Journal of Pharmacological and Toxicological Methods 92 (2018) 43–51

Contents lists available at ScienceDirect

Journal of Pharmacological and Toxicological Methods

journal homepage: www.elsevier.com/locate/jpharmtox

Research article

The evaluation of endpoint variability and implications for study statistical power and sample size in conscious instrumented dogs

Alan Y. Chiang[a], Brian D. Guth[b,j], Michael K. Pugsley[c], C. Michael Foley[d], Jennifer M. Doyle[e], Michael J. Engwall[f], John E. Koerner[g], Stanley T. Parish[h,*], R. Dustan Sarazan[i]

[a] Eli Lilly and Company, Indianapolis, IN, United States
[b] Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany
[c] Purdue Pharma, LP, Stamford, CT, United States
[d] AbbVie, Abbott Park, IL, United States
[e] Data Sciences International, St. Paul, MN, United States
[f] Amgen, Thousand Oaks, CA, United States
[g] FDA, Washington DC, United States
[h] HESI, Washington DC, United States
[i] Independent, Rhinelander, WI, United States
[j] The Preclinical Drug Development Platform, North-West University, Potchefstroom, South Africa

ARTICLE INFO

Keywords:
Double Latin Square design
Myocardial
Left ventricular dP/dt
Variability
Test sensitivity
Heart rate
Time interval average
Super-interval
Statistical power

ABSTRACT

Introduction: The sensitivity of a given test to detect a treatment-induced effect in a variable of interest is intrinsically related to the variability of that variable observed without treatment and the number of observations made in the study (i.e. number of animals). To evaluate test sensitivity to detect drug-induced changes in myocardial contractility using the variable LVdP/dt_{max}, a HESI-supported consortium designed and conducted studies in chronically instrumented, conscious dogs using telemetry. This paper evaluated the inherent variability of the primary endpoint, LVdP/dt_{max}, over time in individual animals as well as the variability between animals for a given laboratory. An approach is described to evaluate test system variability and thereby test sensitivity which may be used to support the selection of the number of animals for a given study, based on the desired test sensitivity.

Methods: A double 4 × 4 Latin square study design where eight animals each received a vehicle control and three dose levels of a test compound was conducted at six independent laboratories. LVdP/dt_{max} was assessed via implanted telemetry systems in Beagle dogs ($N = 8$) using the same protocol and each of the six laboratories conducted between two and four studies. Vehicle data from each study was used to evaluate the between-animal and within-animal variability in different time averaging windows. Simulations were conducted to evaluate statistical power and type I error for LVdP/dt_{max} based on the estimated variability and assumed treatment effects in hourly-interval, bi-hourly interval, or drug-specific super interval.

Results: We observe that the within-animal variability can be reduced by as much as 30% through the use of a larger time averaging window. Laboratory is a significant source of animal-to-animal variability as between-animal variability is laboratory-dependent and is less impacted by the use of different time averaging windows. The statistical power analysis shows that with $N = 8$, the double Latin square design has over 90% power to detect a minimal time profile with a maximum change of up to 15% or approximately 450 mm Hg/s in LVdP/dt_{max}. With $N = 4$, the single Latin square design has over 80% power to detect a minimal time profile with a maximum change of up to 20% or approximately 600 mm Hg/s in LVdP/dt_{max}.

Discussion: We describe a statistical procedure to quantitatively evaluate the acute cardiac effects from studies conducted across six sites and objectively examine the variability and sensitivity that were difficult or impossible to calculate consistently based on previous works. Although this report focuses on the evaluation on LVdP/dt_{max}, this approach is appropriate for other variables such as heart rate, arterial blood pressure, or variables derived from the ECG.

* Corresponding author at: 1156 15th St., N.W., Suite 200, WA 20005, United States.
E-mail address: sparish@hesiglobal.org (S.T. Parish).

https://doi.org/10.1016/j.vascn.2018.02.009
Received 20 December 2017; Received in revised form 30 January 2018; Accepted 28 February 2018
Available online 02 March 2018

1. Introduction

Safety pharmacology studies are conducted on drug candidates to assess for safety relevant effects when administered at therapeutically relevant or higher doses (ICH S7A, 2001). The assessment of possible effects on the cardiovascular system are frequently conducted in conscious dogs that have been chronically instrumented for the collection of the cardiovascular variables of interest using telemetry which typically includes arterial blood pressure, left ventricular pressure and the electrocardiogram (ECG). The maximal rate of pressure increase in the left ventricular during systole ($LVdP/dt_{max}$) has been shown to be a sensitive variable to assess drug-induced effects on cardiac contractility (Guth et al., 2015). Drugs with both positive (amrinone and pimobendan) and negative (atenolol and itraconazole) inotropic effects, known to produce such effects clinically, were tested in a cross-laboratory evaluation and $LVdP/dt_{max}$ proved to be a robust variable to detect dose-dependent effects of the agents tested. For those studies, each of the laboratories included 8 dogs and studies were conducted using a double Latin square design. The use of 8 dogs was based on the extensive experience of the investigators and limited published data with this type of model; however, ultimately the number of animals for the Health and Environmental Sciences Institute (HESI) supported study was selected subjectively.

With each of the four test compounds studied, one treatment arm was the vehicle used without test article. This is an important treatment arm since the vehicle treatment data was used in this study to evaluate the variability of the collected data within and between animals and across laboratories. We propose herein a methodology for making this assessment that should allow any laboratory to determine the variability of all measured variables. Here we report the evaluation on $LVdP/dt_{max}$, but this approach is appropriate for other variables such as heart rate (HR), arterial blood pressure (BP), or variables derived from the ECG. By defining the variability of each variable assessed, the experimenter can define the test sensitivity of their experimental setting in order to answer the question: what size of a drug-induced effect could have been detected? This is of particular importance for studies concluding that no drug-induced effect was found. Furthermore, since the test sensitivity is also a function of the number of animals included in a study, this approach provides a rational approach for deciding how many animals to include in such a study. This is often mandatory for research scientists to obtain permission from either Institutional Animal Care and Use Committee (IACUC) or governmental agencies (such as the National Institutes of Health, NIH) to conduct this type of non-clinical study.

