Charles M. Beasley, Jr and Roy N. Tamura

What We Know and Do Not Know by Conventional Statistical Standards About Whether a Drug Does or Does Not Cause a Side Effect (Adverse Reaction)

Foreword by John M. Davis

Forward

by

John M. Davis

This book addresses the statistical analysis of side effects (referred to in the specialty of pharmacovigilance as "adverse drug reactions"). The book focuses on estimating the sample sizes needed for interpreting the cause of an adverse event (i.e., a medical complaint by someone taking a drug or placebo that the drug might cause if the person is taking a drug) to be the experimental drug evaluated in a randomized controlled trial (RCT) or a set of RCTs. Patients in both the experimental drug group and the placebo group of an RCT commonly experience and report medical complaints referred to as adverse events. Which adverse events occurring in the experimental drug group should be considered "real" side effects caused by the drug?

Alternatively, are these adverse events occurring in the experimental drug group an amalgam of the base rate of common medical complaints unrelated to the drug, an effect related to participating in the RCT, in which concern about and anticipation of side effects or symptoms of the disease being studied induce the medical complaints, and signs and symptoms caused by the disease being treated? For example, patients treated for COVID-19 might experience loss of taste due to the disease rather than experiencing this phenomenon as a side effect if treated with an experimental drug in an RCT. How does an individual correctly interpret the occurrence of such adverse events? Interpretations by some individuals can sometimes be incorrect.

Much the same considerations apply to rare serious events. For example, four patients died in the Pfizer COVID-19 vaccination trial control group and two in the vaccine group. But four died in each group of the Moderna vaccine COVID-19 trial. How does one make sense of this information? If someone focused on the number of deaths in these trials, comparing deaths in the active treatment group to deaths in the control group, the Moderna vaccine has more deaths relative to its control group than the Pfizer vaccine. The analysis of the difference in deaths with vaccine versus control can be based on two comparisons. The first is the difference in the incidence of deaths between drug and placebo. The second is the ratio of the incidence of deaths on drug to that incidence on placebo. The Pfizer vaccine had two fewer deaths compared to control, relative

to the Moderna vaccine. The Pfizer vaccine had 50% of the deaths compared to control, relative to the Moderna vaccine. If a layperson considers these numbers, the Pfizer vaccine might appear to them to be safer. However, statistically, the numbers of deaths with the two vaccines compared to control are not significantly different, even when 30,000 subjects were included in both the vaccine groups and the control groups for both vaccine trials.

The authors consider the sample sizes necessary to prove that an adverse event is caused by a drug to essentially the same standard as proving a drug is efficacious. They consider the statistics of estimating sample sizes in the calculations involved in both proving a drug does and does not cause a side effect. These sample sizes grow inordinately large as the incidence of an event in both the drug and placebo group decline together and/or the background rate of the adverse event increases.

Dr. Beasley, a psychiatrist, and Dr. Tamara, a statistician, worked at Eli Lilly and Company for more than 20 years managing the statistical issues involved in measuring and interpreting side effect data and are respected by both industry and academia for their expertise, comprehensive knowledge and integrity. The book begins with the statistical section, which deals with the calculations of estimating sample size and interpretation of statistical inference in such work. They discuss common side effects, uncommon side effects, rare side effects and extremely rare side effects. This discussion contextualizes several types of side effects with a detailed discussion of inferring causation. For example, does olanzapine directly cause diabetes?

The book's coverage of these topics and their details is an excellent introduction to statistical significance, power analysis and sample size estimation issues for anyone analyzing RCTs, including RCTs in multiple therapeutic areas. Providing medical examples places the statistical discussion in real-world problem areas, making it a somewhat difficult read. I would highly recommend this book to those individuals interested in these complex issues. This book is technical, directed toward specific statistical analysis and inference about causation. However, the book also considers a more significant, more general matter of how seriously to take adverse events that might or might not be side effects. The authors suggest two factors that should influence the extent to which such adverse events should be seriously considered. The first factor is their clinical significance. The second factor is the extent to which the demonstration that the adverse event might be a side effect conforms with the standard required for a demonstration of efficacy.

I want to discuss these matters in the context of behavioral economics as exemplified by the work of Daniel Kahneman and described in his books *Thinking, Fast Slow* (2011) and *Noise: A Flaw in Human Judgment* (2021). Kahneman, a psychologist, won a Nobel Prize in economics based on the implications for the field of economics in these books. He postulates two types of thinking: System 1 (fast thinking) and System 2 (slow thinking). Since System 1 takes place in the blink of an eye, is automatic, effortless and cannot be turned off, you cannot help being influenced by the biases that automatically come to mind. System 2 thinking takes effort and concentration over minutes, hours, days or years.

System 1 is what we use most of the time and works reasonably well, but cognitive errors happen ubiquitously, such as availability bias, confirmation bias and loss aversion. System 1 thinking is not conducive to good mathematical thinking. System 1 deals poorly with rare and uncommon events, generally overweighting them, especially if the event is vivid. As a result, rare events are sometimes overweighted, are, conversely, sometimes ignored. The human brain is not good at quantifying and conceptualizing uncommon events. The inability to quantify and conceptualize uncommon events is complicated by the uncertainty and interpolation from limited quantitative data.

This difficulty with System 1 thinking is significant from a clinical perspective. It can cause serious harm: patients who overweigh the occurrence of adverse events (whether rare or not) are more likely to be noncompliant regarding pharmaceutical interventions that benefit most patients. Also, reviewers of data, be they the sponsors of RCTs, academics or regulators, can be lead astray in their interpretations through System 1 thinking. Should the same standards of statistical proof be used to assess the efficacy and determine the adverse events reported in an RCT (or set of RCTs) as side effects? The book provided the tools for making calculations to estimate the probability of whether a side effect is caused by a drug, not caused by the drugs or somewhere in-between.

In *Noise: A Flaw in Human Judgment*, Kahneman and his co-authors discussed the role of noise and bias in decision-making, which in this case, would involve decisions about what adverse events are side effects and the choice of clinically used drugs used when comparing alternative medicines from a safety perspective. Empirical research shows a large amount of variability in decision-making and many major and minor biases can influence a given decision, causing noise

and systemic biases. One can research which factors drive a decision, assess whether they are valid and use feedback to improve decision-making to align with sound science and statistics. Reduction of noise can also considerably improve the decision-making process and can be done immediately – no long-term follow-up is needed. Most importantly, reducing noise can improve decision-making when it involves a value judgment, even though different people might have different values. This existence of cognitive bias is supported by a massive number of controlled trials and other evidence. As said above, Beasley's and Tamura's book provides tools to understand better the statistical principles involved in optimizing the estimates of the probability that a given adverse event is caused, not caused by a drug or causation is uncertain.

Should risk and benefits be weighted equally? Of course, patients should decide for themselves. But their doctors, in shared decision making, play a crucial role in shaping patient decisions. Because of this, the standards of information in medical journals and the Food and Drug Administration labeling have an important impact on the doctor's recommendation. Drug labeling influences clinical decisions about drug use. Therefore, should institutions such as the Food and Drug Administration shift their current practices to a more detailed and nonbinary description of the probability that adverse events are side effects by describing the probabilities that listed side effects are, in fact, side effects? Obfuscating the answer to this question has considerable influence on the knowledge base of a given drug. This question of how best to describe individual side effects in drug labeling differs from the value judgment of how much weight to assign to a given side effect in describing a drug's risk-benefit profile. More specifically, drug labeling should list the percentage of patients experiencing adverse events believed to be side effects in drug versus placebo groups and the statistical significance level and/or confidence interval for this drug-placebo comparison. There should be some indication of the probability and confidence around that probability that adverse events listed as side effects are genuinely side effects. Listed adverse events could then be grouped as proven side effects, possible but unproven side effects or unlikely to be side effects (but listed in the drug label due to an abundance of caution due to the clinical significance of the adverse event).

Furthermore, some adverse events might be improved by a drug. Calling something a "side effect" or "adverse drug reaction" implies it is caused by the drug when it may not be. Knowing the probabilities, based on the cumulative controlled database for the drug that an adverse event

listed as a “side effect” / “adverse drug reaction” along with the clinical significance of the adverse event, would clarify the picture, decrease the noise and allow for a more balanced assessment of risk versus benefits. Such practice would improve medication decisions.

Drug companies vigorously promote their drugs, but side effects do not receive as much focus. Availability bias may favor the drug unless a side effect is newsworthy. The discovery of a side effect and, better still, treating it should not be considered less important than any other medical discoveries.

There is still much to be done to improve this process of drug labeling. Using the same standard of proof for side effects as efficacy and describing those adverse events of potentially significant clinical consequence would be a major step toward developing a balanced description of risk. As the authors point out, some adverse events that do not meet the efficacy standard of proof must be described as possible side effects due to their clinical significance (e.g., potentially fatal, potentially leading to permanent disability). This book provides a statistical text, but more importantly, it raises these critical issues.

References:

Kahneman D. Thinking, Fast and Slow. New York: Farrar, Status and Giroux; 2011.

Kahneman D, Sibony O, Sunstein CR. Noise: A Flaw in Human Judgment. New York: Little, Brown; 2021.

Table of Contents

Preface

To understand this book's purpose, the reader must first understand one specific requirement for an experimental drug to be approved by the United States Food and Drug Administration (FDA) to treat a medical disorder (i.e., become an FDA-approved drug). This crucial specific requirement is the amountof 'proof' required to show that the drug treats the medical disorder effectively. This amount of 'proof' is different across various classes of medical disorders, the potential size of the population that might be treated with the drug, the expected length of treatment with the drug, and other factors.

What follows in the next two paragraphs is well known to health care professionals and those with training in research methods and basic statistics but might be unfamiliar to others reading this book.

Our focus is on a drug that would treat a non-life-threatening disorder for a lengthy period. This lengthy period can be for the rest of a patient's life after treatment was begun. For such a drug, the amount of 'proof' required is a set of at least two randomized controlled trials (RCTs) conducted in precise ways and with a specific outcome.[1] The RCTs compare the experimental drug to placebo control. For some medical disorders, only the experimental drug or the placebo control are given to the patients taking part in an RCT. For some medical disorders, most patients would already be taking an approved drug for the medical disorder, and t is considered medically inappropriate to discontinue an ongoing treatment. An experimental drug or placebo is added to continued treatment with the approved drug in an RCT for these disorders.

As part of the RCT design, something is defined as a primary measurement instrument for change in the medical disorder's severity. For example, if the drug is being developed to treat hypertension, systolic and diastolic blood pressures define the disorder's change. If the medical disorder is schizophrenia, a standard indicator would be the PANSS-Total score. A demonstration of efficacy for the experimental drug would be greater improvement with the experimental drug than with the placebo control. For some medical disorders, greater improvement is simply a greater improvement in mean change on the primary measurement instrument. Some change on

[1] In some cases of a drug in this class, only one RCT is required. Only one is required if the results of that trial demonstrate 'overwhelming efficacy' (described in more detail in Chapter 2 – Definition of 'Proof' of a specific drug effect (and 'proof' of the absence of a specific drug effect) for the drug.

the instrument is defined as a clinically meaningful change for other medical disorders. The experimental drug group must have more subjects meeting this criterion for clinically significant improvement than the placebo-control group.

Statistical tests are applied to the RCT's results, comparing the change with the experimental drug to the change with placebo control. Statistical tests produced confidence intervals (CIs) and or p-values. Suppose the experimental drug results in greater improvement in the medical disorder, and the CI and/or p-value from the statistical test of the greater amount of improvement to have only a 5% or less probability of being a chance occurrence. In that case, the RCT is interpreted as 'proving' that the experimental drug is effective. The p-value representing this probability is $p \leq 0.05$. A CI can also represent this probability. Throughout the book, we describe the results of statistical tests in terms of p-values.

The primary purposes of the book are to:

1. illustrate the sample sizes needed to infer with reasonable medical certainty that a drug causes a side effect when using the same standard as required for 'proving' efficacy required by FDA to approve a drug for the treatment of a medical disorder;
2. illustrate the sample sizes needed to infer with reasonable medical certainty that a drug does not cause a specific side effect, although the medical problem that could have been a side-effect was observed while the drug was administered to a group of subjects in an RCT when using the same standard as required for 'proving' efficacy required by FDA to approve a drug for the treatment of a medical disorder; and
3. clarify for all relevant parties what these sample sizes mean for the medical problems listed in regulatory product labeling as 'adverse drug reactions' (a technical term suggesting that the medical problems are side effects) for the quality of the 'proof' that these medical problems are side effects;
4. demonstrate to all relevant parties who read and use product labeling that for some of the medical problems listed as 'adverse drug reactions' that their status as side effects has not been 'proven' using the same standard as required for 'proving' efficacy required by FDA to approve a drug for the treatment of a medical disorder;
5. emphasize to all relevant parties that because some medical problems listed as 'adverse drug reactions' lack 'proof' of being side effects using the same standard as required for

> ‘proving’ efficacy required by FDA to approve a drug for the treatment of a medical disorder, the possibility exists that some of these medical problems are not side effects.

This book is not arguing that any medical problem associated with a drug that lacks proof of being a side effect using the same standard as required for ‘proving’ efficacy required by the FDA to approve a drug for the treatment of a medical disorder should not be listed as an ‘adverse drug reaction’. There are reasons why some medical problems lacking this robust magnitude of ‘proof’ of side effects status should be listed as ‘adverse drug reactions’. This point should be kept in mind and is hopefully made clear throughout the book. Chapter 10, a Postscript, was written to emphasize this point at the book’s conclusion.

Hopefully, this book improves the understanding of the ‘adverse drug reactions’ sections of product labeling documents and leads to improved use of these documents in many contexts. The most important context is patient care but ranges through others, including malpractice litigation against prescribing health care providers and product liability litigation against pharmaceutical manufacturers.

Other research methods can supply high-quality data and results useful in ‘proving’ whether a medical problem is a side effect. These other methods are most relevant when an RCT of sufficient size and/or length could not be conducted to study some potential side effects. Some of these methods are discussed in the book chapters and online, interactive discussions about the book chapters. These discussions were initially posted individually on an internet discussion forum. The discussions are included in the book and follow the book chapters.

There were two stimuli for this book. Both stimuli grew out of an online discussion posted on the online discussion site that posted the book chapters, but a series of postings that preceded the book chapters was the first stimulus.

The International Network for the History of Neuropsychopharmacology (INHN) website (www.inhn.org) is an online discussion site. INHN issues multiple postings each week about topics relevant to history and matters of current interest in neuropsychopharmacology. The INHN website is an interactive vehicle for discussing and debating topics and ideas pertinent to neuropsychopharmacology. Any person can read the postings on the website and the archives of all postings. Writing and posting to the site, including comments and questions about others' postings, is limited to members of INHN.

This book builds on topics I briefly addressed in an online exchange with Barry Blackwell on INHN (references in order of their posting: Blackwell, 2016; Blackwell, 2017a; Beasley, 2017; Blackwell, 2017b; Beasley, 2018). Barry's posting, Corporate Corruption in the Psychopharmaceutical Industry, was the first posting in the dialogue. My response posting of 2018 to Barry contains the material that provided one of the two stimuli for this book.

In this 2018 response to Blackwell, I supplied sample sizes for a hypothetical study. The purpose of providing these sample sizes was to illustrate the difficulty in 'proving' that an adverse event (AE, any undesirable change in a person's medical/health status [referred to simply as a medical problem above]) observed in a placebo-controlled, randomized clinical trial is an adverse drug reaction (ADR, an AE directly or indirectly caused by the drug in the study).[2] The hypothetical study was of sufficient sample size to provide 80% power to 'prove' that an AE is an ADR. Causation would be tested under the null hypothesis that the AE was coincidental and not caused or contributed to by the drug. The standard for 'proving' that the AE is an ADR and not simply a coincidental AE is comparable to the standard required to gain regulatory (United States Food and Drug Administration [FDA]) approval for a claim of efficacy for a drug intended to treat a non-life-threatening disorder, administered on a long-term basis[3].

The background incidence (incidence of cases where the drug under study does not cause or contribute to the cause of an AE; the incidence that would be observed in a placebo-treated control group) has a substantial influence on the sample sizes required to 'prove' that an AE is an ADR. In placebo-controlled studies, this background incidence of the AE would be observed in both the drug-treated and the placebo-treated groups. If the drug did cause cases of the AE (ADR cases), the incidence of the ADR cases would be added to the background incidence for the total incidence observed in the drug-treated group.

---

[2] Adverse event is another technical term used to refer to a medical problem that occurs during treatment with a drug that might or might not be a side effect (technically an adverse drug reaction). The two technical terms, adverse event and adverse drug reaction are used throughout the rest of the book and additional components of their definitions are provided in Chapter 2

[3] The sample sizes we provided were for a single study. However, as noted above, FDA usually requires two studies with statistically significant demonstrations of efficacy to grant marketing approval for a drug intended for use as described in product labeling for a drug.

The sample size calculations were for an ADR with an incidence of 1 in 1,000 (0.1%) subjects treated with the investigational drug.

Background incidence on the required sample sizes has a considerable influence on sample size requirements, especially if an ADR does occur but with low incidence. Therefore, I considered two background incidences, virtually zero (0%) occurrence in a substantial placebo-treated population (the probable background incidence [cases not caused by some drug or toxic exposure] for the life-threatening dermatological disorder of toxic epidermal necrolysis), and 50 in 1,000 (5.0%) in a placebo-treated population (a realistic background incidence of myocardial infarction in an older diabetic population in a study of one to two years).

In this response (Beasley, 2018), the sample sizes were enormous, particularly with the 5% background incidence. With a background incidence of 5% and an ADR incidence of 0.1%, the study would need to be of sufficient total sample size to prove that the difference between 5.1% and 5.0% was not a chance observation. A definitive study where the background incidence of the AE of interest is 5%, and the AE incidence as an ADR in the drug-treated group is 0.1%, would be impossible to conduct due to the required sizes in the experimental groups.

I neglected to say that smaller sample sizes with specific study outcomes would be sufficient to 'prove' that the AE is an ADR. Over the months after writing and posting my response (Beasley, 2018), I decided it was necessary to acknowledge that smaller sample sizes would be sufficient with specific outcomes in this hypothetical study and explain these outcomes. The first stimulus was my perceived need to be more objective, balanced, and comprehensively discuss what I had said about these sample sizes.

The second stimulus for the book is based on my belief that many parties have a poor understanding of the credibility of drug labeling of ADRs. This stimulus was noted when I described the purpose of the book. This misunderstanding would be especially relevant for health care providers prescribing drugs, members of the media, legislators, attorneys representing both plaintiffs and defendants in both malpractice actions brought against individual practitioners, as well as product liability actions brought against pharmaceutical companies. I wanted to produce a book that would explain the scientific basis for better understanding and hopefully improve it.

The book's stimuli are also described in a 'response' to a 'comment' by Barry Blackwell about my 'response' to Edward Shorter. In his comment, Shorter expressed interest in data evolution and

analyses addressing the relationship between olanzapine and diabetes mellitus. This ‘response’ to Blackwell is Part d of the first ‘chain’ of ‘comments/’questions’ and ‘responses’/’replies’ that all follow the chapters and References for the chapters (see the Table of Contents).

Both my education and work experience influenced the book. My undergraduate education was in psychology and computer science. With psychology, I completed courses in experimental design and basic statistics. With computer science, I completed courses at the graduate level in discrete mathematics. Before medical school, I worked for one year as a research programmer with an artificial intelligence project developing machine learning. The following year, I designed a database management system and analysis tools for an EEG evoked potentials laboratory. My residency training was in general psychiatry.

I began my career with Eli Lilly and Company immediately after completing my residency and was with Lilly for 27 years and 10 months. At Lilly, I had responsibilities for designing and interpreting the results of many RCTs along with my statistician colleagues. I also designed multiple complex *post-hoc* analysis plans that incorporated data from multiple RCTs and other data sources to address the question as to whether a specific AE was or was not an ADR for a specific drug. During my last 12 years at Lilly, I designed standardized analysis systems adopted as analysis standards across multiple drug development programs. I am familiar with using multiple statistical software systems to compute sample sizes for various RCT analysis methods. Advising on sample sies for RCTs has been one part of my consulting work following my Lilly retirement.

However, I am not a formally trained statistician at a doctoral or master's level. Although I performed the sample size computations presented in the book, I asked for the aid of my coauthor, Roy Tamura, Ph.D., now an academic statistician with whom I had the privilege of working at Lilly for 20 years. Roy independently validated all my sample size computations, read all the chapters, and made valuable suggestions for changes in the text incorporated in the chapters’ first postings.

I reviewed the chapters, initial postings, edited them to improve readability, correct typographical errors, and made what I believed were necessary changes in content.

I most extensively reviewed my response to Edward Shorter’s comment, which addressed the evolution of data and study of what I consider to be the most relevant research on whether second-

generation antipsychotics, specifically olanzapine, cause the ADR of diabetes mellitus. This response forms the first of the seven chains that follow the chapters.

This response is a lengthy summary of the evolution of Lilly's complex analyses investigating the relationship between olanzapine and treatment-emergent hyperglycemia/diabetes. My response also reviews the hyperglycemic clamp studies and the euglycemic-hyperinsulinemic clamp studies (the gold-standard methods for assessing pancreatic β-cells' ability to respond to increased peripheral glucose with the appropriate production and release of insulin into the systemic circulation and assessing the ability of peripheral tissues to take up glucose as well the suppression of glucose production by the liver during increases in plasma glucose). I edited this response to improve its understanding. In this 'response', I updated a hypothetical study's description to include an additional subject group. I also corrected what I considered my errors in describing several studies' assessments and results in the original posting. Some of what I considered errors resulted from insufficient detail in my descriptions. This response to Shorter involves extremely complex study methods and is longer than the combined book chapters.

I wrote all responses/answers to the comments/questions alone without Roy's review.

I added one reference to my response in the second chain, begun by Edward Shorter's comment about using the dechallenge-rechallenge method to determine if an AE is an ADR. I suggest that a formal 'N-of-1' study is a superior method to the dechallenge-rechallenge method. I use the disorder of cyclic neutropenia as an example where dechallenge-rechallenge might lead to the false conclusion that a patient with cyclic neutropenia was experiencing neutropenia as an ADR. I added a reference to this disorder to the two references included in this chapter's original posting.

I performed limited editing of what was written by others in their comments/questions and further comments on my responses/replies. I ensured that the formatting was consistent. Finally, I confirmed and edited, if necessary, the references and expressed them in a consistent style.

Any errors of fact or errors due to omission are solely my responsibility as the first author.

I thank my wife, Rebecca L. Bushong, M.D., for her many hours assisting in validating the 17 tables in the book and reviewing each chapter as they were written and a complete review of the final draft before submission for publication. I also thank Tom Ban and Olaf Fjetland for their

editorial reviews of each chapter and my responses to comments and replies to questions before they appeared originally as postings on the INHN website.

Barry Blackwell provided me with the motivation to author the book, as described above, based on my perceived need to clarify my comments on his work Corporate Corruption in the Psychopharmaceutical Industry (inhn.org., September 1, 2016). Tom Ban encouraged me to author the book and kept me on course when the work was frustrating. Edward Shorter and Barry Blackwell stimulated my review of the data relevant to the association between second-generation antipsychotics, specifically olanzapine, and the treatment-emergence of diabetes mellitus. I dedicate this small book to them.

Charles M. Beasley, Jr., M.D.
January 4, 2021

Introduction

Hopefully, a description of the book's organization helps the reader follow the books' contents, especially the online, interactive discussions that follow the book's chapters. The book's first nine chapters were separate postings every two weeks on the INHN website. Therefore, what are now Chapters 1-9 were posted between November 29, 2018, and April 4, 2019. I wrote the postscript to clarify my belief on what standards should be applied when deciding AEs that should be included as ADRs in product labeling. The full text (Chapters 1-9), except for the Postscript (Chapter 10), was posted as a collation on November 21, 2019. The Postscript was posted before the posting of the collation on October 24, 1919.

