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)

Outline

This commentary builds on topics Beasley briefly addressed in his response (Beasley 2018) to Blackwell's reply (Blackwell 2017) to an earlier comment of Beasley (Beasley, 2017) in response to Blackwell's essay about Corporate Corruption in the Psychopharmaceutical Industry (Blackwell 2016). The primary purposes of the commentary are to:

1. illustrate the sample sizes required to infer with reasonable certainty that some adverse medical event is caused by a drug; and
2. illustrate the sample sizes required to infer with reasonable medical certainty that some adverse medical event, while possibly observed during administration of a drug, is not caused by the drug.

We focus on adverse medical events that are infrequently observed in temporal association with the administration of a drug and are likely to be medically serious. The

point that is made in illustrating these sample sizes is that for such adverse medical events the inference that a drug caused or did not cause the events is not based on robust empirical evidence. Furthermore, obtaining such robust medical evidence would be a practical impossibility.

The commentary progresses in sections as follows:

1. A section that provides definitions of technical terms that have a precise meaning in the domain of drug safety/pharmacovigilance as these terms will be used in the commentary.
2. An introductory section that restates our purposes and briefly describes some complexities of the time course of observation of an adverse medical event over time that is caused by a drug. While these complexities can complicate a correct analysis of whether such an event is or is not caused by a drug, we address the simplest case in sections that follow.
3. A section that discusses the variability that can occur when a subset of a population of interest is selected for inclusion in a study in terms of what would be observed in the total population compared to the subset. Such variability is an important topic as it is relevant to an understanding of sample size computations. As a special case of this variability, we discuss what can be inferred when no events or outcomes of interest are observed in a subset of a population of interest that is embodied in the statistical Rule-of-3.
4. A section that discusses sample sizes in studies where the objective is being able to infer that an effect occurs under the assumption that the effect does not occur.
5. A section that discusses sample sizes in studies where the objective is being able to infer that an effect does not occur under the assumption that the effect does not occur.
6. A section that illustrates the extreme rarity of events that would be of interest in the assessment of the safety of a drug. This section provides context for understanding the incidence of an event associated with a drug that is used in our sample size calculations.
7. A section that discusses regulatory requirements for drug exposure (number of patients) in development programs for drugs used on a long-term basis in the treatment of disorders that are not acutely life-threatening. This section further discusses what regulatory authorities acknowledge regarding the limitations of

such sample sizes in determining with reasonable certainty what events are caused by a drug before its approval.

8. A section that briefly enumerates some of the methods used to attempt to determine events caused by a drug, both before and after its approval, which are not as robust as a study or set of studies, using appropriate controls.

1. Definition of Terms Used in this Document

- **Adverse Event: (AE)** – an adverse or untoward medical event (complaint, symptom, sign, syndrome, disorder, disease) that occurs or worsens in temporal association with a study treatment (investigational drug or control [placebo or active drug]) or during any period of observation without treatment in a randomized clinical trial (RCT). An AE might be etiologically related to a treatment or an incidental observation with an etiology other than treatment.
- **Adverse Drug Reaction: (ADR)** – an AE where there is "reasonable evidence" that the AE was etiologically related to treatment (investigational drug or control). To the best of our knowledge, "reasonable evidence" has never been operationally defined or even quantified by any regulatory entity or drug safety organization, including:
    - U.S. Food and Drug Administration (FDA) or other national regulatory agencies;
    - International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH – a group of major worldwide drug regulatory agencies and pharmaceutical manufacturers' associations);
    - Council for International Organizations of Medical Sciences (CIOMS – a nongovernmental organization set up by WHO and UNESCO that works with ICH to establish standards and methods of evaluating drug safety.

  "Reasonable evidence" might be the medical equivalent of the legal standard of "preponderance of evidence" that is quantitatively well defined (>50%). However, it might be some quantity ≤50%. ADRs are identified based on the totality of relevant available data. The most robust data are provided by placebo-

controlled RCTs and meta-analyses of multiple such RCTs. But, prospective and retrospective epidemiological studies, post-marketing surveillance and multiple other sources of data contribute to sponsors' and regulatory bodies' decisions about what AEs are ADRs and should be identified as such in product labeling. Even if "reasonable evidence" was quantitatively well-defined, the judgment of the magnitude of the totality of data and analyses relevant to whether an AE is or is not an ADR would remain a subjective opinion, at least for "uncommon" AEs (see definition below). In some cases, an ADR can be attributed to a drug treatment (or the potential for a specific ADR is considered a strong possibility) in product labeling even if the AE has not been observed with that drug treatment (e.g., all dopamine antagonist antipsychotics are potentially associated with the ADR of neuroleptic malignant syndrome [NMS]). The potential for this ADR will appear in product labeling, in the Warnings and Precautions Section of a US label for all drugs in this class. If NMS had not been observed at the time of approval, the Warnings and Precautions text related to NMS is likely to include that caveat. Pharmacological class effect (a supposition rather than an empirical finding) is the basis for believing that there is "reasonable evidence" that a dopamine antagonist causes or contributes to the development of NMS.

- **Incidence categories of ADRs (and AEs observed in a clinical trial):**
    - Very common: ≥ 1/10, 10%, 0.1000
    - Common (frequent): ≥1/100, 1%, 0.0100 to <1/10, 10%, 0.1000
    - Uncommon (infrequent): ≥1/1,000, 0.1%, 0.0010 to <1/100, 1%, 0.0100
    - Rare: ≥1/10,000, 0.01%, 0.0001 to <1/1,000, 0.1%, 0.0010
    - Very rare: <1/10,000, 0.01%, 0.0001
- **"Proof" of a drug effect (and proof of absence of a specific effect):** The standard of proof for a binary categorical outcome (in our case of interest the occurrence of an AE that might be an ADR) is based on a difference in incidences or a ratio of incidences observed in well designed, prospective, RCTs (or meta-analysis of multiple RCTs). If the difference or ratio, analyzed with proper statistical methods, is significant ($p \leq 0.05$), the results are interpreted as 'proof' of an effect. For "proof"' of efficacy the regulatory standard, at least that of FDA for potential drugs intended to treat non-life-threatening disorders, is generally two RCTs with inferential results of $p \leq 0.05$. If statistical significance is overwhelming in a single trial (e.g., $p < 0.001$ in the single trial/analysis and/or the

trial could be randomly split into two trials/analyses multiple times, and analyses of the split samples would consistently result in $p<0.05$) one trial might be sufficient.

