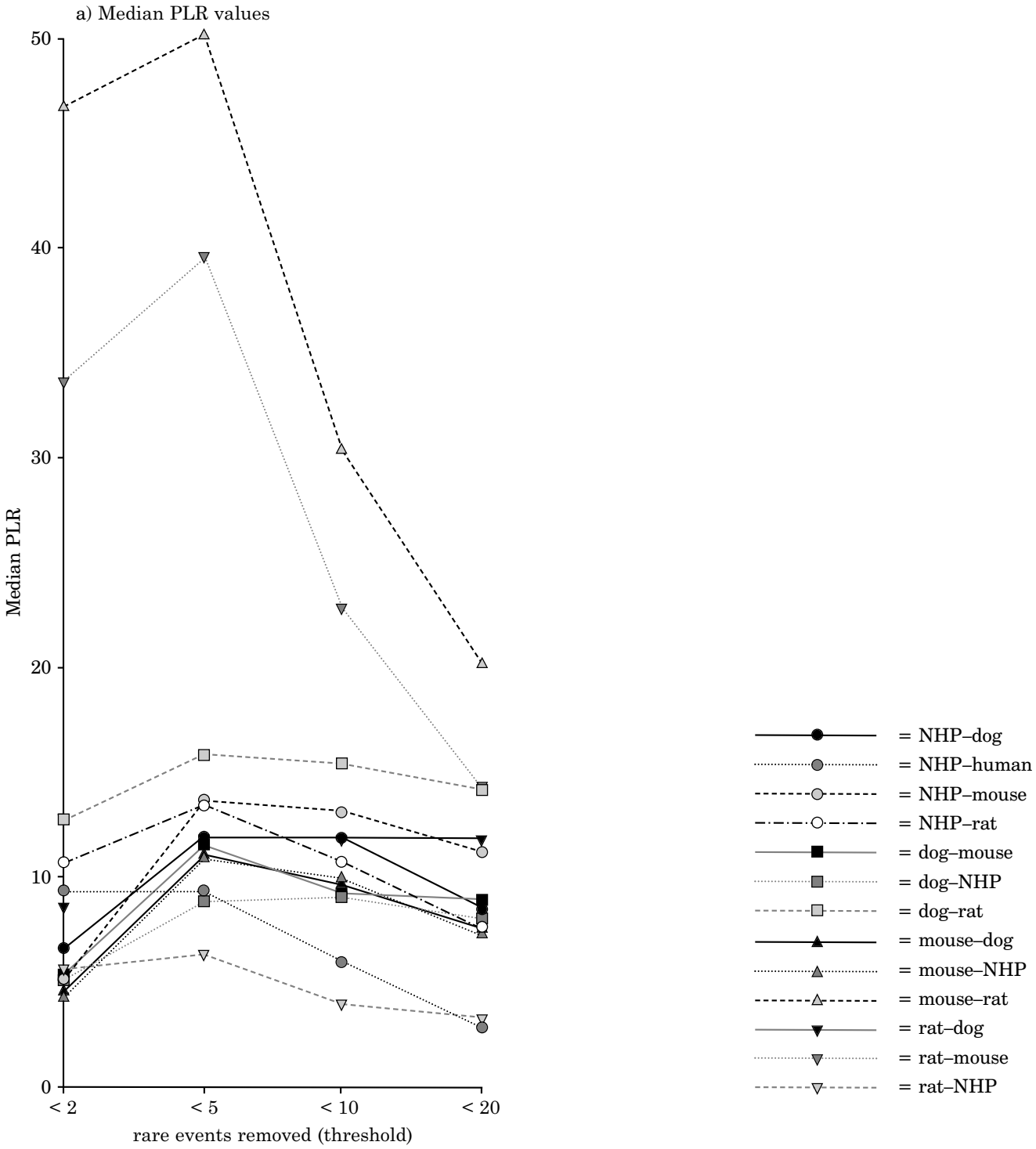
Table 2: continued

b) The number of classifications of adverse effects for each species pair, as used in this analysis, *with rare events (fewer than two, five, ten, or twenty in either or both species) removed from the data set*