2. Materials and methods

2.1. Test facilities

Studies were performed by 6 independent companies and data were reported previously (Guth et al., 2015). Each individual study was subject to the local guidelines in terms of the vivarium conditions, study conduct and animal use approval procedures. All participating institutions have warranted strict adherence to all applicable animal use regulations in the conduct of these studies. Although efforts were made to harmonize testing procedures and conditions, the local animal use regulations were always prioritized should any conflicts have arisen during the conduct of the study.

2.2. Experimental animals

All participating laboratories used purpose bred beagle dogs acquired from a vendor within their geographic region (North America or Europe). Some laboratories used only male dogs and other laboratories used both males and females. The source and sex of the dogs used by the various laboratories were reported previously (Guth et al., 2015).

Most animals had been used previously during the conduct of safety pharmacology studies but were healthy and free of any residual test article at the start of the study. At one laboratory the animals were naïve at the study onset. No animals were required to be euthanized in the context of this study. After an appropriate recovery period following surgery or washout period after receiving a drug, animals were subjected to a standard clinical pathology examination to evaluate their health status according to local procedures (typically including blood cell counts, serum electrolytes and biochemistry parameters indicative of kidney and liver function) and were qualified for use in further studies.

2.3. Telemetry instrumentation

Each participating laboratory used one of three commercially available implantable large animal telemetry systems; PhysioTel™ model D70-PCTP (Data Sciences International, St. Paul, MN), PhysioTel™ Digital model L21 (Data Sciences International, St. Paul, MN), or ITS model T27 (Konigsberg Instruments, Monrovia, CA).

Regardless of the telemetry system used, all dogs were instrumented to monitor aortic BP, left ventricle pressure (LVP), the ECG, body temperature and activity. Note, however, that body temperature and activity endpoints were not evaluated during the conduct of the study. All methods related to the surgical preparation of animals, telemetry implants and recording systems employed, and drugs evaluated are found in Guth et al. (2015) and Pugsley et al. (2017).

2.4. Study design

Four different treatments were administered to each dog in the order prescribed by a randomly generated double Latin square design over four treatment days at each test site with an appropriate washout period between days (Guth et al., 2015). The washout period was a minimum of 72 h between treatment days. The double Latin square study design combines two randomly generated 4×4 Latin squares (Sarazan et al., 2011). See Appendix A for an illustration of Latin square designs.

The food provided was withdrawn approximately 2 h before dosing in the morning and reintroduced in the afternoon, which was well after the anticipated time to peak drug concentration (T_{max}) of the tested drug. The study dosing technicians were not blinded to treatment; however, the studies were conducted by the same technicians within each laboratory under standard GLP procedures. Best practices for animal handling were implemented to minimize any potential bias in telemetry data collection and analysis.

2.5. Data collection and analysis

2.5.1. Raw data (signals)

Digital LVP, aortic BP and ECG signals were continuously acquired from at least one hour prior to dosing through 24 h post dose on each study day. Sampling rates were ≥ 500 Hz for LVP and ECG signals and ≥ 250 Hz for BP signals which is adequate for the frequency content of each of these signal types (Sarazan, 2014). Digital raw data files were archived to electronic media and retained at each individual study site for future analysis as agreed upon within the HESI Cardiac Safety Technical Committee.

2.5.2. Derived data (variables)

Various derived variables were calculated from output of digital acquisition units at each study site. However, for the purpose of this evaluation, only $LVdP/dt_{max}$ data were used. A similar evaluation could be performed with any of the additional variables measured as previously reported (Pugsley et al., 2017).

Derived data were calculated for every cardiac cycle and the results were collapsed into 10-min mean values for analysis. These mean

values were further averaged into various time intervals (0.5, 1, 1.5, 2, 2.5, 3, 3.5, and 4 h) plus an additional pre-specified ("large summary") set of "super-intervals" defined by Guth et al. (2015).

The hourly intervals were derived based on the averages of 6 of the 10-min mean values, resulting 24 intervals during the 24-h post-dosing period. Other fixed time intervals were derived similarly. The super-intervals used for each compound were defined by a data evaluation subteam prior to conduct of the statistical analysis. The selection of super-intervals was intended to limit variability associated with ambulatory dog cardiovascular assessments and avoids disturbances associated with dosing, changes in light cycle or at the time of blood sampling for drug exposure confirmation. Each compound was treated individually, selecting intervals from the average of LVdP/dt_{max} across the laboratories that tested a given compound.

2.5.3. Between- and within-animal variability

The use of various time averaging windows for the derived LVdP/dt_{max} from each site and its impacts on variability were statistically evaluated. Only the vehicle data was used in this evaluation. Let y_{ijk} be the averaged LVdP/dt_{max} measured from the k-th time interval of the j-th animal in the i-th experiment of each site during the vehicle treatment period, where: $i = 1, \ldots, M$, $j = 1, \ldots, N$, $k = 1, \ldots, K$, M = the number of studies per site, N = the number of animals in each study, and K = the number of total time intervals used for analysis. Between- and within-animal variability was assessed based on the following linear mixed effect model (Littell, Pendergast, & Natarajan, 2000):

$$y_{ijk} = \mu_i + e_{ij} + \varepsilon_{ijk} \quad (1)$$

where e_{ij} are independent and identically distributed as a normal with mean 0 and variance σ_b^2 (between- or inter-animal variability), and ε_{ijk} are independent and identically distributed as a normal with mean 0 and variance σ_e^2 (within-animal, or intra-animal variability). In order to assess the source of variability associated with different time averaging windows, model (1) was fitted for the following time averages: every 0.5, 1, 1.5, 2, 2.5, 3, 3.5 or 4 h of the 10-min mean values, corresponding to 48, 24, 18, 12, 10, 8, 7, or 6 time points respectively during the 24-h post-dosing period. The super-interval was not included in this evaluation because the defined intervals were irregular and compound dependent. A standard deviation (SD) for each variance component and corresponding coefficient of variation (100% × SD/mean) were derived to characterize variability. A step-by-step procedure to estimate between- and within-animal variability and SAS codes used are provided in Appendix B.

