POWER CALCULATIONS for LLNA Protocols

1.0 LLNA:BrdU-ELISA

During their review of the LLNA: BrdU-ELISA test method, some members of the ICCVAM LLNA expert peer review panel requested information on statistical power vs. number of animals used for this assay. They wanted know how many animals would be adequate for detecting the associated threshold stimulation index for a positive response (e.g., SI > 3).

This required power calculations to determine the number of animals needed to demonstrate statistical significance control and treatment groups. According, Dr. Haseman was provided vehicle control data (spectrophotometer absorbance values) from 11 different experiments with the same vehicle in order to establish the variability among these animals. Within each experiment, there were four animals and three replicates per animal. For each animal, the three replicates were averaged, and then the four individual animal means were averaged ("a mean of the means") to obtain overall control means and standard deviations for that experiment. The data were also log transformed and the transformed data were averaged. The summary statistics are given in **Table 1-1**.

Table 1-1 Summary of the Control Absorbance Data for the LLNA: BrdU-ELISA

Experiment	Original Scale		Log Scale	
	Mean	SD	Mean	SD
1	0.0676	0.0051	-2.70	0.077
2	0.1197	0.024	-2.14	0.221
3	0.1068	0.0425	-2.29	0.367
4	0.0982	0.0216	-2.34	0.212
5	0.0696	0.0275	-2.73	0.410
6	0.0766	0.0329	-2.64	0.457
7	0.0687	0.0062	-2.68	0.092
8	0.4833	0.0681	-0.74	0.151
9	0.4516	0.110	-0.82	0.249
10	0.2479	0.1425	-1.52	0.590
11	0.2252	0.1044	-1.58	0.491

Several comments on the data:

- Note that there is considerable study-to-study variability. For example, note that if Experiments 8 and 9 were actually a “treatment”, then it would be declared active relative to most if not all of the first 7 control groups (treated/control ratio >3).

- There is much less within-study variability. Note also that on the original scale, the SD tends to increase with increasing means. This suggests that a log transformation will help to stabilize the variability, which in fact was the case.
- Another important advantage of taking logs is that the apparent variable of interest is the ratio of the treated to control response. Testing the null hypothesis that this ratio is one is equivalent to testing the null hypothesis that the difference in the logs is zero, which was the test chosen for the power calculations.

The first step in the power calculation was to use the data from the 11 experiments to derive a representative mean and SD for the control response. Although alternative approaches are certainly possible, only the mean mean and mean SD were calculated for simplicity (on the log scale). These were mean=-2.02 and SD=0.302. The corresponding control mean on the original scale is 0.133.

Three hypothetical changes to the decision criteria when then evaluated: a tripling of the control response (on the original scale), a doubling of the control response, and a 1.3-fold increase in the control response. Although more elegant tests may be possible, I chose to base my power calculations on a simple one-sided Student's t test applied to the log-transformed data. The calculations that are given below assume the same design that was used in the 11 experiments (i.e., three replicates per animal). I focused on an N of 4, but also looked at other sample sizes as well. The results are summarized in **Table 1-2** assuming a control response of -2.02 (log scale) and an SD of 0.302.

Table 1-2 Treatment Group (Rx) Response Relative to Controls

Parameter	3-fold Increase	2-fold Increase	1.3-fold Increase
Mean Rx response	0.399	0.266	0.173
Log (mean Rx response)	-0.92	-1.32	-1.76
Difference from control (log scale)	1.10	0.70	0.26
Difference/SD	3.64	2.32	0.88
Power for N=4	99%	80-90%	<50%
Other power	95% (N=3)	95% (N=5)	50% (N=8)
Other power		50-80% (N=3)	80% (N=16)
Other power			90% (N=22)

Therefore, four animals per group with three replicates per animal is sufficient to detect a three-fold increase in the control response and would likely (with reasonable power) detect a two-fold increase (an additional animal would give 95% power; N=3 would be more problematic). However, it would not be realistic to expect to detect a 1.3 fold increase in the control response without a significant addition of animals. Slight changes in the underlying assumptions would not change the results of these power calculations in any meaningful way.

2.0 LLNA: BrdU-FC

This set of power calculations is based on vehicle control data (flow cytometry BrdU absorbance values) 64 experiments with four to five animals per experiment. Separate power calculations were carried out for five different vehicles. There were four additional experiments with other vehicles, (acetone, PEG 400 and 1% L92/dH20) but since these vehicles involved only one or two studies, there was insufficient data to carry out a meaningful power calculation. The data are summarized in **Tables 2-1** to **2-6**.

Table 2-1 Summary of the Control Data for the LLNA: BrdU-FC (DAE Vehicle)

Experiment	Original scale		Log scale	
	Mean	SD	Mean	SD
1	11564.6	7776.85	8.9124	1.3722
2	7420.2	2387.47	8.8702	0.3228
3	4949.4	2273.08	8.4040	0.5330
4	8169.4	3838.27	8.8964	0.5612
5	18143.0	5594.13	9.7644	0.3316
6	7860.6	6780.59	8.6538	0.9457
7	11551.2	4883.84	9.2772	0.4474
8	7524.6	5591.07	8.7500	0.6241
9	17610.8	14954.73	9.5542	0.6937
10	22822.4	11361.37	9.9076	0.6001
11	3759.25	2862.25	7.9983	0.8003
12	14580.2	5268.96	9.5270	0.4045

Table 2-2 Summary of the Control Data for the LLNA: BrdU-FC (AOO Vehicle)

