An analysis of the use of animals in predicting human toxicology and drug safety: a review

Dr Elisabeth Harley
Understanding Animal Research

Summary

Two papers published in the journal Alternatives to Laboratory Animals (ATLA) in 2013 and 2014 present a series of analyses designed to evaluate the usefulness of animals as test subjects in preclinical drug development. The work was funded by the British Union for the Abolition and Vivisection (BUAV) and the Fund for the Replacement of Animals in Medical Experiments (FRAME), and authored by scientific advisor to the BUAV Dr Jarrod Bailey, BUAV CEO Michelle Thew, and former president of FRAME Dr Michael Balls. Based on the results of their analyses the authors conclude that dogs, mice, rats and rabbits are highly unreliable at predicting toxicity outcomes. Here we evaluate their methodology, their findings and conclusions.

The methods used in these studies are established diagnostic tools that are widely used to evaluate medical tests, and are appropriate to the questions that the authors ask. Our concern is that the authors' conclusions are not fully supported by the results of their data analysis.

In summary, our concerns are:

- Both studies show that **dog, mouse, rat and rabbit tests are extremely good at detecting which compounds will be toxic to humans**. However, despite the strong statistical significance of these results the authors dismiss them as unreliable without an adequate explanation.

- In the 2013 paper the authors attempt to discredit an existing method of test evaluation (the positive likelihood ratio) in favour of their own method (the positive likelihood ratio). These two metrics are not mutually exclusive. Rather they are different ways of examining the relationship between outcomes in humans and outcomes in animals. It is unclear why the authors would wish to do this, unless it was to cast unfair doubt on previous, similar work.

It is of paramount importance that the laws governing drug development are based upon sound scientific evidence. However, the reason to use animals during preclinical testing is to ensure that potentially dangerous compounds do not end up being given to human patients. Within these two studies (Bailey et al 2013, 2014) there are too many details omitted or not adequately explained, which cast significant doubts upon the conclusions drawn by the authors.

Testing for toxicity

For their analysis Bailey et al (2013, 2014) obtained data for 2,366 drugs, for which adverse drug reactions (ADRs) had been tested for in both humans and animal models, either dogs, mice, rats or rabbits (table 1). For each ADR, e.g. kidney failure, they identified the numbers of drugs that caused ADRs in both humans and animals (true positives); that caused ADRs in animals but not humans (false positive); that did not cause ADRs in animals but did in humans (false negative); and that did not cause ADRs in either dogs or humans (true negative).

It is important to note that this data is heavily biased towards true negative effects, as the drugs involved had been cleared to enter human trials.

From this data it is possible to calculate a series of diagnostic metrics that measure an animal model's ability to identify toxic and non-toxic outcomes in humans (see appendix 1).

Results

Using the likelihood ratio analysis Bailey et al (2013, 2014) found that PLRs were generally high in all four animal species, but iNLRs very low (table 2).

Table 1. *Results of the positive and negative likelihood ratio analyses*

	PLR		***iNLR***
Dogs			
Median	28.43	Median	1.10
Range	4.68-548.71	Range	1.01-1.94
Rats			
Median	253	Median	1.82
Range	24-2360	Range:	1.02-100
Mice			
Median	203	Median	1.39
Range	23-2361	Range	1.03-50
Rabbits			
Median	101	Median	1.12
Range	13-1348	Range	1.01-1.92

In these analyses a likelihood ratio of greater than 1 indicates that an animal test does contribute significant insight into whether compound will be toxic or non-toxic to humans. Values that are very close to 1 have very little or no significance.

Bailey et al (2013, 2014) argue that *"The critical observation for deciding whether a candidate drug can proceed to testing in humans in the absence of toxicity in tests on animals"*. By this logic the results of the iNLR analyses are more important in determining the usefulness of animal models in predicting human responses.

In their conclusion Bailey et al (2013) provide a quantitative illustration to demonstrate that the predictive value of dog tests is barely greater than chance. They posit a hypothetical candidate compound that has a 70% probability of "freedom from ADRs". Using the median iNLR value calculated by the study the probability "that the compound will also show no toxic effects in humans will have been increased by the animal resting from 70% to 72%".

This outcome is obtained through a series of calculations that use the likelihood ratio and the pre-test probability that a compound will be toxic or non-toxic to a human (in this example 70%) to calculate the post-test probability, or the additional probability contributed by the dog test.

