

## Current nonclinical testing paradigm enables safe entry to First-In-Human clinical trials: The IQ consortium nonclinical to clinical translational database



Thomas M. Monticello<sup>a,\*</sup>, Thomas W. Jones<sup>b</sup>, Donna M. Dambach<sup>c</sup>, David M. Potter<sup>d</sup>, Michael W. Bolt<sup>e</sup>, Maggie Liu<sup>f</sup>, Douglas A. Keller<sup>g</sup>, Timothy K. Hart<sup>h</sup>, Vivek J. Kadambi<sup>i</sup>

<sup>a</sup> Comparative Biology and Safety Sciences, Amgen, Thousand Oaks, CA 91320, USA

<sup>b</sup> Eli Lilly and Company, Indianapolis, IN 46285, USA

<sup>c</sup> Safety Assessment, Genentech, South San Francisco, CA 92056, USA

<sup>d</sup> Drug Safety Research and Development, Pfizer, Groton, CT 06340, USA

<sup>e</sup> Drug Safety Research and Development, Pfizer, Cambridge, MA 02139, USA

<sup>f</sup> IQ Consortium, Washington, DC 20005, USA

<sup>g</sup> Preclinical Safety, Sanofi, Bridgewater, NJ 08807, USA

<sup>h</sup> GlaxoSmithKline, King of Prussia, PA 19406, USA

<sup>i</sup> Nonclinical Development Sciences, Blueprint Medicines, Cambridge, MA 02139, USA

### ARTICLE INFO

#### Keywords:

Animal testing  
Translational  
Nonclinical  
Clinical  
Safety  
Concordance

### ABSTRACT

The contribution of animal testing in drug development has been widely debated and challenged. An industry-wide nonclinical to clinical translational database was created to determine how safety assessments in animal models translate to First-In-Human clinical risk. The blinded database was composed of 182 molecules and contained animal toxicology data coupled with clinical observations from phase I human studies. Animal and clinical data were categorized by organ system and correlations determined. The  $2 \times 2$  contingency table (true positive, false positive, true negative, false negative) was used for statistical analysis. Sensitivity was 48% with a 43% positive predictive value (PPV). The nonhuman primate had the strongest performance in predicting adverse effects, especially for gastrointestinal and nervous system categories. When the same target organ was identified in both the rodent and nonrodent, the PPV increased. Specificity was 84% with an 86% negative predictive value (NPV). The beagle dog had the strongest performance in predicting an absence of clinical adverse effects. If no target organ toxicity was observed in either test species, the NPV increased. While non-clinical studies can demonstrate great value in the PPV for certain species and organ categories, the NPV was the stronger predictive performance measure across test species and target organs indicating that an absence of toxicity in animal studies strongly predicts a similar outcome in the clinic. These results support the current regulatory paradigm of animal testing in supporting safe entry to clinical trials and provide context for emerging alternate models.

### 1. Introduction

Major goals of nonclinical toxicology testing in drug development are to ensure human safety, aid in establishing a starting clinical dose, and identify potential organs of toxicity along with safety biomarkers that can be monitored in the clinic. The nonclinical data package required for a First-In-Human (FIH) trial includes animal toxicology studies that are designed to characterize potential toxic effects under the conditions of the supported clinical trial, as reported in the Investigational New Drug (IND) or Clinical Trial Application (CTA) and

the Investigational Brochure (IB) (EMA, 2012).

The purpose of the IB is to present the chemical, pharmaceutical, pharmacological, toxicological, and clinical information on the drug candidate to the investigator conducting the clinical trial. Nonclinical safety scientists interpret the body of toxicological data obtained from *in vitro*, mechanistic and animal safety studies to determine the potential risks that are then identified in the IB to inform the clinician. A potential risk is defined as an untoward occurrence for which there is some basis for suspicion of an association with the medicinal product of interest but where this association has not been confirmed; for example,

\* Corresponding author at: MS 29-2-A, 1 Amgen Center Drive, Thousand Oaks, CA 91320, USA.  
E-mail address: [tmontice@amgen.com](mailto:tmontice@amgen.com) (T.M. Monticello).

<http://dx.doi.org/10.1016/j.taap.2017.09.006>

Received 31 July 2017; Received in revised form 2 September 2017; Accepted 7 September 2017

Available online 08 September 2017

0041-008X/ © 2017 Elsevier Inc. All rights reserved.

a nonclinical toxicology finding that has not yet been observed in clinical studies (EMA, 2012).

Currently, the conduct of animal toxicology studies is based on historical precedence and International Council for Harmonization (ICH) recommendations, centered on the assumption that the animal model of choice and the design of the toxicology study provide value in identifying potential human hazards in the clinic that help ensure patient safety (Mangipudy et al., 2014). Limited publications exist, however, that scientifically address correlations between observed toxicities in animal models to adverse events in the clinic and the methodologies in assessing the concordance have varied (Olson et al., 2000; Tamaki et al., 2013). Furthermore, there has been debate regarding the value of animal models in both drug discovery and development (Shanks et al., 2009; van Meer et al., 2012; Bailey et al., 2013; Hartung, 2013; Greek and Menache, 2013; Mangipudy et al., 2014; Bailey et al., 2015). Numerous emerging academic, government and industry efforts are aimed to refine or replace animal studies, and this analysis may help prioritize gaps and set the standard for this work.

A principal goal of our analysis was to interrogate the relevance of nonclinical study design and interpretation of the animal toxicology data to identify and characterize hazards (target organs), provide risk evaluation and manage those risks. To accomplish this, we developed a database that is prospective, in that the potential safety risks based on animal data and reported in the regulatory dossier, were followed through the completion of phase I trials. As phase I trials are generally < 4-weeks in duration with the goals of establishing pharmacokinetics, pharmacodynamics and possible adverse events, we compared nonclinical and clinical datasets with similar endpoints and study duration. While phase I trials may not always dose humans to exposures equal to the exposures obtained in the animal studies at which effects were observed, the comparison of hazards can still be useful. The creation and analysis of this database aligns with the 2011 FDA strategic plan to advance regulatory science and modernize toxicology to enhance product safety (FDA, 2011).

## 2. Materials and methods

### 2.1. Study design

An industry-wide nonclinical to clinical translational database was created and analyzed to determine how nonclinical safety assessments in animal models translate to First-In-Human (FIH) clinical risk. The blinded database was composed of 182 molecules and contained FIH-enabling animal toxicology and safety pharmacology data (rodent, dog, and/or nonhuman primate (NHP) also referred to as monkey), coupled with the clinical observations from completed phase I human studies. Contingency tables were populated by classifying true positive, false positive, true negative and false negative events. Concordance statistical analytics were applied based on species and organ system.

### 2.2. Biopharmaceutical company contributions to the database

Biopharmaceutical company sponsors were requested to submit molecules that had completed phase I clinical trials from the period of 2006 to 2011. Our database excludes drug candidates that were dropped from development prior to FIH. Sponsors were instructed to submit the molecules chronologically in order to avoid bias in therapeutic indication, modality or clinical outcome. The sponsors extracted the data from the IB and IND/CTA. The data submitted were based on the contributing sponsor's interpretation of the level of concern for a particular safety signal (based on the animal data), which determined what was provided in the IB as a potential safety risk.