	Number of classifications used in analysis when 'Rare Observations' discounted				
Species pair	**< 2**	**< 5**	**< 10**	**< 20**	**Total**
NHP–human	159	37	9	3	16,757
NHP–dog	214	86	43	19	2241
NHP–rat	237	64	25	3	8515
NHP–mouse	247	89	46	16	4572
Dog–rat	638	312	166	76	8950
Dog–mouse	471	223	108	54	5295
Rat–mouse	1500	688	304	131	9704

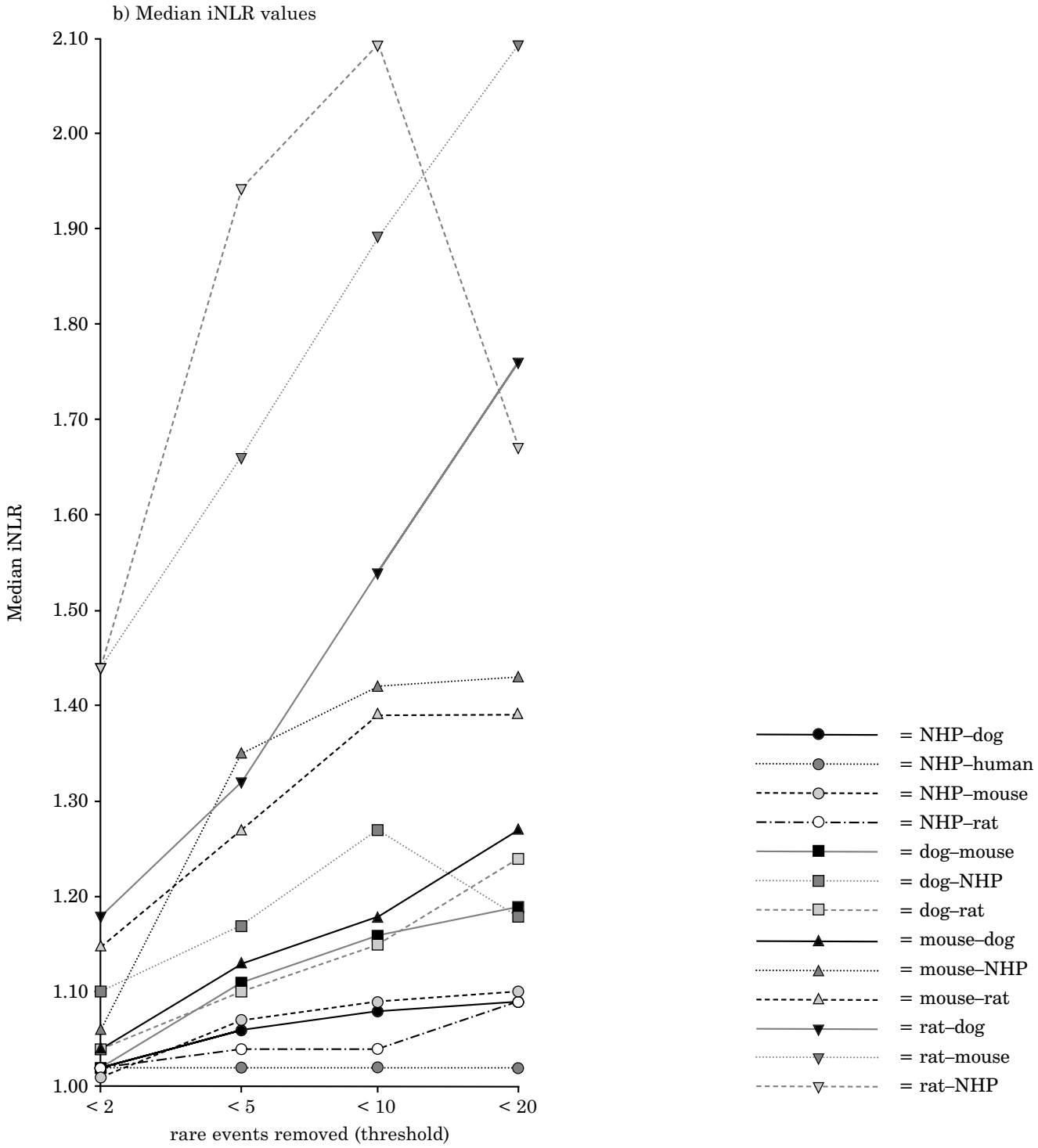
The number of observations for which data were discounted is shown at the head of the columns. As the numbers were identical for each species pair, regardless of the 'direction'; (see Table 2a legend), only one pair is shown, which may be assumed to be 'bi-directional'. The most stringent approach (< 10 and < 20) can be considered illustrative only; for practical purposes, < 5 has been used previously, and is used by Instem (the data provider) as a cut-off point. The relatively low numbers of 'non-rare' observations for species pairs involving NHPs (and indeed the absolute low numbers of observations when, for example, less than 20 observations are discounted) is evidence of the difficulty in accessing NHP toxicology data, of which little is publicly-available.