	Calculation	*Example*
Pre-test odds	= (Pre-test probability/(1 − Pre-test probability))	0.7/(1 – 0.7) = **2.33**
Post-test odds	= Pre-test odds × Likelihood ratio	2.33 x 1.01 = **2.57**
Post-test probability	= Post-test odds / (Post-test odds + 1)	2.57/(2.57 +1) = **0.72**

So when looking at true negative results – when a compound is non-toxic in both dogs and humans – testing the compound on dogs does not contribute a significant amount of additional certainty that a compound will not be toxic.

Both papers note that the PLR values were high *"implying that compounds that are toxic in dogs are likely also to be toxic in humans."* However they go on to dismiss the PLR results because *"the PLRs vary considerably (range 4.7 – 548.7)… the reliability of this aspect of canine models cannot be generalised or regarded with confidence."*

Problems

Prevalence

The likelihood ratio method used in both studies (Bailey et al 2013, 2014) is entirely appropriate for the questions asked. However, in the 2013 study the authors go to some lengths to discredit an alternate, and widely used, method for assessing whether animal models show similar drug responses to humans. The positive predictive value (PPV) and its negative counterpart (NPV) are the proportions of toxic or non-toxic outcomes that are observed in both animals and humans (see Appendix 1). Bailey et al (2013) note that:

"…PPVs [are] dependent on the prevalence of toxicity in compounds, and thus an inappropriate measure of the reliability of the test with any specific compound."

Prevalence is the **occurrence of a specific human toxic effect within a set of drugs**. For example, in the dataset used by Bailey et al (2013), the prevalence for kidney neoplastic disorder is 0.015, meaning that this effect occurs in humans only 1.5% of the time .

The likelihood of obtaining a true positive outcome – the positive predictive value – is strongly associated with prevalence ($r^2_{434} = 0.4219$, $p < 0.01$; fig. 2). For high values of prevalence, for example tissue level effects on bodily fluid (prevalence 0.61), the likelihood of obtaining a true positive result is very high. However, for low values of prevalence such as kidney failure (prevalence 0.132), the likelihood of obtaining a true positive result becomes very variable and far less reliable.

Bailey et al (2013) argue that this relationship makes PPV an *"inappropriate measure of the reliability of the test with any specific compounds"*. To further discredit the usefulness of PPV they compare ranked values of PPV against their corresponding ranked likelihood ratio values (fig. 3.), claiming that this shows the *"misclassifications and misplaced assumptions of the accuracy of canine data for the prediction of human adverse drug reactions".*

It is worth noting that even though their appears to be a random distribution of points in fig. 3, as noted by the authors, there is in fact a weak but statistically significant positive association between the two variables ($r^2_{434} = 0.0090$, $p = 0.0474$).

What the authors fail to note is that because of the effect of prevalence on PPV outcomes it is common practice to adjust the value of PPV to take into account this. As a result it is unclear why the authors feel that it is important to discredit PPV as a statistical method. **The likelihood ratio and the positive predictive value are not mutually exclusive metrics**; they simply ask different questions about the relationship between dog and human toxicity outcomes.