Experiment	Original scale		Log scale	
	Mean	SD	Mean	SD
1	2328.4	1566.27	7.5122	0.8425
2	17079.0	9402.10	9.6138	0.5903
3	11277.6	6872.04	9.1858	0.5980
4	17932.8	14014.27	9.3336	1.2341
5	8187.6	4714.16	8.8978	0.5121
6	34472.5	10504.11	10.4082	0.3370
7	14813.0	5897.59	9.5208	0.4876
8	14020.8	9854.00	9.2056	1.0883
9	19897.2	11461.51	9.7562	0.6043
10	17975.8	3813.69	9.7756	0.2400
11	6631.8	5725.49	8.4558	0.9473
12	15472.2	8093.26	9.5202	0.5829
13	8749.4	5702.84	8.8432	0.8431
14	11794.6	2858.56	9.3484	0.2688
15	20898.6	10979.71	9.7754	0.7342
16	10648.0	1927.73	9.2612	0.1749
17	16180.0	7711.57	9.5848	0.5393
18	6204.6	3877.74	8.5434	0.7277
19	9628.8	5075.28	9.0446	0.5858
20	7637.6	4022.84	8.8060	0.6072

Table 2-3 Summary of the Control Data for the LLNA: BrdU-FC (DMSO Vehicle)

Experiment	Original scale		Log scale	
	Mean	SD	Mean	SD
1	11892.8	4239.52	9.3338	0.3499
2	17427.0	7999.14	9.6654	0.5283
3	8148.75	3707.66	8.9220	0.4842
4	8031.4	1939.59	8.9676	0.2428
5	40758.25	12831.56	10.5765	0.3238
6	28371.8	14171.47	10.1586	0.4781
7	46420.8	18065.75	10.6844	0.3918
8	24726.0	5326.84	10.0974	0.2146
9	14027.4	3476.44	9.5208	0.2729
10	15314.5	9320.34	9.5210	0.5276
11	13386.0	5516.88	9.4284	0.4399
12	24955.6	9786.46	10.0250	0.5643
13	19335.2	7644.20	9.8158	0.3544
14	41366.4	14242.19	10.5892	0.3088
15	26519.8	10408.41	10.1218	0.4048
16	52644.0	17384.31	10.8276	0.3306
17	21824.8	9779.87	9.9156	0.4243
18	21865.4	9182.90	9.8892	0.5617
19	29371.2	6978.60	10.2632	0.2539
20	22575.4	9225.93	9.9564	0.4170
21	11929.2	6187.36	9.2744	0.5411
22	22382.6	8667.60	9.9672	0.3325
23	22221.0	15029.10	10.1200	0.4161
24	17486.2	4157.51	9.7444	0.2531

Table 2-4 Summary of the Control Data for the LLNA: BrdU-FC (DMF Vehicle)

Experiment	Original scale		Log scale	
	Mean	SD	Mean	SD
1	5728.8	3829.90	8.3252	1.1704
2	16018.4	4502.49	9.6438	0.1034
3	11607.4	9643.83	9.0312	0.8762
4	35928.2	25375.35	10.2938	0.4949

Table 2-5 Summary of the Control Data for the LLNA: BrdU-FC (ETOH Vehicle)

Experiment	Original scale		Log scale	
	Mean	SD	Mean	SD
1	4096.2	2343.60	8.2070	0.5064
2	6636.5	4310.69	8.6040	0.7779
3	18806.4	5220.25	9.8122	0.2697
4	6920.4	3307.72	8.6970	0.6828

Table 2-6 Average Means and Standard Deviations for Each Vehicle

Vehicle	N	Original Scale Averages		Log-transformed Scale Averages		Converted Control Mean[1]	Maximum Difference[2]
		Mean	SD	Mean	SD		
DAE 433	12	11329.6	6131.05	9.043	.6364	8459	29-fold
AOO	20	13591.5	6703.74	9.220	.6273	10093	15-fold
DMSO	24	23457.6	8969.54	9.891	.3924	19753	4-fold
DMF	4	17320.7	10837.89	9.324	.6612	11198	14-fold
EtOH	4	9114.9	3795.57	8.830	.5592	6837	7-fold

[1] Anti-log of the log transformed scale average (used in the power calculations).
[2] Maximum difference among animals within an experiment using this vehicle.

Note the large SD for every group except for the DMSO control. The power calculations given in **Tables 2-7** to **2-12** are based on a one-sided p<0.05 Student's t test applied to the log-transformed data (just as in the previous power calculations; for completeness, the power calculations are included for the acetone vehicle as well, although only two experiments used this vehicle, as noted above). ***It should be noted that these calculations make the additional assumption that any "treatment effect" produced will have essentially the same SD (on a log-transformed scale) as the control data, i.e., that the treatment will change only the mean response and not the variability.***

Table 2-7 Treatment Group (Rx) Response Increase Relative to Controls for DAE 433

	3.0-fold increase	2.5-fold increase	2.0-fold increase	1.5-fold increase	1.3-fold increase
Mean Rx response	25377	21147.5	16918	12688.5	10996.7
Log (Mean Rx response)	10.142	9.959	9.736	9.448	9.305
Difference (log scale)	1.099	0.916	0.693	0.405	0.262
Difference/SD	1.73	1.44	1.09	0.64	0.41
Power for N=5	nearly 80%	50-80%	<50%	<50%	<50%
Power for N=4	50-80%	50%	<50%	<50%	<50%
Power for N=3	50%	<50%	<50%	<50%	<50%
Other Power	95% (N=9)	95% (N=12)	95%(N=19)	95% (N=54)	95% (N>100)
Other Power	90% (N=7)	90% (N=10)	90% (N=15)	90% (N=43)	90% (N>100)

