The Concordance between Nonclinical and Phase I Clinical Cardiovascular Assessment from a Cross-Company Data Sharing Initiative

Lorna Ewart⁎,1, Mike Aylott†,2, Mark Deurinck‡, Mike Engwall§, David J. Gallacher¶, Helena Geys††, Philip Jarvis⁎,2, Haisong Jujj, Derek Leishmanjjj, Louise Leongjjjj, Nick McMahon†, Andy Mead#, Phil Milliken†, Willi Suter‡, Ard Teisman¶, Karel Van Ammel¶, Hugo M. Vargas§, Rob Wallis**, and Jean-Pierre Valentin⁎,2

⁎AstraZeneca R&D Mölndal, Pepparedsleden 1, 431 83, Mölndal, Sweden, †GlaxoSmithKline, Park Road, Ware, Hertfordshire, SG12 ODP, UK, ‡Novartis Pharma AG, PO Box, CH-4002, Basel, Switzerland, §Amgen, Inc, One Amgen Center Drive, Thousand Oaks, CA 91320, ¶Janssen Research & Development, a division of Janssen Pharmaceutica NV, Turnhoutseweg 30, B-2340 Beerse, Belgium, ††Novartis Institutes for BioMedical Research, One Health Plaza, East Hanover, NJ 07936, jjjEli Lilly and company, Indianapolis, IN 46285, jjjjAssociation of the British Pharmaceutical Industry, 105 Victoria Street, London SW1E 6QT, UK, #Pfizer Inc., Eastern Point Road, Groton, CT 06340 and **Safety Pharmacology consultant, Sandwich, Kent, UK

1To whom correspondence should be addressed. E-mail: lorna.ewart@astrazeneca.com.
2Present addresses: Consultant, St Albans, Hertfordshire, UK; Novartis Pharma AG, PO Box, CH-4002, Basel, Switzerland; UCB Pharma SA, Chemin du Foriest – B-1420 Braine-l’Alleud, Belgium.

ABSTRACT

It is widely accepted that more needs to be done to bring new, safe, and efficacious drugs to the market. Cardiovascular toxicity detected both in early drug discovery as well as in the clinic, is a major contributor to the high failure rate of new molecules. The growth of translational safety offers a promising approach to improve the probability of success for new molecules. Here we describe a cross-company initiative to determine the concordance between the conscious telemetered dog and phase I outcome for 3 cardiovascular parameters. The data indicate that, in the context of the methods applied in this analysis, the ability to detect compounds that affect the corrected QT interval (QTc) was good within the 10–30x exposure range but the predictive or detective value for heart rate and diastolic blood pressure was poor. These findings may highlight opportunities to refine both the animal and the clinical study designs, as well as refocusing the assessment of value of dog cardiovascular assessments beyond phase 1. This investigation has also highlighted key considerations for cross-company data sharing and presents a unique learning opportunity to improve future translational projects.

Key words: cardiovascular system; concordance; phase I; safety pharmacology; statistical analysis; translational safety

Drug discovery and development is essential to deliver innovative drugs to patients. Developing new drugs is becoming increasingly difficult, regulatory hurdles are becoming increasingly higher and aversion to adverse effects, from patients, regulators, and payers, is increasing. It has been calculated that only 32% of compounds entering phase II make it into...
phase III (Hay et al., 2014) and just 12% of potential drugs ultimately make it onto the market (Munos, 2009). Nonclinical and clinical safety-related terminations together contribute substantially to the overall attrition within drug development (Cook et al., 2014; Kola and Landis, 2004). The cardiovascular system is frequently cited as a principal cause of this attrition (Guengerich, 2011; Redfern et al., 2010), despite strict regulatory guidelines (ICH S7A, ICH S7B, and ICH M3). Translational safety science is a growing practice that aims to further understand the concordance between nonclinical and clinical models (Sung et al., 2003) and it offers the potential to improve drug development productivity by guiding risk assessment and decision making. Failure to accurately predict safety can be potentially harmful to humans, or could result in the discontinuation of potential new drugs.