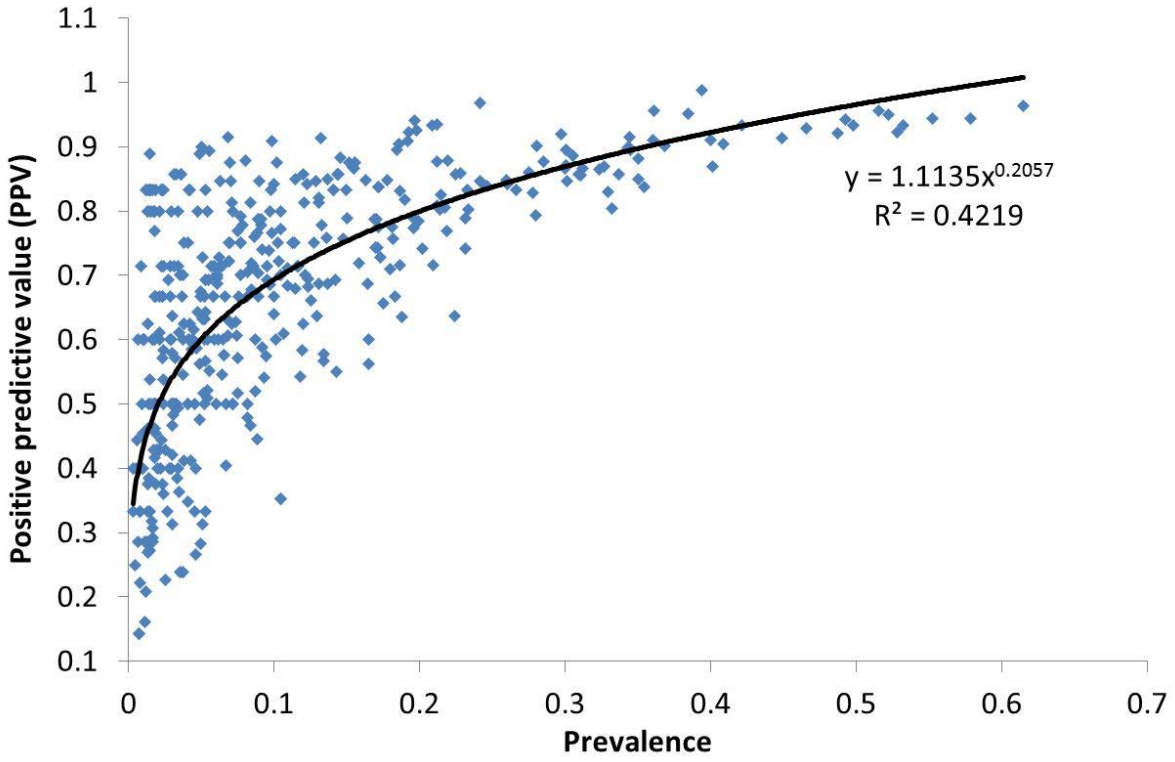

Fig. 2. *The relationship between prevalence and PPV for the 436 toxic effects within Bailey et al (2013)'s dataset. Each data point represents data obtained from the same 2,366 drugs for a specific biological effect e.g. kidney failure.*

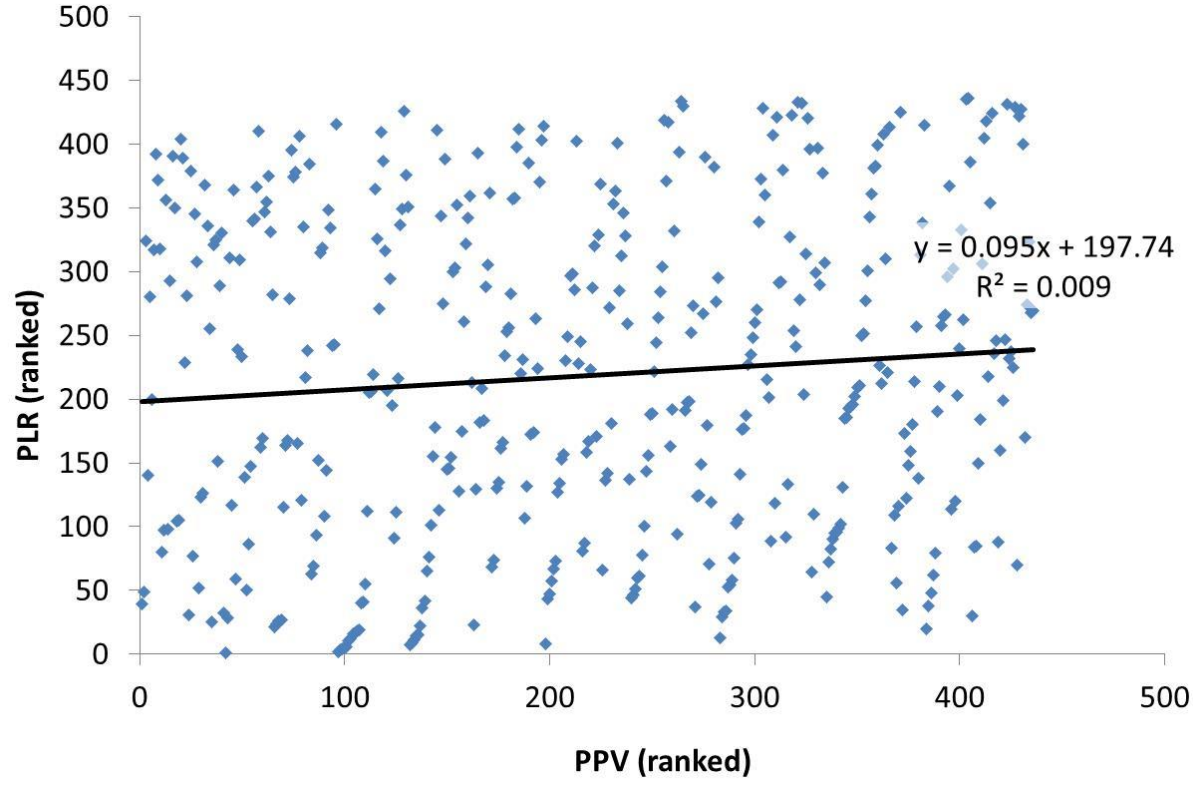

Fig. 3. *Ranked values of PPV plotted against ranked values of PLR. Reproduced from Bailey et al (2013).*

PPV

If we examine the occurrence of true positive and true negative outcomes within both the 2013 and 2014 datasets, we find that both values are high (table 3). This suggests that animals and humans show the same outcomes, whether toxic or non-toxic, in the majority of cases.