Table 2-8 Treatment Group (Rx) Response Increase Relative to Controls for AOO

	3.0-fold increase	2.5-fold increase	2.0-fold increase	1.5-fold increase	1.3-fold increase
Mean Rx response	30279	25232.5	20186	15139.5	13120.9
Log (Mean Rx response)	10.318	10.136	9.913	9.625	9.482
Difference (log scale)	1.098	0.916	0.693	0.405	0.262
Difference/SD	1.75	1.46	1.10	0.65	0.42
Power for N=5	80%	50-80%	<50%	<50%	<50%
Power for N=4	50-80%	50%	<50%	<50%	<50%
Power for N=3	50%	<50%	<50%	<50%	<50%
Other Power	95% (N=9)	95% (N=12)	95%(N=19)	95% (N=52)	95% (N>100)
Other Power	90% (N=7)	90% (N=10)	90% (N=15)	90% (N=42)	90% (N>100)

Table 2-9 Treatment Group (Rx) Response Increase Relative to Controls for DMF

	3.0-fold increase	**2.5-fold increase**	**2.0-fold increase**	**1.5-fold increase**	**1.3-fold increase**
Mean Rx response	33594	27995	22396	16797	14557.4
Log (Mean Rx response)	10.422	10.240	10.017	9.729	9.586
Difference (log scale)	1.098	0.916	0.693	0.405	0.262
Difference/SD	1.66	1.39	1.05	0.61	0.40
Power for N=5	50-80%	50-80%	<50%	<50%	<50%
Power for N=4	50-80%	50%	<50%	<50%	<50%
Power for N=3	50%	<50%	<50%	<50%	<50%
Other Power	95% (N=10)	95% (N=12)	95%(N=21)	95% (N=63)	95% (N>100)
Other Power	90% (N=8)	90% (N=10)	90% (N=17)	90% (N=48)	90% (N>100)

Table 2-10 Treatment Group (Rx) Response Increase Relative to Controls for ETOH

	3.0-fold increase	**2.5-fold increase**	**2.0-fold increase**	**1.5-fold increase**	**1.3-fold increase**
Mean Rx response	20511	17092.5	13674	10255.5	8888.1
Log (Mean Rx response)	9.929	9.746	9.523	9.236	9.092
Difference (log scale)	1.099	0.916	0.693	0.406	0.262
Difference/SD	1.97	1.64	1.24	0.73	0.47
Power for N=5	80-90%	50-80%	50%	<50%	<50%
Power for N=4	80%	50-80%	<50%	<50%	<50%
Power for N=3	50-80%	50%	<50%	<50%	<50%
Other Power	95% (N=7)	95% (N=10)	95%(N=15)	95% (N=42)	95% (N=100)
Other Power	90% (N=6)	90% (N=8)	90% (N=12)	90% (N=33)	90% (N=80)

Table 2-11 Treatment Group (Rx) Response Increase Relative to Controls for DMSO

	3.0-fold increase	**2.5-fold increase**	**2.0-fold increase**	**1.5-fold increase**	**1.3-fold increase**
Mean Rx response	59259	49382.5	39506	29629.5	25678.9
Log (Mean Rx response)	10.990	10.807	10.584	10.297	10.153
Difference (log scale)	1.099	0.916	0.693	0.406	0.262
Difference/SD	2.80	2.33	1.77	1.03	0.67
Power for N=5	95-99%	95%	80%	<50%	<50%
Power for N=4	90-95%	80-90%	50-80%	<50%	<50%
Power for N=3	80-90%	50%	50-80%	<50%	<50%
Other Power			95%(N=8)	95% (N=22)	95% (N=49)
Other Power			90% (N=7)	90% (N=17)	90% (N=39)

Table 2-12 Treatment Group (Rx) Response Increase Relative to Controls for ACE

	3.0-fold increase	2.5-fold increase	2.0-fold increase	1.5-fold increase	1.3-fold increase
Mean Rx response	25881	21567.5	17254	12940.5	11215.1
Log (Mean Rx response)	10.161	9.979	9.756	9.468	9.325
Difference (log scale)	1.098	0.916	0.693	0.405	0.262
Difference/SD	1.70	1.42	1.07	0.63	0.41
Power for N=5	50-80%	50-80%	<50%	<50%	<50%
Power for N=4	50-80%	50%	<50%	<50%	<50%
Power for N=3	50%	<50%	<50%	<50%	<50%
Other Power	95% (N=9)	95% (N=12)	95%(N=20)	95% (N=56)	95% (N>100)
Other Power	90% (N=7)	90% (N=10)	90% (N=16)	90% (N=45)	90% (N>100)

It is important to understand that the primary factor that influences power (in addition to sample size) is the variability in response among control animals in a given study: the greater the variability, the lower the power. Using this assay, for four of the five vehicles, the variability among animals is so great, that it is unlikely that even a 3-fold increase in response will be detected statistically, with 3-5 animals per group. For example, if the controls show a range of variability similar to that seen in the first DAE 433 study, ranging from 694 to 20171, a 29-fold difference, how realistic would it be to expect to detect a much smaller (3-fold) increase in the treated group response relative to the response seen in that control group?