Although the identification of novel drug targets takes advantage of a wide range of data sets (eg, genomics, molecular fingerprints etc.), the nonclinical assessment of both efficacy and safety for new drugs remains highly dependent on the use of animals. In spite of their pivotal role in this process, there is no accepted method to assess the ability of animal models to predict human outcome and/or to detect potential safety hazards, despite several previous consortia investigations (Igarashi et al., 1995; Olson et al., 2000; van Meer et al., 2012) or concerns from regulators (Shah, 2008). Igarashi et al. (1995) analyzed data from 104 new drugs that were approved in Japan and established a positive association between animal data and human adverse reactions but the associations were based solely on published data points for analysis, this article also discusses the advantage of a wide range of data sets (eg, genomics, molecular fingerprints etc.).

In the year 2001 or later (ie, after the introduction of the ICHS7 guidelines). The reasons for choosing this animal model was 3-fold: (1) cardiovascular safety is often cited as a leading contributor to safety-related attrition and an improved method for detection and prediction of such liability will improve human risk assessment; (2) a relatively large data set should be available because, in accordance with ICH S7A guidelines, all small molecules are required to be tested in a non-rodent species, usually dogs, to assess their potential effects on the cardiovascular system prior to phase I clinical study; and (3) the design of such studies (ie, as defined by the ICHS7 guidelines) was considered to be relatively standard across companies, enabling the pooling of data and the generation of a larger data set. Each participating company also shared study designs of the conscious dog telemetry model and the phase I study. All clinical and nonclinical studies were performed in compliance with good clinical practice (GCP) or good laboratory practice (GLP), respectively, and with the respective international and national regulatory guidelines and directives. The dog telemetry studies were generally performed using ascending, parallel or cross-over designs and in compliance with the recommendations of the BVAAFE/FRAME/RSPCA/UFAW Joint Working Group on Refinement (Hawkins et al., 2004; Morton et al., 2003) and local animal welfare regulations.

Exposure data (ie, Cmax in nonclinical and clinical studies) was also recorded. In the dog telemetry study, these data were collected for each dose level tested (typically 3). In some cases, exposure was not assessed in the telemetry study therefore free plasma exposure data were extrapolated from other dog studies which used the same or similar dose levels and routes of administration. For the clinical exposure, where there was no change on any of the parameters of interest, the exposure at the highest dose level only was recorded. Where there was an effect on a parameter (ie, increase or decrease), 2 exposures were recorded (1) the exposure at the lowest dose level causing the challenges and limitations of such approaches in order to facilitate future projects in this area.

**MATERIALS AND METHODS**

Data collection. The effect (eg increase, decrease, or no change) of a small molecule on hemodynamic parameters (diastolic blood pressure [DBP] and heart rate [HR]) and ECG intervals and complexes (HR corrected QT interval [QTc], QRS, and PR) as documented in the study director conclusion (conscious dog telemetry) or as judged by the clinical report author to be test article related and noted in their conclusion, was shared for 113 small molecule compounds. Systolic blood pressure data were not collected in an attempt to simplify and thus ensure consistency in sharing data across companies. Systolic blood pressure has both a cardiac and vascular input, whereas DBP represents primarily a vascular input. The authors believe that by capturing the HR data the cardiac component was accounted for within the analysis. This qualitative approach (eg, the analysis and conclusion of the expert performing the study) was taken in preference to a quantitative approach because it was considered that this also provided the most consistent assessment between the animal model and the clinical study and furthermore accounted for differences in the way that the studies were performed across each of the companies.

Each small molecule included in the analysis had been tested in the conscious telemetered dog and had also completed single ascending dose phase I studies. The authors agreed that in order for the dog telemetry data to be accepted into the analysis, it should have been conducted in the year 2001 or later (ie, after the introduction of the ICHS7 guidelines). The reasons for choosing this animal model was 3-fold: (1) cardiovascular safety is often cited as a leading contributor to safety-related attrition and an improved method for detection and prediction of such liability will improve human risk assessment; (2) a relatively large data set should be available because, in accordance with ICH S7A guidelines, all small molecules are required to be tested in a non-rodent species, usually dogs, to assess their potential effects on the cardiovascular system prior to phase I clinical study; and (3) the design of such studies (ie, as defined by the ICHS7 guidelines) was considered to be relatively standard across companies, enabling the pooling of data and the generation of a larger data set. Each participating company also shared study designs of the conscious dog telemetry model and the phase I study. All clinical and nonclinical studies were performed in compliance with good clinical practice (GCP) or good laboratory practice (GLP), respectively, and with the respective international and national regulatory guidelines and directives. The dog telemetry studies were generally performed using ascending, parallel or cross-over designs and in compliance with the recommendations of the BVAAFE/FRAME/RSPCA/UFAW Joint Working Group on Refinement (Hawkins et al., 2004; Morton et al., 2003) and local animal welfare regulations.