2. Introductory comments

For several years, Charles Beasley has had an interest in what RCTs that support approval of a potential new treatment tell us, with a robust degree of scientific certainty ("prove" – see Part 1), about possible ADRs associated with treatment and what possible ADRs are not associated with treatment? Current designs and practical limitations on the size and length of time over which an RCT can be conducted influence what an RCT can "prove." With what incidences must an AE occur in association with an investigational treatment and control treatment to "prove" that the AE is an ADR for the investigational treatment under consideration? What size would studies need to be conducted to "prove" that a rare AE is an ADR? The sample size requirements for deciding what distinguishes ADRs from among AEs and "proving" either the presence or absence of any given potential ADR is the essence of what we are discussing.

The hypothetical case on which we focus is that of a highly uncommon ADR with an incidence of 1 per 1,000 persons treated (0.001 or 0.1%), the low boundary of "uncommon" events. If the incidence is 1 in 1,001 subjects, the event would be "rare." However, just because such an ADR is highly uncommon, this does not mean that it will not be experienced by a considerable number of individuals during the commercial life of a widely prescribed drug for disorders common in the general population. As Beasley said in his earlier response to Blackwell (2018), if some 20,000,000 individuals are treated with a drug (and that number might be higher by several multiples), the ADR with an incidence of 1 per 1,000 would occur in 20,000 persons. The successful drug will become generic and more people would be treated with more persons experiencing the ADR.

The majority of what we say below about complexities deals with simple incidence (events/person) for the 0.1% of individuals who experience the hypothetical ADR. However, the distribution of time to experience the ADR can have a substantial impact on the extent to which a specific study design, sample size, and analysis can influence the "proof" of the presence or absence of an ADR. Even rare ADRs, with enough individuals

treated, might show three patterns of distribution of time to occurrence (temporal patterns of occurrence):

1) early in treatment (acute toxicity) – a curve of the cumulative incidence over time would rise rapidly and then taper off (sigmoidal / Gompertz function pattern);
2) later in treatment – with increasing incidence in later epochs of time (delayed toxicity with increasing exposure [can be due to drug exposure accumulation or a lag between acute exposure that is toxic and the manifestation of the toxicity, e.g., myocardial infarction and ischemic stroke due to acceleration of atherosclerosis]) – a curve of the cumulative incidence over time would reflect an initial linear rise followed by exponential rise after some lag time; and
3) random occurrence with equal distribution across time of treatment – a curve of the cumulative incidence over time would be linear with a slope dependent on incidence during the period of observation.

The rate of occurrence (event/person-time [e.g., number of ADRs / 100-patient-years of treatment]) and the temporal pattern of occurrence are two of the multiple factors that complicate "proving" the presence or absence of an ADR. These two related factors would be important considerations in discussing limitations of attempts at such "proof." In Installments 4 and 5 we discuss sample sizes required for "proving" that an observed AE is or is not an ADR. These sample sizes for "proving" that an AE is an ADR apply best to temporal pattern of occurrence #1 above (especially if there is a short lag time between initiation of treatment and first occurrence of the ADR) for the AE of interest. An ADR with temporal pattern of occurrence of #2 above would generally result in the requirement for longer periods of observation (a longer RCT) than temporal pattern #1 and therefore require additional subjects to begin a definitive RCT in order to account for subjects discontinuing the RCT/observation prior to the planned end of observation and the more frequent occurrence of the ADR. A relative infrequent or rare ADR, occurring with temporal pattern #3 would also require a longer period of observation in a definitive RCT. Therefore, the sample sizes discussed in Installment 4 that focus on "proving" that an AE is an ADR should be considered conservative estimates for ADRs that would only be observed late in treatment, with an accelerating rate of occurrence after some relatively lengthy period of observation or in a random pattern over time but very infrequently overall. Additionally, any pattern of occurrence that is a change as a function of time might require special statistical techniques (beyond comparing incidences or assessing the ratio of incidences) to "prove" presence or absence of the

ADR. Therefore, RCTs required to address the complexity of changes in the rate of occurrence of an ADR over time are likely to require larger numbers of subjects beginning such an RCT (because human subjects discontinue participation). There is one final caveat regarding patterns over time: as events become rarer, they generally appear to be randomly distributed over time, and there are never a sufficient number of cases observed to discern a temporal pattern within RCTs of practical size, even if a pattern exists. Rare AE occurrences, of which the majority are ADR occurrences, will generally appear with temporal pattern #3 unless sufficiently large number of subjects are observed to discern temporal pattern #2 when that is the pattern of occurrence.

3. lack of occurrence of an AE in an RCT (Rule-of-3) and impact on sample size calculations

An RCT or set of RCTs samples only a subset of the entire population of interest as subjects. Interpretations of the results of RCTs are then extrapolated to the entire population of interest and this is the very essence of clinical research. Even with the use of the best methods of random allocation of subjects to the treatments in an RCT, the observations in the RCT (within treatments and between treatment differences) can differ from what would be observed if the entire population of interest was studied in the RCT. The statistical "Rule-of-3" (Eypasch 1995; Hanley 1983) addresses the potential difference between what is not observed in a subset sample compared to what would be observed if the entire population of interest (or another subset) was to be studied in an RCT.