Figure 1: PLRs and iNLRs for all species pairs, for all four thresholds below which rare events were removed from the data set



Graphs a and b show the median PLR and iNLR values, respectively. For each species pair, the first point on each line (i.e. < 2) shows the median PLR or iNLR for the data set with rare events below the first threshold removed (fewer than two observations; see Results *section); the second point (i.e. < 5) shows the median PLR or iNLR for the data set with rare events below the second threshold removed (fewer than five observations; see* Results *section); and so on. Removing rare events progressively from < 2 to < 20 has little effect on most of the LR values. The exceptions are for the PLRs of the rat–mouse pairs, which roughly halve in value; and rat–mouse and rat–dog pairs for the iNLR values, most of which increase only marginally. Notably, the NHP–human values, which indicate the degree of evidential weight NHP tests provide for human toxicity, are one of the lowest for PLRs, and the lowest for iNLR.*

Figure 1: continued



Graphs a and b show the median PLR and iNLR values, respectively. For each species pair, the first point on each line (i.e. < 2) shows the median PLR or iNLR for the data set with rare events below the first threshold removed (fewer than two observations; see Results *section); the second point (i.e. < 5) shows the median PLR or iNLR for the data set with rare events below the second threshold removed (fewer than five observations; see* Results *section); and so on. Removing rare events progressively from < 2 to < 20 has little effect on most of the LR values. The exceptions are for the PLRs of the rat–mouse pairs, which roughly halve in value; and rat–mouse and rat–dog pairs for the iNLR values, most of which increase only marginally. Notably, the NHP–human values, which indicate the degree of evidential weight NHP tests provide for human toxicity, are one of the lowest for PLRs, and the lowest for iNLR.*

Table 3: Median LRs and ranges for each species pair

Species pair	PLR (median)	iNLR (median)	PLR range	iNLR range
a) Values for the complete data set				
NHP–human	0	1.00	605	2.00
NHP–dog	0	1.00	3027	1.01
Dog–NHP	0	1.00	3027	9.73
NHP–rat	0	1.00	1513	1.00
Rat–NHP	1.18	1.00	3027	20.93
NHP–mouse	0	1.00	1009	3.00
Mouse–NHP	0	1.00	3027	4.89
Dog–rat	0	1.00	3027	1.68
Rat–dog	0	1.00	3027	6.63
Dog–mouse	0	1.00	3027	3.97
Mouse–dog	0	1.00	3027	2.04
Rat–mouse	14.70	1.00	3027	27.86
Mouse–rat	11.47	1.00	3027	10.98
b) Values calculated with rare events (< 5) removed from the data set				
NHP–human	9.39	1.02	605	0.09
NHP–dog	11.97	1.06	3027	0.39
Dog–NHP	8.81	1.17	96	1.98
NHP–rat	13.40	1.04	60	0.19
Rat–NHP	6.28	1.94	56	5.55
NHP–mouse	13.81	1.07	173	0.40
Mouse–NHP	11.11	1.35	173	2.31
Dog–rat	15.90	1.10	173	1.21
Rat–dog	11.83	1.32	173	6.56
Dog–mouse	11.64	1.11	129	3.57
Mouse–dog	11.23	1.13	146	1.60
Rat–mouse	39.43	1.66	480	8.28
Mouse–rat	50.30	1.27	472	7.92