### 2.3. Nonclinical data

A common excel worksheet template was created and provided to

each sponsor; data was codified to maintain confidentiality as to the identification of the sponsor and drug candidate. Sponsors were requested to enter animal toxicology data (general toxicology and in vivo safety pharmacology data) that included drug candidate modality, preclinical species used in safety testing (e.g. mouse, rat, dog, non-human primate), therapeutic indication, duration of animal study, duration of recovery period (if applicable), route of administration, dosing schedule and, if applicable, the type of in vivo safety pharmacology study (e.g. CNS, respiratory or cardiovascular). Sponsors provided each potential safety liability obtained from the IB/IND/CTA with the supporting animal toxicology and safety pharmacology data that included in-life observations, complete blood count (CBC) data, clinical chemistry data, unique biomarker data (if applicable) and histopathology data. Toxicokinetic data were also requested that included the area under the curve (AUC) and the maximum serum concentration (C<sub>max</sub>) at the corresponding dose level at which the potential safety liability was identified in the animal study. For each potential safety liability listed in the IB, the sponsor indicated if that safety liability was observed in the completed phase I trial. All in vivo studies were conducted after review by the respective Institutional Animal Care and Use Committee and adhered to the NIH Guide for the Care and Use of Laboratory Animals.

### 2.4. Phase I clinical data

Sponsors were requested to provide clinical data in the excel datasheet that included clinical observations (e.g. adverse events, other findings, etc.) and corresponding pharmacokinetic data at which the clinical observations were observed. The sponsor entered whether the clinical observation was also listed in the IB as a potential safety risk based on the animal data. Trials were designed according to the current Declaration of Helsinki and conducted in accordance with Good Clinical Practice Guidelines. Written informed consent was obtained from each participating patient before study entry.

### 2.5. Data warehouse

For each molecule (IND/CTA) a single datasheet containing the animal and clinical data was uploaded by the sponsor to a password protected common server (cloud-based data warehouse, Amazon web services). No proprietary information on the molecule (e.g. company identification number, structure) was submitted. The data warehouse was managed and maintained by the IQ Consortium Secretariat that also served as an independent third party that maintained the anonymity of the blinded data provided by each sponsor. Each uploaded datasheet had a unique codified case number generated that identified the sponsor and molecule. Only the IQ Consortium Secretariat knew the identity of the sponsor.

### 2.6. Database evaluation based on species and organ categories

The potential safety liabilities identified with the animal toxicology data (mouse and rat as rodent; nonrodent characterized as beagle dog, NHP and pig) and the clinical observation data from completion of phase I were categorized by organ system. The following organ system categories were used based on Olson et al. (2000): liver, hematology, gastrointestinal, cardiovascular, nervous system, biochemical (e.g. clinical chemistries), renal, pulmonary, musculoskeletal, cutaneous, ophthalmology and reproductive. Data from the reproductive category was collected but not analyzed as the clinical outcome in phase I was not reported. An organ system category of 'other' was used, in general, for constitutional symptoms that could not be categorized by organ system (e.g. nausea, headache, dizziness, fatigue, insomnia, anxiety, discomfort, euphoric mood). The nonclinical to clinical findings were cross-classified using contingency tables (Table 1). A contingency table may also be referred to as a "2 × 2 table", a "confusion matrix" (Ting,

**Table 1**

The contingency table (A) in our database summarizes the agreement between a non-clinical animal model finding (positive or negative) and the respective clinical finding (positive or negative). (B) Definition of the counts in each cell of the contingency table.

A		
	Clinical result	
	Clinical outcome Positive	Clinical outcome Negative
Nonclinical result		
Animal test positive	True positive	False positive
Animal test negative	False negative	True negative

B	
True positive (TP)	The identified nonclinical safety liability based on animal data was observed in the clinic.
True negative (TN)	Lack of an identified nonclinical safety liability based on animal data and no clinical safety liability observed.
False positive (FP)	An identified nonclinical safety liability based on animal data was not observed in the clinic.
False negative (FN)	No identified nonclinical safety liability based on animal data but a clinical observation was identified in the clinic.

2010) or a “concordance table”.

## 2.7. Statistical animal to human concordance parameters

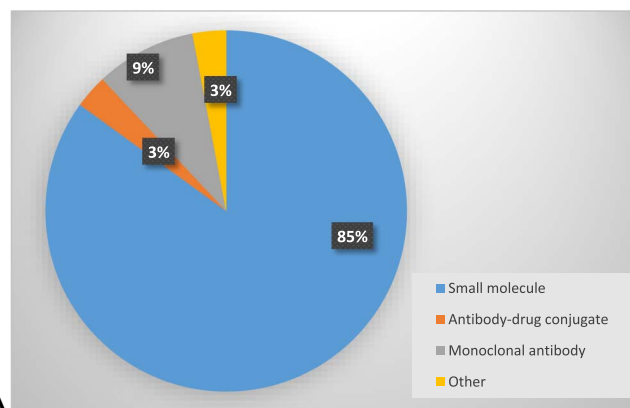
There is no single statistical measure that is best in comparing the concordance of animal data to humans. The concordance parameters we evaluated to determine the potential impact of animal testing in predicting clinical outcomes are presented in Table 2. The most commonly cited measures of test performance, often used in the context of diagnostics, are sensitivity and specificity, which determine how well a test classifies known positive (sensitivity) and negative (specificity) outcomes. In isolation, these parameters offer limited insight into the predictive performance of a test and the calculated values are dependent on the definition used to establish the threshold or “cut-point” for a positive test result. For example, in the current analysis, the definition of a positive nonclinical signal of concern was any adverse effect as determined by the sponsor, that was significant enough to be included in the IB. When the definition of a positive test result is made more stringent, requiring a stronger nonclinical signal, sensitivity will decrease as specificity increases.

We placed the most emphasis on positive predictive value (PPV) and negative predictive value (NPV) as they are more aligned with non-clinical to clinical safety translation, that is, we observe a result in an animal study that we deem important and provide that potential safety liability in the regulatory dossier that helps advise the clinician on what

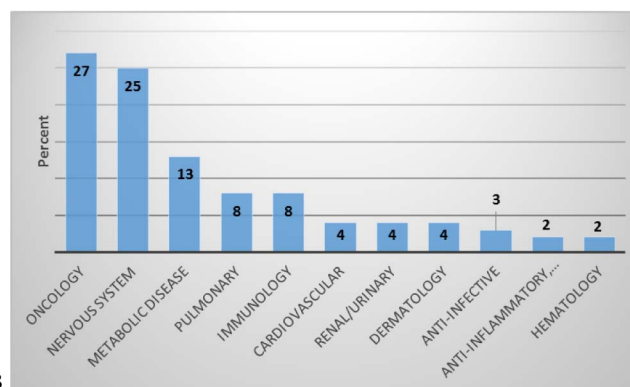
**Table 2**

Concordance parameters evaluated to determine impact of animal testing in predicting clinical outcomes.