Because each chapter of the book appeared as a separate posting on the INHN website every two weeks, Chapter 1 outlines the chapters to follow and resembles a table of contents for the chapters that follow. The title of Chapter 3 is Introductory Comments and covers material contained in the Preface above. Chapter 3 also describes the hypothetical experiment and the assumed background incidences that were used in the sample size calculations that provided one of the stimuli for this book

Several postings/chapters generated online 'comments' or 'questions' from other persons who have access to post on INHN. I 'responded' to most 'comments' and 'replied' to 'questions'. In some cases, the postings/chapters generated 'comments' or 'questions' from multiple parties. The 'comments'/'questions' and 'responses'/'replies' constitute the book's contents following the chapters and a single set of References for all chapters. The meanings of 'comment', question', 'response', 'reply, and 'chain' are explained in the second and fourth paragraphs below.

As the postings/chapters appeared over 17 weeks, 'comments'/'questions' followed by my 'responses'/'replies' were posted between the postings/chapters.

Each 'comments'/'questions' and 'responses/'replies' set about specific book chapters or topics are grouped, irrespective of their posting date, and are referred to as 'chains'. This organizational scheme was chosen to enhance an understanding of the logical flow for a specific topic. If 'comments'/'questions' and 'responses'/'replies' were included in their postings' temporal order, these materials could not be followed in a logical flow.

These chains are ordered by the posting date of the first ‘comment’/’question’ in the chain. The ‘response’/’reply’ to that first comment appears next in the chain. ‘Comments’/’questions’ and ‘responses’/’replies’ to those additional ‘comments’/’questions’ appear in order of the posting date for each subsequent ‘comment’/’question’ within the chain. The additional ‘responses’/’replies’ follow immediately after those ‘comments’/’questions’.

As initially posted, the labels of ‘comment’ and ‘question’ were not used consistently across the INHN postings. Also, the labels ‘response’ and ‘reply’ were used inconsistently across these postings. For all the book’s references and the chains of interactions directed at the book’s authors, I have labeled anything that was not an explicit question a ‘comment’ and labeled the one explicit question a ‘question’. I have labeled all text responding to a ‘comment’ a ‘response’, and the text answering the one ‘question” a ‘reply’.

I have given brief descriptive titles to each chain intended to describe the focus of what is being discussed in the chain. This brief title is preceded by a list of authors and designations of the authors’ contributions as comments/questions or responses/replies. Hopefully, this organizational schema helps the reader follow the material's logical flow after the book chapters.

1. Outline

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

6. illustrate the sample sizes needed to infer with reasonable medical certainty that a drug causes some adverse medical event ('prove' an effect);
7. illustrate the sample sizes needed to infer with reasonable medical certainty that an adverse medical event, while observed during administration of a drug, is not caused by the drug; ('prove' the absence of an effect); and
8. clarify for all relevant parties what these sample sizes mean for the adverse medical events listed in regulatory product labeling as adverse drug reactions for the quality of the 'proof' of the adverse drug reaction status of these adverse medical events.

We focus on adverse medical events infrequently observed in temporal association with drug administration that are likely to be medically serious (e.g., are fatal, are life-threatening, can lead to complicated and prolonged hospitalization, are potentially permanently disabling). The point made in illustrating these sample sizes is that the inference that a drug causes or does not cause an AEs is often not based on robust empirical evidence for such adverse events. Furthermore, obtaining such robust medical evidence would be a practical impossibility.

The book progresses in subsequent chapters following this Outline chapter (Chapter 1) as follows:

2. a chapter that provides definitions of technical terms that have a precise meaning in the domain of drug safety/pharmacovigilance that are used in the book;
3. an introductory chapter that restates our purposes and briefly describes some complexities of the time course of observation of an adverse medical event 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 most straightforward case in chapters that follow);
4. a chapter that discusses variability that can occur when a subset of a population of interest is selected for a study compared to what would be observed in the total population if it were studied (such variability is an essential topic as it is relevant to an understanding of the sample size computations and as a particular case of this variability, we discuss what

can be inferred when no events or outcomes of interest are observed when only a subset of a population of interest that is studied, embodied in the statistical Rule-of-3);

5. a chapter that discusses sample sizes in studies where the objective is to infer that an effect occurs under the assumption that the effect does not occur (i.e., determine that the drug causes an event);
6. a chapter that discusses sample sizes in studies where the objective is to infer that an effect does not occur under the assumption that the effect does not occur (i.e., determine that the drug does not cause an event);
7. a chapter that illustrates the extreme rarity of events that would be of interest in assessing the safety of a drug (provides the context for understanding the incidence of an event associated with a drug used in our sample size calculations);
8. a chapter that discusses regulatory requirements for drug exposure (number of patients) in development programs for drugs used on a long-term basis to treat disorders that are not acutely life-threatening and regulatory authorities' acknowledgment of the limitations of such sample sizes in determining with reasonable certainty what ADRs a drug causes before its approval;
9. a chapter briefly describing methods used 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 proper controls; and
10. a postscript chapter that supplies clarifying comments.

2. Definition of Terms Used in this Book

- **Adverse Event: (AE)** – an adverse or untoward medical event (complaint, symptom, sign, syndrome, disorder, disease) that occurs or worsens in temporal association with 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 with '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 quantified by any regulatory entity or drug safety organization, including:
    - the U.S. Food and Drug Administration (FDA) or other national regulatory agencies;
    - the 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' member companies); or
    - the 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 a quantity of ≤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. However, prospective and retrospective epidemiological studies, post-marketing surveillance, and multiple other data sources contribute to sponsors' and regulatory bodies' decisions about what AEs are ADRs and should be included 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 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 the neuroleptic malignant syndrome [NMS]). This ADR's potential is contained in product labeling (Warnings and Precautions section) of the US labels 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 NMS development.

- **Incidence categories of ADRs (and AEs observed in a clinical trial):**
  - Very common (frequent): ≥ 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 specific drug effect (and 'proof' of the absence of a specific drug 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.

3. Introductory Comments

For 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 (i.e., 'prove' – see Chapter 2), 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 an RCT 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 in studies of the sizes currently conducted? What would be the size of studies need to 'prove' that a rare AE is an ADR or 'prove' that a rare AE is not 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 we focus on is a highly uncommon ADR with an incidence of 1 per 1,000 persons treated (0.001 or 0.1%). This incidence of events (AEs and ADRs) is the lower boundary of the incidence range for '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 a considerable number of individuals will not experience it 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 substantially affect the extent to which a specific study design, sample size, and method of 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 the 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 with sufficient numbers of subjects followed for sufficient, often lengthy periods; and
3. random occurrence with equal distribution across the 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 occurrence rate (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 are essential considerations in discussing the limitations of attempts at such 'proof'. In Chapters 5 and 6, we discuss sample sizes needed 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 the temporal pattern of occurrence #1 above for an ADR (especially if there is a short lag time between initiation of treatment and the first occurrence of an ADR case of the AE). An ADR with the temporal pattern of occurrence of #2 would usually result in the requirement for prolonged periods of observation (a longer RCT) than temporal pattern #1 and therefore require more subjects to begin a definitive RCT to account for subjects discontinuing the RCT before the planned end of observation and the more frequent occurrence of the ADR during later treatment. A relative infrequent or rare ADR occurring with temporal pattern #3 would require an even more extended observation period in a definitive RCT. Therefore, the sample sizes discussed in chapter 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 a prolonged period of observation or in a random pattern over time but very infrequently overall.

Additionally, any occurrence pattern that is a change as a function of time might require special statistical techniques (beyond comparing incidences or assessing the ratio of the incidences) to 'prove' the ADR's presence or absence. Therefore, RCTs needed to address the complexity of

changes in the ADR rate over time are likely to require larger numbers of subjects beginning such an RCT because human subjects discontinue participation, as previously said.

There is one final caveat regarding patterns over time. As events become lower in incidence, they often appear to be randomly distributed over time. Therefore, there are never enough cases observed to discern a temporal pattern within RCTs of practical size, even if a pattern exists. When most rare AE occurrences are ADRs, the AEs/ADRs appear to occur with temporal pattern #3 unless a sufficiently substantial number of subjects are included to discern temporal pattern #2 when that is the pattern of occurrence.

4. Potential Sampling Error in an RCT, What We Learn from the Lack of Occurrence of an AE in an RCT (Rule-of-3), and Impact of a Potential Sampling Error on Sample Size Calculation Results

An RCT or set of RCTs samples only a subset of the entire population of interest as subjects. Interpretations of RCTs' results are then extrapolated to the entire population of interest, which is the essence of clinical research. Even with the best methods randomly allocating subjects to the treatments in an RCT, the observations in the RCT (within treatment changes and between treatment differences or differences in changes) can differ from what would be observed if the entire population of interest were studied in the RCT. The statistical 'Rule-of-3' (Hanley, 1983; Eypasch, 1995) 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) were 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 causes 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 with a 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 ADR occurrences, 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 sampled the entire population perfectly, we might observe one case of this ADR. However, if we sample only 1,000 subjects, we are highly likely to obtain a sample where the ADR incidence differs from the incidence if the entire population was studied. 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 not observing an outcome when only a subset of the population of interest is studied.

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 the AE "Bad-Thing", we can conclude with 95% probability that the true incidence of AE "Bad-Thing" is at most <1/334 subjects (this 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 subjects.

- Extrapolation: We are studying only a subset of the population of interest, and our sample might have an incidence of the ADR of interest 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 the AE 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 substantial number of subjects studied (e.g., ≥100) (Jovanovic, 1997).

Note that the two variants of the 'Rule-of-3' only address not observing a solitary 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 entire population of interest compared to what would be observed if the entire population of interest were studied is essential in understanding the sample size computation results that Beasley provided in response to Blackwell (2018). Sample size computations consider the potential for what is observed (incidence of an AE) in the sample selected to deviate from what would be observed if the entire population of interest were 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 results in p-values <0.05 if the experimenter was:

1. sufficiently lucky to guess the incidences with the experimental treatment and control treatment that would be observed if the entire population were studied (with 50% assigned to the active treatment and 50% assigned to the control treatment;
2. used these incidences in #1 for the treatment and control groups in sample size calculations; and
3. sufficiently lucky to select:
    a. a subset receiving the experimental treatment with the observed incidence equal to or greater than the expected incidence if the entire population were treated; and/or
    b. a subset receiving the control treatment with the observed incidence equal to or less than the expected incidence if the entire population received control treatment;

In other words, smaller sample sizes than those obtained with an 80% or 90% power sample size computation are sufficient to 'prove' that an AE is an ADR if one is lucky enough to guess the observed outcome incidences and uses these incidences in the sample size calculations. However, one might not get lucky with sampling and miss 'proving' that an AE is an ADR without sample sizes that supply 80+% power even if one is lucky to guess incidences for both treatment and control groups using the entire population. Even with perfect guessing of the actual incidences in the treatment and control groups using the entire population, there is no guarantee that the true difference would be proved with any statistical power if only a subset of the population is studied. To be guaranteed of detecting this true difference, the entire population would need to be included in the study. Sample sizes resulting from 95% power should have a 95% probability of detecting this true difference. For very infrequent or rare AEs that are ADRs, even when there is a low background incidence of the AE, studies with 95% power are a practical impossibility[4].

[4] If the incidence of the AE that is an ADR is quite high (e.g., substantially greater than 10%) and the background incidence does approach 0%, a study with 95% power might be a practical possibility.

5. 'Proof' of the Presence of an ADR (Statistically Significant Excess in the Experimental Group Compared to the Control Group): 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 (Beasley, 2018)[5] and computed the sample sizes for 51% power employing PASS 15.0.6 (PASS, 2017) software. We then performed the conventional inferential test (Fisher's Exact, 2-sided) employing NCSS 12.0.5 (NCSS, 2018) software. The results illustrate that one might get lucky and 'prove' an AE is an ADR with fewer subjects than the number of subjects needed by any power ≥51% (see Table 1 below[6]).

The sample size below for 80% power is somewhat less than Beasley reported in response to Blackwell because for this work below, 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 a long runtime. Some of the sample size computation results in Table 1 required six days. The computations were 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 actual incidences are 1 in 1,000 with drug and 0 in subjects treated with placebo.

---

[5] All sample size computations results presented in this book were computed in PASS (2017) and were then validated in nQuery Advanced (2018).

[6] To perform the sample size computation, one cannot use a 0.0 incidence for the control group in PASS 15 software (but 0 events can be used in a Fisher's Exact inferential test computation within NCSS 12 software). We set the incidence for drug at 0.001 ($1x10^{-3}$, 0.01%, 1 in 1,000) and the incidence for control at $1.0x10^{-15}$ that is effectively 0.0 in the sample size computation.

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 smaller than the sample size of 9,742 previously reported by Beasley in his response to Blackwell (2018). However, sample sizes of 7,905 per treatment group are still large sample sizes and are practical impossibilities in RCTs evaluating psychiatric medications. These are the sample sizes with a treatment-placebo comparison. These are not the sample sizes for a study where the experimental treatment and placebo are added to standard active treatment. This latter design would commonly result in background incidences of many AEs of interest due to the potential to observe ADRs as responses to the standard active treatment and/or drug-drug interactions between the experimental and standard active treatments. Therefore, the latter design would usually require even larger sample sizes. The latter design is considered the only ethical design for many medical disorders.

Acknowledging that under certain circumstances, a sample size with only 51% power (or even lower power) supplies improved transparency, building on Beasley's response (2018). However, for any hypothesis explicitly tested in an RCT, the power used in computing the sample size is commonly 80%. The power used might be higher if a critical hypothesis is being investigated, as implied above. It is also not common to compute sample sizes using binomial enumeration because of the time needed 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 massive 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 a cardiovascular [CV] drug, diabetes drug, or seizure maintenance treatment drug [where the add-on design is used exclusively]), that number of subjects exposed usually 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 formal meta-analysis would consider differences across the RCTs and differences in study size to compute the inferential statistical result. Therefore, a proper meta-analysis usually 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 are for a single RCT.

Also, as Beasley (2018) pointed out, there is often a background incidence (events not caused by the drug under study observed in both the treatment and control groups) 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%[7] would be observed in the control group and the drug group due to causes other than drug) with an additional 0.1% (therefore, total observed event incidence of 0.5% with control vs. 0.6% with drug) observed in the drug group due to drug causation/contribution. In this scenario, the 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 incidences of 1% and 2%, expressed as a percent, is 1% (2% compared to 1%),

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

while the ratio, expressed as a percent of the two incidences 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 critical.

Additionally, with low incidence AEs, inferential analyses are most often conducted using multiple RCTs where it is likely that the AE of interest is not observed (0% incidence) in one of the treatment arms being compared, and in a portion of the RCTs in none of the treatment arms in a study. Both outcome cases complicate using 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 and has one or more arms with >0% incidence, a small incidence must 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 not observed in any treatment arms being compared, the entire RCT is excluded from the meta-analysis. In such a case (no event with any treatment in a study), significant amounts of meaningful data are 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 these meta-analyses of rare events of interest (Tian, Cai, Pfeffer, et al., 2009). In assessing 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, Dornseif, Bossomworth, et al., 1991; Beasley, Ball, and Nilsson, 2007). However, it is 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 Log-rank 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 most often used for AEs when performing a noninferiority analysis ‘proving’ the absence of an effect (i.e., the absence of an ADR) described in more detail in the next chapter.

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 analyses 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 comparator 'known' not to have an ADR of interest – incidence due to control approaching 0]). The following were additional specifications for each procedure:

- Fisher's Exact Test:
    - Test drug observed incidence: 0.001 ($1.0x10^{-3}$, 1 in 1,000, 0.01%)
    - Control observed incidence: $1.0x10^{-15}$ (cannot set to 0.0 for sample size computation)
    - Computation by binomial enumeration (where computed sample size for both treatment groups ≤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 AE 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[8])
        - 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

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

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 observation period. 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 the RCT) to observe 100% of the incidence of events for the two treatments reflected in the probabilities of an event for each treatment. Early discontinuations 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, it still grows to 4,078.

Sample sizes are smaller with a Cox Proportional Hazards Model analysis. However, with either of these inferential test methods, the required sample sizes are large. If multiple studies are used in a meta-analysis (most often required for assessing a very uncommon AE), the total sample size increases, as mentioned briefly. To assess 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 'proving' a treatment effect) are particularly challenging.

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 needed to 'prove' efficacy for drugs intended to treat

non-life-threatening conditions unless there is overwhelming evidence of efficacy in a single RCT. This replication requirement is an excellent, conservative requirement protecting against a Type 1 error in a single RCT from a rigorous scientific perspective. From our perspective, the assertion that any AE is an ADR with full scientific rigor would require the same level of evidence as required for an efficacy assertion. We are not suggesting labeling of ADRs should require the same degree of 'proof' as needed 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 clearly understand 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.

To 'prove' a hypothesis (that a drug causes a rare ADR), one needs many subjects. The sample sizes 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 a critical outcome were being studied, even greater statistical power would be desirable.

6. 'Proof' of the Absence of an ADR (Statistically Significant Noninferiority of the Experimental Group Compared to the Control Group): Sample Size Requirements

Chapter 5 discussed 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 do not reject the null hypothesis of no difference (a study intended to demonstrate a drug effect but failing to do so). The correct interpretation of such an outcome is that the study did not demonstrate the effect of interest. To infer such an outcome 'proves' a lack of the effect of interest is an incorrect interpretation of the outcome. Unfortunately, even amongst physicians, this incorrect interpretation is often inferred from such results. The correct interpretation of an RCT where a null hypothesis of no difference is not rejected is essential for correctly interpreting both efficacy results and AE observations.

Suppose our interest is in 'proving' absence. In that case, a noninferiority inferential test (Mauri and D'Agostino, 2017) with the null hypothesis of a difference between groups is used, and we conclude that no difference exists between groups if that null hypothesis is rejected at the $p \leq 0.05$ level of significance (≤ 0.025 in some cases) (Mauri and D'Agostino, 2017)[9]. There is an essential difference between the conventional inferential test of a difference and the noninferiority inferential test. There is no necessity in the conventional test to define a meaningful difference (except when computing sample sizes). However, in the noninferiority inferential test, defining a difference between treatments considered 'no difference' (i.e., a not clinically meaningful difference) is necessary. This difference cannot be set to "0" because sample sizes would need to be infinity if the acceptable difference is "0". A slight difference must be considered acceptable and not clinically relevant in noninferiority tests, and one can never completely exclude (statistically) a slight excess with test drug versus the comparator.

[9] The authors describe five possible interpretations (Figure 1 in their manuscript) 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 the simultaneous testing of multiple hypothesis, including noninferiority and superiority and the paradoxical but possible interpretation of both noninferiority and inferiority simultaneously.

As suggested above, we are concerned that there are persons who interpret not 'proving' (failing to reject the null hypothesis of no difference in a classical inferential test) an effect as equivalent to 'proving' the absence of an effect, especially if the study intended to 'prove' the 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 study's power was ≥95%, then failure to reject the null hypothesis might offer a degree of evidence of lack of difference (i.e., lack of difference 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 measurement of the effect of interest) where the sample size was prospectively computed based on 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.

Again, however, the correct, formal interpretation of an RCT outcome described in the paragraph above is simply that the RCT failed, not the absence of an 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 that 'prove' a 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 the 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' the absence of an effect, one designs a noninferiority (to placebo) study. As noted above, one must declare a non-zero excess with the 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. 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 meaningful potential ADR (with our incidence of 1 in 1,000), one might think the ratio might be set at 1.10 (1.1-fold, maximum of 10% excess with the drug) or even 1.05. However, there is

precedent (discussed below) for excess incidence with the drug compared to control up to <30%, based on the upper bound of the 95%CI for the ratio, and still declare noninferiority for the drug. If the upper bound of the 95%CI does not exceed 1.3-fold, incidence with the drug is always close to incidence with control.

In many cases where this upper bound does not exceed 1.3-fold, the drug's incidence is less than the incidence with control. Frequently the incidence with the drug is less than the incidence with placebo. Furthermore, in some cases, with that ratio of 1.3, the drug is non-inferior to control and superior to control.

As noted, Mauri and D'Agostino (2017) define several different outcomes of a non-inferiority trial, with both non-inferiority of the experimental treatment to the control treatment and superiority of the control treatment to the experimental treatment. This paradoxical outcome results from a very narrow CI for the ratio of the two treatment outcomes with the upper bound of the CI below the value that allows an interpretation of non-inferiority and an interpretation of a significant difference between the treatments favoring the control treatment. A Thorough QT Study's (Beasley, Mitchell, Dmitrienko, et al., 2005) results are an example of this outcome. In the majority of non-inferiority studies, superiority of the control over the experimental drug would not be tested as it was in the Beasley, et al. study (2005)

A non-inferiority analysis is mandated for non-insulin hypoglycemic agents to treat diabetes mellitus and is codified in an FDA Guidance to Industry (CDER, 2008)[10]. Sponsors developing

[10] We are aware of at least three studies required by FDA for potential drugs seeking regulatory approval with requirements that are noninferiority studies comparing test drug to placebo. The so-called Thorough QT Study (required for virtually all potential drugs) compares the maximum mean changes 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 defines 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,

such drugs must 'prove' that a drug candidate does not cause serious cardiovascular outcomes that would most likely be due to accelerated atherosclerosis development, 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 hemorrhagic strokes and sometimes including TIAs). Hospitalization for unstable angina, hospitalization for heart failure (or acute heart failure), 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 considered by the reporting investigator or sponsor to be serious for any other reason). These summaries were posted on the sponsor's website, showing the results of the sponsor's studies. These SAEs were described with a term (a label from a regulatory dictionary [MedDRA] used for reporting AEs that can be a complaint, sign, symptom, syndrome, or specific diagnosis).

Unfortunately, SAE reports sometimes inaccurately characterize the SAEs and/or provide an incorrect term/label for a given SAE. A blinded, expert review committee does not necessarily scrutinize these SAE reports to decide the correct term/label for an AE. What was reported by an investigator, required to report such an SAE to the sponsor within 24-hours if fatal or life-threatening and otherwise within seven days of learning of the SAE, 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 SAE. Therefore, the data that Nissen used were not necessarily accurate. Second, events were very infrequent and were not

---

2001; HAP: Chen and Bonson, 2013) while the MACE study compares proportions and there is less explicit empirical basis for the noninferiority with the MACE study. The FDA Guidance Document that specifies the margin cited considers two long-term studies of intensive vs. standard diabetes therapy (UKPDS, 1998a; UKPDS, 1998b) that reported CIs for multiple adverse cardiovascular outcomes in drafting this Guidance.

reported in some treatment groups in Nissen's multiple studies, and 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 to compare incidences was such that not all studies could be used. Those with no event of interest in any treatment group could not be used. Ten of 48 studies had no reports of myocardial infarction, and 25 of 48 studies had no reports of death from cardiovascular causes, the two outcomes being 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 needed to be added to the treatment group with zero actual 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 the data from all 48 available studies where zero incidence is highly informative and would have been a preferred method.

Tian and colleagues (Tian, Cai, Pfeffer, et al., 2009) developed and used a method to analyze the dataset used by Nissan more appropriately. 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 an excessive cost and time burden on companies developing treatments for diabetes, discouraging development, and multiple academic groups have questioned its need 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 to the criticism of this analysis requirement, at least one author has recently espoused the position that the studies and analyses 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/analysis 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 (or an individual ADR). 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 the 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 to test with a p-value of ≤0.025 for noninferiority. To assess the 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); Computation is Based on the Number Events Observed in the Control Group

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 supply examples of noninferiority (to placebo) analyses evaluating MACE events. The analyses' design allowed subsequent testing for superiority after demonstrating non-inferiority (Zinman, Wanner, Lachin, et al., 2015 studying empagliflozin; and Neal, Perkovic, and Mahaffey, 2017 studying canagliflozin). These analyses demonstrated noninferiority and superiority in both cases. The analyses' SAPs were written to allow testing for superiority after a result was indicative of noninferiority. Both manuscripts reported results of meta-analyses.

The empagliflozin manuscript employed a hierarchical-testing approach in the following order: noninferiority for the primary outcome (MACE: death from CV events, nonfatal myocardial infarction excluding silent myocardial infarction and 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 the analyses. The sample size was computed based on the assumption of a hazard ratio of 1.0 with the requirement to observe the upper bound of the 2-sided 95.02% CI for the observed hazard ratio to be <1.3. The required confidence interval resulted from an adjustment of the 95% because the data were previously submitted to the FDA. The 1-sided p-value for this upper bound of the 2-sided CI was 0.024. A power of 90% required 691 events[11] 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. The observed hazard ratio with empagliflozin was 0.86 (95.02%CI: 0.74 – 0.99) for the primary outcome. The p-value for non-inferiority was <0.001, and the p-value for superiority was 0.04. 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 in the empagliflozin analyses.