Thus, based on these data, the only way to assure decent power for this assay is to use DMSO as the vehicle. If this vehicle is used, there is an excellent chance of detecting a 2.5-fold or a 3-fold increase in response if 4 or 5 animals per group are used. Another advantage of using DMSO is that the variability within a study among control animals is very reproducible, and thus predictable. In 24 studies using DMSO as the vehicle, the within study variability among animals never exceeded a 4-fold difference. DMSO does not show the wild fluctuations seen for the other vehicles in which one experiment can show a 29-fold variation among control animals and the next experiment show only a 2-fold variation.

It should be noted that the mean control response using the DMSO vehicle is much greater than the mean control response using the other vehicles, so a 3-fold increase relative to the DMSO control reflects a much larger actual dosed group response than a 3-fold increase relative to a smaller control response. For example, the mean DMSO control response is almost a 3-fold increase relative to the mean EtOH control response, so a 3-fold increase relative to a DMSO control group would be almost a 9-fold increase relative to the EtOH control group.

Finally, regardless of vehicle, it is unlikely that the assay can detect statistically a 2-fold or less increase in response, with only 3-5 animals per group.

3.0 LLNA: DA

This analysis was based on vehicle control data (ATP levels) from 18 different experiments. Within each experiment, there were three or four animals and two replicates per animal. For each animal, the two replicates were averaged, and then the individual animal means were averaged to obtain overall control means and SD's for that experiment. The data were also log-transformed data and then averaged. The summary statistics are given in **Table 3-1**, and **Table 3-2** summarizes the average means and standard deviations for each vehicle.

Table 3-1 Summary of the Control Absorbance Data for the LLNA: DA

Experiment	Original scale		Log scale		Vehicle
	Mean	SD	Mean	SD	
1	4410	752.6	8.381	0.1801	AOO
2	3871	343.8	8.258	0.0890	AOO
3a	3014	435.9	8.003	0.1475	AOO
3b	6674	1526.6	8.785	0.2384	DMSO
4a	2580	517.7	7.838	0.2229	AOO
4b	3465	888.9	8.124	0.2737	DMF
5	5168	4579.3	8.260	0.8812	AOO
6	3528	1880.8	8.040	0.6654	AOO
7	1509	455.0	7.275	0.3666	AOO
8	2668	1019.7	7.835	0.3804	DMF
9	2077	95.0	7.638	0.0452	AOO
10	3129	848.7	8.023	0.2537	AOO
11	2818	567.4	7.928	0.2010	AOO
12	2151	376.9	7.662	0.1740	AOO
13	1611	423.7	7.362	0.2419	ACE
14	3362	736.3	8.103	0.2083	AOO
15a	10204	2765.9	9.203	0.2727	DMSO
15b	4907	656.4	8.491	0.1422	AOO
16	2710	822.5	7.875	0.2716	ACE
17	64899	18696.8	11.047	0.3063	DMSO
18	2894	954.5	7.932	0.3165	AOO

Table 3-2 Average Means and Standard Deviations for Each Vehicle

Vehicle	N	Original Scale Averages		Log-transformed Averages		Converted Control Mean[1]
		Mean	SD	Mean	SD	
AOO	14	3244	942.9	7.988	0.2781	2945
DMSO	3	27259	7663.1	9.678	0.2725	15968
ACE	2	2160	623.1	7.619	0.2568	2036
DMF	2	3066	954.3	7.980	0.3271	2920

[1] Used in power calculations and based on log-transformed scale average.

Clearly, the DMSO vehicle produces responses totally inconsistent with the other three vehicles (which are reasonably similar among themselves). The power calculations given in **Tables 3-3** to **3-6** are based on a one-sided p<0.05 Student's t test applied to the log transformed data (just as in the previous power calculations). ***It should be noted that these calculations make the additional assumption that any "treatment effect"***

produced will have essentially the same SD (on a log-transformed scale) as the control data, i.e., that the treatment will change only the mean response and not the variability. The data in the table above are consistent with this assumption, since although the mean response for the DMSO vehicle is a sizable increase over the mean response for the other controls, the underlying variability (on a log scale) is very similar. The power calculations are summarized below by vehicle.

Table 3-3 Treatment Group (Rx) Response Increase Relative to Controls for AOO

Parameter	3-fold Increase	2.5-fold Increase	2.0-fold Increase	1.5-fold Increase	1.3-fold Increase
Mean Rx response	8835	7362.5	5890	4417.5	3828.5
Log (mean Rx response)	9.086	8.904	8.681	8.393	8.250
Difference from control (log scale)	1.098	0.916	0.693	0.405	0.262
SD of the difference from control	3.95	3.29	2.49	1.46	0.94
Power for N=5	99%	99%	95%	50-80%	<50%
Power for N=4	99%	95-99%	90%	50%	<50%
Power for N=3	95%	90-95%	80%	<50%	<50%
Other power				95% (N=11)	95% (N=25)
Other power				90% (N=9)	90% (N=20)

Table 3-4 Treatment Group (Rx) Response Increase Relative to Controls for ACE

Parameter	3-fold Increase	2.5-fold Increase	2.0-fold Increase	1.5-fold Increase	1.3-fold Increase
Mean Rx response	6108	5090	4072	3054	2646.8
Log (mean Rx response)	8.717	8.535	8.312	8.024	7.881
Difference from control (log scale)	1.098	0.916	0.693	0.405	0.262
SD of the difference from control	4.28	3.57	2.70	1.58	1.02
Power for N=5	99%	99%	95-99%	50-80%	<50%
Power for N=4	99%	99%	90-95%	50%	<50%
Power for N=3	99%	95%	80-90%	<50%	<50%
Other power				95% (N=10)	95% (N=23)
Other power				90% (N=8)	90% (N=18)