Table 3. *Results of the positive and negative predictive value analyses (uncorrected for prevalence)*

	PPV		***NPV***
Dogs			
Median	0.70	Median	0.94
Range	0.14-0.99	Range	0.48-0.99
Rats			
Median	0.91	Median	0.98
Range	0.17-0.99	Range:	0.61-1.00
Mice			
Median	0.93	Median	0.96
Range	0.33-0.99	Range	0.38-1.00
Rabbits			
Median	0.99	Median	0.99
Range	0.67-0.99	Range	0.70-0.99

These values remain high when corrected for prevalence, the importance of which is discussed in the previous section.

In dogs, for example, a predictive value analysis shows that true positives occur around 70% of the time, suggesting that the dog test is actually very good at predicting human toxic outcomes.

While the range of PPVs is high (0.14 – 0.99), only 76 of the 436 PPV values fall below 50%, suggesting that in the majority of cases, dog tests outcomes match those of humans. True negatives occur 94% of the time, with only one of the NPV values falling below 50%, again suggesting that the results of dog and human tests are often similar.

For the other species, four out of 404 mouse PPVs and zero out of 275 rabbit PPVs fell below 50%. In the rat data 359 PPVs out of 610 fell below 50%.

The fact these results suggest that animals are in fact capable of accurately predicting human responses, at odds with the conclusions drawn by the authors (Bailey et al 2013, 2014), might explain why the authors are so keen to discredit PPV and NPV as valid test metrics.

PLR interpretation

In both papers, Bailey et al (2013, 2014) choose to focus their analysis and interpretation on the inverse negative likelihood ratio test (iNLR), and tend to discredit and ignore its positive counterpart (PLR), as discussed earlier (see 'Results').

Detecting potentially toxic compounds before they are given to humans during clinical trials is of paramount importance during drug development. This is why animals are used as part of regulatory safety testing. The fact that all the models studied produced such high PLRs (table 1) implies that, far from being as flawed as the authors suggest, **animal models are extremely good at detecting effects that may be harmful to humans.**

The PLR values do vary widely as noted by the authors, but a simple additional analysis, described previously, shows that even at the lowest PLR values animal tests can add enormous support to our understanding of whether a compound will be toxic to humans or not.

Again using dogs as an example, and using the median PLR median value of 28.43, we can see that the contribution of the dog tests to the post-test probability is large. In other words, conducting a test on dogs can dramatically increase our knowledge of a compound's potential effects in humans.

Pre-test probability	*Pre-test odds*	*Post-test odds*	***Post-test probability***
10%	0.11	3.16	**76%**
20%	0.25	7.11	**88%**
30%	0.43	12.18	**92%**
40%	0.67	18.95	**95%**
50%	1.00	28.43	**97%**
60%	1.50	42.65	**98%**
70%	2.33	66.34	**99%**
80%	4.00	113.72	**99%**
90%	9.00	255.87	**100%**

If we repeat those calculations for the lowest value of PLR obtained (4.67), the contribution of the dog tests to detecting toxic effects is still high.

Pre-test probability	*Pre-test odds*	*Post-test odds*	***Post-test probability***
10%	0.11	0.52	**34%**
20%	0.25	1.17	**54%**
30%	0.43	2.01	**67%**
40%	0.67	3.12	**76%**
50%	1.00	4.68	**82%**
60%	1.50	7.02	**88%**
70%	2.33	10.92	**92%**
80%	4.00	18.72	**95%**
90%	9.00	42.12	**98%**

This analysis would suggest that far from being unreliable, dog tests do have some importance for detecting toxic effects. Bailey et al (2013) note in their methodology that *"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 a compound under test will be toxic to humans."* For dogs all the PLR values calculated are greater than 1, with the vast majority (408/436) being greater than 10.

Conclusions

The likelihood ratio and predictive value analyses from both studies (Bailey et al 2013, 2014) can be summarised as follows:

1) Animal tests show the same toxic and non-toxic outcomes as human tests for the same drugs in over 50% of cases on average;
2) Animal tests **can** significantly increase our ability to predict toxic outcomes in humans, but **cannot** significantly increase our ability to predict non-toxic outcomes in humans.

