Charles M. Beasley, Jr and Roy Tamura: What We Know and Do Not Know by Conventional Statistical Standards About Whether a Drug Does or Does Not Cause a Specific Side Effect (Adverse Drug Reaction)

4. "Proof" of the presence of an ADR (significant excess compared to control): sample size requirements

As we have said, such "proof"is generally based on an inferential statistical test. If our interest is in proving presence, a conventional inferential test with the null hypothesis of no difference between groups is usedand we conclude that a difference exists between groups if the null hypothesis is rejected at the $\alpha \leq 0.05$ level in a 2-sided test.

We have gone back to Beasley's primary example (incidence of 1 in 1,000 with drug and no occurrence without drug) from his response to Blackwelland computed the sample sizes for 51% power employing PASS 15.0.6 software (Beasley 2018; PASS 2017).We then performed the conventional inferential test (Fisher's Exact, 2-sided) employing NCSS 12.0.5 software (NCSS 2018).[i] The results illustrate the point that one might get lucky and "prove" an ADR with fewer subjects than the number of subjects necessitated by any power ≥51% (see Table 1 below).[ii]

The sample size below for 80% power is somewhat lower than Beasley reported in response to Blackwellbecause for this work we used the binomial enumeration method of computation, rather than a normal approximation method of computation, for sample sizes up to 100,000 (Blackwell 2018). Binomial enumeration computation provides exact results but requires long runtime (some sample size computations required six days performed on an Intel i7-6700K CPU @ 4.00GHz with 32 GB RAM system). As can be seen, ~6,000 (~51% power) subjects per treatment group are sufficient to get a result of nominal statistical significance with perfect sampling, but 5,000 is insufficient when the true incidences are 1 in 1,000 with drug and 0 in an infinity of subjects with placebo.

Table 1: Demonstration of p-Value with Sample Sizes based on Two Prospective Power Requirements with Study Outcome as Prospectively Estimated

Fisher's Exact Test, 2-sided (α=0.05)						
Sample Size Computation (binomial enumeration)			Inferential Test Results with ~51% Power			
Event Incidence (with drug)	Sample Size / Treatment (80% Power)	Sample Size / Treatment (51% Power)	Events with Treatment	Events with Placebo	Sample Sizes Used	p-Value
1:1,000	7,905	5,730	6	0	6,000	0.0312
			5	0	5,000	0.0624

The sample size of 7,905 per treatment group required to obtain 80% power with a 2-sided Fisher's Exact Test is lower than the sample size of 9,742 previously reported by Beasley in his response to Blackwell (2018). However, a sample size of 7,905 per treatment group is still a large sample size and a practical impossibility in RCTs evaluating psychiatric medications. This sample size with placebo, not added on to another treatment, as control would be particularly difficult.

The discussion of the sample size resulting in 51% power and how that sample size is adequate to achieve conventional statistical significance with precise estimation of what will be observed in an RCT and the sample sizes computed with binomial enumeration offer full transparency building on Beasley's response (2018). However, for any hypothesis that is being explicitly tested in an RCT, the power is generally 80% and might be higher if a particularly important hypothesis is being investigated. Also, it is not common to compute sample sizes using binomial enumeration because of the time required if the sample size is expected to be large.

Fisher's Exact Test is the classical inferential test applied to "proving" a difference with small incidences being compared. While huge drug and control (placebo) sample sizes (about 5,000 – 10,000 subjects) for each treatment might be obtained in some development programs (not for a psychiatric drug, but for a cardiovascular [CV] or diabetes drug), that number of subjects exposed generally would not be obtained in a single RCT but in multiple RCTs. The results from the multiple RCTs would be combined in a meta-analysis. A proper meta-analysis would consider differences across the RCTs and differences in study size to compute the inferential statistical result. A proper meta-analysis, therefore, generally requires an increased subject number for any given power relative to the number of subjectsrequired in a single,

prospective, large RCT. For simplicity, however, the computations above and those below will be for a single RCT.

Also, as pointed out by Beasley (2018), there is almost always some background incidence of any AE of interest. Required sample sizes become even larger because of such background incidence in inferential tests intended to "prove" difference (null hypothesis of no difference). Beasley provided the example of an event with a 0.5% background incidence (i.e., an incidence of 0.5%[iii] would be observed in the control group and the drug group due to causes other than drug) with an additional 0.1% (0.5% vs. 0.6%) observed in the drug group due to drug causation / contribution. In this scenario, sample size per treatment grows to 87,851 for 80% power with a 2-sided Fisher's Exact Test, when computed with normal approximation.