Measure	Definition	Formula
Sensitivity (SEN)	The proportion of positive clinical findings that had positive nonclinical findings	$TP / (TP + FN)$
Specificity (SPE)	The proportion of negative clinical findings that had negative nonclinical findings	$TN / (TN + FP)$
Positive predictive value (PPV)	The proportion of positive nonclinical findings that had positive clinical findings	$TP / (TP + FP)$
Negative predictive value (NPV)	The proportion of negative nonclinical findings that had negative clinical findings	$TN / (TN + FN)$
Clinical positive prevalence (PRV)	Proportion of positive clinical findings	$(TP + FN) / (TP + FP + FN + TN)$
Odds positive “pre”	Overall odds of a positive clinical finding	Prevalence positive / (1 – prevalence positive)
Likelihood ratio positive (LR +)	The increase in odds of a positive clinical finding that is due to knowledge of a positive nonclinical finding	Odds Positive “post” / odds positive “pre”, or equivalently: sensitivity / (1 – specificity)
Inverse Likelihood Ratio Negative (iLR –)	The increase in odds of a negative clinical finding that is due to knowledge of a negative nonclinical finding	Odds Negative “post” / odds negative “pre”, or equivalently, specificity / (1 – sensitivity)



A



B

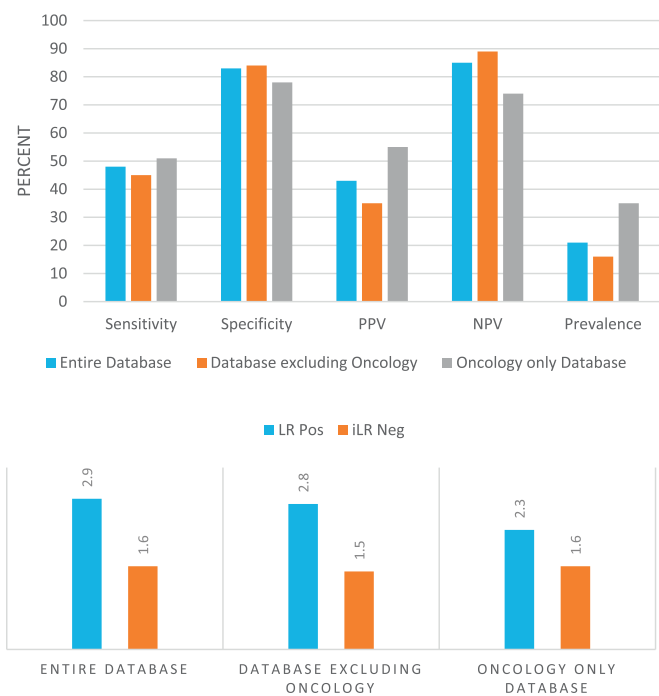
Fig. 1. Modality (A) and therapeutic distribution (B) of the database. Drug candidate submissions consisted of small molecules, antibody-drug conjugates, monoclonal antibodies or ‘other’ (not provided to database). The oncology indication was the highest therapeutic category of database submissions followed by nervous system and metabolic diseases.

may occur in the clinic. We also determined likelihood ratio positive (LR +) and inverse likelihood ratio negative (iLR –), as statistical parameters not influenced by clinical positive prevalence. In our setting, LR + is the ratio of the probability that a molecule that tests positive in the animal will result in an adverse clinical outcome versus the probability that a compound that tests positive in the animal will not cause an adverse clinical outcome. Inverse likelihood ratio negative (iLR –) is defined as the inverse of what is typically presented as the “likelihood ratio negative”, so that values > 1 reflect improved odds of successfully predicting a negative clinical finding. The “pre” and “post” terminology in the description of likelihood ratio (LR) is for consistency with the use of these measures with diagnostic tests, where

**Table 3**  
Concordance results by organ category and species.

Organ category	Prev (%)	Species	SEN (%)	SPE (%)	PPV (%)	NPV (%)	LR +	iLR –
Cutaneous	22	Rodent	9	97	40	82	2.8	1.1
		Dog	14	98	50	91	9.1	1.1
		NHP	11	98	67	70	4.2	1.1
Musculoskeletal	8	Rodent	10	95	14	92	1.8	1.1
		Dog	14	100	nd <sup>a</sup>	91	inf <sup>b</sup>	1.2
		NHP	0	95	0	93	0	0.9
Biochemical	11	Rodent	50	94	54	93	8.8	1.9
		Dog	25	91	14	95	2.8	1.2
		NHP	33	92	43	88	4.2	1.4
Nervous system	18	Rodent	74	84	52	93	4.5	3.2
		Dog	69	82	52	90	3.8	2.6
		NHP	63	98	83	94	31.9	2.6
Cardiovascular	23	Rodent	3	94	17	75	0.6	1.0
		Dog	87	62	50	91	2.3	4.7
		NHP	20	84	20	84	1.2	1.0
Hematology	14	Rodent	59	83	36	92	3.4	2.0
		Dog	60	89	46	93	5.2	2.2
		NHP	25	86	22	88	1.8	1.1
Liver	16	Rodent	65	69	33	89	2.1	1.7
		Dog	50	73	27	88	1.8	1.3
		NHP	27	94	50	85	4.4	1.3
Ophthalmology	4	Rodent	0	98	0	96	0	1.0
		Dog	0	100	nd	94	nd	1.0
		NHP	0	100	nd	97	nd	1.0
Pulmonary	5	Rodent	17	89	7	95	1.5	1.1
		Dog	0	99	0	96	0	1.0
		NHP	33	100	nd	97	inf	1.5
Renal	5	Rodent	43	88	19	96	3.7	1.5
		Dog	50	91	33	95	5.4	1.8
		NHP	nd	92	0	100	nd	nd
Gastrointestinal	33	Rodent	25	84	43	69	1.5	1.1
		Dog	54	74	52	76	2.1	1.6
		NHP	48	97	91	77	18.1	1.9
'Other'	71	Rodent	3	33	9	14	0.0	0.3
		Dog	0	36	0	15	0.0	0.4
		NHP	6	50	25	16	0.1	0.5

<sup>a</sup> Not determined.  
<sup>b</sup> Infinity.



**Fig. 2.** Database concordance parameters excluding oncology indication. PPV, positive predictive value; NPV, negative predictive value, LR Pos, likelihood ratio positive, iLR Neg, inverse likelihood ratio negative.