Statistical planning and analyses for canagliflozin were comparable to those used in the empagliflozin manuscript, but there was no adjustment of required p-values (Neal, Pervock,

[11] 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.

Mahaffey, et al., 2017). The sample size needed for 90% power was determined to be 688 events[12]. 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 albuminuria, 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. These steps were taken to maximize the data quality used in the respective analyses. There were 31.5 MACE events per 1,000 subject-years with placebo and 26.9 MACE events per 1,000 subject-years with canagliflozin in the canagliflozin analyses.

In both drug development programs, an event of interest adjudication committee, blinded to treatment, reviewed all records pertinent to each event (MACE event) to make a final determination of what each reported event represented clinically (term/label) and whether it was a MACE event as defined in the prospective protocol. The need for all records and methods to obtain these records would have been put in place prospectively before each RCT initiation. These measures were in place to maximize the precision of the identification of MACE cases.

The two real-world examples above emphasize the magnitude of effort and, therefore, expense needed to 'prove' the absence of a specific set of events in a population with an increased risk of such events (Zinman, Wanner, Lachin, et al., 2015; Neal, Pervock, Mahaffey, et al., 2017). 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 explicitly describe those AEs 'proven' 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

[12] 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.

parties, AEs that have been identified as ADRs with reasonable medical certainty. Therefore, the information above about 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 show 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 safety labeling's primary intent, it can be critical to a sponsor trying to 'prove' that some AE described as an ADR by some party is not an ADR for that given drug.

We should be cautious about what we believe about what a drug does and does not do from a safety perspective and fully understand the robustness, or lack thereof, of such attributions' supportive data.

7. Incidences of AEs of Real-World Interest and Limitations on 'Proof' of Presence or Absence of an ADR

How relevant 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 to clinical reality? What would be the relevance of our hypothetical example to both prescribers and patients who might suffer a significant (i.e., life-threatening or fatal) ADR? Aplastic anemia and the spectrum of Stevens-Johnson Syndrome (SJS) - toxic epidermal necrolysis (TEN) afford a context for considering our example's relevance.

A major international study of agranulocytosis and aplastic anemia was conducted under the WHO's sponsorship. This study's 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). About 25-40% of aplastic anemia cases are considered due to exogenous exposures (drugs, toxic substances) or other external factors, but most are considered idiopathic with no identifiable etiology (Kaufman, Kelly, Issaragrisil, et al., 2006). Therefore, the rate of aplastic anemia (~2-3 / per million-person-years) is much lower than the 1 in 1,000 incidence in our example. Furthermore, some background incidence of aplastic anemia would be expected due to idiopathic factors and exposures to substances other than a test drug, further increasing sample sizes needed to 'prove' causation by a drug.

Stevens-Johnson Syndrome (SJS) and toxic epidermal necrolysis are (TEN) the extreme manifestation of the continuum of the clinical diagnoses of erythema multiforme (EM) – SJS – TEN, although some authorities consider EM to be a separate entity. All three clinical diagnoses do share some characteristic features of the histopathological finding of epidermal necrolysis.

A large UK epidemiological study reported the combined SJS-TEN rate 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 (the range for individual years across four years) and rates of 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 rate of SJS – TEN is then in the range of ~6.5 per million-person years.

In contrast to aplastic anemia, SJS – TEN is primarily due to exogenous exposures. Therefore, the background rates of SJS-TEN could approach zero in a study if the study could be conducted with:

1. the study subjects receiving the investigational drug receiving no other drugs during the study;
2. the control subjects receiving placebo receiving no drugs during the study; and
3. both treatment groups avoiding exposure to other substances that might cause SJS – TEN. Because of the difference in the presumed background incidences or rates, fewer total subjects would be required to 'prove' SJS-TEN is an ADR for a drug than to 'prove' aplastic anemia is an ADR. However, a definitive study for either disorder would be completely impractical.

Aplastic anemia and SJS – TEN show that our hypothetical study of a drug causing an ADR with an incidence of 0.1% (1 in 1,000) is of practical interest. ADRs with even much lower incidences would be of interest to persons taking medications and the persons prescribing medications.

Definitive 'proof' that a drug is associated with an ADR or that a drug is not associated with a specific ADR is virtually impossible given the practical limitations affecting the conduct of human RCTs when:

1) the incidence of an associated ADR is less than approximately 2-3%; and
2) when experimental treatment and placebo-control treatment sample sizes are below several hundred subjects per treatment group.

A high background incidence increases the required sample sizes. These sample sizes or even larger sample sizes would be common with depression and anxiety disorders studies. 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 substantially 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 an ADR's presence or absence for a drug in these therapeutic areas, the studies are generally conducted as a 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 the definitive interpretation of safety observations.

8. Regulatory Requirements for Investigational Treatment Exposure in Development Programs and their Implications for 'Proof' of the Presence or Absence of an ADR

To what extent are regulatory authorities aware of these 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 multiple-dose, clinical studies, 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 (CDER, 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 Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) recommendations/requirements and apply to a wide range of potential drugs across various disorders. For some disorders, the potential drug can be tested against a placebo, while in many disorders, the potential drug can only be tested as an add-on to single, standard therapy compared to a 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 'proven' and those that could be definitely ruled out. However, the Guidance (CDER, 1995) offers the following suggestions on what these exposure requirements can and cannot detect:

> "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 AEs, for example, those occurring in less than 1 in 1000 patients."

The phrase "well-characterized" is not expressly defined. It would seem to us to convey more than merely 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 a control that reaches conventional

statistical significance in a proper inferential test. We would hope there to be some reasonable estimate of the AE incidence that combines AEs due to the background (could be learned from epidemiological literature for many disorders) with those that are ADRs with a reasonable degree of difference in incidences or ratio of incidences between active treatment and control to believe that the AE is reasonably likely to be an ADR. In short, we suggest that it should be possible to offer some quantitative guidance in product labeling as to which AEs listed as ADRs have evidence of ADR status that approximates the evidence for their efficacy claim. The ability to offer such quantitative guidance would require epidemiological data describing comorbidities in large populations treated with the drug being labeled.

In a later Guidance Document addressing "Premarketing Risk Assessment" (CDER and CBER, 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."

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 determining 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.

9. Practical Alternatives to 'Proof' of the Presence or Absence of an ADR: The Need for the Best Assessment Possible as Quickly as Possible of the AE / ADR Profile of a Marketed Drug

Statisticians and data scientists, industry, 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 (mainly 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 AE 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 most often 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. Interested parties should be mindful of the nature of the analyses that lead to the attribution of ADR status to all AEs that are not 'common' and the potential uncertainty of such attribution. All interested parties should also clearly understand the virtual impossibility of 'proving' by a conventional gold standard what is or is not an ADR associated with a drug, except with some 'common' ADRs.

It cannot be emphasized enough that for AEs that might or might not be ADRs but are of low incidence, it can be impossible to 'prove' that the AE is or is not an ADR for a drug based on the RCTs that are conducted to 'prove' that the drug is efficacious. The best that we can do in the future is develop more robust prospective epidemiological studies that are begun soon after a drug is launched. By more robust, we mean studies with vast numbers of subjects, extended exposure time frames, and rigorous prospective methods for identifying with reasonable clinical certainty the AEs of interest that are ADRs. 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 more rapidly with a lower probability of false positive and false negative attribution.

10.Postscript

In considering various comments in response to our work and thinking a bit more about its contents, we concluded that our position about labeling AEs might not be sufficiently clear. Although our position was stated multiple times above, we believe it was not stated with sufficient force and clarity. This postscript is an effort to remedy this potential fault.

We want to clarify our position about labeling an AE, observed in temporal association with administration of a drug, as an ADR where 'proof' that the AE is an ADR is of a lesser standard than the 'proof' required to receive an efficacy claim.

Much of our work in this book was intended to make it clear that many AEs observed in temporal association with administration of a medication that might be included in product labeling lack 'proof' of being ADRs of comparable robustness to the robustness of 'proof' needed to obtain regulatory approval for an efficacy claim. We went to considerable lengths to demonstrate the matter from a statistical perspective. It might be easy to conclude that our position is that without comparable 'proof' that an AE is an ADR required for an efficacy claim, the AE should not be included in product labeling.

Our primary intent was to broaden the understanding of the statistical realities and quality of the evidence pertaining to infrequent-rare AEs identified as ADRs in product labels. Our concern is that many persons who might read a product label might believe that any AE included in a product label has been conclusively 'proven' to be an ADR. Robust 'proof' that AEs included in product labeling are ADRs is sometimes lacking. All persons who read product labeling (or derivatives of product labeling that can be found on many internet sites, including subscription medical services) and use product labeling for any purpose should have a clear understanding of what has been robustly 'proven' and what has not.

We unequivocally believe that some AEs that have been observed in temporal association with medication are of such clinical significance due to their actual outcome (e.g., death, permanent disability, lengthy and costly hospitalization) or potential outcome that they should be included in product labeling, even with only a modest (in some cases very modest) amount of evidence suggesting they are ADRs for a given drug. We also believe that such labeling should offer the reader guidance about the magnitude and quality of evidence supporting the hypothesis that the AE is an ADR if it appears in a medication's product labeling (see Chapter 8). We believe such

information is essential when that magnitude of 'proof' is minimal and quality marginal. The rationale for the inclusion of that AE in labeling should be briefly explained in such cases.

The following example, intended to illustrate our position, deals with a medication marketed in several international regulatory venues for a non-psychiatric indication.

During the medication's development, AEs were observed that could be grouped clinically on a spectrum of clinical severity and severity of the outcome (analogous to but not necessarily the spectrum of erythema multiforme, Stevens-Johnson Syndrome, and toxic epidermal necrolysis). In at least one international regulatory venue, these several AEs are described in product labeling in several paragraphs in sections of the label intended to describe more clinically serious AEs that are possible ADRs.

Multiple placebo-controlled trials in two indications had been completed and analyzed before regulatory submission for review and potential approval. These trials extend well beyond the standard length of 6-8 weeks for placebo-controlled, Phase 3 studies with psychiatric disorders (e.g., Major Depressive Disorder; Schizophrenia; Generalized Anxiety Disorder; Bipolar I Disorder, Manic Episode). More than 3,400 subjects were included in the placebo-controlled phases of these studies. Also, these trials included open-label, active medication-only extension phases. At the last time analyses of this database were conducted, one trial was completed after regulators reviewed the trials for potential approval. To be thorough in the analyses, they were conducted comparing incidence differences, incidence ratios, and odds ratios for active medication versus placebo. Multiple non-exact (e.g., Chi-square) and exact plus bootstrap inferential methods were used for the analyses to provide sensitivity analyses of the observations.

The incidence of combined events with the medication was approximately 1.25% and with placebo approximately 0.75%. All AEs in the spectrum were combined in one set of analyses, and all studies across all indications were combined. Depending on the inferential method, the exact plus bootstrap methods that supplied p-values resulted in p-values in the range of 0.2022 to 0.2264. Those methods that supplied only confidence intervals (CIs) resulted in 95% CIs that ranged from (-0.0117 – 0.0017) to (-0.0117 – 0.0030) for comparisons of differences and ranged from (0.30 – 1.29) to (0.32 – 1.70) for comparisons of ratios.

In one of the indications, the difference in incidences and ratios suggested a slightly larger disparity between drug and placebo for observations of these AEs. Within this indication, the exact methods

that supplied p-values resulted in p-values in the range of 0.3063 to 0.4450. Those methods that supplied only confidence intervals (CIs) resulted in 95% CIs that ranged from (-0.0174 – 0.0042) to (-0.0181 – 0.0072) for comparisons of differences and ranged from (0.20 – 1.62) to (0.23 – 3.65) for comparisons of ratios. The larger, non-significant p-values and 95% CIs for this indication with a greater imbalance in incidence than the combined indications result from the sample sizes for the one indication being smaller than those for the combined indications, the disparity between incidences being modest.

Most notably, however, for the AE of greatest clinical severity and most easily confirmable as an AE in this continuum, all the exceedingly small number of cases occurred during placebo treatment within the indication with the greatest disparity between medication and placebo.

The one trial not available in data reviewed at submission did not appear to alter the incidence with medication compared to placebo.

Finally, these analyses were conducted based on a simple pooling of all available studies. Across the two indications (all indications were approved), there were differences in the AE incidences between medication and placebo. However, the conventional interpretation of the inferential results across indications would be consistent.

By conventional statistical standards, the interpretation of these results would be that observed outcomes were most likely due to chance rather than drug effect. The most severe outcome in the spectrum of outcomes was associated exclusively with placebo treatment. However, there was a slight excess incidence of the least severe AE with medication. During the open-label, medication-only extension phases, more AEs in this continuum were reported with medication.

A non-inferiority analysis (potential for 'proving' lack of drug effect) was not conducted with these data because there is no well agreed upon margin of excess with a drug for this AE spectrum or its least clinically significant specific AE that would still allow a conclusion of non-inferiority and a slight excess incidence with medication was observed.

We believe that it would be appropriate to describe this spectrum of AEs in product labeling as events to which the prescriber should be alert, that there was a numerical excess of the least serious manifestation (but still an important and potential AE) with the drug, and that all the extremely few cases of the most ominous manifestation were with placebo. Additionally, some quantification

of the likelihood of the observed data being due to chance or drug should be included as these observations would customarily be viewed as due to chance.

In the product labeling for this medication in the regulatory venue we are discussing, it was acknowledged that all occurrences of the AE of the most severe outcome in this spectrum were during placebo treatment. In this regulatory venue, all AEs (with or without robust 'proof' of being ADRs) are described with the English language label of "Adverse Drug Reactions". This label for events included in product labeling (be they AEs or AEs with robust 'proof' of being ADRs) is found across multiple regulatory venues. To describe an AE or set of AEs with the magnitude of 'proof' of being ADRs, as in the example above, with the label "Adverse Drug Reaction" potentially conveys implications that are not supported by the data. Some AEs are of such clinical significance that they should be included in product labeling if there is the slightest excess with a drug compared to placebo or other, softer evidence (evidence from sources other than controlled clinical trials) of the potential for these AEs being ADRs. However, when the evidence is weak and the AE is described out of an abundance of caution, it should be clearly stated that it cannot be concluded that the AE is an ADR by conventional interpretive standards.

Product labeling needs to serve the intent of "first do no harm" and help the prescriber accurately understand the degree of evidence supporting the potential for doing harm by prescribing a given medication. Not treating a patient with a medication based on a potentially inaccurate understanding of the probability of that medication doing harm is itself potentially harmful. The relevant evidence can grow over time.

References

Albaugh VL, Judson JG, She P, Lang CH, Maresca KP, Joyal JL, Lynch CJ. Olanzapine promotes fat accumulation in male rats by decreasing physical activity, repartitioning energy and increasing adipose tissue lipogenesis while impairing lipolysis. *Mol Psychiatr* 2015; 16:569-581.

Beasley CM Jr. Comment on Barry Blackwell's Corporate corruption in the psychopharmaceutical industry. inhn.org.contoversies. March 23, 2017.

Beasley CM Jr. Response to Barry Blackwell's response to Charles Beasley's comment on Barry Blackwell's Corporate corruption in the psychopharmaceutical industry. inhn.org.controversies. January 12, 2018.

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; 27:682-686.

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; 303:685-692.

Beasley CM, Mitchell MI, Dmitrienko AA, Emmick JT, Shen W, Costigan TM, Bedding AW, Turick MA, Bakhtyari A, Warner MR, Ruskin JN, Cantilena LR, Kloner RA. The combined use of ibutilide as an active control with intensive ECG sampling and signal averaging as a sensitive method to assess the effects of tadalafil on the human QT interval. *J Am Coll Cardiol* 2005; 46:678-687.

Blackwell B. Corporate corruption in the psychopharmaceutical industry. inhn.org.controversies. September. 1, 2016.

Blackwell B. Corporate corruption in the psychopharmaceutical industry (revised). inhn.org.controversies. March 16, 2017a.

Blackwell B. Response to Charles Beasley's comment on Barry Blackwell's Corporate corruption in the psychopharmaceutical industry. inhn.org.controversies. July 13, 2017b.

Blackwell B. Response to Charles Beasley's response to Barry Blackwell's response to Charles Beasley's comment on Barry Blackwell's Corporate corruption in the psychopharmaceutical industry. inhn.org.controversies. May 3, 2018.

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; 311:619-620.

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; 137:1240-1257.

Hahn MK, Chintoh A, Remington G, Teo C, Mann S, Arenovich T, Fletcher P, Lam L, Nobrega J, Guenette M, Chon T, Giacca A. Effects of intracerebroventricular (ICV) olanzapine on insulin sensitivity and secretion in vivo: an animal model. *Eur Neuropsychopharmacol* 2014; 24:448-458.

Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA* 1983; 249:1743-1745.

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

International Agranulocytosis and Aplastic Anemia Study. Incidence of aplastic anaemia: the relevance of diagnostic criteria. *Blood* 1987; 70:1718-1721.

Jovanovic BD, Levy PS. A look at the rule of three. *Am Stat* 1997; 51:137-139.

Kang YW, Ko YS, Kim KY, Sung C, Lee DH, Jeong E. Trends in health-related behaviors of Korean adults: study based on data from the 2008-2014 Community Health Surveys. *Epidemiol Health* 2015 Sep 29; 37:e2015042. doi: 10.4178/epih/e2015042. eCollection 2015.

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; 81:65-67.

Kowalchuk C, Teo C, Wilson V, Chintoh A, Lam L, Agrwal SM, Giacca A, Remington GJ, Hahn MK. In male rats, the ability of central insulin to suppress glucose production is impaired by olanzapine, whereas glucose uptake is left intact. *J Psychiatr Neurosci* 2017; 42:424-431.

Ludbrook J, Lew MJ. Estimating the risk of rare complications: is the 'rule of three' good enough? *Anz J Surg* 2009; 79:565-570.

Malik M. Problems of heart rate correction in assessment of drug-induced QT interval prolongation. *J Cardiovasc Electrophysiol* 2001; 12:411-420.

Mauri L, D'Agostino RB Sr. Challenges in the design and interpretation of noninferiority trials. *NEJM* 2017; 377:1357-1367.

Mauri L, D'Agostino RB Sr. Noninferiority Trials. *NEJM* 2018; 378:304-305.

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; 377:644-657.

Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *NEJM* 2007; 356:2457-2471.

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

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; 34:173-180.

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; 10:275–281. 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; 352:837-853.

UKPDS Group. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). *Lancet* 1998b; 352:854-865.

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 Biologics Evaluation and Research (CBER) Center for Drug Evaluation and Research (CDER). Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products: Guidance for Industry. December 2019. https://www.fda.gov/media/133660/download. Accessed April 1, 2021.

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.

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.

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; 8:315-825.

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; 373:2117-2128.

Comments and Questions with Beasley's Responses and Replies

1. Edward Shorter's Comment on the Outline (Chapter 1), Followed by Beasley's Response, Followed by Barry Blackwell's Comment on Beasley's Response, Followed by Beasley's Response to Blackwell: Assessment of the Relationship Between Olanzapine and Diabetes Mellitus as an Example of the Complexities in the Assessment of Infrequent ADRs with Relatively High Background Incidences

a. Edward Shorter's Comment on the Outline (Chapter 1)

Readers of this website will look forward with special interest to Charles Beasley's comments, particularly on the issue of side effects and their measurement, given that in his long tenure at Eli Lilly, he often confronted these issues on an almost daily basis. In the late 1990s there was an intense in-house discussion about possible hyperglycaemia, weight gain and diabetes associated with olanzapine and much of this correspondence has, in connection with discovery in litigation, now become part of the public record. In these exchanges, Alan Breier and Dr. Beasley come across very much as the in-house investigators committed to the high road of science and one hopes that in the coming instalments of this thread, Dr. Beasley might illustrate his points with references to some of this material.

April 25, 2019

b. Charles Beasley's Response to Shorter

Olanzapine and Diabetes Mellitus, Evolution of Data – Illustrating the Difficulties in Identification of ADRs

First, we want to thank Prof. Shorter for his interest in our book and his willingness to comment. We have not addressed understanding the relationship of second-generation antipsychotics (or specifically olanzapine) with glucose homeostasis dysregulation and diabetes mellitus in our writing. However, Prof. Shorter's specific interest in diabetes mellitus afforded us a reason to review the evolution of studies of olanzapine and dysregulation of glucose homeostasis after the early 2000s when our responsibilities within Eli Lilly and Company were shifted away from olanzapine. Other than the studies described below sponsored by Lilly and the study of Ader, Kim, Catalano, et al. (2005), for which an abstract appeared several years earlier, we were not previously familiar with any of the work we summarize and compare.

Also, Professor Shorter's interest allows us to illustrate two difficulties in detecting (with reasonable medical certainty if not 'proving' by the standard required for regulatory approval of an efficacy claim) an ADR briefly discussed in our writing and another difficulty we did not discuss. We did not discuss this third difficulty because our focus was on AEs that might or might not be ADRs that can be assessed based exclusively on events only described with text. Such events are binary entities. An investigator writes that a subject has Stevens-Johnson Syndrome (not present before entering a study) during study participation. That subject has

shifted their binary state from the state of absence of Stevens-Johnson Syndrome to the state of the disorder's presence, at least per the investigator.

With diabetes mellitus as an example, there are objective numerical data collected systematically (collected for every subject at specified times as prospectively specified in the study protocol) throughout a study (e.g., fasting glucose, random glucose, HbA1c, fructosamine, urine glucose) available to determine the accuracy of text attribution of such an AE to a subject. These objective data also allow attribution of such an AE to a subject even if an investigator has not described an AE in text. Additionally, the incidences of abnormalities of such objective numerical data can be compared between treatments to contribute to the determination of whether the test drug is causing an ADR as identified by values considered to be indicative of pathology rather than merely relying on the comparison of incidences of text attributed AEs that are treatment-emergent.

As stated above, our earlier work addressed AEs, where the only data sources for analysis are the text-based reports of occurrences of AEs. Continuing with the example of Stevens-Johnson Syndrome, there are no laboratory data collected systematically throughout a study that would allow confirmation of the presence or absence of Stevens-Johnson syndrome for an individual subject for whom an investigator recorded the AE of Stevens-Johnson syndrome or to identify the presence of the disorder even if the investigator did not record it as an AE. Although objective data (such as tissue biopsy with microscopic examination) might be obtained to confirm or refute the text attribution of an AE to a subject by an investigator, for most AEs, there are no systematically collected data that allow identification of an AE based exclusively on those objective data that are usually numerical. While having objective numerical data to aid in identifying true AEs is usually helpful, this data source can introduce specific difficulties, as we describe below.

The first difficulty in determining ADRs discussed in this book is when an AE that is an ADR has its onset well after the drug's initiation. The negative effect of this difficulty on identifying ADRs increases as the onset of the ADR increases relative to beginning study treatment and if the onset rate increases over time. A study might require an impractical number of subjects on both drug and placebo, followed for an impractical length of time (especially for placebo treatment) to detect even slim evidence of a difference between groups in either the incidence of text-described AEs or subjects with systematically collected numerical data meeting objective criteria for an AE. It is likely that special analytical time-based methods would be required to be most sensitive to differences in incidence if there is a time delay in onset, as also briefly discussed. Such methods are not routinely applied to all AEs recorded nor to treatment-emergent instances of values of objective safety parameters considered probably indicative of pathology (values above the upper reference limit or below the lower reference limit for some laboratory analyte or diagnostic procedure that results in numerical data). Such specialized analyses sensitive to delayed onset of ADRs and changes in ADRs' rates are often applied on only a for-cause basis when more routine incidence-based analyses (difference in incidences or the ratio of

incidences) suggest differences between treatment groups. Such a time delay is expected for most cases of diabetes mellitus that might be ADRs to second-generation antipsychotics; this is underscored when we discuss olanzapine analyses specifically.

The second difficulty impacting the detection of ADRs that we discussed in more detail is the difficulty detecting differences between treatment groups when there is a relatively high incidence of new-onset cases of an AE that are not ADRs but are entirely independent of the drug treatment (background incidence) compared to the incidence of new-onset cases that are ADRs. Chapter 5 of our work illustrated the difficulty of a relatively high background incidence complicating analyses. We believe this difficulty is relevant to finding excess diabetes mellitus in the development databases for second-generation antipsychotics, or at least for the agent with which we are most familiar, olanzapine.