Exposure data (ie, Cmax in nonclinical and clinical studies) was also recorded. In the dog telemetry study, these data were collected for each dose level tested (typically 3). In some cases, exposure was not assessed in the telemetry study therefore free plasma exposure data were extrapolated from other dog studies which used the same or similar dose levels and routes of administration. For the clinical exposure, where there was no change on any of the parameters of interest, the exposure at the highest dose level only was recorded. Where there was an effect on a parameter (ie, increase or decrease), 2 exposures were recorded (1) the exposure at the lowest dose level causing the...
effect and (2) the exposure at the preceding dose level (ie the highest dose level that had no effect). On occasion, these exposures were different for each parameter (HR, DBP, QTc) because the compound differentially affected the individual parameters. Exposure data represent the maximum free drug concentration in plasma (free Cmax) in the dog and the human.

Within this investigation, there was no reinterpretation of the data collected because the authors were respectful of the expert’s conclusions. Data were contributed by each company on an ease of extraction basis, hence it is assumed that there was no intended selection bias with respect to therapeutic area. Study outcomes were not considered when identifying which studies to include. The data were assumed to represent multiple therapeutic areas and to span a reasonable chemical diversity. Consequently, the data set will contain compounds that have different definitions of a safety risk (eg oncology compounds will be different to those intended for cardiovascular metabolic disease). It was recognized that the data were naturally biased toward drugs that make it into phase I testing, but the application of the outcome of the analysis is intended to be applied only to similar data sets, and therefore this bias is considered a strength of the current approach.

Data classification. In this investigation, binary classification was used to classify the effects of the small molecules. For each exposure multiple, the effects of the small molecule were classified as a true positive (TP) (the compound caused the same directional effect of an increase or decrease in the dog telemetry model as it did in phase I), a false positive (FP) (the compound caused directional effect of an increase or decrease in a parameter in the dog telemetry study but did not cause a change in this parameter in the phase I study), a false negative (FN) (the compound did not change the parameter of interest in the dog telemetry study but caused an increase or a decrease in the phase I study), and a true negative (TN) (the compound did not cause any change in the parameter in both the dog telemetry study and the phase I study). The direction of change in the animal model was always the same in the clinical study but for two compounds, the direction of effect in the dog changed at the highest exposure level. This had no impact on the analysis because the compounds were FP in both cases. The number of TP, FP, TN, and FN compounds in the data set analyzed is presented in Table 1.

Statistical analysis. Standard statistical approaches (sensitivity and specificity) for diagnostic tests were then applied. Sensitivity represents the proportion of small molecules that affected the parameter of interest which were correctly identified by the dog telemetry model (Altman and Bland, 1994a; Loong, 2003). Sensitivity was calculated by dividing the number of TP by the sum of the TP and FN (Figure 1). Specificity, for a model, was calculated by dividing the number of TN by the sum of the TN and FP (Figure 1). To obtain an estimate of the variation in the data set, 95% confidence limits were calculated. Fisher’s exact test was applied where p < 0.05 was considered to be statistically significant.

Positive predictive value (PPV) and negative predictive value (NPV) provide information on the likelihood of the model correctly predicting the outcome, and may have greater utility when applying the learning’s from this analysis to novel data sets where only the nonclinical data are available. That is, given the data from the dog telemetry model, the predictive values allow the researcher to estimate the probability of a finding in the phase 1 study. It should be noted that predictive values depend critically on the prevalence of positives in the clinical outcome, and therefore this variable has been factored into our analysis (Altman and Bland, 1994b; Loong, 2003). The PPV of a model is the number of TP divided by the sum of the TP and FP, whereas the NPV is the number of TN divided by the sum of the TN and FN (Figure 1). Actual prevalence data to calculate negative and PPVs are very rarely available. However, prevalence data at a macro level for attrition rates at key stages of