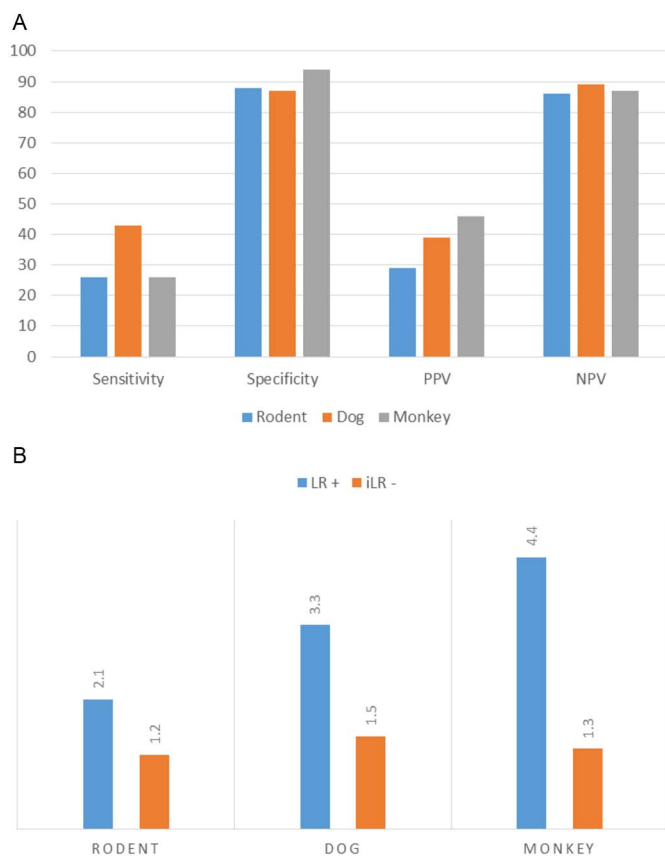
the focus is on the increased predictivity after (i.e., “post”) observing results from a diagnostic test. LR values > 1 increase the probability of the outcome. LR from 2 to 5 yield small increases in the post-hoc probability, from 5 to 10 moderate increases and above 10, large increases (Grimes and Schulz, 2005). As a general rule, a test is considered “diagnostic” in predicting a positive outcome when the LR + is ≥ 10 or for predicting a negative outcome when the iLR – is ≥ 10.

2.8. Concordance of database excluding oncology indication

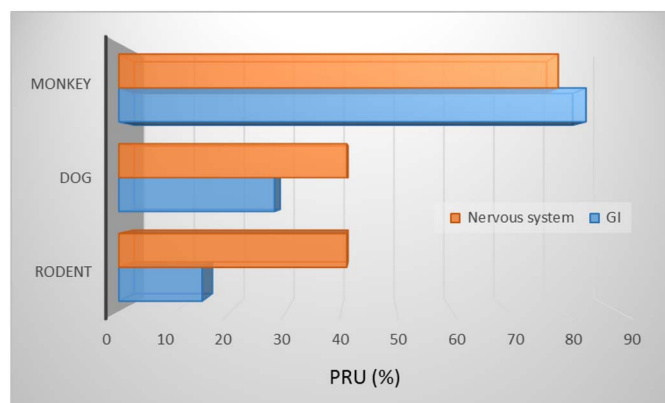
Oncology clinical trials enroll cancer patients in contrast to phase I trials for most other indications, which are conducted in ‘healthy’ volunteers. Reporting of adverse events in oncology trials is suboptimal and characterized by substantial selectivity and heterogeneity resulting from concurrent disease in the oncology clinical trial population (Sivendran et al., 2013). In order to estimate this effect, we also analyzed the database excluding the oncology indication for statistical concordance.

2.9. Determining proportionate reduction in uncertainty (PRU)

Introduced by Coulthard (2006), the proportionate reduction in uncertainty (PRU) is a measure of the reduction in diagnostic uncertainty of a test. Prior to the FIH-enabling toxicology studies, our prediction for the probability of an adverse effect in phase I is simply the clinical positive prevalence. Hence, the uncertainty regarding the potential for an adverse effect in phase I is equal to (100% – Prevalence). Following the animal toxicology study results, the updated



**Fig. 3.** Concordance parameters by test species evaluated. A. sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV). B. Likelihood ratio positive (LR +), inverse likelihood ratio negative (iLR -).



**Fig. 4.** Proportionate reduction in uncertainty (PRU) as a measure of the clinical utility of animal model testing: GI and nervous system examples. A large degree of uncertainty was resolved by the animal testing.

prediction for the probability is captured by the PPV ( $100\% - \text{PPV}$ ), which represents a reduction in uncertainty equal to ( $\text{PPV} - \text{Prevalence}$ ). Finally, to estimate what proportion of uncertainty is resolved, the following calculation is used:  $\text{PRU} = (\text{PPV} - \text{Prevalence}) / (100\% - \text{Prevalence})$ . We calculated the PRU for the monkey, dog and rodent for the gastrointestinal (GI) and nervous system organ categories as examples, to evaluate an alternative approach in examining the potential value of animal testing and human safety.

### 2.10. Evaluation of false positives and their impact on accuracy

As liver toxicity is an important cause of clinical attrition, the

relevance of animal models to detect liver injury has been questioned (Atienzar et al., 2016). It was out of scope for this initial concordance assessment to investigate animal to human exposure margins for all of the potential false positives in the database, however, an assessment was completed for the liver category. Exposure data (AUC at the dose eliciting the adverse liver signal in the animal and the highest clinical AUC exposure achieved in phase I for that molecule) were evaluated for the false positives in the liver category. We elected to categorize exposure margins (fold-difference from animal AUC to human AUC) as either animal AUC < 5-fold human AUC or animal AUC > than 5-fold human AUC. There were a total of 40 liver category false positives out of the total 132 datasets (oncology indication excluded); 32 of the 40 liver false positives were in the rat (80%); 16 of the 40 were in the dog (40%), and 3 of the 40 were in the monkey (8%). Seven liver false positive cases lacked complete animal and/or human supplemental exposure data and were excluded from exposure-fold analysis (rat only,  $n = 4$ ; rat and dog,  $n = 2$ ; dog only,  $n = 1$ ). Of the 33 liver false positives with supplemental AUC animal and human exposure data, there were 27 cases of rat false positives (82%), 14 dog false positives (42%), and 3 monkey false positives (9%).

### 2.11. Evaluation of false negatives and their impact on accuracy

Another confounding factor for the overall accuracy of our animal to human concordance is the impact of false negatives. In our database, the more subjective clinical adverse events (e.g. nausea, headache, fatigue, insomnia, anxiety) were categorized as the ‘other’ category. As animal models would not be expected to predict such constitutional signs, we performed statistics on the database with the ‘other’ category excluded.