2.5.4. Statistical power analysis

Statistical power analysis was conducted by simulating expected treatment effects compared to the corresponding vehicle data. We focused on the analysis of LVdP/dt_{max} to illustrate the statistical power evaluation; other variables can be evaluated similarly. The statistical model for LVdP/dt_{max} analysis was described by Guth et al. (2015) and can be expressed as follows (Chiang & Wang, 2015):

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + t_l + b_l x_{ijk} + (\alpha t)_{il} + (\beta t)_{jl} + (\gamma t)_{kl} + e_{ikl} + \varepsilon_{ijkl} \quad (2)$$

where

y_{ijkl} is the l-th post-baseline LVdP/dt_{max} measurement of animal j in period (day) k receiving dose i, with $i, k = 1, .., 4$, $j = 1, .., N$, and $l = 1, .., T$,
μ is the overall mean,
α_i, β_j, γ_k, and t_l describe the main effects for dose, animal, period and time, respectively,
x_{ijk} is the baseline LVdP/dt_{max} for animal j receiving dose i in period k,
b_l is the random slope for each time point, and
$(\alpha t)_{il}$, $(\beta t)_{jl}$, and $(\gamma t)_{kl}$ are the interactions of treatment group, animal and period with time, respectively.

N is the total number of animals per study in each site, with $N = 4$ or $N = 8$ representing the use of a 4 × 4 single or double Latin square design, respectively. See Appendix A for an illustration of the designs. T is the total number of time points in the analysis for each study; $T = 24$, $T = 12$, or $T = 5$, representing the use of 1-h, 2-h, or drug-specific super interval, respectively. Parameter constraints that allow for a unique solution of the main effects and interactions are implicit. The within-animal correlations across time are specified by the random effects designated by e_{ikl}'s, while ε_{ijkl}'s are measurement errors. It was assumed that all e_{ikl}'s and ε_{ijkl}'s are independent and normally distributed, with $e_{ikl} \sim N(0, \sigma_e^2)$ and $\varepsilon_{ijkl} \sim N(0, \sigma^2)$. The variance components constitute a "compound symmetry" covariance structure for the measurements from the same animal across time points. See, for example, Keselman, Algina, and Kowalchuk (2001) and Appendix A of Chiang, Smith, Main, and Sarazan (2004).

Data from Site 2 were used to illustrate the power analysis as its variability was close to the estimated median variability of six sites (see 3.1). First, the estimates of between- and within-animal variability for LVdP/dt_{max} and their 95% confidence intervals (CIs) using 1-h, 2-h, or drug-specific super-interval were calculated. Each study data set was simulated conditional on its estimated variability and the following assumed high-dose treatment effects of interest:

- 5 mg/kg (high-dose) amrinone-like effect,
- profile A, which is approximately 80% of the high-dose amrinone effect, and
- profile B, which is approximately 75% of profile A effect.

Three time-averaging intervals were considered (Fig. 1):

- hourly,
- bi-hourly, and
- the super-interval defined for amrinone.

Mid-dose treatment effects were assumed to be approximately 50% of the high-dose treatment effects. Low-dose treatment effects were assumed to be the same as vehicle. We also assumed no period effect was present. An individual animal data vector was then simulated assuming a multivariate normal distribution with a mean vector from one of the treatment profiles and the covariance matrix from the estimated between- and within-animal variability. False positive rates were assessed when no treatment effect was present. Model (2) was fit to the simulated data and the simulation process was repeated for 2000 times in each of the three time-averaging intervals. The procedure was then repeated using $N = 8$ and $N = 4$.

A positive finding was concluded if

- an overall dose-response trend test is significant at the 0.05 level,
- the dose-response trend test is significant at the 0.05 level for an individual time point when there is strong evidence of treatment-by-time interaction (p-value < 0.01), or
- a significant overall F-test at the 0.05 level for non-monotonic dose response.

The additional interaction test was used (if significant) to trigger multiple testing at each individual time point, in order to reduce false-positive findings. Type I error rate of the statistical testing procedure was evaluated from simulated vehicle data under the null hypothesis (H0). The number of positive findings divided by the number of simulations yields the type I error under H0. Statistical power of the statistical testing procedure was evaluated from based on the simulated treatment effects shown in Fig. 1 under the alternative hypothesis (H1). The number of positive findings divided by the number of simulations is