A 2-sided Fisher's Exact Test (testing a ratio of incidences) is not the only inferential test that can be applied to proportions (incidences) in two groups being compared. The incidence difference (incidence with drug - incidence with placebo) can be tested. This alternative to testing the ratio is important when dealing with small incidences. When dealing with single digit incidences expressed as percentages, the difference between a difference and a ratio can be striking. The difference between an incidence of 1% and 2%, expressed as a percent is 1% (2 - 1), while the ratio, expressed as a percent is 200% (2 / 1), and the excess incidence, expressed as a percent of the lower incidence is 100% ([2 - 1] / 1). The results of inferential tests based on differences versus ratios can be different and sample size computations for a given power can result in different sample sizes. As observed incidences (used in inferential tests) and hypothesized incidences (used in sample size computations) decrease, these differences in computational results can become more important. Additionally, because with low incidence AEs, inferential analyses are most often conducted using multiple RCTs where it is likely that the AE of interest will not be observed (0 incidence) in one of the treatment arms being compared, and in some of the RCTs in none of the treatment arms. Both cases complicate the use of such a study in the meta-analysis using the ratio of incidences. If an RCT has a 0 incidence in one or more arms being compared but an arm with >0 incidence, a small incidence needed to be addedwhere the actual incidence is 0 to use the RCT in the meta-analysis when analyzing the ratio of proportions. When the AE of interest is observed in none of the treatment arms being compared, the entire RCT is excluded from the meta-analysis. In such a case, significant amounts of meaningful data are then disregarded. If the difference in incidences is

used for analysis, both difficulties can be avoided,and all actual data can be used. Techniques are evolving that improve on these meta-analyses of rare events of interest (Tian, Cai, Pfeffer et al. 2009). In the assessment of safety with psychiatric drugs, this problem was highlighted by the analysis of suicidal behaviors and completed suicides in the original study of this potential ADR in the fluoxetine depression database (Beasley, Ball and Nilson2007; Beasley, Dornseif, Bossomworth et al. 1991). However, it is very uncommon for regulators to focus on analyses based on incidence differences and we do not include computations for sample sizes for analyses of incidence differences below.

With a long-term, large study, survival analysis can be used. While a simple Logrank Test is often used for survival data, a Cox Proportional Hazards Model with an analysis of the Hazard Ratio would often, if not most commonly, be employed with survival data. Also, the Cox Proportional Hazards approach is generally used for AEs when performing a noninferiority analysis "proving"absence of an effect (i.e., the absence of an ADR) as is described in more detail in a section below.

Table 2 below shows sample sizes for a classical inferential test (null hypothesis: no difference – "proving" that an AE is an ADR if the null hypothesis is rejected) using Fisher's Exact Test and a Cox Proportional Hazards Model analysis for the 51%, 80%, 90%, and 95% power. In all cases, α=0.05, there is an equal allocation of total subjects to two groups (test drug, control [placebo or active "known"to not have ADR of interest – incidence due to control approaching 0]). The following were additional specifications for each procedure:

- Fisher's Exact Test:
    - Test drug observed incidence: 0.001 ($1.0x10^{-3}$, 1 in 1,000, 0.01%)
    - Control observed incidence: $1.0x10^{-15}$ (cannot set to 0.0 for sample size computation)
    - Computation by binomial enumeration (where computed sample size for both treatment groups $\leq$100,000, otherwise normal approximation used)
    - Addition of 0.0001 (PASS authors' recommendation) to 0 cells only
    - No adjustment for subjects discontinuing early – assume all subjects observed through sufficient time to observe the "adverse event"of interest if it would occur
- Cox Proportional Hazards Model
    - Test drug probability of an event: 0.001

- Control probability of an event: 0.00005 (5 per 1,000,000, 0.005%, 5.0×10^{-5}; hazard ratio of 20 – minimum control probability of event / maximum hazard ratio that allowed for PASS computation with at least 1 event observed in the treatment group[iv])
    - 51% power: estimated 0.08 events with control and 1.67 with the test drug
    - 80% power: estimated 0.17 events with control and 3.33 with the test drug
    - 90% power: estimated 0.22 events with control and 4.46 with the test drug
    - 95% power: estimated 0.28 events with control and 5.52 with the test drug

Table 2: Sample Sizes Required for Assessing a Hypothesis that Drug Does Have an Effect (Null Hypothesis of No Effect)

Power	Fisher's Exact Test (binomial enumeration)	Cox Proportional Hazards Model
51%	5,730	1,673
80%	7,905	3,332
90%	9,273	4,461
95%	10,511	5,517

The Cox Proportional Hazards Model analysis sample sizes are the best cases for such an analysis. The software does not allow for the inclusion of a censoring rate for the treatments that can differ between treatments. Furthermore, the software assumes sufficient time of observation (length of RCT) to observe 100% of the incidence of events for the two treatments that are reflected in the probabilities of an event for each treatment. Early discontinuations will occur, especially for RCTs that have lengths that extend for multiple years. More realistic sample sizes for the Cox Proportional Hazards Model analysis can be computed by reducing the expected observed hazard ratio. For example, with a power of 80% and a hazard ratio of only 10, the sample size for each treatment group increases to 5,640 from 3,332 and for a hazard ratio of 15, still grows to 4,078.