Table 3-5 Treatment Group (Rx) Response Increase Relative to Controls for DMF

Parameter	3-fold Increase	2.5-fold Increase	2.0-fold Increase	1.5-fold Increase	1.3-fold Increase
Mean Rx response	8760	7300	5840	4380	3796
Log (mean Rx response)	9.078	8.896	8.672	8.385	8.242
Difference from control (log scale)	1.098	0.916	0.692	0.405	0.262
SD of the difference from control	3.36	2.80	2.12	1.24	0.80
Power for N=5	99%	95-99%	90%	50%	<50%
Power for N=4	95-99%	90-95%	80%	<50%	<50%
Power for N=3	90-95%	80-90%	50%	<50%	<50%
Other power				95% (N=15)	95% (N=35)
Other power				90% (N=12)	90% (N=28)

Table 3-6 Treatment Group (Rx) Response Increase Relative to Controls for DMSO

Parameter	3-fold Increase	2.5-fold Increase	2.0-fold Increase	1.5-fold Increase	1.3-fold Increase
Mean Rx response	47904	39920	31936	23952	20758.4
Log (mean Rx response)	10.777	10.595	10.371	10.084	9.941
Difference from control (log scale)	1.099	0.917	0.693	0.406	0.263
SD of the difference from control	4.03	3.37	2.54	1.49	0.97
Power for N=5	99%	99%	95%	50-80%	<50%
Power for N=4	99%	95-99%	90%	50%	<50%
Power for N=3	95%	90-95%	80%	<50%	<50%
Other power				95% (N=11)	95% (N=24)
Other power				90% (N=9)	90% (N=19)

Therefore, using three to five animals per group (and two replicates per animal), there is a very high probability that a 2.5-fold and a 3-fold increase will be detected and a good chance that a 2-fold increase will be detected, regardless of vehicle. However, detecting a 1.3 to 1.5-fold increase may be too much to expect with only three to five animals per group.

Note that all four vehicles produce similar power profiles. This is because the transformed SDs in the table above are all very similar; if they were identical, so would be the power profiles. However, the actual magnitudes of the treated group responses for a given power will differ from vehicle to vehicle because the control responses themselves differ (especially for DMSO). For example, a 3-fold increase in the control response for the ACE vehicle would be an increase from 2036 to 6108, and would be detected with approximately a 99% probability. However, a 6108 treatment response relative to the AOO vehicle would only be approximately a 2-fold increase and would be detected with only a 95% probability. A treatment response of 6108 for the DMSO vehicle would actually be far below the DMSO control response.

Finally, Experiments 5-8 produced notably more variability (among and within animals) than the other experiments. I cannot help but wonder if these four studies were done at a different lab than the others. If so, then the power specific to that lab would be notably lower than that currently reported, while the power associated with the other experiments would be increased slightly if the four experiments were excluded.

4.0 Traditional LLNA

These control data come from three different labs, but the same vehicle was used. The raw data are decays per minute (dpm) from a scintillation counter. Within each experiment, there were five animals and one replicate per animal. For each animal, the five animals were averaged to obtain overall control means and SD's for that experiment. The log-transformed data were also averaged. The summary statistics are given in **Table 4-1**.

Table 4-1 Summary of the Control DPM Data for the LLNA

Experiment	Original scale		Log scale		Range of responses	Lab
	Mean	SD	Mean	SD		
1A	443.4	233.86	5.976	0.5531		1
1B	410.2	100.30	5.994	0.2421		1
1C	462.2	172.26	6.078	0.3874		1
1D	397.8	92.64	5.968	0.2092		1
1E	466.8	154.26	6.104	0.3262		1
1F	352.6	118.53	5.826	0.3211		1
1G	333.0	167.74	5.702	0.5336		1
2A	487.8	164.01	6.142	0.3649		2
2B	729.2	314.07	6.496	0.5214		2
2C	586.6	279.96	6.296	0.4252		2
2D	618.4	103.27	6.416	0.1644		2
2E	487.4	80.26	6.178	0.1585		2
2F	304.1	208.62	5.402	0.9937		2
2G	309.4	110.19	5.686	0.3512		2
3A	330.5	145.26	5.706	0.5184	137.67 to 515.98	3
3B	288.5	229.15	5.338	1.0113	42.13 to 654.45	3
3C	152.5	31.78	5.008	0.2275	103.56 to 189.17	3
3D	296.2	126.07	5.604	0.4820	131.13 to 447.97	3
3E	215.3	149.44	5.104	0.9148	38.62 to 437.46	3

Power calculations were carried out for each lab separately and for all labs combined. The summary statistics are given below.

Table 4-2 Average Means and Standard Deviations for Each Vehicle

Lab	N	Original Scale Averages		Log-Transformed Scale Averages		Converted Control Mean[1]
		Mean	SD	Mean	SD	
1	7	409.4	148.51	5.950	0.3675	383.8
2	7	503.3	180.05	6.088	0 .4256	440.5
3	5	256.6	136.34	5.352	0.6308	211.0
All 3	19	403.8	156.93	5.843	0.4582	344.8

[1] Used in power calculations and based on log-transformed scale average.

The power calculations given in **Tables 4-3** to **4-6** are based on a one-sided p<0.05 Student's t test applied to the log-transformed data (just as in the previous power calculations). ***It should be noted that these calculations make the additional assumption that any "treatment effect" produced would have essentially the same SD (on a log-transformed scale) as the control data (i.e. that the treatment will change only the mean response and not the variability)***.