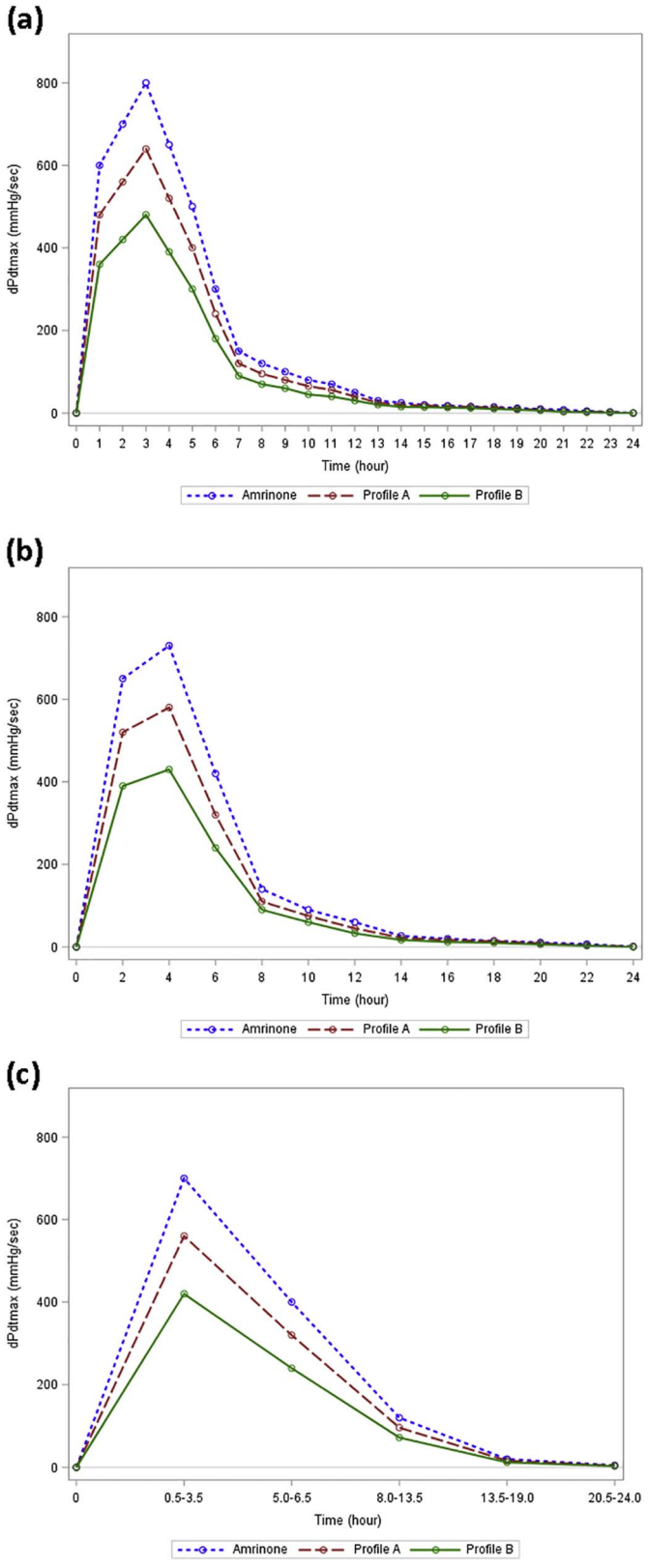

Fig. 1. The effect profiles used for power calculation of LVdP/dt_{max}: 5 mg/kg (high-dose) amrinone, profile A is approximately 80% of the high-dose amrinone effect, and profile B is approximately 75% of profile A effect: (a) hourly intervals, (b) bi-hourly intervals, (c): super-intervals.

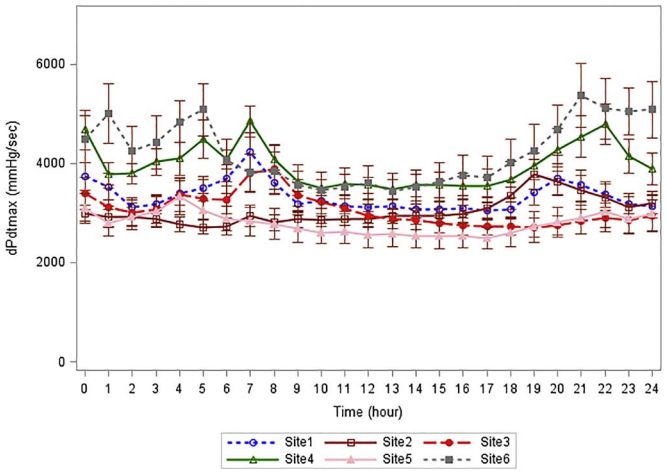

Fig. 2. Hourly averages of LVdP/dt_{max} in vehicle treated conscious instrumented Beagle dogs.

the statistical power under various treatment effect scenarios of H1. A step by step procedure to simulate study level data for power analysis is described in Appendix C.

3. Results

3.1. Between- and within-animal variability

The vehicle LVdP/dt_{max} data from each site are derived and averaged in three time windows: hourly, bi-hourly, and super-intervals calculated from the 10-min mean values. Without loss of generality, the super-intervals are derived based upon the pharmacological effect of amrinone. The mean vehicle data and its 95% confidence intervals at each time point for each site are presented in Figs. 2–4. It is observed that vehicle data from Site 4 and Site 6 tend to have larger mean values, as well as variability across time points. A wider confidence interval at each time point is also observed in Site 4 and Site 6, compared with other sites; however, the coefficient of variation may be offset by its large mean value. The source of variability is further estimated using the variance component approach described in model (1).

Within-animal variability and between-animal variability of LVdP/dt_{max} are evaluated in different time averaging windows to assess the impact of time averaging window in variability. The results are summarized in Fig. 5 for within-animal variability and between-animal variability. Overall, Site 6 has larger within- and between-animal

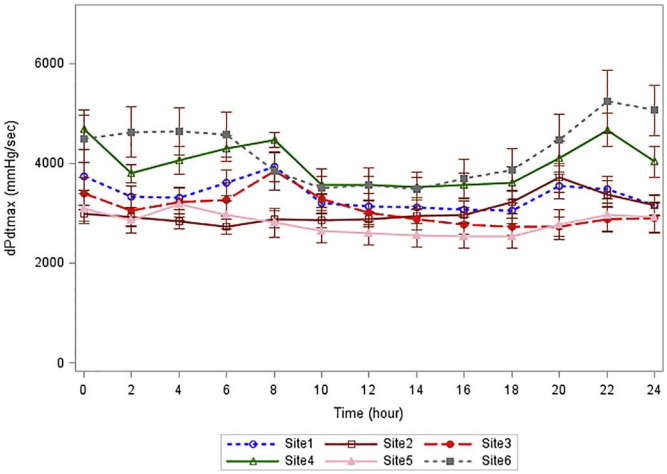

Fig. 3. Bi-hourly averages of LVdP/dt_{max} in vehicle treated conscious instrumented Beagle dogs.

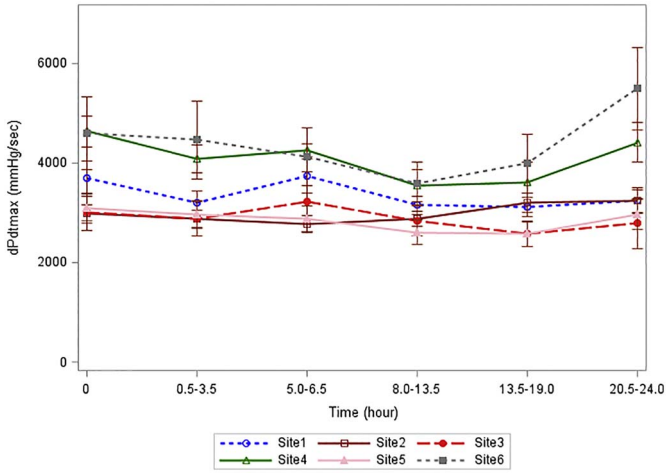

Fig. 4. Measurements of LVdP/dt_{max} based on super-interval in vehicle treated conscious instrumented Beagle dogs. The super-intervals illustrated here are defined as follows: 0.5–3.5, 5.0–6.5, 8.0–13.5, 13.5–19.0, and 20.5–24 (in hours after dosing).