## 3. Results

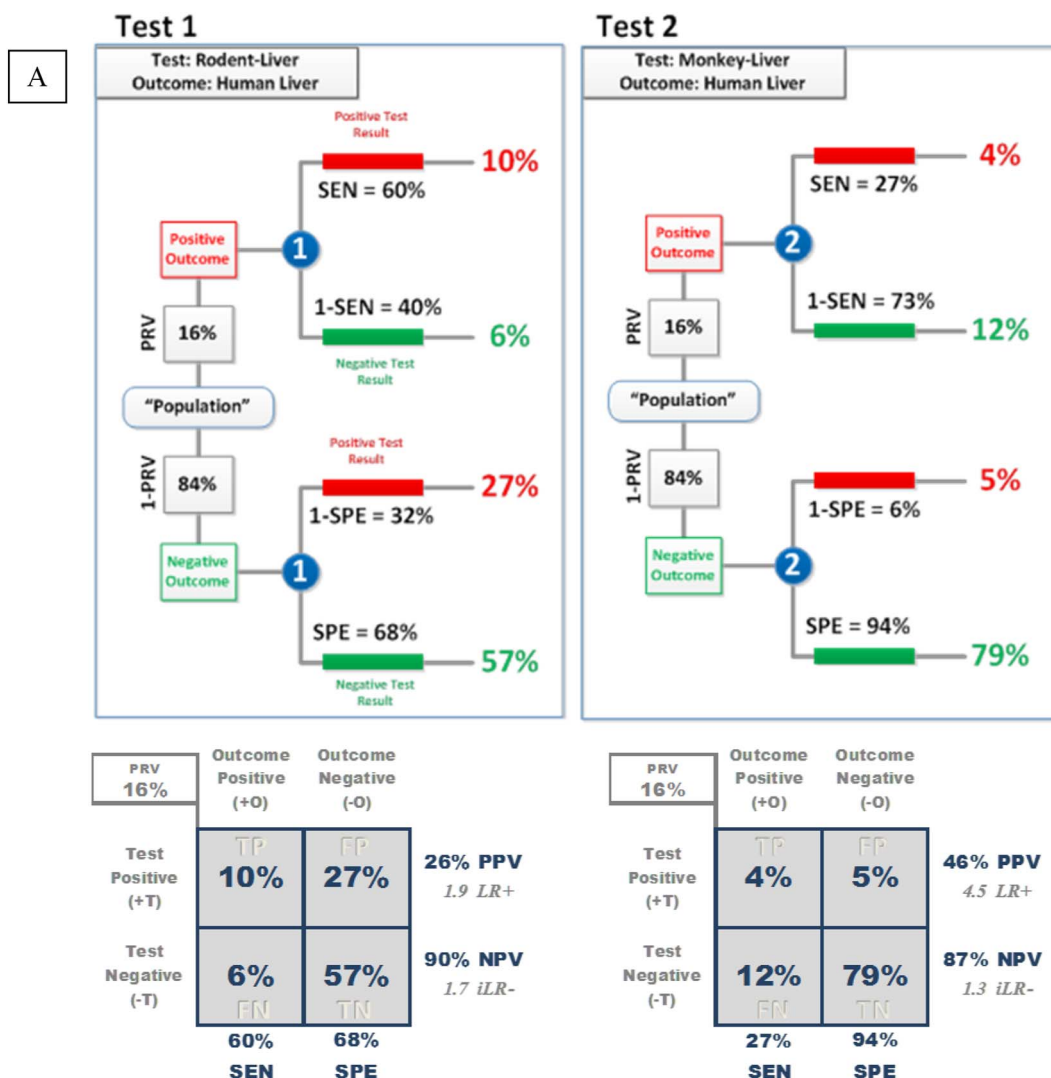
### 3.1. Database characteristics

Eighteen IQ consortium member company sponsors submitted datasets from 182 regulatory submissions that had completed phase I. Guidance on conducting nonclinical safety testing for the registration of new drug candidates is provided by the ICH (ICH S6, 1997; ICH S9, 2009; ICH M3(R2), 2009; ICH S6 (R1), 2011). Aligned with the high percentage of small molecule submissions in which a rodent and non-rodent animal test species is recommended by ICH guidance, the combination of rodent and beagle dog was the most common paradigm used in a drug development program (54%); the rodent and NHP combination was utilized in 32% of the programs, and the rodent and pig combination was used for 3% of the submissions. The use of three species (rodent, dog and NHP) for a particular molecule was rare (2%). Single species, i.e. only the NHP or only the pig, were used in 8% and 1% of the submissions, respectively.

The modality and therapeutic distribution of the database is presented in Fig. 1. For the time period of our database entries (2006–2011), Novel Drug Approvals by the FDA (Mullard, 2017), were composed of 19% biologics license applications (BLA, e.g. large molecule) and 81% new drug applications (NDA, e.g. small molecule), indicating our database was slightly under weighted in large molecule submissions. Of the 182 cases submitted to the database, 7 cases were negative in both the animal testing and in Phase I, another 8 cases were negative in Phase I but had positive animal target organ findings.

### 3.2. Clinical positive prevalence

Prevalence (pre-test probability), as we defined for our database, represents the probability that a molecule being advanced into phase I clinical development will result in an adverse clinical outcome (prior to any knowledge from nonclinical testing). Prevalence will depend on how the positive clinical outcome is defined. In our database, a positive



**Fig. 5.** Dual-species approach to concordance enhances animal testing performance. A: The probability trees and corresponding contingency tables for the prediction of human liver effects based on nonclinical testing using either rodent (Test 1) or monkey (Test 2) individually. However, since the resulting human risk assessment incorporates findings from both the rodent and nonrodent, it is important to consider the performance of the two tests in combination. B: The stacked probability tree for the prediction of human liver effects based on the combined nonclinical testing results of the rodent (Test 1) and monkey (Test 2). The contingency tables for both the “and” view as well as the “or” view are also shown. PRV, prevalence, PPV, positive predictive value, NPV, negative predictive value, LR +, likelihood ratio positive, iLR -, inverse likelihood ratio negative, SPE, specificity, SEN, sensitivity, TP, true positive, FP, false positive, FN, false negative, TN, true negative.

clinical outcome was any clinical signal observed in phase I that was reported in the database by the sponsor. No specific criteria were established for the sponsors in their reporting of abnormal findings or clinical symptoms, nor the requirement for an adjudication process for the clinical data submitted from the phase I studies. It is recognized that typical language in the informed consent documents emphasize that subjects report any changes in their health, however minor (Emanuel et al., 2015) and that the clinical event data can be subjective (Granger et al., 2008). A meta-analysis of phase I trial results have reported that 84% of adverse events were mild, 24% of such were unrelated to the study drug (Emanuel et al., 2015). The clinical positive prevalence for our database was 21%.

For any test, changes in prevalence are directly correlated to changes in PPV and inversely correlated to changes in NPV, so when prevalence is low, the PPV will also be low, even if both sensitivity and specificity are high (Altman and Bland, 1994). To illustrate the importance of prevalence in determining the predictive utility of non-clinical testing, we can consider the application of a more stringent definition of a positive human outcome. For example, if the definition of a positive human outcome was restricted to that of a serious adverse

drug reaction (e.g. death, life-threatening, hospitalization, disability or permanent damage), the prevalence would be reduced. At this lower prevalence, the PPV would decrease and the NPV would increase which would have an important effect regarding which of these predictive measures will be interpreted as performing strongest. To realize the true value of PPV or NPV as measures of performance, it is important to be cognizant of the prevalence of the outcome to be determined. The low prevalence identified in our dataset for clinical outcomes emphasizes the importance of the NPV of nonclinical animal models in ensuring patient safety.