a) While of interest, as almost all the values show the evidential weight provided by an animal test for toxicity in another species to be zero (the exceptions being the rat–mouse pairs), the conclusions must be affected by the preponderance of rare observations in the data set. For example, 97% of the classifications were discounted for the NHP–human species pair (see Table 2), due to extremely low numbers of observations in one or more species. For this reason, and to minimise any associated bias, rare events were removed from consideration, and LRs recalculated (see Table 3b, and graphs in Figure 1).

b) The PLRs were generally quite high, particularly for the rat–mouse pairs, suggesting that compounds showing toxicity in one species are also likely to be toxic in another. This includes the NHP–human species pair. Notably, however, these are significantly lower than median PLRs for the dog, mouse, rabbit and rat with respect to humans, reported in our previous analyses. However, in common with these previous reports, high ranges, with no obvious pattern of toxicity, suggest the reliability of this aspect cannot be generalised or regarded with confidence. Crucially, median iNLRs were substantially lower, supporting the view that, when toxicity appears to be absent in one species, this result provides essentially no evidential weight to the likelihood of lack of toxicity in any other species. As for the PLRs, this includes NHPs and humans.

they may go some way toward suggesting what might be expected from interspecies comparisons that would suggest a high degree of evidential weight provided by one species for another.

Discussion and Conclusions

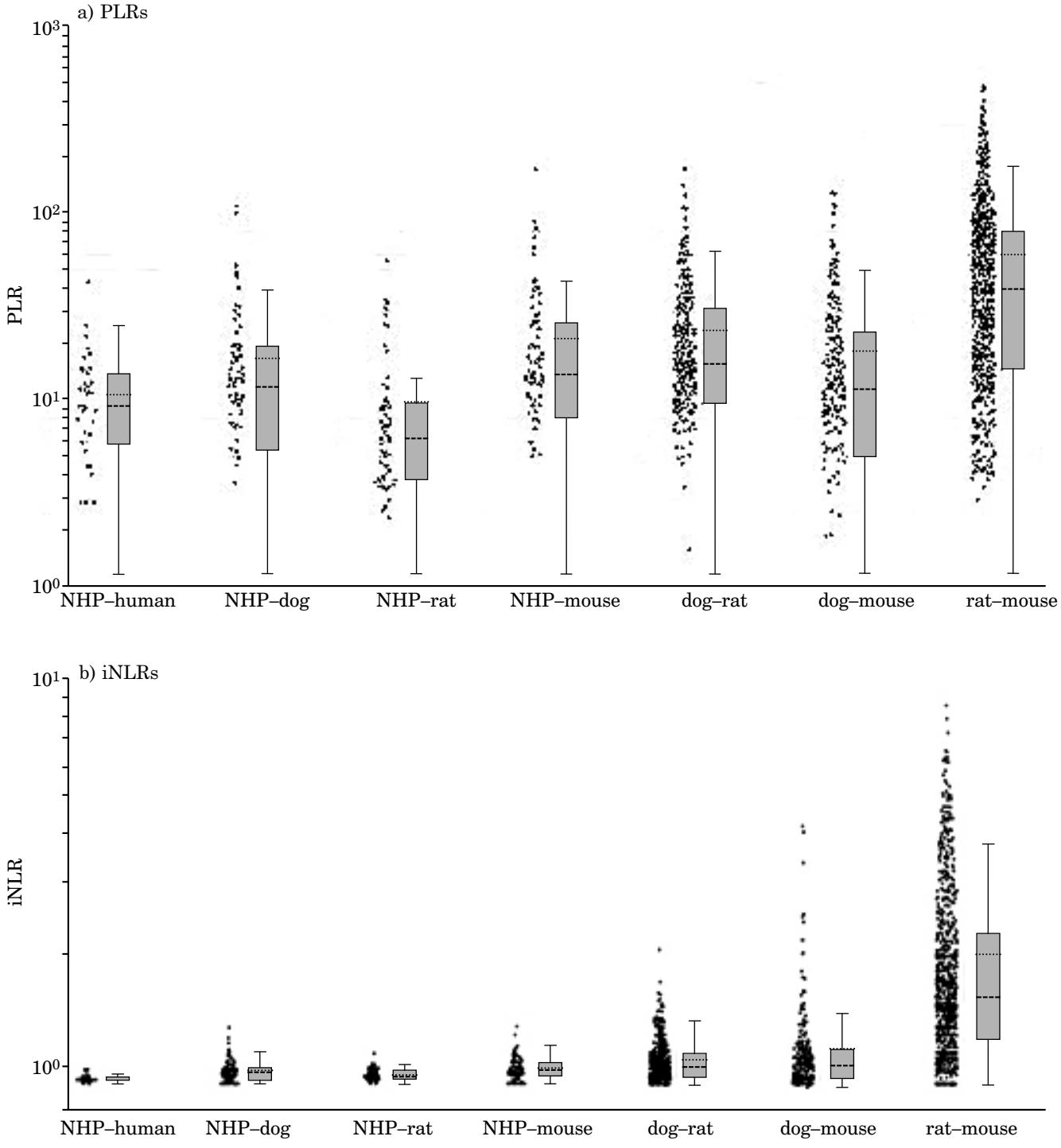
This analysis of animal toxicology in preclinical drug development, complements our other two recently-published evaluations of toxicity tests in rats, mice, rabbits and dogs, and their relevance to humans (8, 9). All were sorely needed, and inspired by the lack of scientific evidence to support such animal testing of new human drugs, and were intended to augment previous studies suggesting there is, in fact, evidence to the contrary (see 8, 9).