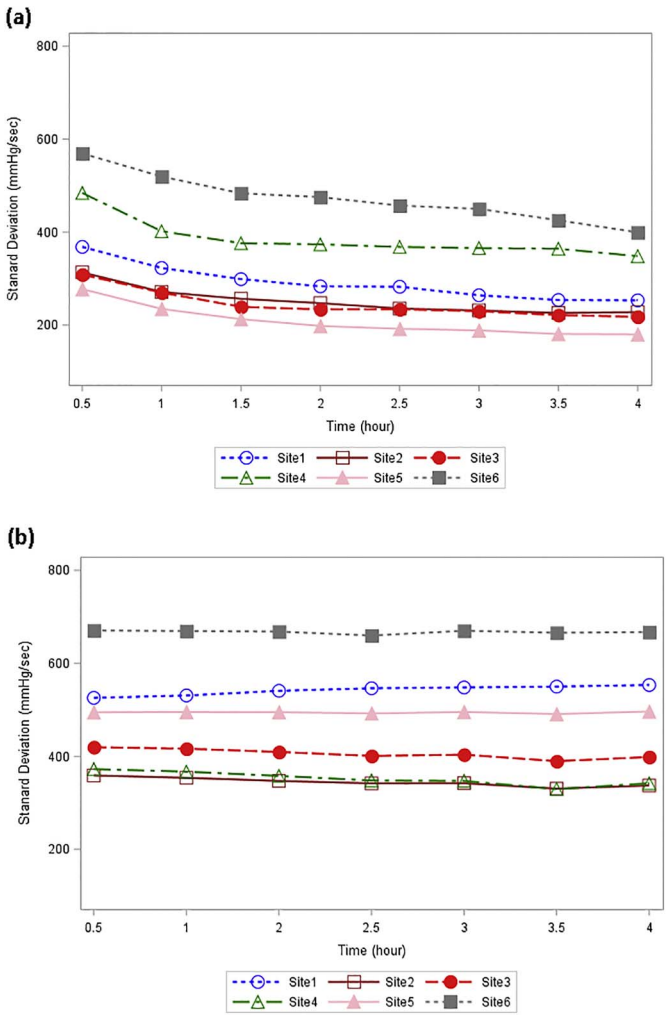

Fig. 5. Comparison of variability of LVdP/dt_{max} in different time averaging windows: (a) within-animal, (b) between-animal. The variability is expressed in standard deviation.

variability. The large variability of LVdP/dt_{max} from Site 4 observed in Figs. 2–4 can be attributed to within-animal variability as the between-animal variability is deemed to be small compared to its peers. While the within-animal variability can be reduced by as much as 30% from

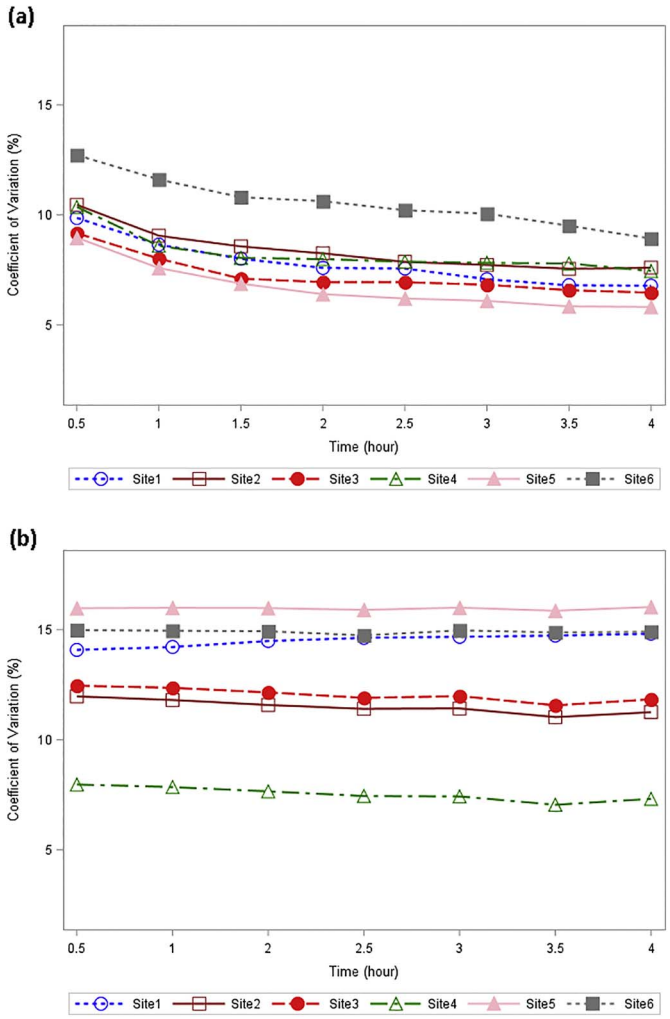

Fig. 6. Comparison of variability of LVdP/dt_{max} in different time averaging formats: (a) within-animal, (b) between-animal. The variability is expressed in coefficient of variation.

the use of 0.5-h intervals to that of 4-h intervals, the between-animal variability remains fairly consistent across different time averaging windows.

The mean-normalized variability, or the coefficient of variation, is shown in Fig. 6. After normalization, data from Site 2 appears to be representative of the median between- and within-animal variability among the six study sites in the dataset. Hence Site 2 data was used to illustrate the comparison of variability in different time averaging windows and statistical power analysis. Table 1 provides a detailed listing of vehicle LVdP/dt_{max} using 1-h, 2-h, or drug-specific super-interval. The super-interval here is based on amrinone. Estimates of between- and within-animal variability for LVdP/dt_{max} and their 95% CIs using these three time averaging windows are shown in Table 2. Using the super-interval approach, the within-animal variability can be reduced by as much as 20% and the between-animal variability can be reduced only by 5%.