The third difficulty, not addressed previously in our writing, is a 'noisy parameter' or a low 'signal-to-noise ratio' when dealing with numerical data relevant to identifying an AE, a difficulty quite familiar to engineers. Here, 'noisy' refers to a large magnitude of unexpected and unexplained within-subject and between-subject variability across time. In the domain of drug safety, nothing illustrates this problem more clearly than the assessment of group changes in QTc. In Thorough QT Studies, the signal of interest is the maximal mean difference in change from baseline of QTc length between drug and placebo. A maximum mean difference of >5-10 ms is considered to suggest a drug effect and an increased risk of experiencing a clinical ADR. However, normal within-subject beat-to-beat variability can be 25 ms even with optimal recording and measurement techniques and collection under optimal conditions (Malik and Camm, 2001). Venous blood glucose concentrations that are most useful for assessing glucose homeostasis and the diagnosis of diabetes mellitus are intended to be collected in a between 10-12 hour fasting state. The difficulties in obtaining such fasting values in subjects with schizophrenia when values are on an outpatient basis should be readily appreciated by most clinicians who treat such patients. As a result of the collection procedures in the studies described below and other factors, the observed glucose values, especially in analyses of long-term data with olanzapine, were very 'noisy'.

To understand the evolution of understanding of the relationship between olanzapine, as a specific example of second-generation antipsychotics, and diabetes mellitus, it is essential to review the evolution of the numerical diagnostic criteria (complete diagnostic criteria include symptom criteria and confirmation of numerical values) for that disease (Kumar, 2016). There were substantial changes in these criteria during the period in which the early second-generation antipsychotics (e.g., risperidone, olanzapine, quetiapine, ziprasidone) were initially developed and evaluated (the mid-1980s through mid-1990s):

- World Health Organization (WHO) 1980 criteria:
    - Diabetes mellitus (DM)
        - Fasting glucose: ≥7.8 mmol/L (140 mg/dL), or

        - Post-glucose glucose tolerance test load of 75 g of glucose: ≥11.1 mmol/L (200 mg/dL)
    - Impaired glucose tolerance (IGT)
        - Fasting glucose: <7.8 mmol/L (140 mg/dL) and post-glucose tolerance test load of 75 g of glucose: ≥7.8 mmol/L (140 mg/dL) but <11.1 mmol/L (200 mg/dL)
    - Impaired fasting glucose (IFG)
        - Not defined
- American Diabetes Association (ADA) 1997 criteria and WHO 1999 criteria:
    - DM
        - Fasting glucose: ≥7.0 mmol/L (126 mg/dL), or
        - Post-glucose tolerance test load of 75 g of glucose (or post-prandial/non-fasting): ≥11.1 mmol/L (200 mg/dL)
    - IGT
        - Fasting glucose: <7.0 mmol/L (126 mg/dL), if measured, and post-tolerance test load of 75 g of glucose: ≥7.8 mmol/L (140 mg/dL) but <11.1 mmol/L (200 mg/dL)
    - IFG
        - Fasting glucose ≥6.11 mmol/L (110 mg/dL) but <7.0 mmol/L (126 mg/dL) and post-glucose tolerance test load of 75 g of glucose: <11.1 mmol/L (200 mg/dL), if measured
- ADA 2003:
    - Modified criteria for IFG
        - Fasting glucose: ≥5.5 mmol/L (100 mg/dL) but <7.0 mmol/L (126 mg/dl) and post-glucose tolerance test load of 75 g of glucose: <11.1 mmol/L 200 mg/dL), if measured

- ADA 2010 and WHO 2011
    - Incorporated HbA1c into diagnostic criteria
        - DM – HbA1c: ≥6.5%
        - IGT and IFG – HbA1c: 5.7%-6.4%

While the diagnostic criteria for diabetes mellitus based on fasting values decreased, the criteria based on a random value, unlikely to be fasting, have not changed. To be precise, the random value should only be applied to a value obtained at least two hours after ingesting an oral 75 g glucose load in a glucose tolerance test. Still, by inference, any value observed ≥11.1 mmol/L (200 mg/dL) two hours or more after a meal or heavy carbohydrate load would strongly suggest diabetes mellitus. By current diagnostic criteria, values between 7.8-11.1 mmol/L (140-200 mg/dL) at any time at least two hours following such ingestion would strongly suggest a degree of impaired glucose homeostasis.

1. Olanzapine Clinical Trial Data: An Illustration of the Impact of 'Noise' Combined with the Impact of Delayed Onset and Relatively High Background Incidence

Olanzapine clinical development Phase III studies were conducted between 1991-1995 with data analyses and preparation of regulatory submission documents occurring in 1995, a time at which diabetes mellitus was diagnosed based on a fasting plasma glucose ≥7.8 mmol/L (140 mg/dL) and HbA1c was not considered valid for diagnosis (and was not collected as a routine laboratory analyte during development studies). The initial development program for risperidone had occurred several years earlier, and the development programs for quetiapine and ziprasidone had occurred in the same period as that for olanzapine. Quetiapine and ziprasidone received US regulatory approvals shortly after the approval of olanzapine.

The initial development program for olanzapine for the treatment of psychosis (psychosis was the indication for which olanzapine was initially approved in the US; the indication was subsequently changed in the US to schizophrenia per FDA) included five studies with extensions:

1. placebo-controlled and haloperidol-controlled, three variable doses of olanzapine (5±2.5 mg/d, 10±2.5 mg/d, 15±2.5 mg/d), six weeks, inpatient with the transition to outpatient. For subjects showing an adequate response, a one-year continued double-blind extension was available; the extension was further extended to indefinite until approval (Beasley, Tollefson, Tran, et al. 1996a);
2. 1 mg/d (pseudo-placebo) controlled and haloperidol-controlled, three variable doses of olanzapine (5±2.5 mg/d, 10±2.5 mg/d, 15±2.5 mg/d), six-weeks, inpatient with the transition to outpatient. For subjects showing an adequate response, a one-year continued double-blind extension was available; the extension was further extended to indefinite until approval (Beasley, Hamilton, Crawford, et al., 1997);
3. placebo-controlled, two fixed doses of olanzapine (1 mg/d and 10 mg/d), six-weeks, inpatient with the transition to outpatient; all subjects who completed at least three weeks and were still substantially symptomatic could switch to open-label olanzapine with an

indefinite extension until approval; the open-label extension was also available to subjects completing the study (Beasley, Sanger, Satterlee, et al., 1996b);

4. haloperidol-controlled, one variable dose of olanzapine (5-20 mg/d), six-weeks, outpatient, or inpatient; an indefinite open-label extension until approval was available to subjects (Tollefson, Beasley, Tran, et al., 1997); and
5. placebo-controlled, one variable dose of olanzapine (2-8 mg/d), six weeks, for subjects with psychotic symptoms and a diagnosis of Alzheimer's Disease; this study was not intended to support registration for an indication of psychosis with Alzheimer's Disease but was intended to study at least 100 subjects ≥65 years of age (few subjects in this age group with schizophrenia enter randomized clinical trials) (Unpublished).

In 1995, at the time of data analyses for initial submission, the following numbers of subjects (not all subjects assigned to treatment contributed data to every analysis; these numbers are approximations for each analysis) were available for analyses:

- Olanzapine (not including 1 mg/d dose) vs. placebo: 248–118
- Olanzapine (not including 1 mg/d dose) vs. haloperidol (up to six weeks): 1,796–810
- All olanzapine (including 1 mg/d dose): 2,500 with some subjects treated for more than four years

The following summaries of analyses performed are based on our recall in December 2018 as neither the source documents with the results of the analyses nor the original US Prescribing Information (the term used by the FDA for product labeling) olanzapine were available to review at the time this was written.

The potential effects on glucose homeostasis must be considered in the context of any changes in body weight, especially where those changes are probably increases in adipose tissue (especially visceral adipose tissue). Weight increases with olanzapine were well characterized in the initial development studies and described in the original US Prescribing Information. Our recollection, in which we are quite confident of its accuracy, is that this Prescribing Information noted that approximately 50% of subjects in long-term treatment with olanzapine with a median exposure of approximately six months gained ≥10% body weight. Based on the manuscripts specifically reporting weight changes with olanzapine and haloperidol, in Study 2 above, haloperidol-treated subjects experienced a mean decrease in weight, and in Study 4 above experienced a mean increase in weight of only 0.02 kg with 4.6% losing ≥7% weight compared to only 2.5% of olanzapine-treated subjects experiencing such weight loss. The differences between olanzapine and haloperidol for both mean change (an increase of 1.88 kg with olanzapine) and proportions gaining ≥7% were both statistically significant in Study 4. The manuscripts describing the short-term study results of studies 1 and 3 did not include weight change data.

The weight gain data available at the time of approval and included in the US Prescribing Information are important in interpreting the glucose data and their analyses for the initial development studies and a subsequent set of analyses conducted using a more extensive set of data performed in the late 1990s. Additionally, these weight gain changes would allow for a

conclusion that olanzapine would be temporally associated with some emergent diabetes mellitus (differs from definitive 'proof'), especially with long-term use, on the part of a reasonably informed clinician. It should be well known that a significant increase in weight, particularly due to increased adipose tissue, is a risk factor for diabetes mellitus.

The analyses of glucose values at the time of submission (conducted between January–August 1995) were interpreted as not suggesting an alteration in glucose homeostasis associated with olanzapine treatment. The analyses included consideration of both mean changes from baseline to the endpoint and the emergence of both high and low outlier values. These analyses considered both placebo and haloperidol comparisons with simple pooling of all direct comparative data, excluding the 1 mg/d dose of olanzapine. Additionally, all data (comparative and non-comparative, open-label) for olanzapine (including the 1 mg/d dose) were similarly analyzed and considered. While some analytes (e.g., CPK) were subjected to more complex analyses based on the results of the initial set of analyses, the initial set of analyses of glucose values were not interpreted as suggesting the need for such additional analyses. What might be considered a potentially more objective review of the data and analyses results (i.e., by the FDA) can be considered concordant with the Lilly interpretation as no additional analyses were requested of Lilly by FDA. We recall that there were low incidences (but not zero incidences) of treatment-emergent AEs described with the terms "hyperglycemia" and "diabetes mellitus" during olanzapine treatment in these initial development studies. If this is correct, they would likely have been included in lengthy lists of AEs that might or might not be ADRs in the Prescribing Information. However, specific Prescribing Information text required by regulatory authorities did not discuss glucose, glucose homeostasis, or diabetes mellitus in any dedicated section or any detail.

Based on spontaneous AE reports received by Lilly and published case reports and case series, Lilly undertook an extensive set of analyses of all available clinical trial data in the late 1990s. Additional data were available with even more extended periods of treatment relative to the data available in 1995. This work's results were presented in several public scientific forums, with the first presentation by Beasley, Berg, Dananberg, et al. in 2000. The analyses' interpretation was that the analyses failed to support the hypothesis of an association between olanzapine treatment and the development of hyperglycemia or diabetes mellitus. This interpretation is highly limited; these analyses could not be interpreted as supporting the hypothesis of a lack of association between olanzapine and the development of hyperglycemia or diabetes mellitus. The data were complex, particularly considering the 1997 ADA numerical criteria (see above, changed since the 1995 initial data analyses) relevant to diabetes mellitus and impaired glucose homeostasis. Many subjects' baseline values were elevated. For the group, as well as individual subjects, variability over time was much greater than expected. The magnitude of variability was so larger that some diabetologists who reviewed the analyses' results questioned the values' veracity. One source probably contributing to the variability was the potential for sample collection intended to be in the fasting state, often being in a non-fasting state and potentially

within the two-hour post-consumption window during which even values ≥200 mg/dL are challenging to interpret.

Subsequent analyses of the Lilly clinical trial database (after the late 1990s' analyses) with multiple additional studies that included HbA1c measurements are best summarized in the US Zyprexa® Prescribing Information updated in 2010 copied verbatim below.

> *"*Olanzapine Monotherapy in Adults – In an analysis of 5 placebo-controlled adult olanzapine monotherapy studies with a median treatment duration of approximately three weeks, olanzapine was associated with a greater mean change in fasting glucose levels compared to placebo (2.76 mg/dL versus 0.17 mg/dL). The difference in mean changes between olanzapine and placebo was greater in patients with evidence of glucose dysregulation at baseline (patients diagnosed with diabetes mellitus or related AEs (present at baseline before receiving olanzapine), patients treated with antidiabetic agents, patients with a baseline random glucose level of ≥200 mg/dL, and/or a baseline fasting glucose level ≥126 mg/dL). Olanzapine-treated patients had a greater mean HbA1c increase from baseline of 0.04% (median exposure 21 days) than a mean HbA1c decrease of 0.06% in placebo-treated subjects (median exposure 17 days). In an analysis of 8 placebo-controlled studies (median treatment exposure 4-5 weeks), 6.1% of olanzapine-treated subjects (N=855) had treatment-emergent glycosuria compared to 2.8% of placebo-treated subjects (N=599). Table 2 shows short-term and long-term changes in fasting glucose levels from adult olanzapine monotherapy studies."

Table 2: Changes in Fasting Glucose Levels from Adult Olanzapine Monotherapy Studies (Table Included with the Text Above)

			Up to 12 Weeks Exposure		At Least 48 Weeks Exposure	
Laboratory Analyte	**Category Change (at least once) from Baseline**	**Treatment Arm**	**N**	**Incidence**	**N**	**Incidence**
Fasting Glucose	Normal to High (<100 to ≥126 mg/dL)	Olanzapine	543	2.2%	345	12.8%
		Placebo	293	3.4%	NA[1]	NA[1]
	Borderline to High (≥100&<126 to ≥126 mg/dL)	Olanzapine	178	17.4%	127	26.0%
		Placebo	96	11.5%	NA[1]	NA[1]

[1] Not applicable

> "Olanzapine Monotherapy in Adolescents – The safety and efficacy of olanzapine have not been established in patients under the age of 18 years. In an analysis of 3 placebo-controlled olanzapine monotherapy studies of adolescent patients, including those with Schizophrenia (6 weeks) or Bipolar I Disorder (manic or mixed episodes) (3 weeks), olanzapine was associated with a greater mean change from baseline in fasting glucose levels compared to placebo (2.68 mg/dL versus -2.59 mg/dL). The mean change in

fasting glucose for adolescents exposed at least 24 weeks was 3.1 mg/dL (N=121). Table 3 shows short-term and long-term changes in fasting blood glucose from adolescent olanzapine monotherapy studies."

Table 3: Changes in Fasting Glucose Levels from Adolescent Olanzapine Monotherapy Studies (Table Included with the Text Above)

			Up to 12 Weeks Exposure		At Least 24 Weeks Exposure	
Laboratory Analyte	**Category Change (at least once) from Baseline**	**Treatment Arm**	**N**	**Incidence**	**N**	**Incidence**
Fasting Glucose	Normal to High (<100 to ≥126 mg/dL)	Olanzapine	124	0%	108	0.9%
		Placebo	53	1.9%	NA[1]	NA[1]
	Borderline to High (≥100&<126 to ≥126 mg/dL)	Olanzapine	14	14.3%	13	23.1%
		Placebo	13	0%	NA[1]	NA[1]

[1] Not applicable

We could not conduct inferential analyses on the mean change in plasma glucose concentration data extracted from the US Zyprexa® Prescribing Information updated in 2010 in the two paragraphs above because the standard deviations were not provided. However, the categorical (outlier) shifts in Tables 2 and 3 could be analyzed by simply pooling the multiple studies and not appropriately adjusting for differences among the studies.

Table 4: 2-Sided Fisher's Exact Test p-Values – Categorical Change in Fasting Glucose, Olanzapine vs. Placebo

Age Group	Change Category	p-Value
Adults	Normal to High	0.3653
	Borderline to High	0.2213
Adolescents	Normal to High	<0.0001, less with Olanzapine
	Borderline to High	0.2222

These aggregated placebo-controlled data from 2010 Prescribing Information included larger sample sizes than the initial submission data, but the length of treatment where a comparison to placebo could be made was still extremely short. Median exposure times were of a length such that HbA1c changes are not relevant. The mean change data for glucose might have demonstrated statistically significant differences with a greater mean increase with olanzapine. Statistical significance would almost certainly be the case with the adolescent mean change results. In the general case, mean change differences (more with the drug than placebo) support the potential for a mechanistic causative process. Application of inferential statistical methods to mean changes in numerical safety data is prone to type 1 error (false positive identification of a difference) without adjustment for the multiplicity of comparisons that can be made. More than

30 laboratory analytes are measured, and vital signs and anthropomorphic characteristics add to this number in most development programs for potential new drugs. However, it is a difference in the incidence of treatment-emergent outliers that are more informative regarding a clinically significant process and less prone to type 1 error when subjected to inferential analyses. In the four treatment-emergent outlier comparisons (change from normal to high [diabetic], change from borderline [prediabetic] to high [diabetic]) in both adults and adolescents, the only olanzapine-placebo comparison difference to reach statistical significance was a shift from normal values to high values in adolescents with the greater incidence associated with placebo. Our overall impression is that these 2010 data more clearly suggest an adverse change in glucose homeostasis in temporal association with olanzapine than were the earlier data at the time of submission. However, with these clinical trial data in isolation, it would be difficult to conclude definitively that olanzapine causes diabetes mellitus directly or indirectly. These results best illustrate the combined problems of insufficient length of comparative treatment data for ADRs, with an infrequent incidence of occurrence, combined with a high background incidence of the AE (same event as the ADR), and further combined with 'noisy' data (high variability). Note that one out of 53 subjects in the adolescent group, in a period of up to only 12 weeks, shifted from a normal baseline glucose value to a glucose value in the diabetic range for at least one measurement. We do not know about the variability in any of the subjects' glucose values, and with these data representing up to 12 weeks of treatment and glucose measurements being obtained potentially weekly, the value ≥126 mg/dL could be a single value out of 12. Therefore, these data might underscore the potential problem of variability in data used to identify an AE, the 'noise-to-signal' ratio problem.

Lilly extended its research efforts regarding glucose homeostasis and olanzapine in several ways following the late 1990s' analyses of available clinical trial data. Although HbA1c had not yet become a standard for diagnosing diabetes mellitus and was considered only helpful for assessing average glycemic control in patients with diagnosed diabetes mellitus, HbA1c was added as a standard laboratory analyte (contributing to the analyses above) collected in studies, and fructosamine was also added. Fructosamine is an analyte comparable to HbA1c, but while HbA1c assesses average glucose changes throughout several months, fructosamine assesses these changes throughout several weeks. Measurement of both analytes served as an attempt to address the 'noise-to-signal' ratio problem with venous glucose measurements, where many of these values were likely for samples collected in a non-fasting state.

Additionally, Lilly conducted three glucose clamp studies, one to primarily assess the release of insulin from the pancreas and two to assess the body's sensitivity to insulin. Lilly also conducted a mixed mean tolerance test in conjunction with one glucose clamp study to evaluate insulin release and insulin tolerance.

2. Clamp Studies and Mixed Meal Tolerance Tests

Three types of studies are discussed below: two types of clamp studies and the mixed meal tolerance test (MMTT). The first type of clamp study is the hyperglycemic clamp study, the gold standard for assessing the pancreas' (β-cells') capacity to produce and release insulin appropriately in the face of exposure to glucose. The hyperglycemic clamp study can estimate whole-body insulin sensitivity (see next paragraph), but not as precisely as the hyperinsulinemic-euglycemic clamp study. Krentz, Heinemann, Hompesch (2015b) provide a review of methods of assessing β-cell function, including the hyperglycemic clamp study.

The second type of clamp study is the hyperinsulinemic-euglycemic clamp study, the gold standard for assessing the body's tissues ability to respond appropriately to insulin – take the glucose into tissues (liver, muscle, and fat) and for those tissues that can produce and release glucose, inhibit this production and release. The liver produces glucose through gluconeogenesis and glycogenolysis. The kidney can also produce glucose through gluconeogenesis, but the liver accounts for 80% of this production (Krentz, Heinemann, Hompesch, 2015a). Tissue uptake of glucose is peripheral insulin sensitivity. Suppression of hepatic (and renal) glucose production is hepatic (and renal) insulin sensitivity. The combination of the two represents whole-body insulin sensitivity. In contrast to whole-body insulin sensitivity, the separate hepatic and peripheral insulin sensitivity measurements require radio-tracer labeled glucose infusion. In subjects with normal glucose homeostasis, the concentration of plasma insulin required to suppress hepatic glucose production maximally is substantially less than required to maximally stimulate skeletal muscle uptake of glucose (Krentz, Heinemann, Hompesch, 2015a).

Although the hyperglycemic clamp study evaluates insulin production in response to increased glucose, a more functional assessment of insulin production quantifies this production relative to whole-body insulin sensitivity. Therefore, the most precise assessment of insulin production's adequacy requires a hyperglycemic clamp study to assess insulin production and a hyperinsulinemic-euglycemic clamp study to determine whole-body insulin sensitivity to assess the magnitude of insulin production relative to whole-body insulin sensitivity.

The third type of study is the MMTT that can assess glucose changes and insulin changes in response to a standardized meal. This test allows for the estimation of both pancreatic insulin production and release as well as insulin sensitivity, although without the precision of the two clamp studies. The MMTT is being discussed because it was included in one research effort sponsored by Lilly. It was also included along with a hyperinsulinemic-euglycemic clamp study in work by another research group discussed below, and results from both study types were combined in one assessment of changes in glucose homeostasis. In this non-Lilly research, the MMTT was used to quantitate insulin production with an increase in glucose.

Several variants of the clamp studies briefly described above have been used by researchers evaluating the potential effects of olanzapine on glucose homeostasis and are included in our summaries of clamp studies evaluating olanzapine presented below. Additionally, when summarized studies include additional evaluation methods such as oral or intravenous glucose tolerance tests, these methods and their results are described.

Additional background information on the methods, the analytes measured, and the parameters computed in these studies helps understand the descriptions of research results that follow and the substantial inconsistencies in results. In the discussions below, we focus on the analytes insulin (and C-peptide that is more informative regarding β-cell function in rodent studies) and glucose. However, other analytes of potential interest, such as free fatty acids, glucagon, GLP-1, were analyzed in some studies.

The hyperglycemic clamp study assesses the adequacy of insulin production and release based on the absolute magnitude of insulin produced and released. Alternatively, as noted above, insulin production and release can be adjusted for whole-body insulin sensitivity (requiring a hyperinsulinemic-euglycemic clamp study or some other procedure to measure or estimate whole-body insulin sensitivity). This adjustment might be particularly important if assessing insulin production and release before and after a treatment that might change whole-body insulin

sensitivity. If this whole-body insulin sensitivity has been changed, it would be essential to know if a corresponding compensatory change in insulin production and release occurred. As noted, whole-body insulin sensitivity can also be assessed in the hyperglycemic clamp study but with potentially less precision than in the hyperinsulinemic-euglycemic clamp study because insulin is not being clamped. Insulin concentration will or should be rising in response to increased glucose exposure. In contrast, in the hyperinsulinemic-euglycemic clamp study, both insulin (rate of infusion, not necessarily concentration) and glucose (concentration) are clamped (held stable).

Both types of clamp studies are begun in a fasting (basal) state where stable plasma insulin and glucose concentrations are expected. An automated system frequently assays plasma glucose in both studies and adjusts the glucose infusion rate to maintain its desired concentration.

In the hyperglycemic clamp study, glucose is usually initially infused as a bolus, resulting in a rapid rise to a specified glucose concentration (a concentration believed high enough to stimulate maximum insulin production and release). The glucose concentration is generally 180 mg/dL or higher (Krentz, Heinemann, Hompesch, 2015b). Then, glucose is slowly infused at a sufficient rate (glucose frequently monitored and the infusion rate adjusted if necessary) to maintain a constant concentration at the target concentration. The glucose is maintained at this target concentration for a specified period after the expected achievement of maximal, steady-state insulin release. Insulin concentrations and concentrations of other analytes of interest (e.g., C-peptide) are measured frequently, especially early during the study (e.g., at 2 min intervals from 0-10 min, at 15-30 min intervals from 10-120 min, and 20 min intervals from 120-240 min of a 240 min hyperglycemic clamp study – as per one Lilly study protocol) with one target glucose concentration.