### 3.3. Concordance

Overall results of the concordance analysis are presented in Table 3. The PPV varied by organ category and species. A PPV > 80% was observed in the NHP for both the nervous system and gastrointestinal organ categories. The LR + by organ category demonstrated variety in the ability of nonclinical species to predict clinical outcome. LR + values that may be considered to be ‘diagnostic’ were observed in the dog for the cutaneous organ category, and in the NHP for the nervous

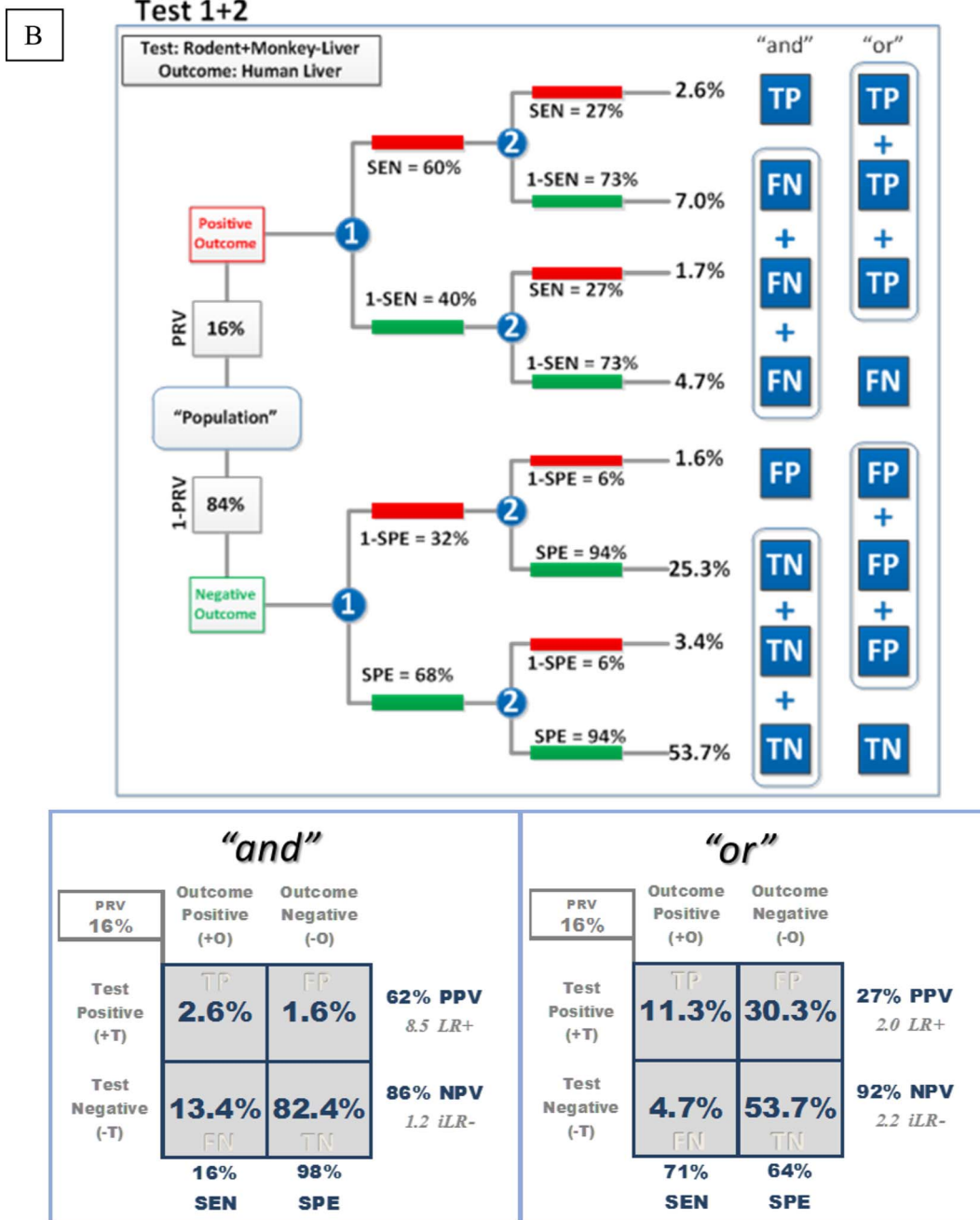


Fig. 5. (continued)

system and gastrointestinal organ categories. All organ categories across all test species, in general, had a very high NPV. The clinical positive prevalence was highest for the gastrointestinal tract (GI) and the 'other' category.

### 3.4. Concordance parameters excluding the oncology indication

Cancer patients may have pre-existing organ toxicities attributed to their concurrent disease that would make it difficult to correctly assign a clinical finding as being treatment-related or not; this could impact the accuracy of ascribing since treatment-relatedness for a clinical finding could be uncertain. An additional concern is that oncology agents are commonly administered to maximum tolerated doses in the clinic and more likely to have a clinical positive outcome (e.g. higher outcome prevalence) that could skew the performance measures based on the higher pretest probability that they would be outcome positive.

As presented in Fig. 2, we found similarity of the sensitivity and specificity for the oncology-only database and the entire database for all other therapeutic indications, demonstrating that the nonclinical safety models perform equally, regardless of the health status of phase I subjects. However, despite these similar test performance characteristics, the impact of the higher clinical outcome prevalence in the case of oncology agents, reflected in the higher PPV and lower NPV, was observed with this cohort. As a result, we excluded oncology agents from subsequent predictive performance analyses to avoid any prevalence-based bias.

### 3.5. Concordance parameters by test species

The NHP, overall, had the highest positive concordance parameters among the test species evaluated (Fig. 3.). Concordance for the pig as a single species was not determined due to the limited dataset. In general,

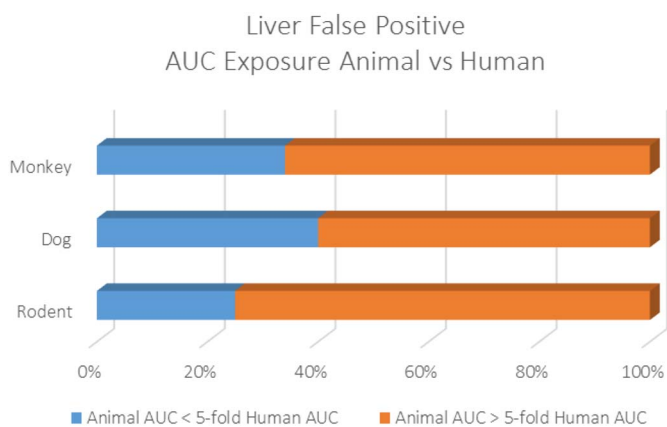


Fig. 6. Animal false positives for the liver category occurred at much higher exposures in the animal than the clinic.

the NPV was equivalent between test species.

### 3.6. Proportionate reduction in uncertainty (PRU)

As an additional approach to define the potential value of animal testing and human safety, we determined the PRU for the nervous system and GI, as organ category examples. As the PRU is calculated from prevalence and either PPV or NPV, it is intended to provide a more intuitive method of quantifying the diagnostic value of a test. Prior to the FIH-enabling toxicology studies, our prediction for the probability of an adverse effect in phase I is simply the clinical positive prevalence. The NHP had the strongest performance in the PRU (Fig. 4) for both the GI and nervous system categories. Even with a low positive clinical prevalence for the nervous system category (18%), nonclinical testing resulted in a strong predictive performance. For these organ category examples, it can be appreciated that a large degree of uncertainty was resolved by the animal testing.