3.2. Statistical power analysis

As indicated 2.5.4, three treatment profiles were considered in the statistical power evaluation. The maximum treatment effects for each of the three profiles are summarized in Table 3. The high-dose amrinone shows a peak increase of approximately 700–800 mm Hg/s or 23–27% from baseline in LVdP/dt_{max}. Profiles A and B assume peak increases of approximately 20% and 15% in LVdP/dt_{max}, respectively.

The statistical power analysis results for Site 2 are summarized in

Table 1
The vehicle LVdP/dt_{max} data based on hourly, bi-hourly or super-interval averages in instrumented Beagle dogs from Site 2.

Hourly		Bi-hourly		Super-interval	
Time (hour)	Mean LVdP/dt_{max} (mm Hg/s)	Time (hour)	Mean LVdP/dt_{max} (mm Hg/s)	Time (hour)	Mean LVdP/dt_{max} (mm Hg/s)
0	2996	0	2996	0	2996
1	2919	0–2	2921	0.5–3.5	2880
2	2924	2–4	2826	5.0–6.5	2772
3	2883	4–6	2721	8.0–13.5	2881
4	2769	6–8	2878	13.5–19.0	3210
5	2714	8–10	2863	20.5–24	3238
6	2728	10–12	2883		
7	2947	12–14	2941		
8	2808	14–16	2968		
9	2874	16–18	3221		
10	2853	18–20	3706		
11	2887	20–22	3380		
12	2878	22–24	3152		
13	2939				
14	2943				
15	2948				
16	2988				
17	3095				
18	3347				
19	3779				
20	3633				
21	3452				
22	3309				
23	3111				
24	3206				

Table 2
A comparison of between and within animal variability across time intervals (vehicle data from Site 2; the Super-Interval was derived based on the profile of amrinone).

Time interval	# of time points	Between-animal		Within-animal	
		Estimate (mm Hg/s)	95% CI (mm Hg/s)	Estimate (mm Hg/s)	95% CI (mm Hg/s)
Hourly	24	353.8	(209.1, 454.6)	271.3	(254.8, 290.1)
Bi-hourly	12	347.1	(202.2, 447.3)	247.3	(226.5, 272.3)
Super-interval	5	335.3	(189.4, 434.7)	217.9	(190.8, 254.0)

Table 3
The maximum treatment effect for each of the 3 profiles in statistical power evaluation: reported vehicle adjusted changes (mm Hg/s) and percent changes from baseline in LVdP/dt_{max} (data from Site 2).

	Amrinone	A	B
Hourly	800 (27%)	640 (21%)	480 (16%)
Bi-hourly	730 (24%)	580 (19%)	430 (14%)
Super-interval	700 (23%)	560 (19%)	420 (14%)

Table 4. It shows that with $N = 8$, the double Latin square design has over 90% power to detect a minimal time profile with a maximum change of up to 15% or approximately 450 mm Hg/s in LVdP/dt_{max}. With $N = 4$, the single Latin square design has over 80% power to detect a minimal time profile with a maximum change of up to 20% or approximately 600 mm Hg/s in LVdP/dt_{max}. A small favorable gain in statistical power is observed in bi-hourly time interval as it is likely attributed to the balance of time averaging (smaller variability observed in larger time windows) and multiple testing (larger time windows result in a small number of time points).

The type I error (false positive rate) for each of the three time averaging windows ranges from 10% to 15%. In general, without multiplicity adjustment, the more time points, the larger the type I error rate. It is not surprising to see that the use of the super-interval approach results in a smaller type I error due to the limited number of time points evaluated.

4. Discussion

The objective of this HESI-sponsored consortium study was to investigate, using formal experimental and analysis methods, the influence of drug-induced changes in LVdP/dt_{max} as an index of cardiac contractility in instrumented Beagle dogs. The scale of this quantitative data set evaluating the acute cardiac effects produced by the action of drugs that exhibit either positive or negative effects consistently across six independent study sites provides a unique opportunity to objectively examine the safety pharmacology study design variables that were difficult or impossible to calculate consistently from previous studies. The study protocol was designed in order to afford researchers an opportunity to use the same experimental design and data collection method in order to ensure a uniform evaluation of maximal rate of pressure increase in the left ventricle during systole (i.e. LVdP/dt_{max}) as a surrogate for the inotropic state of the heart. Importantly, as stated in Guth et al. (2015), all six sites accurately and consistently detected changes with the positive and negative inotropes tested using telemetry-instrumented Beagle dogs in spite of some uncontrollable differences due to animal source, environment, acclimation procedures, and time post-surgery. The source of animal variability, the sensitivity of LVdP/dt_{max} under different time averaging windows and their impacts on statistical power were further evaluated in the present paper. This original approach allowed us to quantify the influence of time averaging window on the sensitivity of within- and between-animal variability and power analysis. Furthermore, this analysis can be utilized to guide investigators to develop a robust study design using the appropriate number of animals supporting a critical component of the 3Rs (Russell and Burch, 1959).

The size of time averaging window is of particular interest because a smaller window size provides opportunities to characterize the pharmacodynamic features of treatment effect over time, while a large window reduces the variability and increases the sensitivity of detecting drug-induced treatment effects. From the HESI-sponsored consortium dataset, the within-animal variability of LVdP/dt_{max} can be reduced by increasing the time averaging window from 0.5-h to a larger time averaging window such as 4-h, while the between-animal variability remains consistent across different time averaging windows. In addition, animal variability among the study sites could be due to the

Table 4
The statistical power for 3 effect profiles (amrinone, profile A and profile B) in LVdP/dt_{max} with N = 8 and N = 4 (data are simulated from Site 2 experiments).

Time interval	# of time points	N = 8			N = 4			Type I error (2-sided)
		Amrinone	A	B	Amrinone	A	B	
Hourly	24	99%	99%	93%	95%	85%	63%	15%
Bi-hourly	12	99%	99%	95%	96%	86%	65%	12%
Super-interval	5	99%	99%	90%	92%	81%	60%	10%

difference in site processes and procedures or the source of animals, their age, acclimation procedures, time post-surgery, or other characteristics. See Table 2 of Guth et al. (2015) for details regarding study site characteristics. Selection of an appropriate time average window is critical and should be determined for each specific study. Multiple factors including pharmacokinetic properties of the test compound (T_{max}, half-life, etc.) and sources of variability (light:dark transitions, room entry, etc.) should be considered. Lengthening the time averaging windows will minimize the impact of within-animal variability and improve statistical sensitivity. In contrast, inappropriately long time averaging windows must be avoided as they can inadvertently blunt the magnitude of an effect and potentially mask a transient or short-lived effect. It should be noted that the statistical evaluation of between- and within-animal variability did not take into account the light: dark transition. A conventional approach has been to block out the data collected during the transition to remove the within-animal variability.