Our previous analyses have applied the apposite statistical metrics of LRs, to data sets of unprecedented scale (though comprising, necessarily, only publicly-available data), to determine the evidential weight for or against the toxicity of a given compound in humans, contributed by data from animal tests. Critically, they found that, when no toxicity was found in the animal tests, these tests contributed very little to no evidential weight to the probability of lack of toxicity in humans. This was the case for tests in dogs, mice, rabbits and rats, and this current study found that the same applies to tests in NHPs, meaning that the five species commonly used in preclinical testing — including dogs and monkeys — fail to contribute evidential weight in this scenario. Though the data set used in the analysis reported in this paper, and therefore as the basis of the NHP–human species pair analysis, was different from that used in our previous two studies, this common theme is summarised in Table 4, which condenses all of our analyses involving human risk of toxicity throughout the three papers. Tests in all five preclinical species, dogs, mice, NHPs, rabbits and rats, provide little to no evidential weight. Indeed, the iNLR value for NHPs is the lowest of all, though the use of a different data set may account for this.

These conclusions are underpinned and bolstered by the other aspect of this paper, which illustrates a similar lack of evidential weight provided by *any* species for lack of toxicity in any other. In other words, these data, and our analyses, suggest strongly that a lack of toxicity in any species cannot be reliably used to imply a probable lack of toxicity in any other species. This is absolutely crucial, because the critical observation for deciding whether a candidate drug can proceed to testing in humans is the *absence* of toxicity in tests on animals, yet our analyses show that this contributes no additional confidence in the eventual human outcome, but at considerable cost in terms of animal welfare and money.