### 3.7. Predictive performance of animal testing using the dual-species method

The predictive performance of animal toxicology testing using the dual-species method demonstrates that a positive finding in both species optimized prediction of a positive human outcome. In our illustration, a “probability tree” provides a way to visualize the path of a new drug candidate through an animal testing paradigm, and to understand the impact of the underlying prevalence of the clinical outcome of interest, as well as the sensitivity and specificity, on the PPV and NPV (Fig. 5). As an example, we used this approach for the liver organ category based on the rodent and monkey liver concordance results. The probability trees for each individual test (e.g. rodent and monkey) are “stacked” and two additional contingency tables are generated depending on how a “positive” combined-test result is defined (Fig. 5). There are two ways to reduce the eight possible combined outcomes to the four cells of the combined contingency table. Both views are useful when considering the combined test results. The first view combines the individual test results using the conjunction “and” meaning that a combined test positive outcome (TP or FP) requires a positive test result in both individual species. Combining results using “and” restricts the definition of an overall test positive outcome and has

Table 4  
False negative impact on test performance: concordance of database excluding the ‘other’ category.

	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Clinical positive prevalence (%)	LR +
Database excluding the ‘other’ category	52	86	36	92	14	3.6

PPV, positive predictive value, NPV, negative predictive value, LR +, likelihood ratio positive.

the effect of reducing sensitivity and increasing specificity of the combined test. This combination view maximizes the PPV and minimizes the NPV relative to the individual species tests.

In contrast, when combining the results of the individual species tests using “or” a combined positive outcome results from a positive test result in either species individually or both species together. This view serves to expand the definition of a combined positive outcome and effectively increases the sensitivity and decreases the specificity of the combination test. In turn, this evaluation method has the effect of minimizing the PPV but maximizing the NPV of the combined test. The power of combined testing is presented by the contingency tables for the prediction of human liver outcomes based on rodent and monkey testing (Fig. 5B). Viewed as individual tests, the strongest positive outcome prediction (prediction of a human liver effect based on the observation of a nonclinical effect) is provided by the NHP at 46% while the strongest negative prediction (prediction of no human liver effect based on no nonclinical liver effect) is provided by the rodent at 90% (Fig. 5A). When the results are viewed as a combined test the “and” view (a liver effect in both nonclinical species) increases the PPV to 62% with a LR + value of 8.5. Even with the impressive reduction in pre-test uncertainty provided by the combined testing approach, the maximum PPV of 62% demonstrates the challenge for any testing paradigm in overcoming a low pre-test probability (in this case a clinical positive liver outcome prevalence of 16%). The “or” view maximizes the prediction of safety (NPV) at 92% when no liver effect is observed in either nonclinical species.

### 3.8. Liver false positives and clinical exposure margins

Nonclinical safety testing is biased towards type 1 errors since ICH guidance recommends a high dose in the animal study that will either achieve a maximum tolerated dose, an exposure saturation, a maximum feasible dose (e.g. solubility limitations) or a mean exposure margin of 50 × to the clinical exposure (ICH M3(R2), 2009). A false positive is defined as a nonclinical liver signal of concern (animal test positive), reported in the IB that did not translate to an adverse human liver effect (clinical outcome negative). In our analysis, the majority of false positive liver signals, across species, occurred at much higher exposures (> 5-fold) than those achieved in the clinic (Fig. 6).

### 3.9. False negative impact on test performance

The ‘other’ organ system category, which generally consisted of subjective, constitutional clinical adverse events (e.g. headache, dizziness, fatigue, insomnia, memory impairment, anxiety) had a very low PPV and NPV (Table 3). Animal models would not be expected to predict such clinical signs. We evaluated the impact on predictive parameters when the ‘other’ category was omitted (Table 4). By omitting these FN entries (type 2 errors), there is an increase in sensitivity, an increase in the LR +, and a greater NPV, as compared to these concordance parameters for the entire dataset (Fig. 2).

## 4. Discussion

The IQ Consortium nonclinical to clinical translational database is unique in that it correlates potential safety liability risks, based on animal data outlined in the regulatory dossier, with the results of phase I clinical observations. A true positive finding was recorded when

effects were reported in both the nonclinical species and human involving the same target organ category, which is consistent with the approach taken by Olson et al. (2000). Because the starting point for the Olson analysis was compounds with adverse human outcomes (prevalence = 100%) their results were limited to an estimate of the sensitivity. Notably absent was information regarding human outcome negative compounds. As a result, there was no way to estimate the prevalence of compounds associated with significant human toxicity. Our database provides information regarding the performance of the three main nonclinical test systems (i.e. rodent, dog and monkey) in anticipating the human outcomes in phase I and provides information regarding the prevalence at which adverse human outcomes were encountered across a number of target organs.

As required by regulatory guidance (ICH M3(R2), 2009), nonclinical toxicology studies are often conducted at high (in some cases excessive) doses in order to identify target organ toxicity. As a result, there is a high rate of false positive signals (nonclinical test positive-human outcome negative). If criteria were applied to either the animal study design (avoid high exposures) or in the interpretation of the animal results (positive findings being only those that occurred within the range of clinical exposure), there would be fewer compounds classified as test-positive. As a consequence of this restricted definition, there would be an overall shift of compounds from the top to the bottom row of the contingency table. This would have the effect of reducing the sensitivity and increasing the specificity of nonclinical safety testing. While this would have the effect of reducing the rate of false positive errors and increasing the PPV, it would come at the cost of increasing the false negative error rate (nonclinical test negative but human outcome positive). The downside of this approach becomes obvious. Given that the goal of animal testing is to provide an assurance of safety prior to phase I, missing or ignoring nonclinical findings would not be prudent.

Even under the current testing paradigm, false negative errors cannot be avoided entirely. In our database, the 'other' organ system, which generally consisted of subjective constitutional adverse events, had a very low PPV and NPV. This is to be expected since these types of clinical events are not reliably demonstrated or detected in animal models (Tamaki et al., 2013). Constitutional adverse events are common in phase I trials and moreover, can often be observed in the placebo group (Emanuel et al., 2015).

The results of our analysis demonstrate that the current testing paradigm has effectively supported safety in Phase I human clinical trials. Animal to human toxicity concordance reporting has been focused on sensitivity (true positives and false negatives). Based on our database, we found that animal studies also provide great value in the PPV for certain species and selected organ categories, but more importantly, our analysis demonstrates that a robust high NPV is quite consistent across test species and target organs. The value of the standard two-species testing paradigm goes beyond simply expanding the breadth of target organs identified. When a similar target organ finding is observed in both test species (rodent and nonrodent), the overall predictive power of nonclinical safety testing is increased. A similar target organ effect in both species increases PPV, while no target organ finding in either species increases NPV.

The NPV is the stronger predictive performance measure that was derived from our database, signifying that a lack of toxicity in animal studies strongly predicts safety in phase I. This robust NPV is largely based on the low clinical positive prevalence observed in our database and in the literature, which can be attributed to the fact that compounds entering clinical development have typically cleared many safety hurdles via extensive *in silico*, *in vitro*, and *in vivo* lead optimization screening activities (Cook et al., 2014; Roberts et al., 2014; Butler et al., 2017).

Finally, the approach the IQ Consortium has taken to create this database demonstrates a viable methodology that fosters pre-competitive, industry-wide sharing of nonclinical and clinical data.

Precompetitive data sharing should be the new standard for the pharmaceutical industry in order to further evaluate the utility of our conventional nonclinical models, support the 3Rs (replacement, reduction, and refinement of animals in research) and help identify predictivity gaps that may be better addressed by alternative models.

