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)

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)[i] 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 non inferiority inferential test, it is necessary to definea 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 non inferiority 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[ii] 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 therapeuticactivities 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 wasused 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 (Hirsbergand Katz 2013; Regier, Venkat and Clo 2016; Smith, Goldfine and Hiatt 2016; Yang, Stewart, Ye and DeMets 2015). In counterpoint, at least one authorhas 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 proportionobserved with control when the proportionobserved 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 non inferiority of AEs ("proof"thatan 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 non inferiority. 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 non inferiority 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[iii]events to occur (rather than subjects studied) based on the assumed hazard ratio and level of statistical significance required. Thus, 4,687 subjectswere 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[iv] events. Hierarchical testing was used in the following order: MACE (deaths from CV events, nonfatal myocardial infarction, non fatal 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 non inferiority, 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.

References:

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.

Guidance for Industry. Diabetes Mellitus: Developing Drugs and Therapeutic Biologics for Treatment and Prevention. U.S. Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER). February 2008.

Hirshberg B, Katz A. Cardiovascular outcome studies with novel antidiabetic agents: scientific and operational considerations. Diabetes Care 2013; 36(Supplement 2):S253-S258.

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. N Engl J Med. 2017;377:1357-67.

Mauri L, D'Agostino RB Sr. Noninferiority Trials. N Engl J Med. 2018;378:304-5.

Neal B, Perkovic V, Matthews DR. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. N Engl J Med. 2017;377:2099.

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

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, Obesity and Metabolism 2018; 20:1089-95.

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; 39:738-742.

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; 10:275–81.

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.

Yang F, Stewart M, Ye J, DeMets D. Type 2 diabetes mellitus development programs in the new regulatory environment with cardiovascular safety requirements. Diabetes, Metabolic Syndrome, and Obesity: Targets and Therapy 2015; 8:315-25.

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. N Engl J Med. 2015;373(22):2117-28.

---

[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.

February 7, 2019