The following is a simple example of the sampling problem and the "Rule-of-3":

Let us say that we are interested in the entire human population and that the truth is that drug X does cause some ADR "Bad-Thing" in 1 in 1,000 persons (and nothing else but drug X causes the AE "Bad-Thing" that in this case is an ADR – background incidence of 0%). As of December 2017, the world's population was estimated at 7.6 billion. If we could somehow study that entire 7.6 billion sample for a sufficient period to observe all occurrences of the ADR, we would observe 7.6 million cases of the AE "Bad-Thing," with all these cases being ADRs. However, if we were to study only 1,000 subjects and we sampled the entire population perfectly, we would observe one case of

this ADR. However, if we sample only 1,000 subjects, we are highly likely to obtain a sample where the incidence of the ADR differs from the incidence in the entire population. While we might observe more than one case of this ADR, we are more likely to observe no cases of this ADR. The "Rule-of-3" addresses the lack of observation of an outcome.

The "Rule-of-3" has two variants relevant to this discussion:

- Precise interpretation: If we study 1,000 subjects and do not observe a single case of AE "Bad-Thing," we can conclude with 95% probability that the true incidence of AE "Bad-Thing" is only <1/334 subjects (AE might or might not be an ADR). The incidence of AE "Bad-Thing" has a 95% probability of being between 0/1000 and 1/333, where 333 is the approximate upper bound of the 95% confidence interval (CI) when 0 events have been observed in 1,000 observations.
- Extrapolation: We are studying only a subset of the population of interest and our sample might have an incidence of the ADR that differs from the incidence in the entire population. Therefore, we would need to study at least 3,000 subjects to have a 95% probability of observing even 1 case of the ADR "Bad-Thing" with a true incidence of 1 in 1,000 (with no cause of the AE other than it being an ADR).

This estimation only applies to cases of 0 observations (Ludbrook 2009) and the simple calculation of the upper bound of the CI is only valid with a relatively substantial number of observations (e.g., ≥100) (Jovanovic 1997).

Note that the two variants of the "Rule-of-3" only address not observing a single case of AE "Bad-Thing" and not "proof" of presence or absence of "Bad-Thing" as an ADR.

The potential difference in what is observed in a subset of the whole population of interest that is studied compared to what would be observed if the whole population of interest was studied, is important in understanding the results of sample size computation that Beasley provided in his response to Blackwell (2018). Sample size computations consider the potential for what is observed (in this case the incidence of an AE) in the sample selected for an experiment to deviate from what would be observed if the entire population of interest was included in the experiment. The result of this adjustment for potential variation between the experimental subset and the entire population is that the sample size for any given power greater than ~50% power will result in p-values <0.05 If the experimenter was:

1) lucky enough to select a subset for which what is observed is equal to what would be observed in the entire population of interest (or greater than the incidence in the entire population);
2) lucky enough to guess the incidences that would be observed;
3) used these incidences in sample size calculations.

In other words, smaller sizes than those obtained with an 80% or 90% power sample size computation will be sufficient to "prove" that an AE is an ADR if one is lucky in guessing the observed outcome incidences and using these in the sample size calculations. However, one might not get lucky with sampling and miss "proving" that an AE is an ADR without a sample size that provides 80+% power even if one is lucky in guessing incidences in the entire population.

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 used and 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 Blackwell and 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[1]). 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[2]).

The sample size below for 80% power is somewhat lower than Beasley reported in response to Blackwell because 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 subjects required 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%[3] 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 also has one or more arms with >0 incidence, a small incidence needed to be added where 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 Nilson 2007; 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 (1.0×10^{-3}, 1 in 1,000, 0.01%)
    - Control observed incidence: 1.0×10^{-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%, $5.0x10^{-5}$; 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[4])
    - 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 (lowest number of subjects) for each power because the calculation does not consider early discontinuation (censoring) from the planned period of observation. The software does not allow for the inclusion of a censoring rate for the treatments and in the actual study, the censoring rates 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.

5. "Proof" of the absence of an ADR (noninferiority compared to control): sample size requirements

We have addressed in Part 4 the difficulties in "proving" that an infrequent or rare AE is an ADR by the standards applied to "proving" efficacy. We now turn to the matter of "proving" that an AE is not an ADR and the related matter of correctly interpreting RCT results that fail to reject the null hypothesis of no difference. The correct interpretation of an RCT where a null hypothesis of no difference was not rejected is essential for the interpretation of both efficacy results and AE observations.

If our interest is in proving absence, a noninferiority inferential test (Mauri and D'Agostino **2017)**[1] with the null hypothesis of some difference between groups is used and we conclude that no difference exists between groups if that null hypothesis is rejected at the $\alpha \leq 0.05$ (≤ 0.025 in some cases) level (Mauri and D'Agostino (2017) There is a very important difference between the conventional inferential test of a difference and the noninferiority inferential test. In the conventional test, there is no necessity to define a meaningful difference (except in determining sample sizes). However, in the noninferiority inferential test, it is necessary to define a difference between treatments that will be considered "no difference" (not clinically meaningful). This difference cannot be set to "0" because sample sizes would then need to be infinity. In noninferiority tests, some slight difference must be considered acceptable and one can never completely exclude (statistically) some slight excess with test drug versus the comparator.

We are concerned that some interpret failing to "prove" (failing to reject the null hypothesis of no difference) an effect as equivalent to "proving" absence of an effect, especially if the study intended to "prove" presence of an effect is well powered (e.g., ~90%). However, this is not the correct interpretation of a $p>0.05$ statistical test result even if the RCT used sample sizes that provided ≥90% prospective power. We would acknowledge that if the power of the study was ≥95%, then failure to reject the null hypothesis might offer some evidence of lack of difference (i.e., lack of difference associated with 95% associated with 95% power). This approximate interpretation of an RCT with a null hypothesis of no difference and an outcome of the analysis with $p>0.05$ applies only to a prospective outcome of interest (e.g., a specific efficacy measurement) where the sample size was prospectively determined based on a 95% power. This approximate interpretation would not be appropriate for multiple outcomes (e.g., the

multiple AEs observed in an RCT) where there was no prospective determination of sample size based on 95% power.