### Conflict of interest

All authors declare no conflict of interest.

### Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

### Acknowledgements

The database was developed under the auspices of the International Consortium for Innovation and Quality in Pharmaceutical Development (IQ), a not-for-profit organization of pharmaceutical and biotechnology companies with a mission of advancing science and technology to augment the capability of member companies to develop transformational solutions that benefit patients, regulators and the broader research and development community. DruSafe is a Leadership Group of the IQ Consortium with the mission to advance nonclinical safety sciences and impact the global regulatory environment. We would like to thank the following biopharmaceutical companies and their scientists that contributed to the IQ Consortium database: Abbvie, Allergan, Amgen, AstraZeneca, Boehringer Ingelheim, Bristol-Myers Squibb, Daiichi-Sankyo, Eisai, Genentech, GlaxoSmithKline, Incyte, Infinity, Janssen, Lilly, Pfizer, Sanofi, Takeda, and Vertex. Also thanks to Qichao Zhu for his work on establishing the cloud-based database and Keith Krisko for managing the database at the IQ Consortium.

### References

- Altman, D.G., Bland, J.M., 1994. Statistics notes: diagnostic tests 1: sensitivity and specificity. *BMJ* 308, 1552.
- Atienzar, F.A., Blomme, E.A., Chen, M., Hewitt, P., Kenna, J.G., Labbe, G., Moulin, F., Pognan, F., Roth, A.B., Suter-Dick, L., Ukairo, O., Weaver, R.J., Will, Y., Dambach, D.M., 2016. Key challenges and opportunities associated with the use of *in vitro* models to detect human DILI: integrated risk assessment and mitigation plans. *BioMed. Res. Intern.* <http://dx.doi.org/10.1155/2016/9737920>.
- Bailey, J., Thew, M., Balls, M., 2013. An analysis of the use of dogs in predicting human toxicology and drug safety. *ALTA* 41, 335–350.
- Bailey, J., Thew, M., Balls, M., 2015. Predicting human drug toxicity and safety via animal tests: can any one species predict drug toxicity in any other, and do monkeys help? *ALTA* 43, 393–403.
- Butler, L.D., Guzzie-Peck, P., Hartke, J., Bogdanffy, M.S., Will, Y., Diaz, D., Mortimer-Cassen, E., Derzi, M.A., Greene, N., DeGeorge, J.J., 2017. Current nonclinical testing paradigms in support of safe clinical trials: an IQ consortium DruSafe perspective. *Regul. Toxicol. Pharmacol.* 87, S1–S15.
- Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., Pangalos, M.N., 2014. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* 13, 419–431.
- Coulthard, M.G., 2006. Quantifying how tests reduce diagnostic uncertainty. *Arch. Dis. Child.* 92, 404–408.
- EMA, 2012. Guideline on Good Pharmacovigilance Practices. Heads of Medicines Agencies (EMA/876333/2011 Rev.1. December 12, 2012).
- Emanuel, E.J., Bedarida, G., Macci, K., Gabler, N.B., Rid, A., Wendler, D., 2015. Quantifying the risks of non-oncology phase I research in healthy volunteers: meta-analysis of phase I studies. *BMJ* 350, 3–11.
- FDA, 2011. Strategic Plan. Advancing Regulatory Science at FDA. August, 2011. [www.fda.gov/regulatoryscience](http://www.fda.gov/regulatoryscience).
- Granger, C.B., Vogel, V., Cummings, S.R., Held, P., Fiedorek, F., Lawrence, M., Neal, B., Reides, H., Santarelli, L., Schroyer, R., Stockbridge, N.L., Zhao, F., 2008. Do we need to adjudicate major clinical events? *Clin. Trials* 5, 56–60.
- Greek, R., Menache, A., 2013. Systematic reviews of animal models: methodology versus epistemology. *Int. J. Med. Sci.* 10, 206–221. Available from: <http://www.medsci.org/v10p0206.htm>.
- Grimes, D.A., Schulz, K.F., 2005. Refining clinical diagnosis with likelihood ratios. *Lancet* 365, 1500–1505.
- Hartung, T., 2013. Food for thought look back in anger – what clinical studies tell us about preclinical work. *ALTEX* 30, 275–291.
- ICH M3(R2), 2009. Guidance on non-clinical safety studies for the conduct of human

- clinical trials and marketing authorization for biopharmaceuticals. Retrieved Jan 2017 from. <http://www.ich.org> (June).
- ICH S6, 1997. Preclinical safety evaluation of biotechnology-derived biopharmaceuticals. Retrieved Jan 2017 from. <http://www.ich.org> (July).
- ICH S6 (R1), 2011. Addendum to ICH S6 guideline: preclinical safety evaluation of biotechnology-derived biopharmaceuticals. Retrieved Jan 2017 from. <http://www.ich.org> (May).
- ICH S9, 2009. Nonclinical Evaluation for Anticancer Biopharmaceuticals. Retrieved Jan 2017 from. <http://www.ich.org> (October).
- Mangipudy, R., Burkhart, J., Kadambi, V., 2014. Use of animals for toxicology testing is necessary to ensure patient safety in pharmaceutical development. *Regul. Toxicol. Pharmacol.* 70, 439–441.
- Mullard, A., 2017. 2016 FDA drug approvals. *Nat. Rev. Drug Dis* Published online 2 Feb. <http://dx.doi.org/10.1038/nrd.2017.14>.
- 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 pharmaceutical in humans and animals. *Regul. Toxicol. Pharmacol.* 32, 56–67.
- Roberts, R.A., Kavanagh, S.L., Mellor, H.R., Pollard, C.E., Robinson, S., Platz, S.J., 2014. Reducing attrition in drug development: smart loading preclinical safety assessment. *Drug Discov. Today* 19, 341–347.
- Shanks, N., Greek, R., Greek, J., 2009. Are animal models predictive for humans? *Phil. Ethics Humanit. Med.* 4. [www.peh-med.com/content/4/1/2](http://www.peh-med.com/content/4/1/2).
- Sivendran, S., Latif, A., McBride, R.B., Stensland, K.D., Wisnivesky, J., Haines, L., Oh, W.K., Galsky, M.D., 2013. Adverse event reporting in cancer clinical trial publications. *J. Clin. Oncol.* 32, 83–89.
- Tamaki, C., Nagayama, T., Hashiba, M., Fujiyoshi, M., Hizue, M., Kodaira, H., Nishida, M., Suzuki, K., Takashima, Y., Ogino, Y., Yasugi, D., Yoneta, Y., Hisada, S., Ohkura, T., Nakamura, K., 2013. Potentials and limitations of nonclinical safety assessment for predicting clinical adverse drug reactions: correlation analysis of 142 approved drugs in japan. *J. Toxicol. Sci.* 38, 581–598.
- Ting, K.M., 2010. In: Sammut, C., Webb, G. (Eds.), *Encyclopedia of Machine Learning*. Springer, US, pp. 209.
- van Meer, P.J.K., Kooijman, M., Gispens-de Wied, C.C., Moors, E.H.M., Schellekens, H., 2012. The ability of animals studies to detect serious post marketing adverse events is limited. *Regul. Toxicol. Pharmacol.* 64, 345–349.