An alternative approach to the hyperglycemic clamp study is to clamp glucose at multiple concentrations (each concentration referred to as a ‘step’); reaching a final glucose concentration expected to elicit a maximum insulin response and assess the insulin response to those various concentrations of glucose and the change in the insulin response across the concentrations (steps) of glucose.

In the hyperglycemic clamp study, stimulation of the β-cells with the same magnitude of hyperglycemia (a constant concentration of plasma glucose) allows comparison of the β-cell response through insulin and/or C-peptide measurement between treatments and between conditions (e.g., pre- and post-treatment). Whole-body insulin sensitivity can be estimated (again, the hyperinsulinemic-euglycemic study is the more precise method for determining insulin sensitivity) based on the amount of glucose infusion necessary to maintain the hyperglycemic target normalized (divided) by the plasma insulin level. The Disposition Index is the parameter that assesses β-cell response where insulin release is adjusted for insulin sensitivity. There are multiple computing methods for the Disposition Index, described in more detail below. However, each computation method is an adjustment of insulin (or C-peptide) output for insulin sensitivity.

In the hyperinsulinemic-euglycemic clamp study, insulin is infused at one or more different rates (steps) and glucose infused to maintain a fixed glucose concentration of a typical fasting magnitude (e.g., 90 mg/dL) or the fasting concentration for individual subjects. Two 120-min steps assess endogenous glucose production inhibition in step 1 and the facilitation of glucose uptake in step 2. A maximal insulin infusion rate would be expected to completely suppress

hepatic (and renal) glucose production and maximize glucose uptake into peripheral tissues (primarily muscle and fat and the liver). A steady glucose infusion rate is achieved and held for some period at one or more insulin steps. The test is begun with an infusion of radio-tracer labeled glucose for a period before beginning the insulin infusion, and radio-tracer labeled glucose is also added to the cold (unlabeled) glucose being infused to maintain euglycemia if hepatic (and therefore potentially peripheral) and whole-body insulin sensitivity are to be measured. At basal steady-state, before beginning the actual clamp study with insulin infusion, circulating glucose is comprised of hepatically produced cold (non-labeled) glucose and radio-tracer labeled glucose being infused. The ratio of unlabeled to radio-tracer labeled glucose over specific periods allows for the computation of hepatic and/or peripheral insulin sensitivity. Multiple alternative computational formulae are all complex. Some references to these formulae are provided below, where these computations are discussed in more detail.

During the hyperinsulinemic-euglycemic clamp study, at steady-state, any cold glucose is a hepatic product plus the cold glucose infused to maintain euglycemia (a known quantity of glucose). The ratio of cold glucose to total glucose represents the ratio of hepatically produced glucose plus infused cold glucose to total glucose. These ratios (that involve total, radio-tracer labeled, and cold glucose), knowledge of the radio-tracer labeled and cold infused glucose amounts / concentrations / rates of infusion, and knowledge of the glucose concentration maintained allow computation of hepatic glucose output and peripheral glucose uptake separately. Whole-body glucose sensitivity is assessed based on the rate of glucose infusion required to maintain euglycemia. This required glucose infusion rate increases as hepatic glucose production decreases and peripheral glucose uptake increases (both changes representing better/improved sensitivity).

During the hyperinsulinemic-euglycemic clamp study, glucose uptake increases for an extended period, and steady-state is not achieved during a typical 120-minute study. However, the last 30 to 60 minutes of a two- to four-hour study are adequate for assessing insulin-mediated glucose uptake (and suppressing endogenous glucose production) (Krentz, Heinemann, Hompesch, 2015a).

Any lack of decrease (or a decrease in the magnitude of decrease before treatment) in hepatic glucose output during hyperinsulinemia would indicate a decrease in hepatic insulin sensitivity, possibly due to the treatment (unless the same was observed with placebo treatment). Likewise, any decrease in peripheral glucose uptake would indicate a decrease in peripheral insulin sensitivity. Assuming complete suppression of hepatic glucose production, at steady-state, the rate of glucose infusion is the rate of peripheral glucose uptake. If hepatic glucose production has not been completely suppressed, then peripheral glucose uptake is the sum of infused glucose and residual hepatic-produced glucose.

In some hyperinsulinemic-euglycemic clamp study protocols, somatostatin is infused to suppress any residual endogenous glucose production by the liver (inhibition of glucagon release). Endogenous insulin production would be expected to be suppressed by an insulin concentration that completely suppresses hepatic glucose production and maximizes peripheral glucose uptake. However, somatostatin also suppresses insulin release and guarantees that only infused insulin (a known concentration and known constant rate) is responsible for peripheral glucose uptake and suppression of hepatic glucose production.

The glucose infusion rate necessary to 'clamp' the glucose level at steady-state is a measure of whole-body insulin sensitivity. The radio-tracer glucose infused before and during the clamp can also allow the separate determination of glucose production (rate of glucose appearance, Ra) and utilization (rate of disappearance, Rd) by a tracer dilution principle. The suppression of glucose production by insulin during the clamp is a measure of hepatic insulin sensitivity. The stimulation of glucose uptake during the clamp is a measure of peripheral insulin sensitivity.

The computation of separate hepatic glucose production and peripheral glucose uptake is conceptually simple but computationally complex, as already noted. The computation is based on the ratio of radio-tracer labeled (infused) to cold (hepatically produced insulin plus infused) and known infusion rates as described above. In practice, the computation requires knowledge of the complex pharmacokinetics of glucose, and alternative methods have been suggested as optimal that have been used across laboratories conducting such clamp studies (Finegood, Bergman, and Vranic, 1987; Molina, Baron, Edelman, et al., 1990). Furthermore, several alternative forms of radio-tracer labeled glucose (e.g., 2-^{3}H, 3-^{3}H, 6-^{3}H, 6,6-^{2}H2, 6-^{14}C) can be used. In some computational methods, multiple radio-tracers are used. The use of alternative computational methods between studies/laboratories might complicate the comparison of study results. Whole-body insulin sensitivity is more straightforward as it is measured by the rate of glucose infusion required to maintain stable euglycemia during steady-state insulin infusion/concentration. Steady-state insulin might be best developed when somatostatin infusion is used.

The MMTT assesses the area under the curve (AUC) for glucose and insulin changes from a fasted state with frequent sampling before and after a standardized breakfast and in some protocols through a post-standardized lunchtime point. Calories are fixed for each subject; the proportions of those total calories from carbohydrates, fats, and protein are standardized across subjects (e.g., carbohydrates – 55%, fats – 30%, proteins – 15%). AUCs are computed for total values and values above the baseline (before the first meal) AUC for glucose and insulin (and other analytes of interest). Additionally, the peak values of these analytes can be determined.

In the clamp studies and the MMTT and other study methods described below, within treatment changes from baseline can be compared, between treatment differences without a baseline assessment can be compared, and between treatment, within treatment changes from baseline can be compared.

The paragraphs above describe the basic methods of these two types of clamp studies and the MMTT. However, there are differences across laboratories in specific details of study conduct. Perhaps more importantly, how parameters that are measured (infusion rates of glucose and/or insulin, glucose concentrations, insulin concentrations, concentrations of other analytes of interest) are then used to compute the parameters of interest (insulin production, insulin production relative to insulin sensitivity, hepatic insulin sensitivity, peripheral insulin sensitivity, whole-body insulin sensitivity) are also different across laboratories. A description of the measured and computed parameters and a high-level overview of some of the computed parameters' different alternatives follows. Some computed parameters are computed in both types of clamp studies, but the computation methods differ between the two types of study and across laboratories. Discussions of measured and computed parameters in the hyperinsulinemic-euglycemic clamp study and the conduct of this study type have been published in the review by Krentz, Heinemann, Hompesch, et al. (2015), and manuscripts by Muniyappa, Lee, Chen, and Quon (2008) and Bergman, Finegood, and Ader (1985) for example. The presentation of

parameters below groups the parameters as insulin-related, glucose-related, and insulin sensitivity-related. Each group of parameters begins on a separate page. However, some parameters in one group are used to compute parameters in another group. For example, insulin-related parameters and glucose-related parameters are used to compute insulin sensitivity-related parameters (the last group). Additionally, the most crucial insulin production/release parameter that assesses this production /release that takes accounts for any change in insulin sensitivity requires an insulin sensitivity parameter. Therefore, the most crucial insulin-related parameter (first group) requires a third group parameter for its computation.

Different parameters have been measured and/or computed across different research groups. Additionally, different abbreviations for the parameters have been used across researchers and their reports. Thus, for comparative purposes, throughout the summaries of studies below, we have attempted to employ the following abbreviation scheme[13]:

Insulin-related parameters:

- Insulin (I): the plasma concentration of insulin measured at a specific time or during a time interval of interest (usually the AUC for a time interval)
    - Multiple insulin concentration measurements contribute to the computation of the AUC of insulin concentrations during an interval of interest
        - The concentrations are weighted in computing the AUC based on the time between the previous concentration measurement and the current concentration measurement
- Change in Insulin (ΔI): the change in the plasma insulin concentration or insulin concentration AUC from the basal state to a step or one step to the next step in a hyperglycemic clamp study
    - Calculated but based on direct measurements of insulin concentrations
- Insulin Response (IR): the AUC of insulin concentrations during a period of interest after beginning the glucose infusion in a step of a hyperglycemic clamp study from which is subtracted the AUC of insulin at the previous step (or basal in the first or only step) [AUC-$I_{stepX+1}$ – AUC-$I_{stepX/basal}$] in a hyperglycemic clamp study
    - This parameter does not consider the adaptation of insulin response to any change in whole-body insulin sensitivity
    - Calculated but based on direct measurements of insulin concentrations
    - An alternative to the computation above is the AUC of insulin concentration during the first 10 minutes of a 1-step hyperglycemic clamp study (see Disposition Index below)
    - Conceptually and computationally equivalent to ΔI
- β-cell Slope: an intermediate variable in the computation of a parameter that adjusts insulin response for any change in whole-body insulin sensitivity
    - Computed using a 3-step hyperinsulinemic-euglycemic clamp study protocol

[13] The list of parameters with their abbreviations cover those used across the majority of studies. However, the list does not include all parameters discussed in the summaries. For example, Teff and colleagues (Teff, Rickels, Grudzia, et al. 2013) computed the C-peptide to insulin ratio in an MMTT to assess hepatic metabolism/clearance of insulin. When uncommon parameter results are presented in the summary of a research project, they are explained within the summary.

        - β-cell slope: the slope of the linear regression line for the values of IR for the multiple steps
    - Used by Richard Bergman's laboratory
- Disposition Index (DI): the parameter that assesses β-cell function (insulin production and output) between two clamp study values for any changes in whole-body insulin sensitivity (ISIw – see insulin sensitivity related parameters below following glucose-related parameters)
    - In general terms, the parameter is computed from the change in insulin (or C-peptide) output in a hyperglycemic clamp study multiplied by whole-body insulin sensitivity (in some cases, possibly peripheral insulin sensitivity)
    - DI = (β-cell slope) * (whole-body insulin sensitivity)
        - The above computation using β-cell slope is the computation used by Richard Bergman's laboratory
        - In the Lilly manuscript (Hardy, Meyers, Yu, et al., 2007) that describes alternative analyses of the Lilly hyperglycemic clamp study (Sowell, Mukhopadhyay, Cavazzoni, et al., 2002), DI = IR * (whole-body insulin sensitivity) with IR being obtained from the first 10 minutes of a 1-step hyperglycemic clamp study (DI = I $AUC_{0\text{-}10}$ * ISIw) and by a second computation (DI = I $AUC_{0\text{-}10}$ * HOMA1-IR) where (HOMA1-IR = (((fasting plasma insulin) * (fasting plasma glucose)) / 22.5))
        - See insulin sensitivity-related parameters below for additional alternatives for computation of whole-body and peripheral insulin sensitivity
- Insulin infusion rate (IIR): the rate at which insulin is infused in a hyperinsulinemic-euglycemic clamp study
    - Generally adjusted for body mass or body surface area
    - Among the insulin-related parameters, this is the parameter relevant to the hyperinsulinemic-euglycemic clamp study

Glucose-related parameters:

- Glucose Concentration at Steady State (GLU_{SS}): the AUC of glucose concentrations during a steady-state period of interest in either clamp study
    - The AUC is computed based on weighting the glucose concentrations as described above for an insulin AUC
- Glucose Infusion Rate (GIR): the rate at which glucose is being infused to maintain either hyperglycemia or euglycemia in either clamp study
    - At steady state, represents whole body (muscle, adipose tissue, hepatic tissue) glucose uptake and suppression of hepatic insulin production
    - Often adjusted for body weight or fat-free body mass when used in the computation of insulin sensitivities
    - Directly measured
- Change in Glucose Infusion Rate (ΔGIR): the change in the GIR from a basal state or step to a step up in either clamp study
- Rate of Glucose Appearance (R_a) in a hyperinsulinemic-euglycemic clamp study:
    - Two alternative definitions appear to have been used across manuscripts

        - The rate at which total glucose is added to the body and is the sum of hepatic glucose production plus infused glucose
            - With this definition, it is used in computing whole-body insulin sensitivity
        - The rate of only hepatically produced glucose (Finegood, Bergman and Vranic 1987)
            - With this definition, it is used in computing a hepatic insulin sensitivity index
- Endogenous Glucose Production (EGP), also referred to as Hepatic Glucose Output (HGO) in a hyperinsulinemic-euglycemic clamp study
    - EGP = (R_a – GIR) by the first definition above
    - EGP = R_a by the second definition above
- Change in Endogenous Glucose Production or Hepatic Glucose Output (ΔEGP) in a hyperinsulinemic-euglycemic clamp study: change in glucose production
    - The change across steps in a clamp study
- Rate of Glucose Disposal (R_d), also referred to by some authors (e.g., Lilly) as "M" (from either clamp study but more precisely determined in a hyperinsulinemic-euglycemic clamp study):
    - The rate at which tissues take up glucose
    - Used to compute insulin sensitivity
    - Will be equivalent to GIR at steady-state in either if there is no endogenous glucose production and equivalent to GIR plus hepatically produced glucose at steady-state if there continues to be endogenously produced glucose
- As described above, EGP (and R_a and R_d) are computed in a hyperinsulinemic-euglycemic clamp study by infusing radio-tracer labeled glucose before and during a study. The study allows the computation of hepatic glucose production that contributes to total glucose. These computations are based on the known quantity of infused glucose, glucose concentration, and hot to cold glucose ratio changes. Multiple methods of computation exist, for example:
    - Finegood, Bergman and Vranic (1987)
    - Molina, Baron, Edelman, et al. (1990)

Insulin sensitivity-related parameters when both hyperglycemic clamp study and hyperinsulinemic-euglycemic clamp study (or other study methods that provide insulin responses to increased glucose and changes in glucose responses to increased insulin) data are available:

- Whole-body insulin sensitivity index (ISIw), most precisely computed in the hyperinsulinemic-euglycemic clamp study:
    - Several alternative computations are very similar
        - ISIw = ΔGIR (also referred to as M) / (ΔI * GLUss)
            - Most common computation
            - It may be most precise to normalize ΔGIR/M for fat-free mass (FFM) (GIR/FFM), but a more common normalization is (GIR/weight); body surface area is sometimes used for normalization
            - Some researchers do not normalize GIR based on either FFM or weight

        - ISIw = ΔGIR /ΔI
            - Because whole-body insulin sensitivity is being compared between pre-treatment and post-treatment states and glucose concentrations that are used in both types of clamp studies would be equivalent in the pre-treatment and post-treatment clamp studies, the GLU_{SS} term would effectively cancel out in the comparison as the two GLU_{SS} values would be equivalent under the assumption that the two clamps were successful in maintaining constant glucose at the target value; this applies to ISIh, and ISIp discussed below as well
        - ISIw = GIR / I
            - This formula computes ISIw that is an absolute value at a step rather than being a change from basal state to a step or from one step to another
            - Lilly studies have computed in this way (referred to as "M / I" in Lilly manuscripts)
        - When determined in a hyperinsulinemic-euglycemic clamp, the ΔI (or I) is influenced by the insulin infusion (especially if a somatostatin infusion is used or a very high rate of insulin infusion is used), but when determined in a hyperglycemic clamp, the ΔI (or I) is influenced by the insulin release elicited by the hyperglycemia
    - ISIw can be approximated in the hyperglycemic clamp study as: ISIw = mean GIR / mean I at steady state in the study
        - In a single-step hyperglycemic clamp study, ΔGIR used in the computation is the absolute rate of glucose infusion during the step
- Peripheral insulin sensitivity (ISIp): ISIp = ΔR_d / (ΔI * GLUss)
    - If EGP is not 0, the EGP adds to GIR to give R_a and at steady-state $R_d = R_a$ if R_a is being defined as total glucose being added to the system rather than just hepatic glucose
- Hepatic insulin sensitivity (ISIh): ISIh = ΔEGP / (ΔI * GLUss)
    - Lilly: ISIh = ΔEGP / $EGP_{(basal)}$ * 100 (the ΔEGP as a percentage of the basal EGP)
        - Lilly reverses the order of subtraction of basal/clamp EGP from the order used by some other research groups
        - As insulin is being infused, even at a higher rate, we would believe that insulin concentrations might be changing within a clamp study, and the absolute basal and step 1 values might be different before and after some treatment; therefore, we believe that these two computational formulas have different results
- ISIw, ISIp, and ISIh might be calculated as absolute values at a step but are more customarily calculated as changes from basal to step 1 in a 1-step study or changes between steps in a multi-step study

The following material illustrates the use of measured parameters in a hyperinsulinemic-euglycemic clamp study plus an MMTT to compute additional parameters relevant to assessing potential glucose homeostasis changes (Teff, Rickels, Grudziak, et al., 2013). Rather than using a hyperglycemic clamp study to evaluate the insulin response to increasing glucose, this group

used an MMTT. A hyperinsulinemic-euglycemic clamp study was used to collect the measured parameters and those computed based on the clamp study data. With measured and computed parameters for insulin response and response to insulin, a comprehensive set of parameters are available or can be computed. Details of the hyperinsulinemic-euglycemic clamp study conduct and the study results are discussed below as one of the summarized studies. This study did not use a hyperglycemic clamp study to determine insulin output in the face of increased plasma glucose but used the MMTT.

Table 5: Hyperinsulinemic-Euglycemic Clam Study Procedure (Teff, Rickels, Grudziak, et al., 2013)

Total Time (min)	Time from Clam Initiation (min)	Activity
0	-120	Radio-tracer labeled glucose 5 mg/kg bolus over 5 min
	-120	Initiate radio-tracer labeled glucose 0.5 mg/kg/min continuous infusion
90	-30	Plasma sample
105	-15	Plasma sample
119	-1	Plasma sample
120	0	Insulin 1.6 μU/kg bolus over 10 min
120	0	Initiate insulin 0.8 μU/kg continuous infusion
120	0	Initiate (unlabeled glucose + ~20% of total infused glucose as radio-tracer labeled) continuous to clamp plasma glucose at 90 mg/dL
150	30	Plasma sample
180	60	Plasma sample
210	90	Plasma sample
240	120	Plasma sample
270	150	Plasma sample
300	180	Plasma sample
330	210	Plasma sample
360	230	Plasma sample
400	240	Plasma sample

Measured variables used in computations:

- F: rate of radio-tracer labeled glucose infusion
- E(t): mean of the radio-tracer labeled glucose at 2 adjacent measurements
- V: the volume of distribution of plasma glucose (40 mL/kg)
- $C_2 + C_1$: the sum of plasma glucose concentrations at times 2 and 1
- $E_1 + E_2$: the radio-tracer labeled glucose contribution to total glucose at times 1 and 2
- $T_2 - T_1$: the difference in $time_1$ and $time_2$

Computations (times relative to initiation of clamp):

- Basal glucose = mean of glucose at -30, -15, -1 min
- Rate of glucose appearance (R_a): $R_a = ((F / E(t)) – ((V * (C_2 + C_1)) / 2) / ((1 + E(t)) * ((E_1 + E_2) / (T_2 – T_1))))$ – used in the computation of EGP and R_d
- Endogenous glucose production during the clamp (EGP_{clamp}) during the clamp study: R_a – GIR – used in computation of ISIh
- Rate of glucose disposal (R_d): $R_d = R_a – ((V * ((C_2 + C_1) / (T_2 - T_1))))$ – used in the computation of ISIp
- Peripheral insulin sensitivity index (ISIp): $ISIp = ((R_{d\text{-}SS} - R_{d\text{-}basal}) / (I_{SS} - I_{basal}) * GLU_{ss})$
- Disposal index (DI): DI = (ISIp * IR [referred to AIR[14]])

[14] AIR is the acute insulin response in the MMTT calculated as the AUC for insulin plasma concentration from the onset of the meal at 2, 4, 6, 8, 10 min

In the following discussion of the Lilly conducted and non-Lilly studies, not all findings are summarized. We focus on food intake, activity level, weight, fat tissue, insulin sensitivity, and pancreatic β-cell insulin production/release. In some instances, we summarize findings for other analytes/parameters.

3.1 Lilly Studies and Analyses

These studies and analyses were initiated in the late 1990s. All Lilly work is presented first, although a supplemental analysis of one study (Hardy, Meyers, Yu, et al., 2007) and the last study (Hardy, Henry, Forrester, et al., 2011) was conducted after an important study in dogs (Ader, Kim, Catalano, et al., 2005).

The hyperglycemic clamp study (Beasley, Berg, Dananberg, et al., 2000; Sowell, Mukhopadhyay, Cavazzoni, et al., 2002) involved placebo (n=18), olanzapine (10 mg/d, n=17), and risperidone (4 mg/d, n=13) administered in parallel to healthy volunteers for 15-17 days. A single concentration (1-step) of glucose was used (200 mg/dL). The time intervals of interest for Lilly were: 1) the first phase of the insulin response (0-10 min); the second phase of the insulin response (10-240 min); and the total insulin response (0-240 min). The steady-state of glucose uptake in response to the maximum insulin response used for computing ISIw was between hours 3 and 4 (last 60 min of 4 hours). Important results are as follows.

Table 6: Change from Pre-treatment to Post-treatment in the Insulin Response (IR [pmol/L], referred to as "I") During the Time Interval of Interest:

	Placebo	Olanzapine	Risperidone
First Phase IR	-4.8 (3.4%)	69.0 (38.7%)[2]	35.4 (30.2%)[3]
Second Phase IR	-82.2 (17.9%)	117.0 (22.3%)[2]	90.0 (22.5%)[3]
TIR[1] (Total IR)	-~80(~13%)	~200 (~29%)[4]	~70 (~13%)[5]
IR At Steady State	-112.8	111.0	81.6

[1] Estimated from figure
[2] $p<0.01$, within group
[3] Inferential test results not reported, described as comparable to olanzapine
[4] $p<0.01$ within group; $p<0.001$ vs. placebo
[5] $p=0.054$ within group; $p=0.014$ vs. placebo

Table 7: Change from Pre-treatment to Post-treatment in the Glucose Infusion Rate (GIR [(mmol/kg)/min] [x10-3], referred to as "M") During the Time Interval of Interest [1]:

	Placebo	Olanzapine	Risperidone
At Steady State	0.3[1]	-2.4[1]	-7.8[1]

[1] No significant change within group

Table 8: Change from Pre-treatment to Post-treatment in the Whole-Body Insulin Sensitivity Index (ISIw, referred to as ["M/I") [x10^{-5}] During the Time Interval of Interest (estimated from GIR [referred to as M] and IR [referred to as "I"]:

	Placebo	Olanzapine	Risperidone
At Steady State	0.92	-4.63[1,2]	-3.7[2]

[1] $p<0.05$, within group
[2] $p\geq0.05$ vs. placebo

Mean weight gain with the three treatments was: placebo-0.5 kg; olanzapine-2.8 kg; risperidone-3.1 kg. Changes with both olanzapine and risperidone were significant ($p<0.01$) within the treatments and significant ($p<0.001$) vs. placebo. Multivariate regression analyses with therapy and BMI as covariates were performed for parameters of interest, including TIR and M/I (ISIw), to assess the potential influence of weight gain on these parameters.