However, the correct, formal interpretation of an RCT outcome described in the paragraph above is simply that the RCT failed, not the absence of effect. The design and prospective Statistical Analysis Plan (SAP) for an RCT must test for noninferiority to control to allow for correct, formal interpretation of results as indicating lack of effect, irrespective of sample size. The RCT could be accompanied with a complex SAP that would allow for sequential testing of multiple and alternative hypotheses (such as first testing a null hypothesis of no difference [potentially "proving" an effect] followed by the testing of a null hypothesis of a difference [potentially "proving" lack of an effect]). The SAP could include adjustment of α for the multiple testing without rejection of the null hypothesis in the first test in the sequence. Such SAPs would allow simultaneous tests for both an effect and lack of effect.

To "prove" absence of an effect one designs a noninferiority (to placebo) study and as noted above one must declare some non-0 excess with drug, usually expressed as a ratio of incidences in the case of binary outcomes for individual subjects such as AEs (or "response" for efficacy) as clinical equivalence. The excess incidence with the drug could be expressed as a difference rather than a ratio and the observed difference rather than the observed ratio tested but, in the concrete, required study example described below, the ratio of incidences is tested. For a clinically important potential ADR (with our incidence of 1 in 1,000), one might think that the ratio might be set at 1.10 (maximum of 10% excess with the drug) or even 1.05 (5% excess with the drug). However, there is precedent (discussed below) for an excess incidence with the drug of any magnitude <30%, based on the 95% CI for the observed ratio, above the incidence observed in the control group and still declare noninferiority for the drug. With any magnitude of excess <30% as the maximum estimated from the CI, the actual observed excess incidence with drug in the study will be less than 30% because the upper bound of the 1-sided (in some cases of such a study possibly a 2-sided) 95% CI[1] around the ratio of incidences cannot be ≥ 1.3 for drug:control. In many, if not most cases, the observed ratio with drug to placebo will be less <1 for the upper bound on that CI to be <1.3. Furthermore, in some cases with that ratio of 1.3, the drug will be not only non-inferior to control, but also superior first potential outcome in a noninferiority trial - real examples provided below) (Mauri's and D'Agostino 2017).

This analytical requirement is mandated for hypoglycemic agents for the treatment of diabetes mellitus and is codified in an FDA Guidance to Industry (CDER 2008). Sponsors developing such drugs must "prove" that a drug candidate does not cause serious cardiovascular outcomes that would most likely all be due to accelerated development of atherosclerosis, grouped under the acronym MACE (Major Adverse Cardiac Events). There are multiple definitions of MACE, but the events always included are: 1) all cardiovascular AEs with an outcome of death (sometimes includes all outcomes of death when the cause cannot be determined); 2) myocardial infarction; and 3) stroke (ischemic or ischemic and hemorrhagic and sometimes including TIA). Hospitalization for unstable angina, hospitalization for heart failure (or acute heart failure) and revascularization and stent placement procedures might be included.

This requirement, established in 2008, grew out of what Beasley believes was a flawed analysis of data for the PPAR drug rosiglitazone, conducted by the cardiologist Steven Nissen (Nissen 2007). Beasley thinks the analysis was flawed for two reasons. First, the data source was study summaries that reported incidences of "Serious Adverse Events" (SAEs) (AEs that are fatal, acutely life-threatening, result in or prolong hospitalization [inpatient], result in permanent disability, are congenital anomalies, are cancer, are deemed by the reporting investigator or sponsor to be serious for any other reason) on the sponsor's website disclosing results of studies. These SAEs were described with a term (a label from a regulatory dictionary [MedDRA] used for reporting AEs that can be a sign, symptom, syndrome or specific diagnosis). Unfortunately, SAE reports sometimes inaccurately characterize the AEs and/or provide an incorrect term/label for a given AE. These SAE reports are not necessarily subjected to scrutiny by a blinded, expert review committee to decide the correct term/label for an AE. What was reported by an investigator, required to report such an event within 24-hours if fatal or life-threatening and otherwise within seven days of learning of the AE, will sometimes not be what would have been concluded by a review committee reviewing all available medical records following all diagnostic and therapeutic activities in association with AE. Therefore, the data that were used by Nissen were not necessarily accurate data. Second, events were very infrequent and were not reported in some treatment groups in the multiple studies used by Nissen. Furthermore, in some studies considered for use, the SAEs of interest were not reported in any treatment arm. Nissen used a ratio of incidences (proportions) for his analysis rather than the difference in incidences. The meta-analytic technique that he used at the time to compare incidences was such that not all studies

could be used (those with no event of interest in any treatment group [10 of 48 reported no myocardial infarction and 25 of 48 reported no death from cardiovascular causes, the two outcomes analyzed separately]). Additionally, because of the technique used, when a study had an event or events of interest in one but not another treatment group used in the comparison, a small incidence needs to be added to the treatment group with actual 0 incidence, as described above. From an analytical method perspective, using the difference in incidences, briefly mentioned above, rather than the ratio of incidences (odds ratio) would have at least allowed use of data from all 48 available studies where 0 incidence is highly informative and would have been a preferable method.

The method developed by Tian et al for meta-analysis was used by the authors to reanalyze the dataset used by Nissan (Tian, Cai, Pfeffer et al 2009). For neither the CV mortality endpoint nor the myocardial infarction endpoint were the results statistically significant. For CV death, the risk difference was 0.063% (95%CI: -0.13%-0.23%; p=0.83). For myocardial infarction, the risk difference was 0.183% (95%CI: -0.08%-0.38%; p=0.27).

This study requirement has placed a significant cost and time burden on companies developing treatments for diabetes, discouraging development, and its need has been questioned by multiple academic groups based on experience with several such analyses results (Hirsberg and Katz 2013; Regier, Venkat and Clo 2016; Smith, Goldfine and Hiatt 2016; Yang, Stewart, Ye and DeMets 2015). In counterpoint, at least one author has recently espoused the position that the studies that evaluate MACE events as an outcome are insufficient to assess the potential for contributing to heart failure (although congestive heart failure is sometimes included in the analyses of MACE events), arrhythmia and microvascular disease with its multiple adverse clinical consequences (Packer 2018). As a patient with Type II diabetes, Beasley is personally very distressed by this obstacle to innovation that also drives up the cost for those new drugs that are developed.