If no one species can add evidential weight in this regard to any other species, then how can any

Figure 2: Box plots showing ranges of PLRs and iNLRs for all seven species pairs, with values for rare observations (< 5) removed



The results summarised in Table 3b are shown in their entirety in these box plots, to aid visualisation and comprehension. For each species pair, the distribution of Likelihood Ratios — a) PLRs, b) iNLRs — is illustrated via 'box and whisker' plots. Standard quantile box plots are shown with a logarithmic scale (log_{10}) to incorporate higher values, with each individual LR value plotted alongside, and 'jittered' for ease of visualisation. The box on each plot is bounded by the lower (25%) and upper (75%) quartiles, and therefore represents the interquartile range, i.e. the 50% of LRs that lie either side of the median value, which itself is shown by the dashed line bisecting the box. The mean value is indicated by the dotted line bisecting the box. The maximum and minimum LRs are shown by the whiskers at the top and bottom of the plots, respectively. Outliers are shown as dots beyond the upper whisker, which are greater than 1.5× the interquartile range above the upper (75%) quartile. These box plots illustrate that, for most of the species pairs, there is a high range of LR values, but that there are high-value outliers, and that most LR values lie toward the lower end of the range.

Table 4: Median iNLRs and ranges for all five preclinical species with regard to humans

	iNLR (median)	iNLR range
Rat	1.82	1.02–100.0
Mouse	1.39	1.03–50.0
Rabbit	1.12	1.01–2.33
Dog	1.10	1.01–1.92
NHP	1.02	1.00–1.08

The median iNLRs were extremely low for all the preclinical species, supporting the view that animals provide very little or essentially no evidential weight to this aspect of toxicity testing. Though it may appear interesting that the value for NHPs is the lowest (and therefore most inferior) of all five preclinical species, it was derived from a different total data set to the other four, which may partly account for this.

non-human species be expected to add evidential weight to what is likely to happen in humans? At present, the main rationale given by those who use, for example, non-rodents such as dogs and NHPs, is that they are not rodents — rodent data are not trusted, so further data are sought from other, non-rodent, species. But our analyses and other data strongly suggest that additional data from other species do not solve the problem of interspecies extrapolation, and in particular, extrapolation to humans.

While the presence of toxicity in one species may sometimes add evidential weight for risk of toxicity in another, the LRs are extremely inconsistent, varying substantially for different classes of drugs and their effects. As we have previously commented, this suggests that this aspect of animal toxicity tests cannot be considered robust and reliable for any specific new drug, particularly when one considers the prior evidence to the contrary, suggesting that animal toxicity tests are poorly predictive of human risk (see our previous related papers [8, 9] for references). At the very least, the fact that there is conflicting evidence in this respect, and that our analyses are, necessarily, based on the relatively limited publicly-available data, demands that further analyses of proprietary data, conducted and/or overseen by industry, are conducted in a transparent, and preferably independent, manner.

As we have argued previously, all of this must have practical and urgent implications for the future use of animals in toxicity testing, especially in the pharmaceutical industry, because poor iNLRs mean that toxic compounds are progressing to testing in humans, only to fail in clinical trials or soon after approval for marketing. This is not conjecture — it is widely acknowledged, and evidenced by adverse drug reactions and new drug failures at record levels, after an inexorable rise over the past two decades (4, 5, 14–20). Further, this is not something that can be 'fixed' by improving the rigour of the animal tests. There is a scientific basis for this level of failure, namely, major interspecies differences in the cytochrome P450 enzymes (CYP), which are involved in the metabolism of more than 90% of drugs (21). This is discussed in our recent papers, but briefly, increasing (and long overdue) comparative analyses of CYPs across species, including dogs and monkeys, are revealing important differences (which are often minor in nature, but significant functionally), with important consequences for the extrapolation of animal data to humans. Indeed, such differences also seem prevalent *within* species, and are a basis for human variability in susceptibility to adverse drug reactions. So, even if, for the sake of argument, any one or a combination of non-human species *could* reliably predict effects in humans — which 'humans' would they be, given the vast scale of human polymorphism?