Table 9: Change from Pre-treatment to Post-treatment in the Insulin Response (IR [pmol/L], referred to as "I") and Change from Pre-treatment to Post-treatment in the Whole-Body Insulin Sensitivity Index (ISIw) [x10^{-5}], referred to as "M/I" During the Time Interval of Interest; Multivariate Regression Analyses Including BMI (adjusting for the change in BMI):

	Placebo	Olanzapine	Risperidone
TIR - SS	-111.6[1]	42.6	-24.0
ISIw - SS	2.8	1.9	2.8

[1] $p<0.05$, within group

This study's results for olanzapine can be interpreted as follows:

1. significantly increased insulin output;
2. significantly decreased ISIw; and
3. weight gain might explain the decreased ISIw, and when accounted for, olanzapine did not decrease ISIw

The hyperinsulinemic-euglycemic clamp study (Beasley, Sowell, Cavazzoni, et al., 2001; Sowell, Mukhopadhyay, Cavazzoni, et al., 2003) included placebo (n=19), olanzapine (10 mg/d, n=22), and risperidone (4 mg/d, n=14) administered to healthy volunteers for ~21 days. As a secondary method of assessment, an MMTT was included in the study.

The clamp study was a 2-step study (insulin infused at 20 µU/m^2/min for 3 hr and 120 µU/m^2/min for 2 hr). Somatostatin was not infused, and radio-tracer labeled glucose was not infused. Steady-state for the two steps was defined as 140-160 min and 240-260 min (20 minutes, excluding the last 20 minutes of each step). Glucose was clamped at 90 mg/dL.

Weight gains with olanzapine (+1.95 kg) and risperidone (+1.6 kg) were significant within the treatments, and both were significantly different from the weight loss with placebo (-0.22 kg).

Table 10: Change from Pre-treatment to Post-treatment in Change within Study

Treatment	Low Dose Insulin		High Dose Insulin	
	ΔGIR ("M", R_d) ((mg/kg)/min)	**ISIw ("M/I")[1,2]**	**ΔGIR (M, R_d) ((mg/kg)/min)[3]**	**ISIw ("M/I")[4,2]**
Placebo	↑[5]	↑	NR	-4.7%
Olanzapine	NR[6,7]	↑	NR	6.9%
Risperidone	↓[8]	NC	NR	-0.7%

[1] Absolute change values are shown in a figure
[2] All p-values, within treatments and between treatments were non-significant
[3] All p-values within treatments and the p-value among treatments were non-significant
[4] Percent change values reported in the text
[5] p=0.019, within treatment
[6] Not reported
[7] p=0.332 vs. placebo; p-value, within treatment not reported
[8] p=0.045 vs. placebo; p=0.215 vs. olanzapine

Olanzapine was associated with slight, non-significant numerical increases in ISIw under low and high insulin steady-state conditions.

Results of the MMTT:

Table 11: Change from Pre-treatment to Post-treatment:

Treatment	Glucose AUC ((mg/dL)/min) * 10^3)		Insulin AUC (μU/min * 10^3)	
	Total	**Above Fasting**	**Total**	**Above Fasting**
Placebo	-0.18	0.32	1.0	1.0
Olanzapine	1.85[1]	0.80	1.4	1.1
Risperidone	0.47	0.50	0.7	1.0

[1.] p=0.033 vs. placebo; p=0.018 within treatment

Weight gain was better controlled with olanzapine and risperidone in this study than in the Lilly hyperglycemic clamp study, but still not wholly controlled; placebo-treated subjects lost weight, as described above. These weight changes introduce a potential confound in the interpretation of the results of the MMTT.

This study's results for olanzapine can be interpreted as follows:

1. based on the hyperinsulinemic-euglycemic study, ISIw was slightly numerically increased (not decreased); and
2. the MMTT suggests the possibility of a slight decrease in insulin sensitivity. The results of these two study types might be viewed as contradictory; Sowell, Mukhopadhyay, Cavazzoni, et al. (2003) concluded: "Nevertheless, results from the euglycemic clamps strongly suggest that the small changes in postprandial glucose and insulin observed during the MMTT in subjects treated with olanzapine or risperidone were not clinically significant and were unlikely to be due to a change in insulin sensitivity.".

Some might wonder about the hyperinsulinemic-euglycemic clamp study's sensitivity with a brief 21-day treatment period and a small number of subjects. As Sowell, Mukhopadhyay, Cavazzoni, et al. (2003) pointed out, comparable studies with β-blockers demonstrate an ~25% decrement in ISIw with 4-8 weeks treatment in 10 healthy volunteers. β-blockers are generally not considered to have a clinically significant impact on glucose homeostasis. Prednisone, 30 mg/d for seven days in 10 subjects, was associated with a 2-fold (50%) reduction in ISIw. The protease inhibitor indinavir was associated with a 34% decrease in ISIw after a single dose. The hyperinsulinemic-euglycemic clamp method is extremely sensitive in the detection of changes in insulin sensitivity.

In response to the Ader, Kim, Catalano, et al. (2005) study with its finding of a decreased pancreatic insulin response to a non-significant decrease in whole-body insulin sensitivity that is discussed below, Lilly (Hardy, Meyers, Yu et al., 2007) conducted additional analyses of the results of its initial hyperglycemic clamp study (Beasley, Berg, Dananberg, et al., 2000; Sowell, Mukhopadhyay, Cavazzoni, et al., 2002).

Hardy and colleagues (Hardy, Meyers, Yu, et al., 2007) analyzed insulin release during the first 10 minutes of the hyperglycemic clamp. The presumption was that the insulin response during this first 10 minutes (when bolus glucose was being administered) is the most sensitive indicator of the pancreatic β-cells' functional adequacy. The following parameters were computed:

- Incremental (change from baseline) in insulin AUC from 0-10 minutes ($AUC_{0\text{-}10}$). Insulin, C-peptide, and glucose were measured every two minutes during this period
- Steady-state ISIw
- Homeostasis model assessment-1 of insulin resistance (HOMA1-IR); calculated from mean baseline glucose and insulin values. HOMA1-IR = (((fasting plasma insulin) * (fasting plasma glucose)) / 22.5)
    - An alternative estimate of ISIw that can be computed without any intervention other than obtaining a venous blood sample when the subject/patient is in the fasting state
    - Lower values indicate greater ISIw compared to higher values
- Glucose disposal index (DI):
    - Computation method 1: (DI = (I $AUC_{0\text{-}10}$ * ISIw))
    - Computational method 2: (DI = (I $AUC_{0\text{-}10}$ * HOMA1-IR))
    - This DI value is the first phase insulin production value multiplied by a whole-body insulin sensitivity value. This DI is not equivalent to Richard Bergman's laboratory's computation of DI, which is the product of multiplying ISIw by the insulin production line's slope across a 3-step hyperglycemic clamp study. Richard Bergman's laboratory's DI computation method was described above, and a result for olanzapine that found a decrement in DI computed by Bergman's method is described below (Ader, Kim, Catalano, et al., 2005). However, as Hardy, Meyers, Yu, et al. (2007) computed, this DI is a parameter that does adjust insulin output for insulin sensitivity.

Table 12: Changes from Pre-treatment to Post-treatment in Relevant Parameters

Treatment	Δ AUC_{0-10} Insulin (pmol/L/min)	Δ AUC_{0-10} C-peptide (pmol/L/min)	ISIw (based on insulin) (((mg/kg)/min)/(pmol/L))	DI - Method 1
Placebo	-3.3	8.9	0.002	0.10
Olanzapine	44.0[1]	39.4	-0.007[3]	-0.18
Risperidone	22.9	75.8[2]	-0.005	0.07

[1] $p<0.05$ vs. placebo; $p<0.05$ within treatment
[2] $p<0.05$ within treatment
[3] $p<0.05$ within treatment

When DI was computed with method 2, the results did not change.

These analyses clearly show a robust pancreatic β-cell response (significant for insulin and directionally consistent for C-peptide) to an initial glucose bolus. Findings served to support Hardy and colleagues' conclusion that olanzapine did not negatively influence pancreatic function with adjustment for any change in ISIw. As additional evidence of lack of any impairment of pancreatic function, Hardy and colleagues pointed out that DI was not significantly changed, and pancreatic function is one component of DI (the other component being a measure of insulin sensitivity). Hardy and colleagues acknowledged that the two components of DI might be interdependent and that with method 1 of computing DI, both components are derived from the same clamp study. However, method 2 of computing DI used an independent measure of insulin sensitivity (based on fasting glucose and insulin) and found no significant DI decrement.

This reanalysis results for olanzapine can be interpreted as follows:

1. olanzapine had no negative effect on β-cell function; and
2. olanzapine did have a within treatment, significant negative effect on ISIw, but this effect did not differ significantly from that with placebo.

To further argue against Ader and colleagues' (Ader, Kim, Catalano, et al., 2005) conclusions regarding the impact of olanzapine on pancreatic function, Hardy and colleagues (2007) pointed out that in the Ader and colleagues' study, DI did not show an actual within-treatment statistically significant decrease. Additionally, the decrease in pancreatic function inferred by Ader and colleagues was based on the olanzapine-treated animals not showing the same increase in DI as animals induced to gain adipose tissue through dietary manipulation (not exposed to olanzapine), based on a statistically significant difference between these two groups. However, Hardy and colleagues pointed out that another group of animals who had fat-induced obesity and had not been treated with a drug showed a 62% decrement in DI as reported by the Bergman laboratory that conducted the Ader and colleagues (2005) study and that the olanzapine-treated and fat-treated dogs received different insulin infusion rates.

Hardy and colleagues acknowledged that a 2004 review published by the American Diabetes Association (2004) and authored by multiple medical societies concluded that olanzapine and risperidone were associated with an increased risk of diabetes. However, Hardy and colleagues stated that the results of this supplemental analysis of the Lilly hyperglycemic clamp study argued: "against a substantial and generalized impairment of insulin secretion with these agents

after short-term treatment." Hardy and colleagues (2007) also stated that the results of the Lilly hyperinsulinemic-euglycemic clamp study did not show "differential effects" on insulin sensitivity "in normal individuals". Hardy and colleagues (2007) acknowledged that limitations of these two Lilly clamp studies included small numbers of subjects (probably not a limitation given observations with other drug classes), short duration of the studies, and the lack of inclusion of subjects with risk factors for the development of diabetes (including excess adipose tissue and schizophrenia itself [literature supports the possibility of this association]).

We believe that it is important to note that in Hardy and colleagues' (2007) supplemental analysis of the hyperglycemic clamp results, olanzapine is the only treatment associated with a slight (non-significant vs. placebo and within treatment) decrement in DI. This DI decrement might hint at some overall decrement in glucose homeostasis where the parameter is dependent on both pancreatic function and insulin sensitivity.

Lilly performed a third clamp study (hyperinsulinemic-euglycemic) (Hardy, Henry, Forrester, et al., 2011). The study included DEXA measurements of whole-body fat mass and fat-free whole-body mass, as well as CT scans to distinguish subcutaneous fat from visceral fat changes in a 12-week (sufficient time to observe an initial change in HbA1c) comparison of olanzapine (n=41 completing both clamp studies, dose-5-20 mg/d) and risperidone (n=33 completing both clamp studies, dose-2-6 mg/d). Placebo-control was omitted as subjects were patients with schizophrenia and the intended treatment period was 12 weeks. While we are not including other human or animal studies in this review that did not include placebo, we include this Lilly study because this is, in part, a response regarding work that Lilly performed. ISIw was adjusted to fat-free mass by adjusting the ΔGIR for fat-free mass ($ISIw_{ff}$ = (ΔGIR / (fat-free mass)) / ΔI). Also, weight was included as a covariate in the analytical model. Notably, this study included (as a supplemental study with separate informed consent) the administration of radio-tracer labeled glucose (3-^{3}H) to assess suppression of EGP/R_a (specific hepatic insulin sensitivity) due to the results (Ader, Kim, Catalano, et al., 2005) summarized below. Methods for estimating hepatic-specific insulin sensitivity can be reviewed in the manuscript. Results for changes in insulin sensitivity and selected other metabolic parameters were as follows:

Table 13: Fasting Metabolic Parameters and Insulin Sensitivity Indices – Change from Pre-Treatment to Post-treatment (Clamp Study Completers – at least 1 insulin step)

	Olanzapine (Low insulin N=41 High insulin N=40)	**Risperidone (Low insulin N=33 High insulin N=30)**
HbA1c (completers)	0.07%[1] (N=36)	-0.04%[2] (N=28)
Fasting glucose (completers) (mg/dL)	5.41[3]	1.62[2]
Basal (fasting) insulin (completers) (μ/mL)	2.60[4]	1.25[2]
ISIh (low insulin step)	-5.28%[1] (N=5)	-4.33%[2] (N=4)
$ISIw_{ff}$ (low insulin step)	-9.0%[1]	-13.2%[5]
ISIh (high insulin step)	9.13%[1] (N=5)	No change[2] (N=4)
$ISIw_{ff}$ (high insulin step)	-10.4%[6]	-2.1%[2]
Total fat body mass (kg)	1.73[7]	1.08%[2]
Total lean body mass (kg)	1.53[7]	0.64[2]
Total weight (kg)	3.90[7]	2.16[8]

[1] p-value non-significant within treatment and vs. risperidone
[2] p-value non-significant within treatment
[3] p=0.007 within treatment; p non-significant vs. risperidone
[4] p=0.002 within treatment; p non-significant vs. risperidone
[5] p=0.047 within treatment
[6] p=0.036 within treatment; p non-significant vs. risperidone
[7] p<0.01 within treatment; p non-significant vs. risperidone
[8] p<0.05 within treatment

In this study without placebo control, at the low insulin dose step, olanzapine was only associated with a numerical decrement in ISIw after adjusting the parameter to fat-free mass and including weight as a covariate in the analytical model. The lack of a statistically significant decrease in ISIw at the low insulin step was maintained even if weight was not a covariate in the analytical model and separately if ISIw was not adjusted for fat-free mass. Olanzapine was associated with a significant decrement in ISIw at the high insulin step, but the authors maintain that the low insulin step is most appropriate for assessing ISIw. The small number of subjects participating in the assessment of ISIh might limit the ability to draw conclusions regarding ISIh. However, Hardy and colleagues argued that the sample size was adequate to detect the decrement noted by Ader and colleagues (2005) with only 10 dogs treated with olanzapine.

This study's results for olanzapine can be summarized as follows:

1. based on the hyperinsulinemic-euglycemic study, ISIw was slightly numerically decreased (and significantly so at the high insulin step; however, the low insulin step ISIw results might be better for assessing ISIw that supports the conclusion of no effect on ISIw.

The Hardy, Henry, Forrester, et al. (2011) work concludes the discussion of Lilly's work published in this area. However, subsequent work with olanzapine has been performed by multiple other laboratories.

3.2 Non-Lilly Clamp Studies and Mixed Meal Tolerance Tests

In 2005, Ader and colleagues (Ader, Kim, Catalano, et al., 2005) published a significant and complex study mentioned above. This work was conducted in the laboratory of Richard Bergman, a prominent researcher in diabetes who, along with colleagues, has made substantial contributions to the development of methods for assessing both pancreatic function and insulin sensitivity. The study is not without potential caveats concerning extrapolating results to patients treated clinically with risperidone, olanzapine, or other second-generation antipsychotics associated with significant weight gain as with olanzapine.

Mongrel dogs were studied with a mean weight at baseline of 28.6 kg. The doses were: olanzapine – 15 mg/d (n=10); risperidone – 5 mg/d (n=10); placebo (n=6). All animals were allowed ad libitum access to standardized food during the study. A separate group of six dogs received no treatment but were fed with a food isocaloric to the food fed to the other dogs, but higher in fat content than the dogs receiving comparative treatments. This study group's purpose was to produce a group showing a marked increase in adiposity, as was expected with olanzapine and possibly risperidone. This group was thought to help determine if any alterations in glucose homeostasis that was observed with an antipsychotic that produced weight gain were due to the weight gain (in which case comparable alterations would be expected to be observed in the fat-

fed, obese animals) or were more likely explained by an additional action of the drug. The dogs were treated for 4-6 weeks with a series of pre-treatment and post-treatment examinations.

The dogs were treated at about a 2-fold greater dose on a body-mass basis than what is likely a maximum dose for most patients treated clinically with the two test drugs. Olanzapine-treated dogs received a dose of 0.52 mg/kg/d, based on mean overall weight. Patients weighing a mean of 80 kg (accounting for greater body mass in most patients than a population without schizophrenia) would generally receive a maximum dose of olanzapine of 20 mg/d [0.25 mg/kg/d].

The parameters measured and computed for assessing pancreatic function and insulin sensitivity in Ader and colleagues' hyperglycemic and hyperinsulinemic-euglycemic clamp studies (Ader, Kim, Catalano, et al., 2005) described below differ in some ways from those described above for the Lilly studies and other researchers' studies.

Three different examinations were performed on the dogs on separate days in random order at baseline and after the 4-6 weeks of treatment:

- An abdominal MRI to measure trunk fat (subcutaneous and visceral), normalized to the volume of non-fat tissue as cm^3 / cm^3 of non-fat tissue, expressed as cm^3
- A hyperinsulinemic-euglycemic clamp study using radio-tracer labeled glucose (3-^{3}N) and somatostatin (the radio-tracer labeled glucose and somatostatin were begun 3 hr before beginning the clamp)
    - Along with somatostatin, insulin was infused at 0.15 mU/kg/min to standardize the basal insulin exposure
    - The hyperinsulinemic-euglycemic clamp was a single step with 1 mU/kg/min insulin infusion for 3 hours. The glucose concentration target was not specified in the manuscript other than as "euglycemic"; and
- A hyperglycemic clamp with 3 steps
    - glucose was clamped at three concentrations of 100, 150, 200 mg/dL over a total of 4 hours (60-, 90-, and 90-minutes for the 3 steps)

The important findings were as follows:

Table 14: Changes from the Pre-Treatment Baseline to the Post-Treatment Endpoint (data/results were not provided for cells without values or text)

Parameter Measured	Placebo (N=6)					Fat-Fed (N=6)						Olanzapine (N=10)						Risperidone (N=10)					
	BL[1]	EP[2]	Δ[3]	% Δ	P w/in[4]	BL	EP	Δ	% Δ	P w/in	P OZ[5]	BL	EP	Δ	% Δ	P w/in	P PLC[6]	BL	EP	Δ	% Δ	P w/in	P PLC
Food intake (calories)			inc[7]	inc				inc	inc					inc	inc	0.031	0.82			slight dec[8]	slight dec	0.17	
Anthropomorphic Parameters																							
Bodyweight (kg)			1.5 ±0.3	4.8 ±1.0	0.006	27.5 ±1.5		inc	inc					1.7 ±0.4	5.9 ±1.2	0.001				inc	inc	0.09	
Total abdominal fat (cm3)				27-30[10]	0.042	21.0 ±5.2	35.8 ±7.3	14.8	70.4		0.60	24.9 ±3.2	43.4 ±3.9	18.5	74.3	<0.00001	0.0088	21.9 ±3.0	31.8 ±2.8	9.9	45	0.005	>0.33
Visceral fat (cm3)				27-30[10]	0.046	13.5 ±3.1		7.0 ±3.5	51.9		0.65	13.1 ±1.3	21.8 ±1.1	8.7 ±0.9	66	<0.00001	0.025	11.4 ±0.8	17.3 ±1.2	5.9	52	0.001	>0.33
Subcutaneous abdominal fat (cm3)				27-30[10]	0.044	7.6 ±2.2		7.8 ±3.2	103		0.60	11.8 ±2.0	21.6 ±3.1	9.8 ±1.5	83	0.0001	0.0078	10.5 ±2.4	14.4 ±1.8	3.9	37	0.053	>0.33
Insulin Sensitivity Parameters																							
ISIw ((dL/min)/kg))/(μU/mL)	25.6 ±5.2	28.9 ±6.2	3.3 ±5.8	12.9	0.6			-8.9 ±4.1			0.63			-6.2 ±3.6		>0.1				-6.9 ±5.8		>0.1	
ISIp ((dL/min)/kg))/(μU/mL)	20.1 ±4.2	25.3 ±5.7	5.2	25.9	>0.3							24.3 ±4.4	23.3 ±6.4	-1.0	-4.1	>0.3		24.7 ±6.3	19.9 ±1.9	-4.8	-19.4	>0.3	
ISIh ((dL/min)/kg))/(μU/mL)	5.5 ±1.08	3.3 ±0.4	-2.2	-40.0	0.12							6.1 ±1.0	1.5 ±0.9	-4.6	-75.4	0.009		4.3 ±0.8	3.0 ±0.9	-1.3	-30.2	0.35	
β-cell Function (insulin release) Parameters																							
IR-step1 (μU/mL)			in		NS[11]									inc		0.005				inc		0.015	
IR-step2 (μU/mL)			dec		NS									inc		NS				inc		0.007	
IIR-step3 (μU/mL)			nc[12]		NS									nc		NS				inc		NS	
β-cell function (insulin release) – Relative to Insulin Sensitive Changes Parameters																							
β-cell response (slope) (((μU/mL)/(mg/dL)))					NS	0.74 ±0.21	2.18 ±0.57	1.44	194.5	0.01		1.24 ±0.15	1.07 ±0.25	-0.17	-13.7	0.58		0.64 ±0.11	0.97 ±0.10	0.33	51.6	0.038	
DI (SIclamp * β-cell slope)	33.3 ±7.9	37.9 ±16.7	4.6	13.8	0.80	14.4 ±2.4	32.7 ±9.2	18.3	127	0.053	0.02	35.7 ±4.2	24.8 ±6.6	-10.9	-30.5	0.222		19.8 ±5.0	21.8 ±2.7	2.0	10.1	0.74	

[1] Baseline (pre-treatment) mean
[2] Endpoint (post-treatment) mean
[3] Difference between treatment group endpoint mean and baseline mean

[4] p-value for the test of within treatment change from baseline
[5] p-value for the test of change compared to change with olanzapine
[6] p-value for the test of change compared to change with placebo
[7] Increased
[8] Decreased
[10] Single range for percentage fat increase was reported that applied to total, visceral, and subcutaneous fat stores
[11] Not significant
[12] No to minimal change

The Ader, Kim, Catalano, et al. (2005) manuscript did not describe essential results in an easily readable table as above but included most in lengthy text sections. The authors also switched between describing results as pre- and post-values, as absolute change values, as percent change values, and sometimes only providing a p-value for a within treatment change. Additionally, there were apparent errors in the manuscript. For example, in the Abstract, the increase in subcutaneous fat with olanzapine was described as +106%, while in the text, this increase was described as 83% (the 83% value is apparently the correct value). All of this makes the manuscript challenging to read regarding crucial details. However, we doubt that any of the important findings or conclusions are inconsistent with the data collected.

This study's results for olanzapine can be summarized as follows:

1. while olanzapine did not significantly decrease ISIw, it did result in a significant decline in ISIh;
2. insulin production did increase (significantly in step 1 and numerically in step 2 of the hyperglycemic clamp); still, the increase was not sufficient for the numerical decrease in ISIw observed with olanzapine leading to a conclusion that olanzapine negatively affected β-cell function; and
3. the lack of β-cell adaptation to a decrement in ISIw was not observed with risperidone or fat-fed dogs.

It should be noted that this conclusion regarding insulin production (pancreatic function) was based on parameters computed in Bergman's laboratory and that ISIw with olanzapine did not decrease significantly relative to placebo. Finally, insulin production increased significantly in step 1 and numerically in step 2 of the hyperglycemic clamp study.

The only olanzapine-placebo differences in change from baseline described as statistically significant were greater increases in abdominal total fat, abdominal subcutaneous fat, and abdominal visceral fat with olanzapine. The lack of significant differences from placebo should be carefully considered when assessing Ader and colleagues' conclusions regarding the effects of olanzapine on factors influencing glucose homeostasis. These conclusions were based on within olanzapine-treatment statistically significant changes and absence of significant changes within placebo-treatment. The parameters measured and computed are obviously not subjective experiences described by the dogs to the investigators during measurements and, therefore, any within treatment changes that were significant suggest a drug-related effect. However, the absence of a significant difference from placebo, or at least a strong trend toward statistical significance, raises the strong possibility that the observed within treatment changes with olanzapine could be due to random variability or systematic factors in the study methods other than drug influence.