Table 4-3 Treated Group (Rx) Response Increase Relative to Controls: Lab 1

	3.0-fold increase	2.5-fold increase	2.0-fold increase	1.5-fold increase	1.3-fold increase
Mean Rx response	1151.4	959.5	767.6	575.7	498.94
Log (Mean Rx response)	7.049	6.866	6.643	6.356	6.212
Difference (log scale)	1.099	0.916	0.693	0.406	0.262
Difference/SD	2.99	2.49	1.89	1.10	0.71
Power for N=5	99%	95%	80%	<50%	<50%
Power for N=4	95%	90%	50-80%	<50%	<50%
Power for N=3	90%	80%	50-80%	<50%	<50%
Other Power				95% (N=19)	95% (N=45)
Other Power				90% (N=15)	90% (N=36)

Table 4-4 Treated Group (Rx) Response Increase Relative to Controls: Lab 2

	3.0-fold increase	2.5-fold increase	2.0-fold increase	1.5-fold increase	1.3-fold increase
Mean Rx response	1321.5	1101.25	881.0	660.75	572.65
Log (Mean Rx response)	7.187	7.004	6.781	6.493	6.350
Difference (log scale)	1.099	0.916	0.693	0.405	0.262
Difference/SD	2.58	2.15	1.63	0.95	0.62
Power for N=5	95%	90%	50-80%	<50%	<50%
Power for N=4	90%	80%	50%	<50%	<50%
Power for N=3	80%	50-80%	<50%	<50%	<50%
Other Power				95% (N=25)	95% (N=57)
Other Power				90% (N=20)	90% (N=46)

Table 4-5 Treated Group (Rx) Response Increase Relative to Controls: Lab 3

	3.0-fold increase	2.5-fold increase	2.0-fold increase	1.5-fold increase	1.3-fold increase
Mean Rx response	633.0	527.5	422.0	316.5	274.3
Log (Mean Rx response)	6.450	6.268	6.045	5.757	5.614
Difference (log scale)	1.098	0.916	0.693	0.405	0.262
Difference/SD	1.74	1.45	1.10	0.64	0.42
Power for N=5	80%	50-80%	<50%	<50%	<50%
Power for N=4	50-80%	50%	<50%	<50%	<50%
Power for N=3	50%	<50%	<50%	<50%	<50%
Other Power			95% (N=19)	95% (N=53)	95% (N>100)
Other Power			90% (N=15)	90% (N=43)	90% (N=100)

Table 4-6 Treated Group (Rx) Response Increase Relative to Controls: Combined Labs 1, 2, and 3

	3.0-fold increase	**2.5-fold increase**	**2.0-fold increase**	**1.5-fold increase**	**1.3-fold increase**
Mean Rx response	1034.4	862.0	689.6	517.2	448.24
Log (Mean Rx response)	6.942	6.759	6.536	6.248	6.105
Difference (log scale)	1.099	0.916	0.693	0.405	0.262
Difference/SD	2.40	2.00	1.51	0.88	0.57
Power for N=5	95%	80-90%	50-80%	<50%	<50%
Power for N=4	90%	80%	50%	<50%	<50%
Power for N=3	50-80%	50-80%	<50%	<50%	<50%
Other Power			95% (N=11)	95% (N=29)	95% (N=68)
Other Power			90% (N=9)	90% (N=23)	90% (N=54)

These data show considerable variability, and thus, the power is relatively low. Labs 1 and 2 have a reasonably good (but not great) chance of detecting 3 and 2.5-fold increases if N=4 or N=5 are used. Lesser increases will likely be missed. N=3 also appears to be inadequate.

Lab 3 will likely be unable to detect any increase of 3-fold or less, even with N=5. The best case is a power of approximately 80% for detecting a 3-fold increase with N=5. The reason for the low power is the high within-study variability. For example, 2 of the 5 experiments at this lab had 11-fold and 15-fold differences among the control responses. If the control responses can differ by a factor of 15, how reasonable is it to expect to detect a 3-fold increase in a treatment group with only 3-5 animals? Because of Lab 3's poor performance, the "all labs combined" performance suffers as well (**Table 4-6**).

The power calculations presented above assume that the data will be subjected to some formal statistical test at a pre-specified level of significance (e.g., p<0.05). However, it is also possible for an interpretative strategy to adopt a strict decision rule, such as the following, which I will refer to in this report as the "Ratio Rule":

"Declare the result positive if the ratio of mean treated response to mean control response is greater than 3; otherwise, declare the response negative".

One advantage of the Ratio Rule is that it is easy to understand and to apply and requires no statistical test, simply a calculation of means and a ratio. One disadvantage of the Ratio Rule is that the false positive rate (i.e., the "p value" associated with this decision making strategy) is unknown and will vary from assay to assay, depending upon the underlying variability among animals. The associated power is also unknown.

To investigate this matter further, I looked at the ELISA data again, searching for an example showing approximately 95% power based on a Student's t test, so that I could investigate whether this power could be increased or decreased by application of the Ratio Rule. For the ELISA data, the N=3 case had approximately a 95% power associated with a one-sided Student's t test for detecting a 3-fold increase in response (see **Table 1-2**). To compare this power with the "Ratio Rule", I made the following assumptions/calculations.

I assumed that the mean logged response for ELISA was -2.02 and the mean SD response was 0.302 (as before). The standard error (SE) associated with N=3 is simply the standard deviation divided by the square root of 3 or 0.1744. This SE is the SD we would expect to see among (logged) mean responses based on N=3.