### Table 1. This Table Shows a Summary of the Number of False Positives (FP), True Positives (TP), True Negatives (TN), and False Negatives (FN) Compounds for Each Parameter at Each Exposure Multiple Tested

<table>
<thead>
<tr>
<th>Parameter/ exposure multiple</th>
<th>QTc up to 1X</th>
<th>HR up to 1X</th>
<th>DBP up to 1X</th>
<th>QTc up to 3X</th>
<th>HR up to 3X</th>
<th>DBP up to 3X</th>
<th>QTc up to 10X</th>
<th>HR up to 10X</th>
<th>DBP up to 10X</th>
<th>QTc up to 30X</th>
<th>HR up to 30X</th>
<th>DBP up to 30X</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP</td>
<td>2</td>
<td>12</td>
<td>5</td>
<td>5</td>
<td>23</td>
<td>14</td>
<td>12</td>
<td>34</td>
<td>22</td>
<td>16</td>
<td>40</td>
<td>29</td>
</tr>
<tr>
<td>TP</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>TN</td>
<td>82</td>
<td>67</td>
<td>79</td>
<td>58</td>
<td>46</td>
<td>56</td>
<td>38</td>
<td>32</td>
<td>38</td>
<td>22</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>FN</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>95</td>
<td>88</td>
<td>88</td>
<td>69</td>
<td>76</td>
<td>75</td>
<td>58</td>
<td>73</td>
<td>65</td>
<td>48</td>
<td>64</td>
<td>55</td>
</tr>
</tbody>
</table>

FIG. 1. An illustrative example for the calculation of sensitivity, specificity, NPV, and PPV.
pharmaceutical development were collected by the KMR Group (Chicago) for the R&D performance 2009 Pharmaceutical Benchmarking Forum (PBF). The PBF integrates R&D productivity data from a group of large pharmaceutical firms including Abbott, AstraZeneca, Bayer Healthcare, Bristol-Myers Squibb, Boehringer Ingelheim, Eli Lilly, GlaxoSmithKline, Johnson & Johnson, Merck Research Labs, Novartis, Pfizer, Roche, and Sanofi-Aventis. PBF companies provide blinded pipeline and productivity information through a third-party administrator (KMR Group, Chicago, Illinois). For the period relevant to our study, 2004–2008, the overall proportion of compounds moving between nonclinical and phase I clinical study was approximately two-thirds. On the assumption that up to one-half of the one-third of compounds (ie, one-sixth) may be stopped for cardiovascular-related reasons, it seemed reasonable to use the observed clinical prevalence in this data set (0.10) as a lower bound for the prevalence and the clinical prevalence plus one-sixth as the upper bound (0.27). To obtain an estimate of the variation in the data set, 95% confidence limits were calculated. All statistical calculations were performed using PROC FREQ of SASv9.2.

Exposure analysis. The major challenge in any comparison of nonclinical and clinical data is to ensure that the comparisons being made are pharmacologically relevant. In this article, good agreement between a clinical finding and a nonclinical finding was claimed if the same effect was observed nonclinically within a given exposure window of the clinical finding. The values used to define these windows of agreement were 1-fold, 3-fold, 10-fold, and 30-fold (Figure 2). For example, if nonclinical exposure data were obtained within 3-fold of the exposure associated with the clinical effect then that data would be included in the assessment of nonclinical versus clinical agreement for the 3-fold analysis window. Sensitivity, specificity, NPV, and PPV have been calculated using these windows of agreement thus enabling a comparison between effect and exposure.

RESULTS

Characteristics of the data set used. The total data set used in this analysis was compromised of 113 small molecule compounds (Figure 3). The majority of these small molecules (63%) were provided by two companies; however, the remainder was a representative set of the other 5 contributing companies. Phase I studies were mainly conducted in healthy human volunteers with patients being used in phase I studies for 7 compounds.