Irrespective of the wisdom of the regulatory requirement for this study of MACE outcomes for potential new non-insulin anti-diabetic therapies, the study outline establishes the model for "proving" that a drug does not cause a specific group of ADRs. The group of ADRs that might or might not have common underlying pathophysiology in the case of MACE events (e.g., an ischemic cerebral infarction is vastly different compared to a subarachnoid hemorrhage from a pathophysiological perspective).

Table 3 below displays the sample sizes for demonstration of noninferiority of test drug to control ("proof" of absence of effect – null hypothesis is that an effect does occur with the proportion observed with test drug of ≥1.3-fold the proportion observed with control when the proportion observed with control is 1 in 1,000 [0.001, $1x10^{-3}$]). While noninferiority is conceptually a 1-sided test and a 1-sided 95% CI might be used in the inferential test when testing the ratio of incidences, a 2-sided confidence interval is often used as effectively testing at a p-value (α) of ≤0.025 for noninferiority. For assessment of noninferiority of AEs ("proof" that an AE is not an ADR), the Cox Proportional Hazards Model is customarily employed.

Table 3: Sample Sizes Required for Assessing a Hypothesis that Drug Does Not Have an Effect (Null Hypothesis of An Effect with an Observed Ratio ≥ the Ratio Considered to be Clinically Equivalent to No Effect)

Power	**Cox Proportional Hazards Model**	
	1-sided (α=0.025)	**1-sided (α=0.05)**
51%	114,487	81,024
80%	228,049	179,634
90%	305,294	248,823
95%	377,561	314,439

Two published manuscripts provide examples of noninferiority (to placebo) RCTs evaluating MACE events with subsequent testing for superiority (Neal, Perkovic and Mahaffey 2017; Zinman, Wanner, Lachin et al. 2015). These RCTs demonstrated noninferiority. Also, the SAPs for the RCTs were written in such a way that allowed testing for superiority after a result that would be interpreted as indicative of noninferiority. Both manuscripts reported results of meta-analyses. The empagliflozin manuscript employed a hierarchical-testing approach in the order of: noninferiority for the primary outcome (MACE: death from CV events, nonfatal myocardial infarction excluding silent myocardial infarction or nonfatal stroke), noninferiority for the key secondary outcome (the primary outcome plus hospitalization for unstable angina), superiority for the primary outcome and superiority for the key secondary outcome (Zinman, Wanner, Lachin et al. 2015). A Cox Proportional Hazards Model was used for

analyses. A 2-sided p-value (for analysis of superiority) was adjusted to ≤0.0498 as indicative of statistical significance because the data had been submitted to the FDA in a New Drug Application. Noninferiority was declared if the upper bound of the 2-sided 95.02% CI was <1.3, resulting in a p-value for the noninferiority analyses of 0.0249 (comparable adjustment as with the superiority analyses). Therefore, superiority was declared if noninferiority was declared: the upper bound on the 2-sided 95.02% CI for the hazard ratio was <1.0 and the p-value was ≤0.0498. Because a Cox Proportional Hazards Model was used for analysis, the sample size was determined based on the assumption of a hazard ratio of 1.0. A power of 90%, required **691**[1] events to occur (rather than subjects studied) based on the assumed hazard ratio and level of statistical significance required. Thus, 4,687 subjects were included who began empagliflozin and 2,333 subjects were included who began placebo. The analysis included 48 months of treatment observation. For the primary outcome, the hazard ratio was 0.86 (95% CI: 0.74 – 0.99). For noninferiority, the p-value was <0.001 and for superiority was 0.04.

The canagliflozin manuscript also reported the results of a meta-analysis (Neal, Perkovic and Matthews 2017). Statistical analyses were comparable to those used in the empagliflozin manuscript but there was no adjustment of required p-values (Zinman, Wanner, Lachin et al. 2015). The sample size required for 90% power was determined to be **688**[1] events. Hierarchical testing was used in the following order: MACE (deaths from CV events, nonfatal myocardial infarction, nonfatal stroke); death from any cause; death from CV events; the progression of albinuria; and death from CV events plus hospitalization for heart failure. The manuscript does not specify where in the hierarchy superiority for any of the outcomes noted above was tested. There were 5,795 subjects included who began canagliflozin and 4,347 included who began placebo. The analysis included 338 weeks (~80 months) of treatment observation. For the primary outcome, the hazard ratio was 0.86 (2-sided 95% CI: 0.7 – 0.97). For noninferiority, the p-value was <0.001 and for superiority was 0.02.

In both drug development programs an event of interest adjudication committee, blinded to treatment, reviewed all records pertinent to each event (AE) to make a final determination of what each reported event represented (term/label). The need for all records and methods to acquire these records would have been put in place prospectively before each RCT initiation. These steps were taken to maximize data quality used in the respective analyses.

In the empagliflozin analyses, there were 43.9 MACE events per 1,000 subject-years with placebo and 37.4 MACE events per 1,000 subject-years with empagliflozin (Zinman, Wanner, Lachin et al. 2015). The comparable rates in the canagliflozin analyses were 31.5 with placebo and 26.9 with canagliflozin per 1,000 subject-years.

The two real-world examples above emphasize the magnitude of effort and therefore expense required to "prove" absence of a specific set of events in a population with an increased risk of such events (Zinman, Wanner, Lachin et al. 2015). The subject population, therefore, would be expected to have an increased background incidence of MACE events. However, presumably, there would also be a markedly increased risk of the events in the drug-treated group if the drug caused or contributed to the MACE events as ADRs.

Product labeling is not intended to describe explicitly those adverse events that have been demonstrated with reasonable certainty not to be ADRs. Instead, those sections of product labeling that address the safety of the treatment to which the labeling is applicable are intended to identify for the prescriber, and other interested parties, AEs that have been identified as ADRs with reasonable medical certainty. Therefore, the information above regarding sample sizes for noninferiority studies that might "prove" the absence of a specific ADR is of little relevance to the primary task of pharmacovigilance/drug safety monitoring and the development of product labeling. These noninferiority study sample sizes demonstrate the limitations on the robustness of what we know about what a drug does not do from a safety perspective based on the highest quality of evidence for medical decision-making.