Google Scholar and Pub Med searches on the text string ("olanzapine" and ["diabetes" or "glucose"]) performed on December 22, 2018, returned 19,500 and 893 citations, respectively. A Pub Med search on the text string ("olanzapine" AND ["placebo" OR "vehicle"]) AND ("hyperinsulinemic" or "euglycemic") performed on December 23, 2018, returned 11 unique citations (one citation was included twice for a total of 12 citations - the Sowell, Mukhopadhyay, Cavazzoni, et al., 2003). One human and one animal study did not include placebo controls. This search text string within Pub Med was insufficient to return all manuscripts describing a hyperinsulinemic-euglycemic clamp study of olanzapine that included a placebo control as evidenced by the search not returning the Ader, Kim, Catalano, et al. (2005) manuscript, for example. However, this search added to the list of manuscripts we were aware of through prior personal knowledge and other searches. A similar search but using ("olanzapine" AND ["placebo" OR "vehicle"]) AND "hyperglycemic") performed on February 6, 2019, to search for hyperglycemic clamp studies yielded no additional manuscripts. This hyperglycemic search also failed to find the Ader, Kim, Catalano, et al. (2005) manuscript. Many additional studies were conducted in both humans and animals that employed both types of clamp studies but included only active drugs and no placebo/vehicle control.

The text that follows is not an exhaustive review of subsequent studies not sponsored by Lilly that attempted to study olanzapine and the dysregulation of glucose homeostasis. However, the material summarizes all the hyperglycemic and hyperinsulinemic-euglycemic studies that we could find by the searches described above that employed placebo/vehicle control. We believe these studies to be the most important in understanding the phenomenon under discussion or suggesting a hypothesis about the phenomenon.

These summaries are presented first for human studies and then for animal studies, both in chronological order of publication.

We review 12 studies below that included hyperinsulinemic-euglycemic study assessments. Only three of these studies did not employ radio-tracer labeled glucose (Kopf, Gilles, Paslakis, et al., 2012; Boyda, Procyshyn, Pang, et al., 2013; Wu, Yuen, Boyda, et al., 2014) in their protocols. Therefore, there are nine studies described below that could measure ISIw, ISIp, and ISIh. When authors of these nine reports described results that could be interpreted as for only ISIw or ISIp (with or without additional results for ISIh), there was some degree of uncertainty about whether the results represented ISIw or ISIp. As noted above, terminology and computational methods differed across laboratories. The individual study summaries below and the tabular summary of all studies reviewed (Table 17 in Section 4) include our best assessment of which insulin production and insulin sensitivities indices were assessed and these assessments' results. In our summary table (Table 17), we focus on results from the two clamp studies, but some of these works employed alternative methods (e.g., MMTT, OGTT), and it was necessary to consider the results of all the types of methods informing our interpretations of the studies. Additionally, as all studies involved relatively small sample sizes, we interpreted strong trends for a difference from a placebo, and/or within treatment changes to support the hypothesis of an effect.

Kopf and colleagues (Kopf, Gilles, Paslakis, et al., 2012) performed a one-stage hyperinsulinemic-euglycemic clamp study (120 min with glucose clamped at 90 mg/dL) begun 30 min following study treatment administration. A 1-stage hyperglycemic clamp study was begun immediately following the completion of the euglycemic-hyperinsulinemic clamp study. The target glucose concentration in the hyperglycemic clamp study was 180 mg/dL. The

treatment design was a 3-way crossover, a single oral dose of placebo, olanzapine 10 mg, and amisulpride 200 mg administered one hour before the clamp studies. Ten subjects were studied.

In the hyperinsulinemic-euglycemic clamp study, the GIR (referred to as "M") was based on glucose measurements during the last 30 min of the clamp per the manuscript's Clamp Procedure section. However, the manuscript's Calculations section described using the mean GIR over the last 60 min of the clamp to calculate ISIw. ISIw was computed as (ISIw = (((mean GIR) / weight) / (mean I))) where (mean I) was collected during the same period as mean GIR.

In the hyperglycemic clamp study, DI was computed as (DI = (ISIw * (mean C-peptide))) with C-peptide collected at 5 and 10 min after the clamp's initiation glucose bolus.

In the hyperinsulinemic-euglycemic clamp component, ISIw with olanzapine was only slightly numerically less than with placebo. Based on C-peptide (not insulin), the pancreatic β-cell response was virtually identical between olanzapine and placebo. This single, oral dose study without a positive finding is unlikely to be particularly informative regarding glucose homeostasis changes associated with second-generation antipsychotics.

The final human study was reported by Teff and colleagues (Teff, Rickels, Grudzia, et al. 2013). Healthy, non-overweight volunteers (10 per treatment) not engaged in active exercise other than walking were treated for nine days with olanzapine (10 mg/d) associated with weight gain, aripiprazole (10 mg/d) associated with less weight gain, or placebo. Notably, the subjects' activity levels (daily caloric expenditure) were actively encouraged to be maintained at pre-study levels throughout the treatment period. The subjects underwent both a hyperinsulinemic-euglycemic clamp study and an MMTT pre- and post-treatment. Radio-tracer labeled glucose (6,6-$^{2}H_2$) was used in the clamp study. Insulin was initially infused at 1.6 μU/kg for 10 minutes, followed by 0.8 μU/kg for 240 minutes. Glucose was clamped at 90 mg/dL.

The MMT was with a single breakfast meal. The meal consisted of 10 kcal/kg with 45% carbohydrate, 15% protein, and 40% fat calories. Radio-tracer labeled glucose (1-^{13}C) was included in the meal. Parameters of interest were assayed for 330 minutes from the initiation of the meal.

In this pair of studies, the following parameters were measured or computed (see notes below Table 5 above for the formulae used for the computed parameters in the clamp study and MMT):

- At baseline before the pre- and post-treatment clamp study
    - EGP
- From the hyperinsulinemic-euglycemic clamp study
    - GIR
    - R_a
    - EGP
    - R_d
    - R_d / I
- From the MMTT
    - ΔIR (an increase from baseline in insulin AUC during the first 10 minutes after meal initiation)
    - (C-peptide) / I: an index of hepatic metabolism of insulin (lower/decreased values indicate less hepatic metabolism/clearance of insulin)
- From data from both the hyperinsulinemic-euglycemic clamp study and the MMTT
    - ISIp

- DI

The results of the study are summarized in the two tables below:

Table 15: Change from Pre-treatment to Post-treatment

Parameter Measured	Placebo (N=10)					Olanzapine (N=10)						Aripiprazole (N=10)					
	BL[1]	EP[2]	Δ[3]	% Δ	P w/in	BL	EP	Δ	% Δ	P w/in[4]	P PLC[5]	BL	EP	Δ	% Δ	P w/in	P PLC
Energy intake and Output																	
Food intake (calories/day)[4]	3300	3500	200		NS[5]	3750	4200	450		NS		3500	3200	-300		NS	
Activity (steps/day)[4]	6200	6000	-200		NS	10000	9000	-1000		NS		7000	6600	-400			
Body weight (kg)	68.1	68.5	0.4		NS	65.9	66.7	0.8			NS	67.8	67.3	-0.5			.08
Insulin Sensitivity Related																	
EGP_{clamp}[6] (mg/kg)/min			slight dec[7]		NS			slight dec		NS				slight dec		NS	
R_d (mg/kg)/min			slight dec		NS			dec	-26	<0.05				dec	-28	<0.05	
R_d/I (mg/kg)/min			-.007		NS			-.025			<0.01			-.028			<0.01
GIR (mg/kg)/min			slight dec		NS			dec	-21	<0.05				dec	-23	<0.05	
ISIp (mg/kg)/min			slight inc[8]		NS	11.7	9.1	-2.6	-22	<0.05		9.2	7.0	-2.2	-24	<0.05	
Pancreatic Function Related																	
ΔIR_{1-10} (μU/mL)/10min ("AIR" from the MMTT)			slight dec		NS	26.0	38.2	12.2	47	<0.05		23.8	27.4	3.6	15	NS	
DI (ISIp * IR)			slight inc		NS			inc		NS				inc		NS	

[1] Pre-treatment
[2] Post-treatment
[3] Some values estimated from figures
[4] Within treatment
[5] Versus placebo
[6] There was no decrease with any treatment in EGP suppression, indicating that neither active treatment had a negative effect on ISIh
[7] Decrease

[8] Increase

Table 16: Changes from Pre-treatment to Post-treatment in the Changes from Fasting to Post-Prandial-MMTT

Parameter Measured	Placebo (N=10)					Olanzapine (N=10)						Aripiprazole (N=10)					
	BL[1]	EP[2]	Δ[3]	% Δ	P w/in[4]	BL	EP	Δ	% Δ	P w/in	P PLC	BL	EP	Δ	% Δ	P w/in	PLC[5]
Glucose AUC (ng/mL)/360min			1600					850			NS[5]			1600			NS
Insulin AUC (μU/mL)/360min			600	5	NS			4650	73		<0.05			3000	24		NS
ΔIR_{1-10}	See the Previous Table					See the Previous Table						See the Previous Table					
C-peptide AUC (ng/mL)/360min			16		NS			170.0			NS			173.1			NS
C-peptide/insulin AUCs ratio	.12	.10	.02		NS	.18	.10	-.08		<0.05		.13	.10	-.03		<.05	
GLP-1 AUC (ρg/mL)/360min			30					220			<0.05			-85			NS
Glucagon AUC (ng/mL)/360min			-2100					1900			<0.05			-2200			NS

1 Pre-treatment baseline
2 Post-treatment endpoint
3 Some values estimated from figures
4 Within treatment
5 Versus placebo
6 Supplementary data available online

Recall that C-peptide is a cleavage by-product of insulin production (C-peptide links insulin A-chain and B-chain and is cleaved). In general, C-peptide is a marker of insulin production and release, as is insulin itself. Insulin is cleared by hepatic metabolism. Therefore, as the C-peptide/insulin ratio decreases, there is likely a decrease in the hepatic metabolism of insulin that can be a response to decreased insulin sensitivity.

This study's results for olanzapine and their potential implications can be summarized as follows:

1. did not increase postprandial glucose;
2. reduced ISIp;
3. did not reduce ISIh (based on lack of increase in EGP);
4. increased the pancreatic β-cell insulin response, with the reduction in ISIp (DI numerically increased, not decreased, but the increase was not statistically significant);
5. increased prandial glucagon, which might increase prandial/post-prandial glucose;
6. increased GLP-1;
7. decreased hepatic clearance of insulin;
8. the increased GLP-1 and decreased hepatic clearance of insulin could serve as effects that are responses to decreased ISIp and serve to increase ISIp and reduce plasma glucose concentrations, but the increased glucagon would tend to raise glucose concentrations; and
9. the findings above are not explained by weight gain or food intake in the olanzapine-treated subjects.

The methods employed in many animal studies could not be employed in humans. While the relevance of these methods' results to these drugs' effects with human oral ingestion of olanzapine or other second-generation antipsychotics, even in large doses, might be questioned, these studies suggest important mechanistic hypotheses.

Houseknecht, Robertson, Zavadoski, et al. (2007) and colleagues performed a hyperinsulinemic-euglycemic clamp study, sponsored by Pfizer, in female rats. Radio-tracer labeled (3-^{3}H) glucose was used to assess EGP, and in a separate study, (U-^{14}C) 2-deoxyglucose was used to assess glucose uptake into muscle, fat, and liver. At steady-state infusion of insulin, somatostatin (to suppress endogenous insulin release), and glucose, the animals were given single subcutaneous (s.c.) injections of either vehicle (placebo control), or olanzapine, or clozapine, or ziprasidone (multiple doses of each were tested). The test drugs' doses were intended to result in D_2 occupancy in brain tissue comparable to what would be expected in humans receiving the medications clinically. Risperidone was studied but without radio-tracer labeled glucose. These single, acute doses of olanzapine and clozapine significantly reduced ISIw but not ISIp. ISIh was decreased. These findings suggest that the drugs exerted their adverse effect on ISIw, primarily through their adverse effect on ISIh. The finding of this effect after a single dose without any opportunity for a clinically relevant change in fat mass supports the belief in a direct effect.

Martins and colleagues (Martins, Hass, and Obici, 2010) administered vehicle and olanzapine both by intravenous (IV) infusion and by intracerebroventricular (ICV) injection. The treatments were single doses by the two administration routes in separate groups of male rats. The rats were assessed after administration (basal period) and then during a hyperinsulinemic-euglycemic clamp study. IV olanzapine increased EGP (R_a) during the basal period. During the clamp study, GIR and peripheral glucose uptake (R_d) were reduced (suggesting a decrease in ISIw and ISIp), and EGP was further increased relative to vehicle (decreased ISIh). There was also an increase in mRNA levels for hepatic enzymes required for hepatic glucose production. Intracerebroventricular olanzapine resulted in similar changes to IV olanzapine except that there was no decrease in peripheral glucose uptake (R_d), suggesting the absence of a decrease in ISIp with ICV administration. The study suggests that both peripherally circulating olanzapine and

olanzapine in the CNS (presumably without substantial peripheral exposure) can decrease insulin sensitivity, especially ISIh. IV olanzapine was also found to increase phosphorylation of hypothalamic AMPK, and ICV olanzapine increased hypothalamic neuropeptide-Y underscoring the possibility olanzapine has a direct effect on glucose homeostasis through effects in the hypothalamus.

Park and colleagues (Park, Hong, Ahn, et al., 2010) experimented with groups of ovariectomized (OVX) and non-ovariectomized (nOVX) female rats. All the rats were diabetic due to a 90% pancreatectomy. The OVX rats were treated with estrogen replacement or placebo-estrogen. All three groups (OVX with estrogen replacement, OVX without estrogen replacement, nOVX) were treated for eight weeks with placebo, risperidone 0.5 mg/kg/d, or olanzapine 2 mg/kg/d. All rats were fed high-fat diets. Olanzapine induced more food intake, body weight gain, and fat gain in OVX and nOVX rats. Changes from pre- to post-treatment hyperinsulinemic-euglycemic clamp study results demonstrated decreased ISIh with increased glucose output (and increased hepatic enzymes involved in gluconeogenesis) with olanzapine in both OVX and nOVX rats but to a lesser extent in nOVX rats than in OVX rats. Estrogen replacement in the OVX rats attenuated the decrement in ISIh. These changes with olanzapine were not observed with risperidone.

Albaugh and colleagues (Albaugh, Judson, She, et al., 2011) conducted perhaps the most elaborate study conducted in male rats. After two days of administration of vehicle, olanzapine 4 mg/kg/d or olanzapine 10 mg/kg/d, assessments of acute administration effects were performed. Assessments were also performed after chronic administration of vehicle or olanzapine titrated up to 12 mg/kg/d over 14 days. Chronic treatment was continued for various times depending on the assessment performed (see the summary of results below). The multiple assessments were performed at various times during chronic treatment. The overall study included the following assessments and procedures:

- Locomotor activity
- Actual energy expenditure using indirect calorimetry
- Weight
- Body composition (adiposity and lean mass) using NMR
- An oral glucose tolerance test (OGTT)
- An insulin tolerance test (ITT) following an acute intraperitoneal (IP) administered dose of insulin
- A hyperinsulinemic-euglycemic clamp study (with methods to allow measurement of tissue-specific [multiple tissues] glucose uptake) and tissue-specific (e.g., hepatic) insulin sensitivity
- Adipose tissue fatty acid uptake
- Tissue lipogenesis
- An isoproterenol challenge test assessing hepatic glucose response and adipose tissue response (the measurement of glycerol and free fatty acids release) to assess lipolytic activity

Multiple metabolic-related analytes (e.g., glucose, insulin, C-peptide, free fatty acids) were measured during the assessments. A number of these assessments compared a vehicle-treated group to a drug-treated group at the time of assessment rather than a change from pretreatment to assessment time between the two treatments. Different cohorts of olanzapine-treated and vehicle-treated animals were used for different assessments, both acute and chronic.

In the hyperinsulinemic-euglycemic clamp studies, the basal/fasting period was 120 minutes. During this basal period, radio-tracer labeled glucose (3-^{3}H) was infused to measure basal hepatic EGP. The insulin infusion was a single step of 1 µU/kg/min infusion (following an initial bolus of 75 µU/kg) for 180 min. Glucose was infused, and glucose was clamped at 100 mg/dL. The (3-^{3}H) glucose infusion was continued at an increased rate (from 0.2 µCi/min to 0.4 µCi/min). Blood samples were collected at -20, 0, 60, 120, 160, and 180 minutes. As with other summarized clamp studies, the parameters computed to assess insulin sensitivity were conceptually like those used in other laboratories but unique in computation details.

An additional component of the research measured tissue-specific glucose uptake.

This study's results for olanzapine can be summarized as follows:

1. acute treatment (two days):
    a. did not change food consumption;
    b. decreased locomotor activity;
    c. increased energy expenditure during daylight hours without changing it during dark hours;
    d. increased fasting glucose and slightly increased glucose in the OGTT;
    e. did not change insulin in the OGTT;
    f. decreased the response to insulin in the ITT;
    g. decreased whole-body glucose disposal;
    h. did not change ISIh;
    i. decreased ISIw with a decrease in muscle but no change or an increase in adipose tissue;
    j. increased adipose tissue lipogenesis;
    k. decreased lipolysis;
2. chronic treatment:
    a. did not change food consumption (over 21 days);
    b. did not change body weight (over 28 days);
    c. did not change activity (during the third week);
    d. increased body fat and decreased lean body mass (fat mass – weeks 1-5, lean body mass – weeks 3 & 5);
    e. increased glucose in the OGTT (after four weeks);
    f. increased insulin in the OGTT (after four weeks); and
    g. decreased the response to insulin in the ITT (after 6 weeks).

This study suggests that olanzapine directly affects insulin sensitivity based on fasting glucose, the acute OGTT results, and the ITT results. The decrement in insulin sensitivity is not accompanied by a compensatory, acute increased pancreatic insulin response. An acute decrement in activity occurred, but there was an inexplicable lack of decrease in energy expenditure. There was increased fat mass without a change in body weight (probably due to loss of muscle mass). The increased fat mass was likely due to decreased activity, no decrease in food consumption, increased lipogenesis, and decreased lipolysis with chronic exposure. While in the OGTT, after both acute and chronic treatment, glucose increased in both (more after chronic dosing), insulin did not increase after acute dosing but did after chronic dosing. The OGTT cannot be considered a robust test of pancreatic insulin response. Increasing adiposity could augment a direct adverse effect suggested by the acute results.

Girault and colleagues (Girault, Alkemade, Foppen, et al., 2012) reported three experiments conducted in separate groups of male rats. One group of rats was treated with intragastric infusion of a total of 3.66 mg/rat of olanzapine over a 165 min period. A second group was treated with intracerebroventricular infusion of 36.6 µg/rat of olanzapine over a 165 min period. Doses were intended to result in CNS dopamine receptor occupancy of approximately 70%. Comparable groups received vehicle by intragastric and ICV infusion. Experiment 1 assessed total EGP using a radio-tracer labeled glucose (6,6-$^2H_2$) infusion before and during the administration of vehicle or olanzapine by both intragastric and ICV routes of administration. Experiments 2 and 3 were hyperinsulinemic-euglycemic clamp studies using radio-tracer labeled glucose (6,6-$^2H_2$) infusion using different insulin doses. In experiment 2 (low dose insulin experiment), insulin was administered as follows: initial bolus of 7.2 µU/kg/min for 5 minutes followed by 3 µU/kg/min for the rest of the experiment. In experiment 3 (high dose insulin experiment), insulin was administered as follows: initial bolus of 21.6 µU/kg/min for 5 minutes followed by 9 µU/kg/min for the rest of the experiment. The clamp study was conducted during vehicle or olanzapine administration by both intragastric and intracerebroventricular administration routes.

This study's results and potential interpretations for olanzapine can be summarized as follows:

1. Experiment 1 (non-clamp) – EGP/R_a assessment using intragastric olanzapine (vehicle: n=5; olanzapine: n=6)
    a. substantially increased glucose concentration from baseline;
    b. slightly numerically increased EGP;
        i. the increased concentration without a substantial increase in endogenous production suggests less peripheral glucose uptake;
    c. increased corticosterone concentration from baseline;
    d. did not increase insulin concentration;
2. Experiment 2 (hyperinsulinemic-euglycemic clamp) – assessment of ISIp (tissue glucose uptake) and ISIh (EGP) with lower dose insulin using intragastric olanzapine (vehicle: n=8; olanzapine: n=7)
    a. glucose uptake was significantly increased but to a smaller numerical extent than with placebo;
        i. olanzapine marginally decreased ISIp;
    b. resulted in only a slight numerical decrease of EGP relative to the significant decrease observed with vehicle;
        i. olanzapine significantly negatively affected ISIh;
    c. increased corticosterone concentration from baseline;
3. Experiment 3 (hyperinsulinemic-euglycemic clamp) – assessment of ISIp (glucose uptake) and ISIh (EGP) with higher dose insulin using intragastric olanzapine (vehicle: n=8; olanzapine: n=7)
    a. glucose uptake was significantly increased with both olanzapine and vehicle but significantly less with olanzapine than with vehicle;
        i. olanzapine decreased ISIp;
    b. resulted in a significant decrease of EGP, but significantly less with olanzapine than with vehicle;
        i. olanzapine significantly affected ISIh; and
    c. increased corticosterone concentration from baseline.

None of these effects described with intragastric administration of olanzapine were observed with ICV administration.

With low-dose insulin, olanzapine resulted in a larger negative effect on ISIh, while with higher-dose insulin, there was a large negative effect on ISIp.

The overall interpretation of these results could be that olanzapine can have immediate (therefore independent of weight gain) adverse influences on glucose homeostasis due to peripheral (muscle, adipose tissue, liver) effects. These adverse influences are not mediated through central nervous system activity.

Boyda and colleagues (Boyda, Procyshyn, Pang, et al., 2013) performed single-dose studies of multiple doses of intraperitoneal asenapine, iloperidone, or olanzapine compared to vehicle in female rats. Separate experiments were performed with an intravenous glucose tolerance test (IGTT) and a hyperinsulinemic-euglycemic clamp study. For olanzapine, the IGTT was performed with five doses, 0.01, 0.05, 1.5, 5.0, and 10.0 mg/kg. There was no effect of olanzapine on fasting glucose concentrations. These three higher doses, especially the two highest doses, resulted in statistically significantly higher glucose concentrations in the IGTT. There was also a dose-related increase in insulin with olanzapine in the IGTT. In the IGTT, olanzapine resulted in only a trend increase in HOMA-IR values ($p=0.09$). In the hyperinsulinemic-euglycemic clamp study, two doses of olanzapine were used, 1.5 and 15 mg/kg. The lower dose resulted in a slightly but statistically significant, lower GIR required to maintain euglycemia than with vehicle, and the higher dose resulted in a greater statistically significantly lower GIR required to maintain euglycemia. The findings would suggest impaired glucose homeostasis, likely due to impaired ISIw.

Wu and colleagues (Wu, Yuen, Boyda, et al., 2014) conducted a study identical in design to that described above by Boyda, Procyshyn, Pang, et al. (2013) (same laboratory affiliation) except that study drug administration was by s.c. injection rather than intraperitoneal injection. The rats received an olanzapine dose of 10 mg/kg in the IGTT and olanzapine doses of either 1.5 or 15 mg/kg in the hyperinsulinemic-euglycemic study experiment. Lurasidone was also included in these experiments. Results with olanzapine were generally like those observed in the Boyda, Procyshyn, Pang, et al. (2013) work. Olanzapine had no significant effect on fasting glucose in the IGTT. However, in this study, olanzapine was associated with higher concentrations of fasting insulin. Additionally, in this study, an increase in insulin resistance was suggested by the HOMA-IR parameter in the IGTT. In this study's hyperinsulinemic-euglycemic component, the lower dose of olanzapine resulted in a slightly lower GIR required to maintain euglycemia than with vehicle, and the higher dose resulted in a statistically significantly lower GIR to maintain euglycemia. The findings suggest impaired glucose homeostasis, likely due to impaired ISIw.

The authors of the following three manuscripts were affiliated with the same laboratory.

Hahn and colleagues (Hahn, Chintoh, Remington, et al., 2014) studied male rats with a hyperinsulinemic-euglycemic clamp and a hyperglycemic clamp in three separate experiments.

In a preliminary experiment, the researchers determined an ICV dose of olanzapine that was well tolerated and significantly decreased the locomotion induced by a 1 mg/kg intraperitoneal injection of d-amphetamine.