Thirty-eight of these small molecules were TN compounds (ie, there was no change in the parameters of interest in either the dog or the phase I clinical study), 89 had no cardiovascular effect in human at the dose levels tested and 42 had no effects in the dog. In the dog, HR change was the most common finding (49 compounds increased HR and 7 decreased HR, Table 2), followed by changes in DBP (21 compounds increased DBP and 16 compounds decreased DBP, Table 2) and then corrected QTc (23 compounds increased QTc, whereas 8 compounds decreased QTc, Table 2). Eighteen compounds caused a change...
in PR interval (10 compounds decreased the interval duration, whereas 8 compounds increased it, Table 2) and 9 compounds caused a change in QRS duration (6 compounds produced an increase and 3 compounds resulted in a decrease, Table 2). Changes in QTc were the most common phase I finding with 16 compounds altering QTc interval. Fourteen compounds were recorded by the expert running the study as having a clinically relevant prolongation of QTc, whereas a further 2 had a clinically relevant QTc shortening (Table 2). The next most common finding was changes in HR with 11 compounds causing an increase in HR and one causing a decrease in HR (Table 2). Five of the compounds caused an increase in DBP and a further 4 caused a decrease in this parameter (Table 2). Only 3 of the compounds caused an increase in PR interval, whereas QRS interval was increased by only 2 of the compounds (Table 2). In this data set, there were no compounds reported to decrease either PR or QRS interval (data not shown). The 2 compounds with an increase in QRS also had an increase in PR and QTc. Nonclinically 2 of these 3 compounds affected both the hemodynamic and ECG parameters and the other had no dose-dependent effects on any of the nonclinical parameters. Because of the small number of compounds affecting the QRS and PR interval, the data were not analyzed any further.

The data set were also described with respect to the range of exposures explored for each compound. To understand how this is related to the maximum exposure tested in phase I, individual compounds were plotted against exposure (Figure 4). The exposure in human exceeded the exposure range tested in dogs for 20 compounds. Three compounds had a minimum dog exposure value greater than 100-fold of the maximum human exposure obtained, 6 had a minimum dog exposure greater than 30-fold, but less than 100-fold of the maximum human exposure and 11 compounds had a minimum dog exposure greater than 10-fold, but less than 30-fold of the maximum human exposure. Because the minimum human exposure was not captured it was not possible to determine if any of the compounds were exposed in the dog at concentrations far below those tested in the clinic. This heterogeneity within the exposure ranges supported the need for the creation of windows of exposure to maximize the use of the data available for analysis.

Sensitivity and specificity determination. With respect to QTc change, the specificity (Table 3) of the dog telemetry model was generally high, ranging from 0.98 (at exposure multiples of 1x) to 0.58 (at exposure multiples of 30x). In contrast, the sensitivity (Table 3) of the dog telemetry model was generally lower, ranging from 0.00 (at exposure multiples of 1x) to 1.00 (at exposure multiples of 30x). This is not surprising because the number of nonclinical positives (TP and FP) increased (Table 3) as the exposure multiples are increased. As is often observed with diagnostic tests, the specificity begins to decline as the sensitivity increases (Figure 5A). The specificity for HR change (Figure 5B) was not as high as that for QTc but also showed a drop off as exposure multiples increased (0.85 at exposures of 1x compared with 0.29 at exposures of 30X, Table 3). Similarly, sensitivity was lower than specificity but again in line with the increase in TP and FP rate, increased with increasing exposure multiples (0.22 at exposures of 1x compared with 0.75 at exposures of 30X.

![Illustration of exposure ranges tested nonclinically and clinically by the compounds used in this analysis; black triangles represent maximum human exposure, the gray circles represent the lowest exposure in the dog telemetry study and the gray squares represent the highest exposure in the dog telemetry study.](https://academic.oup.com/toxsci/article-abstract/142/2/427/3064909)
Table 4. This Table Shows a Summary of Negative Predictive Value (NPV) and Positive Predictive Value (PPV), with 95% Confidence Limits for Each Parameter at Each Exposure Multiple Tested