While demonstrating noninferiority for an ADR is not critical to the primary intent of safety labeling, it can be critical to a sponsor attempting to "prove" that some AE that has been described as an ADR by some party is not an ADR for that given drug.

We should be cautious regarding what we believe about what a drug does and does not do from a safety perspective and fully understand the robustness of the supportive data for such attributions.

Endnotes

i. The authors describe five possible interpretations (Figure 1) of the results of a noninferiority analysis of an RCT. While all five are potential interpretations, from a conservative analytical design perspective, a primary, single null

hypothesis would be tested (i.e., superiority of the control over drug treatment). Failure to reject the null hypothesis would not permit any additional interpretation to be made without prespecifying some sequential order of testing other hypotheses and/or paying a "statistical penalty" for simultaneous testing of multiple hypothesis, including noninferiority and superiority and the paradoxical but possible interpretation of both noninferiority and inferiority simultaneously.

ii. We are aware of at least three studies required by FDA for potential drugs seeking regulatory requirements that are noninferiority studies comparing test drug to placebo. The so-called Thorough QT Study (required for virtually all potential drugs) compares the mean change from baseline in QTc. The Human Abuse Potential (HAP) Study (required for drugs with CNS activity that are perceived by FDA as having any abuse potential based on pharmacological action) compares mean absolute values (integers with a range of 100). Both studies' analyses employ a 1-sided 95% CI (FDA Guidance does not explicitly state use of a 1-sided CI for the TQT study analysis, but this is the commonly used CI). The boundary of a 1-sided 95% CI is equivalent to the upper bound of a 2-sided 90% CI and therefore is a lesser value. If a 1-sided 95% CI is used and the null hypothesis is rejected, the p-value is ≤ 0.05 while if a 2-sided 95% CI is used, the p-value is 0.025 and define the precision of the estimate because both an upper and lower bound are defined. The Major Adverse Cardiac Events Study ([MACE study] required for non-insulin drugs used to treat diabetes) compares the incidence of a set of AEs based on the ratio of incidences. The FDA Guidance Document that outlines this study and its analysis specifies the use of a 2-sided 95% CI. The major distinctions between the TQT study and the HAP study contrasted with the MACE study is that the TQT and HAP studies compare means of integer values and the differences used as not clinically meaningful have explicit empirical bases (TQT: Malik, 2001; HAP: Chen and Bonson 2013) while the MACE study is comparing proportions and there is less explicit empirical basis for the noninferiority with the MACE study. The FDA Guidance Document that specifies the margin cited reviews of two long-term studies of intensive vs. standard diabetes therapy (UKPDS, 1998a; UKPDS, 1998b) that reported CIs for multiple adverse cardiovascular outcomes in drafting its Guidance.

iii. PASS computes the total number of events for 90% power as 688 with a 2:1 assignment of number of subjects to drug:placebo (drug: 4579; placebo: 2290), and with p=0.0249.

iv. PASS computes the total number of events for 90% as 687 with a 2:1 assignment of number of subjects to drug:placebo (drug: 4579; placebo: 2290) and as 623 with a 1.5:1 assignment of number of subjects to drug:placebo (drug: 6869; placebo: 4579) that approximate the actual ratio in the meta-analysis, with p=0.025.

6. Incidences of AEs of real-world interest and limitations on "proof" of presence or absence of an ADR

How relevant to clinical reality and what would be relevant to both prescribers and patients who might suffer a major (i.e., life-threatening or fatal) ADR is our hypothetical example of an ADR that occurs with an incidence of 1 in 1,000 persons treated but virtually never happens in an untreated population. Aplastic anemia and the spectrum of Stevens-Johnson Syndrome (SJS) - toxic epidermal necrolysis (TEN) afford a context for considering the relevance of our example.

A major, international study of both agranulocytosis and aplastic anemia has been conducted under the sponsorship of the WHO. The first report described rates of occurrence for aplastic anemia ranging across seven sites from 0.6 to 3.1 (adjusted mean: 2.2) per million-person-years (International Agranulocytosis and Aplastic Anemia Study 1987). A more recent report of this study reported a range of rates of cases from 0.7 to 4.1 per million-person-years (Kaufman, Kelly, Issaragrisil et al. 2006). For aplastic anemia, about 25-40% of cases are considered due to exogenous exposures (drugs, toxic substances) or other external factors and the majority are believed to be idiopathic and have no identifiable etiology (Kaufman, Kelly, Issaragrisil et al. 2006). Therefore, with aplastic anemia, the incidence on an annual basis (~2-3 / million) is much lower than the 1 in 1,000 in our example. Furthermore, some background incidence of aplastic anemia would be expected due to idiopathic factors and exposures to substances other than test drug and thus would further increase sample sizes required to "prove" causation by a drug.

Stevens-Johnson Syndrome and toxic epidermal necrolysis are the extreme manifestation of the continuum of the clinical diagnoses of erythema multiforme (EM) – SJS – TEN; all share some characteristic histopathological feature of epidermal necrolysis (there is some disagreement in grouping erythema multiforme as a separate clinical entity or as part of the spectrum). A large UK epidemiological study reported the rate for combined SJS-TEN as 5.6 (95% CI: 5.31-6.30) per million-person-years (Frey, Jossi, Bodmer et al. 2017)). A separate, large national epidemiological study in South Korea reported rates for SJS of 3.96-5.03 per million person-years (range for individual years across four years) and rates for TEN ranging from 0.94-1.45 per million person-years (Kang, Ko, Kim et al. 2015). The UK and Korean results are comparable for rates of combined SJS and TEN. The incidence of SJS – TEN is then in the range of ~6.5 per million-person years. In contrast to aplastic anemia, because SJS – TEN is primarily due to exogenous exposure, background rates could approach 0 if a study could be conducted where study subjects receiving the active investigational drug received no other medications and control subjects receiving placebo received no drugs. Of course, this would be a highly impractical study design, especially given the enormous number of subjects required for definitive assessment

While "proving" by conventional statistical standards that a test drug does or does not cause a specific ADR with an incidence of 1 in 1,000 patients treated and when the background incidence (incidence in a placebo- or active-control group) approaches 0 is difficult, that difficulty will grow by orders of magnitude with aplastic anemia and SJS – TEN.