Using data from Site 2, statistical power was evaluated for LVdP/dt_{max} under the three different treatment effect profiles, three different time averaging windows, and two different sample sizes in Latin square designs. We found that the design has over 90% power to detect a minimal time profile with a maximum change of up to 15% or approximately 450 mm Hg/s in LVdP/dt_{max} with a double 4 × 4 Latin square design of $N = 8$, and 80% power to detect a minimal time profile with a maximum change of up to 20% or approximately 600 mm Hg/s in LVdP/dt_{max} with a single 4 × 4 Latin square design of $N = 4$. These estimates could be conservative because we assume that (1) there is no low-dose effect, (2) the mid-dose effect is approximately 50% of the high-dose effect, and (3) the treatment effect quickly diminishes after the maximal treatment effect. Type I errors were inflated in the simulations, as shown in Table 4 where the range was between 10 and 15%. This likely was due to the hypothesis testing procedure employed in the statistical analysis, insofar as corrections to address multiple independent testing were not incorporated into the study design. The procedure first evaluated the linear dose-response trend at the overall time profile level, as well as at the individual time point if there was strong evidence of treatment effect changes over time. If there was also strong evidence of non-dose-response relationship, a multiplicity adjusted *t*-test is also evaluated (see Fig. 7). Type I error can be minimized by reducing the number of time points evaluated and/or by setting the level (alpha) of nominal significance smaller than 0.05 in each test.

The size of the time averaging window used also plays a key role in power analysis. A smaller time averaging window leads to a larger data variability and a larger number of post-dosing time points, resulting in increased power and inflated type I error. In contrast, a larger time averaging window reduces data variability and sensitivity, and a smaller number of post-dosing time points leads to decreases in power and type I error. Using a larger time averaging window in the data analysis could also lose the unique features of 24-h data collection and limit interpretation of findings. This is a well-known bias and variance trade-off phenomenon (e.g., see Chapter 13 of Box & Draper, 2006).

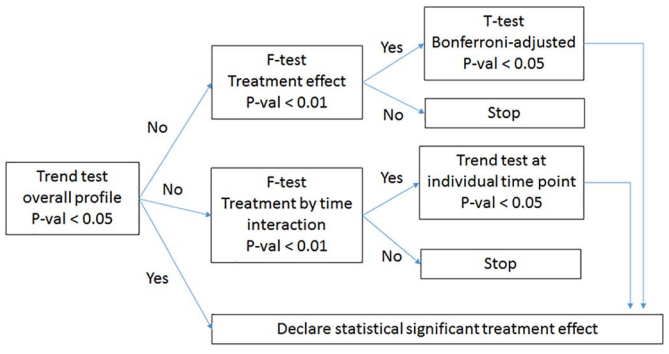

Fig. 7. Statistical testing procedure in evaluating drug induced treatment effect in LVdP/dt_{max}.

While not evaluated in this paper, an alternative approach is to divide 24-h post-dosing into different time phases with 4–5 time points within each phase.

In general, increasing the number of animals included in a given study will allow one to detect smaller treatment-related effects with increased statistical power. The selection of the appropriate number of animals for a given study should therefore be based on the purpose of that study and the treatment size deemed adequate for that purpose. As an example, one might consider including a smaller number of animals in an initial study intended to detect potential cardiovascular effects of a drug candidate. In such cases, the advantage of being able to use a smaller animal number (for instance $N = 4$) might be sufficient to take into account somewhat less test sensitivity and statistical power. On the other hand, definitive cardiovascular safety pharmacology studies with a compound intended for further preclinical and clinical development may warrant using a higher number of animals ($N = 6$–8) to ensure that small treatment-induced effects are not left undetected. This is also important in the case of data drop-out due to loss of an animals or loss of one or more physiological signals from animals. If there is data drop–out in a study with only 4 animals, there may be insufficient data to provide a robust basis for data interpretation. Including eight animals may still be able to provide a robust study outcome even if some data are lost. Furthermore, owing to unreliable *p*-values obtained with small sample sizes, Curtis et al. (2015) expresses concerns of using a size of $N < 5$ per group regardless of the outcome of power analysis.

It is worthwhile to note that statistical power was evaluated under Latin square designs. The simulation procedure can be easily extended to parallel designs (not assessed in this paper). Consider a 4 × 4 Latin square design where 4 dogs each received a vehicle control and 3 dose levels of a compound on four separate dosing days. A parallel design would likely require > 4 dogs per group to achieve the same information. This is because the statistical test of a treatment effect under parallel designs will rely more on between-animal variability, which is larger than within-animal variability (Fig. 5). Given the robustness and ability of the Latin square study design to handle within- and between-animal variability, it would be expected that a parallel study design would require more animals per group to maintain a similar statistical power and detectable change.

5. Conclusion

A key element of complying with the adequacy of experimental design and statistical validity of analysis of drug-induced toxicity is to use a number of animals for a given experiment that provides a data set which adequately addresses the hypothesis being tested and with the predetermined test sensitivity (McGrath & Curtis, 2015); what size of effect should the study be able to detect? This fundamental concept for performing experimental animal studies of high quality has been often forgotten or ignored. Using the HESI consortium data set and the statistical approach described, it is hoped that this will serve to encourage investigators to address these issues prior to beginning any experimental animal study. This requires a definition of the size of effect one would like to detect in a given variable and, with an understanding of the variability of that variable without treatment, one has the basis for a rationale decision on how many animals to include in a study. Only with such an approach can a robust experimental result be obtained without using more animals that are necessary.