It is increasingly clear that the animal testing of human drugs is not fit-for-purpose, and this is especially true in the light of the astounding array of directly human-relevant methods now available to science for testing new drugs (see, for example, 22, 23). In combination with an unprecedented level of public concern over the use of animals in science (24), and the high ethical costs of doing so, we conclude that the preclinical testing of pharmaceuticals in animals cannot currently be justified, ethically or scientifically. At the very least, it is incumbent on the pharmaceutical industry and its regulators to take on board the concerns highlighted by our work, and to augment it via their own studies, using proprietary data that are not publicly-available, as a matter of urgency.

Acknowledgements

The authors are grateful to the Cruelty Free International Trust for funding. They thank Instem Scientific Limited for scientific consultancy, and for the provision of data relating to adverse events in humans and in various animal species.

References

1. Anon. (2004). *Directive 2004/27/EC* of the European Parliament and of 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, 34–79.
2. Anon. (2009). *The 1938 Food, Drug and Cosmetics*

Act. Silver Spring, MD, USA: US Food and Drug Administration. Available at: http://www.fda.gov/AboutFDA/WhatWeDo/History/ProductRegulation/ucm132818.htm (Accessed 29.10.15).
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. 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.
9. Bailey, J., Thew, M. & Balls, M. (2014). An analysis of the use of animal models in predicting human toxicology and drug safety. *ATLA* **42**, 181–199.
10. Altman, D.G. & Bland, J.M. (1994). Diagnostic tests 2: Predictive values. *British Medical Journal* **309**, 102.
11. Grimes, D.A. & Schulz, K.F. (2005). Refining clinical diagnosis with likelihood ratios. *Lancet* **365**, 1500–1505.
12. McGee, S. (2002). Simplifying likelihood ratios. *Journal of General Internal Medicine* **17**, 646–549.
13. 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.
14. Anon. (2010). *Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products — March 2004*. Silver Spring, MD, USA: US Food and Drug Administration. Available at: http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm (Accessed 29.10.15).
15. 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.
16. Bennani, Y.L. (2011). Drug discovery in the next decade: Innovation needed ASAP. *Drug Discovery Today* **16**, 779–792.
17. 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.
18. Hughes, B. (2008). 2007 FDA drug approvals: A year of flux. *Nature Reviews Drug Discovery* **7**, 107–109.
19. Hartung, T. (2009). Toxicology for the twenty-first century. *Nature, London* **460**, 208–212.
20. Aurup, P. (2012). *Er Danmark et Attraktivt Land for Klinisk Forskning?* [Is Denmark an Attractive Country for Clinical Research?; Presentation]. Available at: http://di.dk/SiteCollectionDocuments/Opinion/Sundhed/Høring/Præsentation%20-%20Peter%20Aurup,%20Merck.pdf (Accessed 29.10.15).
21. 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.
22. Anon. (2007). *Petition to the US Food and Drug Administration for Mandatory Use of Non-Animal Methods in the Development and Approval of Drugs and Devices*, 6pp. Available at: http://www.alternatives-petition.org/docs/MAP-Executive-Summary.pdf (Accessed 29.10.15).
23. Anon. (2012). T*he AXLR8 Consortium: Alternative Testing Strategies, Progress Report 2012*, 288pp. Berlin, Germany: AXLR8 Administration. Available at: http://www.axlr8.eu/assets/axlr8-progress-report-2012.pdf (Accessed 29.10.15).
24. Anon. (2010). *Eurobarometer Survey Shows Public Concern on Animal Testing*. London, UK: European Coalition to End Animal Experiments. Available at: http://www.eceae.org/no/category/watching-brief/76/eurobarometer-survey-shows-public-concern-on-animal-testing (Accessed 29.10.15).