Definitive "proof" that a drug is associated with an ADR or that a drug is not associated with some specific ADR of interest is virtually impossible given the practical limitations impacting the conduct of human RCTs when the incidence of an associated ADR is less than some 2-3% when active treatment and placebo control sample sizes are below several hundred subjects per treatment group. For psychiatric disorders, such sample sizes or even larger sample sizes would be common with depression and anxiety disorders. Active treatment and placebo control sample sizes can be smaller with psychotic disorders. For example, with the development program for olanzapine for its initial indication of treatment of psychosis (later restricted to schizophrenia) the total sample sizes that allowed direct comparison with placebo were: olanzapine – 248; placebo – 118. Additionally, these totals were obtained in two separate RCTs. One RCT

compared placebo to olanzapine 5±2.5 mg / d, 10±2.5 mg / d, and 15±2.5 mg / d. The other RCT compared placebo to 1 mg / d and 10 mg / d.

Development programs in other therapeutic areas can be of much greater size. Development programs in diabetes and cardiovascular diseases can easily exceed 5,000 and approach 10,000 subjects treated with the investigational drug. However, complicating the matter of definitive "proof" of presence or absence of an ADR, these studies are generally conducted as drug compared to placebo as an add-on to existing therapies. Therefore, while placebo-controlled, the ongoing treatment (or treatments) with associated ADRs can complicate definitive interpretation of safety observations.

7. Regulatory requirements for investigational treatment exposure in development programs and their implications for 'proof' of presence or absence of an ADR

To what extent are regulatory authorities aware of the limitations? In its 1995 Guidance to Industry addressing the "Extent of Population Exposure to Assess Clinical Safety: For Drugs Intended for Long-term Treatment of Non-Life-Threatening Conditions" (CDER 1995) exposures of 1,500 subjects to one or more doses (in intended multiple dose, clinical studies, generally not including single-dose, Phase 1 studies), 300-600 subjects for at least six months and at least 100 subjects for at least 12 months were specified (Center for Drug Evaluation Research 1995).

Multiple factors (e.g., a preclinical finding that would suggest rare potential toxicity) for individual potential drugs could result in the need for a greater number of exposures in the clinical development program studies.

These requirements were in line with The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) recommendations/requirements and apply to a wide range of potential drugs across a variety of disorders. For some disorders the potential drug can be tested against placebo while in many disorders the potential drug can only be tested as an add-on to single, standard therapy with a comparison to placebo added on to that therapy. With all the potential study variants to which these exposure requirements apply and all the differences in background incidences of events in the general population, the population

with the disorder under study and the standard treatment when an add-on study must be conducted, it would be difficult to make precise statements about the incidence of ADRs that could be definitively ascertained and ruled out. However, the Guidance (CDER 1995) offers the following suggestions on what these exposure requirements will and will not be able to detect (Center for Drug Evaluation Research 1995).

> "It is expected that short-term event rates (cumulative 3-month incidence of about 1%) will be well characterized."

> "The safety evaluation during clinical development is not expected to characterize rare adverse events, for example, those occurring in less than 1 in 1000 patients."

The phrase "well characterize" is not expressly defined. It would seem to us to convey more than simply observing an AE that might be an ADR in the treatment population but in many cases falls short of a difference in incidence from that incidence with control that reaches conventional statistical significance in a proper inferential test. There is likely to be some reasonable estimate of the incidence of the AE that combines AEs due to the background with those that are ADRs with a reasonable degree of difference in incidences to believe to believe that the AE can be an ADR.

In a later Guidance Document addressing "Premarketing Risk Assessment" (CDER 2005), the following is included:

> "Even large clinical development programs cannot reasonably be expected to identify all risks associated with a product. Therefore, it is expected that, even for a product that is rigorously tested preapproval, some risks will become apparent only after approval, when the product is used in tens of thousands or even millions of patients in the general population. Although no preapproval database can possibly be sized to detect all safety issues that might occur with the product once marketed in the full population, the large and more comprehensive the preapproval database, the more likely it is that serious adverse events will be detected during development" (Center for Drug Evaluation Research 2005).

Presumably, the reference to "adverse events" in the last sentence is to AEs that are ADRs. The statement above focuses on identifying ADRs but is equally applicable to the determination of the lack of a specific ADR associated with the drug under development. Here we have tried to quantitate these difficulties and limitations in RCTs, the gold standard for such determinations.

8. Practical alternatives to "proof" of presence or absence of an ADR : the need for best assessment possible as quickly as possible of the AE / ADR profile of a marketed drug.

Statisticians and data scientists, both academic and regulatory, have developed and are continuing to refine methods for working with data from sources other than RCTs. These sources include retrospective and prospective epidemiological studies (especially retrospective studies employing "big data" from evolving large databases possible with electronic medical records), large simple studies including those without a control group, and spontaneous adverse event reporting databases maintained by regulatory agencies where precise knowledge of total persons treated is not available but can be estimated, among other data sources. It can be hoped that these methods result in the reduction in failure to find true ADRs and reduce false attribution of an ADR to a drug. These methods are the ones that generally result in the discovery of very "infrequent," "rare" and "very rare" ADRs associated with a given treatment. However, these methods are more subject to error than those methods used to evaluate efficacy and lack of efficacy. All interested parties should keep in mind the nature of the analyses that lead to the attribution of all but "common" ADRs to a given drug and the potential uncertainty of such attribution. Also, all interested parties should clearly understand the virtual impossibility of "proving" by a conventional gold standard what is or is not an ADR associated with a drug.