As mentioned in the results section of this paper, Bailey et al (2013, 2014) argue that the failure of animal tests to add evidential weight to our ability to predict non-toxic outcomes is the most significant result of their study. They argue that because the animal tests do not contribute any further certainty, not only will many potential drugs that could cause harm to humans proceed to human tests, but that many non-harmful drugs will also be classed as harmful and not developed further.

However, the outcomes of the analyses that they chose to ignore (PLR) or attempt to debunk (PPV) would appear to contradict this conclusion. One of the purposes of animal tests during drug development is to determine whether a candidate drug is safe to be given to a human. Therefore the analysis used by the authors, which shows the high ability of dog tests to show the same toxic outcomes as humans coupled with the strong evidential weight that they add, actually demonstrates the usefulness of dogs as a second species in toxicity testing.

Bailey et al (2013) conclude that *"the predictions [dogs] can provide are little better than those that could be obtained by chance"*. This conclusion is, as we have demonstrated, deeply flawed and based on an assessment of only half of their own results. This is the most serious weakness of their paper.

References

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.

Bailey, J., Thew, M., & Balls, M. (2014). An Analysis of the Use of Animals in Predicting Human Toxicology and Drug Safety. *ATLA* 42, 181 - 199.

Appendix 1: Diagnostic testing metrics

Sensitivity and specificity

Bailey et al (2013) use a simple diagnostic test framework as the basis for their analyses (table 1). The outcome of an animal safety test can be positive (predicting that a drug will be toxic to humans) or negative (predicting that the drug will be non-toxic to humans). These outcomes may or may not match the actual toxicity in humans. Within this setting:

- **True positive:** drugs toxic to humans are correctly identified by animal tests;
- **False positive:** non-toxic drugs are incorrectly identified as toxic;
- **True negative**: non-toxic drugs are correctly identified as non-toxic; and
- **False negative**: toxic drugs are incorrectly identified as non-toxic.

From this data it is possible to calculate two important diagnostic metrics (equations given in table 1). The **sensitivity** describes the ability of the dog test to identify toxic compounds; the **specificity** describes a dog test's ability to identify non-toxic compounds.

Table 2. *The relationship between sensitivity, specificity and predictive values for animal toxicity screening.*

	Toxic in human	**Non-toxic in human**	
Toxic in animal model	True Positive (a)	False Positive (b)	**Positive predictive value:** $\frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$
Non-toxic in animal model	False Negative (c)	True Negative (d)	**Negative predictive value:** $\frac{\text{True Negative}}{(\text{False Positive} + \text{True Negative})}$
	Sensitivity $= \frac{\sum \text{True positive}}{\sum \text{Toxic in human}}$	**Specificity** $= \frac{\sum \text{True negative}}{\sum \text{Non-toxic in human}}$	

However, neither of these metrics can be used in isolation to determine whether an animal test is useful for predicting human outcomes. Bailey et al (2013) note that previous analyses of the reliability of animal tests have used the sensitivity and specificity to calculate the **positive predictive value (PPV)**. This is proportion of all toxic outcomes in humans that are true positives.

There is a fourth metric, not mentioned by Bailey et al (2013), which is the **negative predictive value (NPV)**. This is the proportion of all non-toxic outcomes that are true negatives.

Likelihood ratios

The positive predictive and negative predictive values **identify the probability that the results of a human test will match the results of an animal test**. However, Bailey et al (2013) ask a subtly different question in their study, which requires a different analytical method.

Bailey et al (2013, 2014) assess the **evidential weight** provided by an animal toxicity test. For example, if a potential drug compound has a 40% probability of being toxic to humans Bailey et al (2013, 2014) ask how much additional certainty is added by performing an animal test.

The method used to answer this question is the **likelihood ratio**. As with predictive values, there are both positive and negative likelihood ratios: the **positive likelihood ratio (PLR)** relates to toxic outcomes in humans, while the **inverse negative likelihood ratio (iNLR)** relates to non-toxic outcomes. Both PLR and iNLR are calculated using sensitivity and specificity (fig 1).

Positive likelihood ratio	$= \frac{\text{sensitivity}}{(1\ -\ \text{specificity})}$
Inverse negative likelihood ratio	$= \frac{\text{specificity}}{(1\ -\ \text{sensitivity})}$

Fig. 1. *Equations for calculating the positive and negative likelihood ratios*