<table>
<thead>
<tr>
<th>Parameter/ exposure multiple</th>
<th>QTc up to 1X</th>
<th>HR up to 1X</th>
<th>DBP up to 1X</th>
<th>QTc up to 3X</th>
<th>HR up to 3X</th>
<th>DBP up to 3X</th>
<th>QTc up to 10X</th>
<th>HR up to 10X</th>
<th>DBP up to 10X</th>
<th>QTc up to 30X</th>
<th>HR up to 30X</th>
<th>DBP up to 30X</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPV (un)</td>
<td>0.88</td>
<td>0.91</td>
<td>0.95</td>
<td>0.92</td>
<td>0.90</td>
<td>0.93</td>
<td>0.97</td>
<td>0.91</td>
<td>0.93</td>
<td>1.00</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>CI</td>
<td>0.80–0.94</td>
<td>0.81–0.96</td>
<td>0.88–0.99</td>
<td>0.82–0.97</td>
<td>0.79–0.97</td>
<td>0.84–0.98</td>
<td>0.87–1.00</td>
<td>0.77–0.98</td>
<td>0.80–0.98</td>
<td>0.85–1.00</td>
<td>0.65–0.99</td>
<td>0.68–0.97</td>
</tr>
<tr>
<td>NPV (0.1)</td>
<td>0.90</td>
<td>0.91</td>
<td>0.89</td>
<td>0.81</td>
<td>0.89</td>
<td>0.90</td>
<td>0.98</td>
<td>0.91</td>
<td>0.90</td>
<td>1.00</td>
<td>0.91</td>
<td>0.86</td>
</tr>
<tr>
<td>CI</td>
<td>0.88–0.94</td>
<td>0.86–0.96</td>
<td>0.88–0.95</td>
<td>0.87–0.97</td>
<td>0.80–0.98</td>
<td>0.83–0.98</td>
<td>0.91–1.00</td>
<td>0.77–1.00</td>
<td>0.80–0.99</td>
<td>0.61–1.00</td>
<td>0.61–0.99</td>
<td>0.60–0.99</td>
</tr>
<tr>
<td>NPV (0.27)</td>
<td>0.73</td>
<td>0.75</td>
<td>0.72</td>
<td>0.75</td>
<td>0.72</td>
<td>0.73</td>
<td>0.94</td>
<td>0.75</td>
<td>0.74</td>
<td>1.00</td>
<td>0.76</td>
<td>0.65</td>
</tr>
<tr>
<td>CI</td>
<td>0.68–0.81</td>
<td>0.66–0.89</td>
<td>0.70–0.86</td>
<td>0.66–0.91</td>
<td>0.54–0.92</td>
<td>0.59–0.92</td>
<td>0.75–1.00</td>
<td>0.51–0.99</td>
<td>0.54–0.98</td>
<td>0.32–1.00</td>
<td>0.32–1.00</td>
<td>0.31–0.98</td>
</tr>
<tr>
<td>PPV (un)</td>
<td>0.00</td>
<td>0.14</td>
<td>0.00</td>
<td>0.17</td>
<td>0.08</td>
<td>0.07</td>
<td>0.37</td>
<td>0.11</td>
<td>0.08</td>
<td>0.38</td>
<td>0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>CI</td>
<td>0.00–0.84</td>
<td>0.02–0.43</td>
<td>0.00–0.52</td>
<td>0.00–0.64</td>
<td>0.01–0.26</td>
<td>0.00–0.32</td>
<td>0.16–0.62</td>
<td>0.03–0.25</td>
<td>0.01–0.27</td>
<td>0.20–0.59</td>
<td>0.05–0.26</td>
<td>0.01–0.21</td>
</tr>
<tr>
<td>PPV (0.1)</td>
<td>0.00</td>
<td>0.14</td>
<td>0.00</td>
<td>0.19</td>
<td>0.09</td>
<td>0.10</td>
<td>0.29</td>
<td>0.11</td>
<td>0.11</td>
<td>0.21</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>CI</td>
<td>0.00–0.68</td>
<td>0.01–0.66</td>
<td>0.00–0.55</td>
<td>0.01–0.96</td>
<td>0.00–0.53</td>
<td>0.06–0.65</td>
<td>0.01–0.27</td>
<td>0.00–0.41</td>
<td>0.08–0.44</td>
<td>0.03–0.20</td>
<td>0.00–0.20</td>
<td>0.00–0.20</td>
</tr>
<tr>
<td>PPV (0.27)</td>
<td>0.00</td>
<td>0.35</td>
<td>0.00</td>
<td>0.44</td>
<td>0.24</td>
<td>0.27</td>
<td>0.57</td>
<td>0.29</td>
<td>0.29</td>
<td>0.47</td>
<td>0.28</td>
<td>0.20</td>
</tr>
<tr>
<td>CI</td>
<td>0.00–0.87</td>
<td>0.02–0.87</td>
<td>0.00–0.80</td>
<td>0.03–0.99</td>
<td>0.00–0.59</td>
<td>0.00–0.79</td>
<td>0.19–0.86</td>
<td>0.03–0.55</td>
<td>0.01–0.70</td>
<td>0.22–0.72</td>
<td>0.08–0.45</td>
<td>0.00–0.46</td>
</tr>
</tbody>
</table>