Sample sizes are smaller with a Cox Proportional Hazards Model analysis. However, with either of these inferential test methods, required sample sizes are large. If multiple studies are used in a meta-analysis (generally required for assessment of a very uncommon AE), total sample size increases. For assessment of a very uncommon AE of a clinically significant nature, power >80% would be desirable. Large numbers of subjects treated only with placebo (a

component of the gold standard control treatment for determination of a treatment effect) is a particularly challenging problem.

Additionally, these computations are for a single study. As noted above, at least for an assertion of efficacy, at least two independent findings that reject the null hypothesis of no difference and lead to an interpretation of a drug effect are required to "prove" efficacy for drugs intended to treat non-life-threatening conditions unless there is overwhelming evidence of efficacy in a single RCT. From a rigorous scientific perspective, this replication requirement is an excellent, conservative requirement protecting against a Type 1 error in a single RCT. From our perspective, the assertion that any AE is an ADR with robust scientific rigor would require the same level of evidence as required for an efficacy assertion. We are not suggesting that labeling of ADRs should require the same degree of "proof" as required for an efficacy claim but are describing the nature of the evidence for the assertion of an ADR compared to that for the assertion of efficacy for a given indication.

We believe that clinicians, patients and all other parties should understand the quality of "proof" that any given AE listed as an ADR in lay literature, scientific/clinical reviews and product labeling is an ADR. Additionally, these parties should have a clear understanding of the approximate incidence with which an ADR must occur for the "proof" that the AE is an ADR to be comparable to the standard of "proof" for efficacy.

So, to "prove" a hypothesis (that a drug causes a rare "adverse drug reaction") one needs large numbers of subjects. The sample in the table above (Table 2) for 80% power (a conventional power in high-quality efficacy studies) is 7,905 per treatment group with Fisher's Exact Test (the most conventional analytical method). However, if an important outcome were being studied, even greater statistical power would be desirable.

References:

Beasley CM, Jr. Response to Barry Blackwell reply to Charles M. Beasley Jr.'s comment on Barry Blackwell's Corporate Corruption in the Psychopharmaceutical Industry. inhn.org. Controversies. January 12, 2018.

Beasley C, Ball S, Nilsson M. Fluoxetine and adult suicidality revisited: an updated meta-analysis using expanded data sources from placebo-controlled trials. *J ClinPsychopharmacol* 2007; 27:682-686.

Beasley C, Dornseif B, Bosomworth J, Sayler ME, Rampey AH Jr, Heiligenstein JH, Thompson VL, Murphy DJ, Masica DN.Fluoxetine and suicide: a meta-analysis of controlled trials of treatment for depression. *BMJ* 1991; 303:685-692.

Blackwell B. Response to Charles Beasley's response to Blackwell's reply to Beasley's comment on Blackwell's Corporate Corruption in the Psychopharmaceutical Industry. inhn.org.Controversies. May 3, 2018.

NCSS 12 Statistical Software. 2018. NCSS, LLC.: Kaysville, UT. ncss.com/software/ncss.

nQuery Advanced (8.2.1.0). 2018. Statsols: Cambridge, MA. statsols.com/nquery

PASS 15 Power Analysis and Sample Size Software. 2017. NCSS, LLC.: Kaysville, UT. ncss.com/software/ncss.

Tian L, Cai T, Pfeffer M, Piankov N, Cremieux P-Y, Wei LJ. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 × 2 tables with all available data but without artificial continuity correction. *Biostatistics* 2009; 10:275–281.

[i] All sample size computations results presented in this Commentary, were computed in PASS (2017) were validated in nQuery Advanced (2018).

[ii] To perform the sample size computation, one cannot use a 0.0 incidence for the control group in PASS 15 software (but 0 events can be used in an actual Fisher's Exact inferential test computation within NCSS 12.0.5). We set the incidence for drug at 0.001 ($1x10^{-3}$, 0.01%, 1 in 1,000) and the incidence for control at $1.0x10^{-15}$ that is effectively 0.0 in the sample size computation).

[iii] While in the response to Blackwell by Beasley (2018), the text correctly described the hypothetical background incidence used in computation as 1 case in 200 persons (5 cases per 1000 persons) a typographical error gave the numerical background incidence as "5%" rather than the correct "0.5%".

[iv] While PASS will compute sample sizes where observed incidence in both treatment groups is <1, inferential analysis would not be possible without at least one observed case in the drug treatment group.

January 24, 2019