Acknowledgements

The authors would like to acknowledge the HESI Cardiac Safety Committee Integrative Strategies Working Group members for their intellectual contributions to the study design, compound selection and other key aspects of the studies reported. Additionally, the authors would like to thank individuals who provided additional assistance for this study including: Frank Cools for providing a bioanalysis of the

canine samples dosed with pimobendan and atenolol, Jim Saul for consulting on the statistical analysis plan and QTest Labs for their involvement with the Millennium in-life phase. The HESI consortium includes representatives of the following companies and institutions: AbbVie, Amgen, AstraZeneca, Battelle Memorial Institute, Boehringer Ingelheim, Bristol-Myers Squibb, ChanRx Corporation, Covance, Data Sciences International, Eli Lilly, GE Healthcare, Genentech, GlaxoSmithKline, Hoffman-La Roche, Johnson & Johnson, Lifespan Heart Center, Merck Research Laboratories, Michigan State University, Millennium: The Takeda Oncology Company, MPI Research, National Cancer Institute, NIH, Novartis, Pfizer, Pharmaceuticals & Medical Devices Agency, Purdue Pharma LP., Sanofi, The Ohio State University, University of Miami (FL), US EPA, US FDA, Vertex Pharmaceuticals.

Disclaimer

The opinions presented here are those of the authors. No official support or endorsement by the US FDA and participating companies is intended or should be inferred.

Appendix A. Single and double Latin square designs

An illustration of a double 4-by-4 Latin square design where eight animals are randomly assigned such that each receive a vehicle control and three dose levels of a test compound (denoted by treatment groups 1–4) on four separate dosing days (periods). A single Latin square design consists of only one Latin square instead of two.

Treatment group		Animal				Animal			
		1	7	2	6	5	8	4	3
Period (day)	1	1	2	3	4	1	2	3	4
	2	3	4	2	1	2	3	4	1
	3	4	3	1	2	4	1	2	3
	4	2	1	4	3	3	4	1	2

Appendix B. Evaluating between- and within-animal variability

Step 1: For each animal in each site and each study, average the six 10-min vehicle values within an hour to create the hourly averages. These values are denoted by y_{ijk}.
Step 2: Fit the data to analysis of variance model (1). SAS codes are provided below:

```
PROC MIXED DATA=DATASET COVTEST CL;
     ID STUDY ANIMAL;
     CLASS STUDY ANIMAL HOUR;
     MODEL Y = STUDY STUDY*HOUR / S CL DDFM=KR ALPHA=0.05;
     REPEATED HOUR/SUBJECT = STUDY*ANIMAL TYPE=CS;
     LSMEANS STUDY;
     BY SITE;
RUN;
```

Step 3: Repeat Steps 1–2 for other time averaging windows.

Appendix C. Statistical power simulation procedure

Step 1: Input the hourly mean vehicle data from Table 1.
Step 2: Input the hourly covariance matrix with between- and within-animal variability derived from Appendix B.
Step 3: Generate a multivariate normal distribution based on mean and covariance from Steps 1–2. This simulates a set of hourly vehicle data.
Step 4: Repeat Step 3 to generate a set of hourly low-dose data.
Step 5: Add the 1× or 0.5× of high-dose treatment mean effects illustrated in Fig. 1 to Step 1, and repeat Step 3 to generate a set of hourly high- and mid-dose data, respectively.
Step 6: Fit the data to model (2), and record the *p*-values.
Step 7: Repeat Steps 1–6 for 2000 times.
Step 8: Repeat Steps 1–7 for power analysis based on bi-hourly and super intervals.

References

Anon (2001). ICH S7A: Safety pharmacology studies for human pharmaceuticals. *Federal Register, 66*, 36791–36792.
Box, G. E. P., & Draper, N. R. (2006). *Response surfaces, mixtures, and ridge analyses* (2nd ed.). Hoboken, NJ: John Wiley & Sons Inc.
Chiang, A. Y., Smith, W. C., Main, B. W., & Sarazan, R. D. (2004). Statistical power analysis for hemodynamic cardiovascular safety pharmacology studies in beagle dogs. *Journal of Pharmacological and Toxicological Methods, 50*, 121–130.
Chiang, A. Y., & Wang, M. D. (2015). Incorporating biomarkers into the analysis of preclinical cardiovascular safety studies. *Statistics in Biopharmaceutical Research, 7*, 66–75.
Curtis, M. J., Bond, R. A., Spina, D., Ahluwalia, A., Alexander, S. P. A., Giembycz, M. A., ...

McGrath, J. C. (2015). Experimental design and analysis and their reporting: new guidance for publication in BJP. *British Journal of Pharmacology, 172*, 3461–3471.
Guth, B. D., Chiang, A. Y., Doyle, J., Engwall, M. J., Guillon, J.-M., Hoffmann, P., ... Sarazan, R. D. (2015). The evaluation of drug-induced changes in cardiac inotropy in dogs: Results from a HESI-sponsored consortium. *Journal of Pharmacological and Toxicological Methods, 75*, 70–90.
Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology, 54*, 1–20.
Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine, 19*, 1793–1819.
McGrath, J. C., & Curtis, M. J. (2015). BJP is changing its requirements for scientific papers to increase transparency. *British Journal of Pharmacology, 172*, 2671–2674.
Pugsley, M. K., Guth, B., Chiang, A. Y., Doyle, J., Engwall, M., Guillon, J.-M., ... Sarazan, R. D. (2017). An evaluation of the utility of LVdP/dt_{40}, QA interval, LVdP/dt_{min} and tau as indicators of drug-induced changes in contractility and lusitropy in dogs. *Journal of Pharmacological and Toxicological Methods, 85*, 1–21.
Russell, W. M. S., & Burch, R. L. (1959). *The principles of humane experimental technique.* (Methuen, London).
Sarazan, R. D. (2014). Cardiovascular pressure measurement in safety assessment studies: Technology requirements and potential errors. *Journal of Pharmacological and Toxicological Methods, 70*, 210–223.
Sarazan, R. D., Mittelstadt, S., Guth, B., Koerner, J., Zhang, J., & Pettit, S. (2011). Cardiovascular function in nonclinical drug safety assessment: Current issues & opportunities. *International Journal of Toxicology, 30*, 272–286.