Notes: The NPV and PPV have been calculated as uncorrected (un) or with an assumed prevalence of 0.1 or 0.27. The values in bold indicate statistical significance (p < 0.05 Fisher’s exact test).

DISCUSSION

Predictive animal models offer the potential to reduce compound attrition in clinical development but we need to improve our ability to translate their relevance for efficacy and safety assessment (Chi, 2013). This will drive appropriate application and interpretation of these models, as well as identify opportunities to improve them. Here we describe a cross-company initiative to determine the concordance between the conscious telemetered dog and phase I outcome for 3 cardiovascular parameters.

The analysis showed that there was a good concordance for drug effects on the QTc interval between the dog model and the phase I. At exposure multiples of 10x, sensitivity was 88% and specificity was 76%. The NPV for QTc effects was also very good, and although slightly less impressive, there was also clear value related to the PPV for QTc effects. Taken together, there is a high degree of confidence that compounds which are negative for QTc in the dog will be negative in man and that a positive QTc signal in dog will translate to man. Despite this, as assessed using the methods applied in this analysis, the model’s sensitivity for DBP and HR was low across the exposure multiples and the specificity decreased with increasing exposure.

Sensitivity and specificity are equally important for describing detection of effects, and one should be cautioned against using them independently (Pewsoner et al., 2004). However, by positioning the dog telemetry model just prior to phase I, the appetite for risk is low and thus the model should have a low FP rate (Valentin et al., 2009a,b), or in other words a higher specificity compared with sensitivity. In practice, the level of risk aversion will depend on many factors, for example, the number of project compounds, the therapeutic area, the endpoint, ability to monitor change clinically as well as the degree of change that elicits clinical concern.

There are a number of possible reasons that may explain the apparent lack of concordance for DBP and HR. Each warrant consideration especially for future analyses: (1) effects may be driven by pharmacological mechanisms or metabolites that are species specific (eg, present in dog but not man); (2) inadequate prediction of dosages in man may erode narrow therapeutic indexes; (3) the dog telemetry model is a highly powered study (Guth et al., 2004; Sivarajah et al., 2010; Tattersall et al., 2006) in
FIG. 5. Receiver operating characteristic plots for A, QTc, B, HR, and C, DBP. Each plot was created by plotting the sensitivity against 1 minus the specificity at each of the exposure multiples analyzed (as shown by numbers next to the individual data points). The 95% confidence intervals are shown in light gray for both sensitivity and specificity, whereas the line of unity is indicated with the gray dotted line. Curves that lie to the top left-hand corner in the plot indicate good sensitivity and specificity, whereas curves falling on the line of unity depict poor sensitivity and specificity. The curves indicate that the dog telemetry model has good sensitivity and specificity for QTc but it is less so for HR and DBP changes.

comparison to a typical phase I study (Shah, 2008) and may produce a higher number of FP or conversely, the phase I study may miss effects that might be manifested in a more highly powered clinical study; (4) blood pressure measurement is not equivalent. In dogs, catheters are implanted prior to the study (Leishman et al., 2012), whereas cuff techniques are typically used in humans (Sager et al., 2013). Consequently, blood pressure measurement in dogs is continuous and can occur without any on-study intervention, whereas in clinical practice only "snapshot" measurements are made, and are often a heterogeneous mix of single or triplicate measurements. Additionally, trial participants awareness of the cuff can impact on parameters measured, a phenomenon called the "white-coat" effect (Pickering et al., 1998; Sager et al., 2013); (5) measurement of blood pressure in dogs utilizes a major artery, whereas in humans it is usually measured from a peripheral vessel. This is important because pharmacological agents can exert differential effects on central compared with peripheral vessels (Epstein et al., 2007).