I then enumerated (using the cumulative normal probability distribution at probability intervals of 0.02) the approximate distribution of mean log responses consistent with an SD of 0.1744 and a 3-fold increase in the ratio. That is, I approximated the continuous distribution of both the treated and control responses by a discrete distribution of 50 mean responses, spaced so that each outcome has approximately a 2% probability of occurrence. These two distributions are given below. If you calculate the summary statistics, you will find that the mean of the logged control response is -2.02 and the mean of the logged treated group response is -0.92 (a 3-fold increase on the original scale); both have a SD of 0.1744. Importantly, these are expected mean responses for a group of 3 animals, not individual animal responses, so the range of responses is relatively narrow. These distributions formed the basis of the new power calculations.

Control Mean		Treated Mean		Contribution to power
Logged	Response	Logged	Response	(Control) for detecting
Response		Response		a 3-fold increase
-1.617	.198	-0.517	.596	.02
-1.69	.185	-0.59	.554	.02
-1.73	.177	-0.63	.533	.06
-1.76	.172	-0.66	.517	.08
-1.79	.167	-0.69	.502	.10
-1.81	.164	-0.71	.492	.10
-1.82	.162	-0.72	.487	.14
-1.84	.159	-0.74	.477	.14
-1.85	.157	-0.75	.472	.18
-1.87	.154	-0.77	.463	.20
-1.88	.153	-0.78	.458	.20
-1.89	.151	-0.79	.454	.24
-1.90	.150	-0.80	.449	.24
-1.91	.148	-0.81	.445	.28
-1.92	.147	-0.82	.440	.28
-1.93	.145	-0.83	.436	.32
-1.94	.144	-0.84	.432	.32
-1.95	.142	-0.85	.427	.36
-1.96	.141	-0.86	.423	.36

-1.97	.139	-0.87	.419	.40
-1.98	.138	-0.88	.415	.42
-1.99	.137	-0.89	.411	.42
-2.00	.135	-0.90	.407	.46
-2.01	.134	-0.91	.403	.48
-2.016	.133	-0.916	.400	.50
-2.024	.132	-0.924	.397	.52
-2.03	.131	-0.93	.395	.54
-2.04	.130	-0.94	.391	.56
-2.05	.129	-0.95	.387	.56
-2.06	.127	-0.96	.383	.60
-2.07	.126	-0.97	.379	.62
-2.08	.125	-0.98	.375	.62
-2.09	.124	-0.99	.372	.64
-2.10	.122	-1.00	.368	.68
-2.11	.121	-1.01	.364	.70
-2.12	.120	-1.02	.361	.72
-2.13	.119	-1.03	.357	.72
-2.14	.118	-1.04	.353	.74
-2.15	.116	-1.05	.350	.78
-2.16	.115	-1.06	.346	.80
-2.17	.114	-1.07	.343	.82
-2.19	.112	-1.09	.336	.82
-2.20	.111	-1.10	.333	.84
-2.22	.109	-1.12	.326	.86
-2.23	.108	-1.13	.323	.88
-2.25	.105	-1.15	.317	.92
-2.28	.102	-1.18	.307	.94
-2.31	.099	-1.21	.298	.96
-2.35	.095	-1.25	.287	.98
-2.423	.089	-1.323	.266	.98

Total=25.12
Power = 0.02 x Total or 50.24%

Thus, the power is reduced from 95% to 50% by using the Ratio Rule rather than a one-sided p<0.05 Student's t test, although the "gain" is that the false positive rate is reduced from 5% to essentially zero (note from the distributions given above that the overall range of mean control responses is less than 3-fold, so the false positive rate is essentially zero). This "tradeoff" is typical, even for a formal statistical test. What is needed is a reasonable balance between false positive and false negative rates, and the Ratio Rule seems designed to sacrifice power for the sake of maintaining a low false positive rate.

One way to modify the Ratio Rule to increase its power would be change the critical value of the ratio from 3 to some smaller number such as 2 or 2.5. This would increase power while still keeping the false positive rate low.

For example, by my calculations, if the Ratio Rule applied to the distribution data above was changed from "Ratio > 3" to "Ratio > 2", then the power would be approximately 95%, but the false positive rate would still be low (approximately 0.002).

The 50% power found by enumerating the entire distribution for the example above simply confirms what should be intuitive for the Ratio Rule, namely, that if you have two distributions for which the underlying means differ by a factor of 3, then approximately half the time the ratio of means from sampled data will exceed 3 and approximately half the time it will be less than 3. So it is unnecessary to perform additional power calculations for the Ratio Rule, at least for detecting an underlying 3-fold increase in response. Regardless of the underlying SD (and for that matter, regardless of the number of animals used), the power of the Ratio Rule for detecting a 3-fold increase in response will always be approximately 50%. Of course, if the underlying ratio is greater than 3, then the sample size and underlying variability do become important in the Power Calculations for the Ratio Rule.

The power of Student's t test depends upon the sample size and the underlying variability, but for the various cases considered (see tables above), the power was always well above 50%.

My conclusion is that the "Ratio Rule" has a much lower false positive rate than a formal statistical test, but it also has a much higher false negative rate (i.e., lower power). This reduced power can be considerable, and the Ratio Rule will always show approximately 50% power for detecting an underlying treatment effect that on average shows a 3-fold increase relative to controls. Moreover, the power of the Ratio Rule is less than 50% for detecting increases in the ratio of 2.5, 2, 1.5, or 1.3, but use of the Ratio Rule implies that such increases are likely not biologically important anyway, as discussed in more detail below.