It cannot be emphasized enough that for AEs that might or might not be ADRs but of low incidence, it can be impossible to "prove" that a drug is or is not associated with the potential ADR based on the RCTs that are conducted to prove that the drug is efficacious. Probably the best that we can do in the future is to develop stronger prospective epidemiological studies that are initiated soon after a drug is launched. By stronger, we mean studies with exceptionally large numbers of subjects, extended

exposure time frames and rigorous prospective methods for identifying with clinical certainty AEs of interest. An important and interesting question is: What entity would fund such studies? They would be expensive. Advances in data sciences might make such studies more practical and reduce their costs. Such studies are our best chance of ruling in or ruling out a rare but important potential ADR in a faster time frame with a lower probability of false positive and false negative attribution.

References

Beasley CM, 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; 303:685-92.

Beasley CM, Ball S, Nilsson M. Fluoxetine and adult suicidality revisited: an updated meta-analysis using expanded data sources from placebo-controlled trials. J Clin Psychopharmacol 2007; 27:682-86.

Beasley CM. Charles M. Beasley, Jr's comment regarding Blackwell's essay: Corporate Corruption in the Pharmaceutical Industry. inhn.org.controversies. March 23, 2017.

Beasley CM. Charles M. Beasley, Jr's response to Blackwell's reply. Corporate Corruption in the Pharmaceutical Industry. inhn.org.controversies. January 12, 2018.

Blackwell B. Corporate Corruption in the Pharmaceutical Industry. inhn.org.controversies. Sep. 1, 2016.

Blackwell B. Reply to Beasley's comment. inhn.org.controversies. July 13, 2017.

Chen L, Bonson KR. An equivalence test for the comparison between a test drug and placebo in human abuse potential studies. J Biopharm Stat 2013; 23:294-306.

Eypasch E, Lefering R, Kum CK, Troidl H. Probability of adverse events that have not yet occurred: statistical reminder. BMJ 1995; 311:619-20.

Frey N, Jossi J, Bodmer M, Bircher A, Jick SS, Meier CR, Spoendlin J. The epidemiology of Stevens-Johnson syndrome and toxic epidermal necrolysis in the UK. J Invest Dermatol 2017; 137:1240-57.

Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA 1983; 249:1743-5.

Hirshberg B, Katz A. Cardiovascular outcome studies with novel antidiabetic agents: scientific and operational considerations. Diabetes Care 2013; 36(Supplement 2):S253-S8.

International Agranulocytosis and Aplastic Anemia Study. Incidence of aplastic anaemia: the relevance of diagnostic criteria. Blood 1987; 70:1718-21.

Jovanovic BD, Levy PS. A look at the rule of three. Am Stat 1997; 51:137-9.

Kaufman DW, Kelly JP, Issaragrisil S, Laporte JR, Anderson T, Levy M, Shapiro S, Young NS. Relative incidence of agranulocytosis and aplastic anemia. Am J Hematol 2006; 81:65-7.

Ludbrook J, Lew MJ. Estimating the risk of rare complications: is the 'rule of three' good enough? Anz J Surg 2009; 79:565-70.

Malik M. Problems of heart rate correction in assessment of drug-induced QT interval prolongation. J Cardiovasc Electrophysiol 2001; 12:411-20.

Mauri L, D'Agostino RB Sr. Challenges in the design and interpretation of noninferiority trials. NEJM 2017; 377:1357-67

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

Neal B, Perkovic V, Mahaffey KW, de Zeeuw D, Fulcher G, Erondu N, Shaw W, Law G, Desai M, Matthews DR; CANVAS Program Collaborative Group. Canagliflozin and cardiovascular and renal events in type 2 diabetes. NEJM 2017; 377:644-57.

Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. NEJM 2007; 356:2457-71.

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

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

Packer M. Have we really demonstrated the cardiovascular safety of anti-hyperglycaemic drugs? Rethinking the concepts of macrovascular and microvascular disease in type 2 diabetes. Diabetes Obes Metab 2018; 20:1089-1095.

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

Regier EE, Venkat MV, Close KL. More than 7 years hindsight: revisiting the FDA's 2008 guidance on cardiovascular outcomes trials for type 2 diabetes medications. Clin Diabetes 2016; 34:173-80.

Smith RJ, Goldfine AB, Hiatt WR. Evaluating the cardiovascular safety of new medications for type 2 diabetes: time to reassess? Diabetes Care 2016; 39:738-42.

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. https://doi.org/10.1093/biostatistics/kxn034.

UKPDS Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 1998a; 352:837-53.

UKPDS Group. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). Lancet 1998b; 352:854-65.

U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER). Guideline for Industry: The Extent of Population Exposure to Assess Clinical Safety: For Drugs Intended for Long-term Treatment of Non-Life-Threatening Conditions. ICH-E1A. March 1995. https://www.fda.gov/ohrms/dockets/ac/04/briefing/2004-4068B1_09_ICH-E1A-Guidelines.pdf.

U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER). Guidance for Industry: Diabetes Mellitus – Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes. December 2008. https://www.fda.gov/downloads/Drugs/Guidances/ucm071627.pdf.

U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) Center for Biologics Evaluation and Research (CBER). Guideline for Industry: Premarketing Risk Assessment. March 2005. https://www.fda.gov/downloads/ regulatoryinformation/guidances/ucm126958.pdf.

Yang F, Stewart M, Ye J, DeMets D. Type 2 diabetes mellitus development programs in the new regulatory environment with cardiovascular safety requirements. Diabetes Metab Syndr Obes 2015; 8:315-25.

Yang M-S, Lee JY, Kim J, Gun-Woo Kim, Byung-Keun Kim, Ju-Young Kim, Heung-Woo Park, Sang-Heon Cho, Kyung-Up Min, Hye-Ryun Kang. Incidence of Stevens-Johnson syndrome and toxic epidermal necrolysis: a nationwide population-based study using national health insurance database in Korea. PloS ONE 2016; doi: 10137/journal.pone.1065933.

Zinman B, Wanner C, Lachin JM, Fitchett D, Bluhmki E, Hantel S, Mattheus M, Devins T, Johansen OE, Woerle HJ, Broedl UC, Inzucchi SE; EMPA-REG OUTCOME Investigators. Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. NEJM 2015; 373:2117-28.

November 21, 2019