Ironically, the strength of cross-company data sharing (ie, increased number and diversity of data points) may have also contributed to the wide confidence limits thus affecting the actual predictive value of the model. Whereas each company believes in the predictive value of their own in-house model (Jonker et al., 2005; Parkinson et al., 2013), when entering cross-company consortia limitations arise in the process of creating a
consistent data set. This undoubtedly impacts on the assessment of translation. The following aspects should also be acknowledged in the overall interpretation of the data: (1) the phase I result was the endpoint we were interested in predicting outcome for but these studies are not standardized. They are primarily designed to understand safety and tolerability and consequently are not powered to detect small changes in measured parameters. Hence focusing on the phase I outcome may not relate to other clinical assessments; (2) we respected the expert view from the phase I study and performed no reanalysis but, the expert conclusions are not based on hard “cut off” points, and especially for blood pressure there is a lack of uniform agreement on what constitutes a drug-induced signal (Sager et al., 2013). This introduces inconsistent variances. We took this approach because it was assumed that the expert was best placed to draw conclusions on the data. Operationally, it was not easy for many of the companies to find and access the individual clinical data, due to a lack of a centralized clinical database. When data could be found and accessed, there was a sensitivity around performing new data analysis and interpretation that could differ from the original opinion from which decisions were already made and access to the data itself would have involved potentially seeking the approval from 113 compound development teams/data owners and associated legal clearance across the 7 companies. Therefore, the approach taken was chosen for both scientifically and practical reasons; (3) the “effect” data (clinical and nonclinical) was not numerical data and did not look at individual subjects; (4) the number of clinical positives was low; (5) there were large variations in design of the dog studies and their associated power (eg, parallel versus ascending versus crossover designs), n-numbers used, and standardization of definitions of how to calculate data (eg, beat-to beat bin-size); (6) the free fraction of compound in plasma was used rather than the total amount. By using the free fraction in the exposure analysis, we assume that the free concentration at which findings were observed were not only conserved across species but importantly that free plasma exposure drove the effects found.

Future translational analyses can build on this work and benefit from the learning points. For instance, it would be good to reflect on whether inclusion of known reference agents in the analysis is merited because the drug filtering process of safety pharmacology prior to phase I inherently means that the number of clinical positives are low. There should also be a judgment on whether the use of free drug levels is appropriate. If so, noting how the free concentration has been calculated (ie, based on 1 or 2 concentrations) and how the in vivo values have been extrapolated from these point estimates is important. With the advent of open sharing of clinical trial data, one could also consider comparing dog data in trials beyond phase I, which will be larger and likely to contain patients who may be more susceptible to drug effects, as well as using quantitative data from both nonclinical and clinical studies. Moreover, the conduct of prospective nonclinical studies (such as the ILSI-HESI QT project, Hanson et al., 2006) using agreed protocols that employ best practices in their design (eg, acclimatization, environment, data analysis strategies [including power analyses] and employing a PK/PD component (Leishman et al., 2012)) would reduce the multiple variances in approach across companies. Coupling this to reanalysis of individual human data in a standard way across companies will help to refine the assessment of the ability of our nonclinical assays to predict effects in man.

In conclusion, the data presented here show that the dog cardiovascular model, as currently applied, has value for predicting phase I outcome for QTc but it does not predict phase I outcome for HR and DBP changes, and that a weight of evidence approach should be adopted within decision making for assessing risk in phase I studies. It remains to be seen whether the dog cardiovascular model predicts clinical cardiovascular effects in different studies or populations, and it is important to interpret the data presented here within the context of the data collected. That is, the lack of prediction or identification of DBP and HR changes can only be applied to phase I studies as they are typically performed, within the context of the methods applied in this analysis, and the findings should not be generalized to other endpoints without caution or evidence.

FUNDING

The work presented here was funded by each of the contributing pharmaceutical companies.

ACKNOWLEDGMENTS

The authors would like to thank colleagues within the safety pharmacology and general toxicology groups within their respective organizations and specifically to Russ Bialecki, Berenengre Dumotier, Jack Hammond, Tim Hammond, Mark Holbrook, Pierre Laine, James Milligan, Anders Ottoson, Chris Pollard, Joanna Parkinson, Will Redfern, James Resch, John Sherington, Yves Somers, Martin Traebert, and Bill Zinkand. There are no known conflicts of interest.

REFERENCES