The ultimate objective of a decision strategy is to maximize the ability of an assay's outcome to predict correctly the human response (positive or negative), and to achieve this objective, a formal statistical test may or may not be necessary. It is my understanding that the "Ratio Rule" was not established arbitrarily, but rather was derived empirically, on the basis that 3 was the "cut-off ratio value" that provided the optimal performance of the assay when differentiating "true" human positives from "true" human negatives for one of the assays. It is also my understanding that this Ratio Rule has not been "validated" empirically for all of the various assays to which it is now being routinely applied (ELISA, traditional, DA, FC, etc.).

If the Ratio Rule seems to "work" very well in practice in predicting the human response, that is the ultimate goal, so there may be no need of a formal statistical test, as long as everyone fully understands what the use of such a rule implies. Since, based on the control data provided to me, a false positive outcome is nearly impossible (or at least has a very low probability) using the Ratio Rule, use of this rule implicitly assumes that are some, perhaps even many, compounds that are "true positives" in the assay, but the response that they produce (e.g., a 2-fold or 2.5-fold increase in the treated/control ratio), while detectable statistically, should be considered a negative response, since it is of insufficient magnitude for the compound in question to be positive in humans. Even a 3-fold increase in the ratio of treated to control mean response is considered relatively unimportant, since it will be detected only approximately 50% of the time by the Ratio Rule. Is such a performance acceptable to the scientific community? Are chemicals that are truly active in the assay, but produce a ratio of <3, generally negative in humans and thus can be discounted when a response of this magnitude is observed in an assay? Use of the Ratio Rule assumes that the answer to this question is Yes.

To summarize, use of the Ratio Rule assumes that there are compounds that are statistically positive in the assay (and are actually producing an effect in the assay), but the magnitude of the effect is insufficient for the compound to be positive in humans. Unless this is known with certainty, I personally prefer using a formal statistical test rather than a strict rule, a rule whose performance characteristics (power, false positive rate) are unknown and vary from assay to assay.

5.0 Final Comments and Summary

(1) One result of the data analyses presented above is that it reinforces the need for concurrent control data. **Table 5-1** below shows the variability observed in the mean control responses across experiments. Only for the traditional assay are the results reasonably reproducible. For the other assays, concurrent controls are clearly essential, since the data are so variable across experiments, and I would recommend that concurrent controls be routinely included in the study design of all assays. Note that in many of the assays, a control response in one experiment would clearly be considered "active" relative to the control response in another experiment, since the ratio is far greater than 3.

Table 5-1: Variability in the mean control response across experiments

	Maximum Difference Among Control Means
ELISA	7-fold difference
FC: DAE Vehicle	6-fold difference
FC: AOO Vehicle	15-fold difference
FC: DMSO Vehicle	7-fold difference
FC: DMF Vehicle	6-fold difference
FC: ETOH Vehicle	5-fold difference
DA: AOO Vehicle	3.4-fold difference
DA: DMSO Vehicle	10-fold difference
Traditional: Lab 1	1.4-fold difference
Traditional: Lab 2	2.4-fold difference
Traditional: Lab 3	2.2-fold difference

(2) A related point is that it is important to have individual animal control data, so that

the within-study among-animal variability can be assessed and factored into the data evaluation. Individual animal data are essential if the data are to be evaluated statistically. Even if the “Ratio Rule” is used, individual animal data are highly desirable.

(3) Since a formal statistical test has so much more power than the Ratio Rule, it is definitely of interest to examine those specific compounds for which there are contradictory results, i.e., a statistically verified treatment effect in the assay, but the Ratio Rule criterion is not met. It is important to determine whether these chemicals are positive or negative in the human setting, to understand if these “statistical positives” in the assay are “false positives” or “true positives” in the human setting.

(4) These analyses also reinforce the importance of the choice of vehicle in certain of the assays. For example, DMSO shows a response that is much higher than that seen with the other vehicles. The variability in response among animals may also be dependent upon the vehicle in some experimental settings. Similarly, the data suggest that some labs are better than others in reproducing the control responses among animals within a given experiment (a further argument for the routine reporting of individual animal control data).

(5) The decision to use 4 or 5 animals depends upon whether or not the gain in power achieved by the extra animal is deemed sufficiently important to justify the extra time, effort and cost. In some cases (e.g., the ELISA assay for detecting a 3-fold increase), the extra animal would add little, since the power for N=4 is already 99%. For other assays that show more variability (see tables above), the extra animal may be more important. It is a judgment call.

Importantly, this comparison of sample size is linked to use of a formal statistical test. If the Ratio Rule is used instead, and power is calculated for a true underlying response ratio of 3, then sample size is irrelevant, since the power will always be approximately 50%, regardless of sample size. Although I have not made sample size comparisons for the power of the Ratio Rule applied when the true underlying ratio exceeds 3, the low power of the Ratio Rule in general suggests the use of as many animals as are feasible, so an N of 5 rather than 4 may be important if the Ratio Rule is used.

(6) Finally, the decision whether to use a formal statistical test or the Ratio Rule is beyond the scope of this evaluation. Since the Ratio Rule has notably less power than a formal statistical test, then the “default” approach in my opinion should be to use a formal statistical analysis, unless it can be demonstrated that the “statistical positives” that are identified in the assay but “missed” by the Ratio Rule are in fact negative in humans. If such compounds are in fact negative in humans, it would indicate that the assay is “overly sensitive” and detects effects that are not relevant to humans, and this needs to be understood by the scientific community.

Joe Haseman
2-14-08