In experiment 1 (hyperglycemic clamp study), baseline glucose, insulin, and C-peptide plasma concentrations were obtained. Baseline concentrations were obtained 10 minutes and

immediately before treatment administration. The rats then received vehicle (N=8) or 75 µg olanzapine (N=9) by ICV injection. The clamp study was begun immediately after treatment with a glucose bolus followed by a continuous glucose infusion. Glucose was clamped at 300 mg/dL. Plasma glucose, insulin, and C-peptide concentrations were obtained 10 minutes before treatment administration and periodically over a 90-minute clamp period.

In experiment 2 (hyperinsulinemic-isoglycemic clamp study[15]), radio-tracer (3-^{3}H) glucose administration was begun 90 minutes before the clamp was begun and continued throughout the clamp. The clamp was begun with the insulin infusion (5 µU/kg/min) and a variable rate of unlabeled glucose to maintain a glucose concentration equivalent to an animal's individual, pre-study fasting concentration. Rats received vehicle (N=6) or 75 µg olanzapine (N=6) by ICV injection 90 minutes after the beginning of the clamp. The steady-state period was between 190- and 220-minutes following the initiation of the clamp. Concentrations of the relevant analytes were collected at baseline, 30 minutes and immediately before the clamp began, and every 10 minutes beginning 60 minutes after the clamp began until its completion.

In experiment 3 (hyperinsulinemic-euglycemic clamp study), somatostatin was administered with insulin. Radio-tracer (3-^{3}H) labeled glucose administration was begun 90 minutes before the clamp was begun and continued throughout the clamp. The rats received vehicle (N=10) or 75 µg olanzapine (N=10) by intracerebroventricular injection, and the clamp was immediately begun. Insulin was infused at 3 mU/kg/min along with the somatostatin. The glucose clamp target was 120 mg/dL. The clamp was continued for 120 minutes. Concentrations of the relevant analytes were collected at baseline, immediately before the clamp began, and every 10 minutes beginning 60 minutes after the clamp began until its completion.

Hyperglycemic clamp study parameters:

- GLU
- GIR (adjusted for weight)
- I
- C-peptide
- ISIw = GIR / $(I * GLU)_{SS}$ (although authors describe it as a measure of peripheral insulin sensitivity)
- DI = ISIw * (C-peptide Concentration) (in rats, C-peptide concentration used rather than insulin concentration because the pharmacokinetics of insulin are not sufficiently known to calculate the insulin secretion rate)

Hyperinsulinemic-euglycemic (isoglycemic) clamp study parameters

- GLU
- GIR (adjusted for weight)
- I
- C-peptide

[15] The hyperinsulinemic-isoglycemic clamp study is a variant of a hyperinsulinemic-euglycemic clamp study. In the hyperinsulinemic-isoglycemic clamp study, glucose is clamped at each subject's fasting/basal glucose concentration instead of a fixed concentration for all subjects considered to be euglycemic.

- R_a basal and during the clamp (basal R_a = EGP; during clamp if R_a is referring to total glucose and not just hepatic glucose, then R_a = EGP + infused glucose and EGP = R_a – infused glucose)
- EGP
- R_d

This study's results for olanzapine can be summarized as follows:

1. Hyperglycemic clamp study
    a. decreased GIR;
    a. decreased insulin concentration;
    b. decreased C-peptide concentration;
    c. no change ISIw;
    d. decreased DI;
2. Hyperinsulinemic-euglycemic clamp study (experiments 2 and 3 same results)
    a. No change in GIR;
    b. No change in I;
    c. No change in EGP (ISIh); and
    d. No change in R_d (ISIp).

This study found that an acute, CNS dose of olanzapine reduced pancreatic insulin output without effects on other glucose homeostasis parameters. This finding suggests that olanzapine has a direct (not requiring increased adipose tissue) adverse effect on pancreatic β-cell function mediated through CNS effects.

Remington and colleagues (Remington, Teo, Wilson, et al. 2015) reported the results of a study conducted in male rats intended to determine if metformin would attenuate the adverse effects of acute, oral olanzapine administration. Rats administered olanzapine alone received a single 3 mg/kg s.c. dose immediately before the hyperinsulinemic-isoglycemic clamp study that included radio-tracer labeled glucose (3-^{3}H). The protocol parameters were like those for the Hahn, Chintoh, Remington, et al. (2014) study above, except for the steady-state period relative to beginning the clamp studies. The steady-state period for this study was between 150 and 180 minutes following the initiation of the insulin.

This study's results for olanzapine treatment can be summarized as follows:

1. reduced ISIh based on the higher R_a; and
2. reduced ISIp based on the lower R_d.

Kowalchuk and colleagues (Kowalchuk, Teo, Wilson, et al., 2017) performed a study in the laboratory's male rat model to assess a potential mechanistic contribution to glucose homeostasis dysregulation by olanzapine. Based on their earlier work (Hahn, Chintoh, Remington, et al., 2014) and others' work, the researchers hypothesized that olanzapine disrupts the CNS actions of insulin, leading to disruption of peripheral glucose homeostasis. To evaluate this hypothesis, the researchers employed what they termed a pancreatic-euglycemic clamp study. In this clamp study, insulin (a total of 30 μU during the entire experiment) or vehicle was administered ICV. Radio-tracer labeled glucose (3-^{3}H) was infused peripherally for 90 minutes before beginning and during the clamp study. At 90 minutes, the clamp was initiated. A peripheral infusion of somatostatin (3 μg/kg/min), insulin (1 μU/kg/min), and variable glucose to maintain euglycemia was administered throughout the clamp study. A single, s.c. injection of olanzapine (2 mg/kg, resulting in a 70% occupancy of CNS D_2 receptors) or vehicle was administered immediately

before beginning the clamp study. The clamp study was continued for 120 minutes. Therefore, four combinations of treatments were studied: 1) vehicle+ICV-vehicle; 2) olanzapine+ICV-vehicle; 3) vehicle+ICV-insulin; 4) olanzapine+ICV-insulin.

This study's results for olanzapine treatment can be summarized as follows:

1. olanzapine did not change R_d: No change in ISIp;
2. olanzapine did not change R_a: No change in ISIh;
3. olanzapine+insulin compared to vehicle+insulin: no effect on CNS insulin's capacity to increase R_d - olanzapine did not affect CNS insulin's ability to increase ISIp; and
4. olanzapine+insulin compared to vehicle+insulin: decreased CNS insulin's capacity to decrease R_a - olanzapine reduced CNS insulin's ability to increase ISIh.

We have reviewed 17 studies above. There were many differences in methods, both in the protocols and the computations across these studies for hyperglycemic clamp studies and hyperinsulinemic-euglycemic (or isoglycemic) clamp studies. We will not specifically summarize these differences.

However, it is essential to summarize the inconsistencies among study results that could be due to methodological differences or other unknown factors. The summary focuses on absolute and relative[16] insulin response and insulin sensitivity (hepatic, peripheral, and whole-body). The consistencies and inconsistencies among studies with respect to these results are summarized in the table (Table 17) in Section 4 below. The 17 studies that suggest effects on insulin response and insulin sensitivity provided by methods other than the two types of clamp studies have been included in the individual study summaries above but are not included in Table 17 below. In some instances, data from a non-clamp method were combined with data from a clamp study to assess insulin sensitivity, and in such cases, these results are included

4. Summary of Study Findings: Insulin Production and Insulin Sensitivity

[16] Relative insulin response: absolute insulin output adjusted for any change in insulin sensitivity.

Table 17: Summary of Study Findings Related to Olanzapine and Insulin Sensitivities as well as Pancreatic Insulin Response[1]

Parameter Measured	CNS Effect					Peripheral Effect										Wt[2]
	Acute Dose					Acute Dose					Repeated Dose					
	ISIh	ISIp	ISIw	Insulin Response		ISIh	ISIp	ISIw	Insulin Response		ISIh	ISIp	ISIw	Insulin Response		
				Absolute	Relative				Absolute	Relative				Absolute	Relative to ISIw	
Lilly Studies / Analyses																
Sowell (2002)													Y	N		Y
Sowell (2003)													N[3]			
Hardy (2007)													Y[4]	N	N	
Hardy (2011)											N[5]		N			
Single Studies from Different Laboratories - Human																
Kopf (2012)								N	N							
Teff (2013)											N	Y		N	N	N
Single Studies from Different Laboratories – Non-Human																
Ader (2005)											Y		N	N	Y	N
Houseknecht (2007)						Y	N	Y								N
Martins (2010)	Y	N	Y			N	Y	Y								N
Park (2010)											Y					
Albaugh (2011)						N		Y[6]								N[7]
Girault (2012)	N	N				Y	Y									N
Studies from the University of British Columbia Laboratory – Non-Human																
Boyda (2013)								Y								N
Wu (2014)								Y								N

Studies from University of Toronto Laboratory – Non-Human																
Hahn (2014)	N	N	N	Y	Y											N
Remington (2015)											Y	Y	Y			N
Kowalchuk (2017)	!8	N9														

[1] Y: the study [found] olanzapine to have a negative effect on the parameter; N: the study found olanzapine not to have a negative effect on the parameter; if a cell is empty, the study did not evaluate the parameter

[2] Yes if the study suggested that weight or fat gain could be a variable leading to the adverse effects observed; no if the effects unlikely to be influenced by weight/fat gain (no if findings observed with acute dose, findings not observed with weight gain without olanzapine, findings observed without weight gain)

[3] The hyperinsulinemic-euglycemic clamp study did not support a negative effect on ISIw, but the MMTT did support such a negative effect

[4] ISIw decreased significantly within olanzapine but not with olanzapine compared to placebo; the limited sample sizes resulted in the interpretation of an effect

[5] There was a non-significant numerical decrease in ISIh with olanzapine, and small sample size might suggest an effect; the interpretation is based on the authors' position that with the sample size employed, an effect comparable to that found in the Ader study should have been reproduced

[6] Decreased ISIp in muscle tissue, but glucose uptake increased in adipose tissue

[7] Although a weight change was not found with chronic dosing, a substantial increase in adipose tissue with a loss of muscle mass was found that might have mediated observed negative effects on insulin sensitivity in the OGTT and ITT with chronic dosing only; as the effects on insulin sensitivity with chronic dosing were not obtained from a hyperinsulinemic-euglycemic clamp study, they are not included in the table

[8] Acute IP olanzapine did not reduce EGP during CNS administration of insulin (therefore decreased CNS insulin's ability to increase ISIh)

[9] Acute IP olanzapine did not change peripheral glucose utilization during CNS administration of insulin

It is evident that olanzapine and some other second-generation antipsychotics can be associated with substantial weight gain and that an excess incidence of diabetes mellitus is observed in patients treated with these agents. However, we do not believe that the question has been adequately addressed as to whether these agents have a direct diabetogenic effect in the absence of weight/fat gain. Indirect impairment of glucose homeostasis could be due to one or more factors:

- Increased appetite and/or decreased satiety with ensuing weight/fat gain
- Sedation leading to a decrease in active caloric expenditure with ensuing weight/fat gain

- Decreased basal metabolic rate due to the direct influence of the drug with ensuing weight/fat gain

Direct impairment of glucose homeostasis is possible. Additionally, such direct impairment could result in only a subset of humans due to a discriminating set of genetic factors, environmental experiences, or genetic-environmental interactions.

We believe the question of whether these agents are directly diabetogenic or indirectly diabetogenic, specifically in human subjects, is critically essential in optimal patient treatment and management. Weight gain or lack thereof can be easily monitored and possibly predicted early in treatment (Lipkovich, Jacobson, Caldwell, et al., 2009). Activity level can almost as easily be monitored. Total and regional body fat (but not specifically visceral fat) can be monitored with DEXA that is a relatively quick (15 minute) procedure and not excessively costly at many centers ($50-75). Any individual agent's risk-benefit could be more precisely assessed for individual patients with better knowledge regarding direct or indirect effects.

5. A Hypothetical Study to Resolve the Important Uncertainties

We believe these agents' labeling in 2003 (and subsequent class labeling for any agent approved as an antipsychotic) with a warning regarding diabetes without acknowledging the potential for differential risk among the agents and the potential for the risk to be due to indirect effects effectively curtailed interest in funding these types of studies. Additional studies are necessary to address direct effects that could not be predicted versus indirect effects that could be predicted before selecting a treatment agent or early in treatment by monitoring ongoing changes or lack thereof. Pharmaceutical companies would be the source of and would have a vested interest in funding such studies. However, the required labeling was a major disincentive to the funding of high-quality research in humans. It is virtually impossible to alter, remove or not include (for a new product) class labeling once it has been required. While additional human clamp studies have been performed, they have lacked placebo control and aimed to compare a new agent or agent perceived as lacking a diabetogenic effect to an active agent thought to have the effect, generally olanzapine and sometimes clozapine.

Disruption of glucose homeostasis can be due to one or more factors. These factors include, among others: 1) a decrement in pancreatic insulin production; 2) a decrement in peripheral glucose uptake; 3) excess hepatic glucose production; and 4) changes in several other endogenous substances such as glucagon, free fatty acids, GLP-1, other substances influencing glucose production and/or disposal. It is possible, if not likely, that all such endogenous substances have been identified. If such a disruption in glucose homeostasis is sufficiently substantial and prolonged, diabetes mellitus develops and is a diagnosable medical disorder. If this disorder develops while an individual is taking any medication, then diabetes mellitus is an AE for the drug or drugs the individual is taking.

Diabetes mellitus may be one of the few medical disorders where the probability of being an ADR for a medication can be assessed with good sensitivity and reliability with a clinical pharmacology study. The clamp methods described above are sensitive to detecting early changes in glucose homeostasis parameters before the onset of overt disease. The hyperglycemic and hyperinsulinemic-euglycemic/isoglycemic clamp studies should be able to address the question of a direct or indirect effect that could lead to diabetes mellitus. The clamp studies would be accompanied by measurements of weight, lean body mass, subcutaneous fat, visceral fat, activity, basal (or resting) metabolic rate, and daily total caloric expenditure when olanzapine

is administered orally to healthy adults. While the actual basal metabolic rate is challenging to measure, hand-held devices can measure resting metabolic rate relatively inexpensively.

There is one very important potential limitation to this set of two clamp studies' ability to resolve the question of whether second-generation antipsychotics constitute risk factors for a direct effect that can lead to diabetes mellitus or facilitate the development of diabetes mellitus indirectly through increasing weight/adiposity or some other mediating effect. If a direct effect occurs on a broad population basis, the pair of studies should detect or exclude the effect with sufficient sample sizes in several groups of subjects. However, if a direct effect occurs, but only in a small subset of the general population (a potential mentioned above), the pair of studies could result in a false-negative finding. While this limitation is essential to bear in mind, it would be very beneficial to address a direct or indirect effect on the general population basis.

The following material outlines a hypothetical study that should address whether olanzapine treatment is a risk factor for the development of diabetes mellitus through some direct effect or only through some mediating effect that might be monitored and controlled if necessary, to reduce risk.

Between 20 to 30 subjects per treatment group treated for four weeks should be more than sufficient to detect an effect or lack of effect. The study would be conducted on one or more inpatient metabolic wards. If multiple metabolic wards were to be used, they would all need identical glucose clamp equipment with identical calibration and maintenance protocols. All subjects would be kept on inpatient status throughout the entire study to ensure the maintenance of assigned diets and activity levels described below for separate subject groups.

Before beginning the study treatments, subjects would be continuously observed on an inpatient metabolic ward for four weeks. Bodyweight and adipose tissue (subcutaneous and visceral) would be measured daily. Daily activity (approximate caloric expenditure) would be measured daily and stabilized. Daily caloric would be stabilized with fixed individual diets intended to maintain a constant bodyweight (bodyweight at screening). The diets would be standardized for proportions of protein, carbohydrate, and fat across all subjects.

Given the findings in the studies reviewed, several treatment groups would be important:

- Group 1 – Olanzapine 10 mg/day with no restriction on additional food consumption over the fixed, standardized diet and no mandated activity level – a group expected to gain weight and fat due to any combination of increased caloric intake, decreased activity, and decreased basal metabolic rate
- Group 2 – Olanzapine 10 mg/day with no restriction on additional food consumption over the fixed, standardized diet, but mandated to continue daily activity at the same level as before beginning treatment – a group that might gain weight and fat due to increased caloric intake and/or decreased basal metabolic rate, but not decreased activity
- Group 3 – Olanzapine 10 mg/day mandated to maintain the fixed, standardized diet, but no mandated activity level – a group that might gain weight and fat due to decreased activity and/or decreased basal metabolic rate
- Group 4 – Olanzapine 10 mg/day, mandated to maintain the fixed, standardized diet, mandated to continue daily activity at the same level as before beginning treatment, and mandated to increase activity level if any slight increase in weight or fat began to occur (presumably due to decreased metabolic rate as both caloric intake and activity continued at baseline levels) – a group that would be expected not to gain weight or fat

- Group 5 – Placebo mandated to maintain the fixed, standardized diet, mandated to continue daily activity at the same level as before beginning treatment – a placebo-treated group that would not gain weight or fat
- Group 6 – Placebo with no restriction on additional food consumption over the fixed, standardized diet and no mandated activity level – a placebo-treated group that might gain weight and fat due to being in a restricted environment with decreased activity and/or increased caloric restriction
- Group 7 – Placebo placed on a high-fat diet and no mandated activity level – a placebo-treated group expected to gain weight and fat comparable to the group 1 olanzapine subjects

Group 4 is intended to be a group that would experience a decrease in basal metabolic rate if olanzapine causes such a decrease, but increased activity would adjust total daily caloric expenditure to the baseline level. Some possibility might exist for a change in basal metabolic rate to result in a change in glucose homeostasis even if compensatory, mandated changes in activity and/or diet prevented changes in adiposity in an olanzapine-treated group. Changes in metabolism would likely need to be considered in the statistical model used to analyze the study data.

The study results would be analyzed with a two-step sequential analysis, an analysis for difference, and then, if difference not found, an analysis for non-inferiority.

Important inclusion criteria:

- Age: 20-40
- BMI: 20-25 – only necessary with the fat criteria below to limit muscle mass
- Total body fat and visceral fat between 35^{th}- and 65^{th}-percentiles for age and sex-adjusted norms
- No current or history of a psychiatric disorder
- No current medical illness
- Taking no medications on a chronic basis
- No family history of any form of diabetes mellitus

If possible, to achieve, expert consensus would need to be acquired regarding the optimal conduct of both types of clamp studies, the parameters to be measured, and how to compute those parameters requiring computation. The results of the study would lack good credibility without such consensus.

In the hyperglycemic clamp study, primary considerations would include:

- The number of steps
- The glycemic targets at each step
- What to consider the time of greatest interest for the most meaningful insulin responses
- The time interval at each step during which the system would be considered at steady-state
- How to compute DI – the parameter assessing the adequacy of insulin response adjusted for any change in insulin sensitivity (requires ISIw from the hyperinsulinemic-euglycemic clamp study)
- As this study type is not the gold standard for assessing insulin sensitivity, this study would not be used for this purpose.

In the hyperinsulinemic-euglycemic (or isoglycemic) clamp study, primary considerations would include:

- The radio-tracer labeled glucose tracer to use and when to begin infusion before beginning the clamps
- Euglycemic or isoglycemic
- The number of steps
- The insulin infusion rate at each step
- The time interval at each step during which the system would be considered at steady-state
- How to compute whole-body, hepatic, and peripheral insulin sensitivity
    - The radio-tracer or multiple tracers to use

An additional primary consideration for both types of clamp studies would be any additional metabolic analytes to collect.

Besides the clamp studies, total fat, subcutaneous fat, and visceral fat would be measured at least twice weekly and more frequently during the first one to two weeks of the study's treatment phase to adjust activity, if necessary, in subjects assigned to group 4. A DEXA scan results in radiation exposure slightly less than two days of exposure to natural background radiation. While this is minimal additional radiation exposure annually, this exposure would likely be considered unsafe and unethical for the total fat assessment frequency suggested above. Frequent assessment for the activity level adjustment in group 4 would require MRI assessments of abdominal fat (subcutaneous and visceral) deposits. Small weight changes would offer the most convenient and perhaps most sensitive indicator of this adjustment need. It might be necessary to limit DEXA scans to pre-treatment and post-treatment only; one additional scan at the beginning of the stabilization period would be necessary to exclude subjects based on a proposed inclusion/exclusion criterion based on total body fat.

Diet and/or activity might need to be adjusted for individual subjects to maintain the group statuses described above for individual subjects assigned to those groups. Weight, activity, and resting metabolic rate would be measured daily. The best methods of measuring these fat deposits would require a consensus of experts. While DEXA is probably adequate for total fat by region, CTs and MRIs choice as optimal for quantitating subcutaneous and visceral fat might require discussion. However, given radiation exposure concerns with human subjects, MRI would likely have to be used.

Both clamp studies would be performed at baseline (before treatment), early in treatment, such as when CNS exposure would be expected to reach or be close to steady-state (because of the research suggesting the acute onset of effects), and at the end of four weeks of treatment (endpoint).

Changes in ISIh, ISIp, and ISIw, along with the adequacy of insulin response adjusted for any change in ISIw, would be the primary dependent variables of interest. The primary comparative groups of interest would be: 1) group #3 vs. group #4 (no weight/fat gain in either group); 2) group #1 vs. group #6 (weight/fat gain in both groups). Comparing groups #1 vs. #3 adds information regarding the influence of weight gain vs. no weight gain with olanzapine. A comparison across groups #1, #2, and #3 would increase the knowledge regarding relative contributions of increased food consumption and decreased activity to weight/fat gain (another group treated with olanzapine and food-restricted but not forced to maintain activity would

further assist in this objective). A comparison of groups #4 and #5 serves to assess whether restriction to a closed unit, by itself, facilitates weight gain.

Expert statistical consultation would be required to plan sequential analyses with the suggested multiple comparisons of groups. Maximal differences that would allow declaration of non-inferiority would need to be established *a priori* if olanzapine did not differ from placebo in one or more of the appropriate group comparisons for one or more of the dependent variables of primary or secondary interest. It would be understood that if these differences could not be established with reasonable clinical certainty, then failure to find differences indicating an adverse effect associated with olanzapine could simply not be interpreted. Any strong trend toward statistical significance for an adverse effect with the drug would require reconsideration of the sample size and potentially the need for an additional study.

The study outlined above would be costly. The equipment necessary to perform the clamp studies in humans exists in a limited number of research facilities, and the study would require identical equipment in all participating laboratories. The equipment is costly, requires frequent maintenance, and requires frequent calibration. MRIs are costly. One-hundred-forty to 210 subjects is a large number of subjects for such a study. It would likely be an impossible study to conduct from a practical perspective. Fewer subjects (15 per treatment group) and shorter stabilization and treatment periods (two weeks for each) might be sufficient. Finding subjects willing to consent to be forced with assistance to maintain certain levels of activity daily and eat less when hungry for up to eight weeks would be difficult. Some subjects who consented would likely withdraw consent and discontinue, creating a need for more subjects.

However, we have had experience with a complex Thorough QT study that required more than 120 subjects who screened positive for being CYP-2D6 poor metabolizers and that involved 24-hour continuous, high-fidelity, 12-lead ECG monitoring. Still, the study suggested above would likely be considerably more expensive and challenging to complete than this Thorough QT study.

Finally, this suggested design has limitations, in addition to the limitation of not addressing the limitation of a direct effect in a small subset of the general population mentioned above. The design would not address the questions as to whether a greater differential risk of dysregulation of glucose homeostasis exists between persons taking olanzapine versus those not taking olanzapine among those with diabetes mellitus or with risk factors for developing diabetes mellitus (potentially including having a severe psychotic disorder) compared to those without diabetes mellitus or risk factors for its development. After addressing the question of risk in the absence of known risk factors for developing the disorder and knowing if any excess risk associated with the drug was through direct or indirect and controllable mediators, differential risk in already at-risk persons could be better addressed.

If there were an excess risk with the drug through direct effects, there would likely be little practical need to assess at-risk persons. If there was no excess risk with the drug or any excess risk was mediated through indirect effects, at-risk persons could be studied while controlling those indirect effects.

7. A Postscript Caveat and Apology

A large quantity of numerical data was transcribed directly into this work and extracted from diverse types of figures and then transcribed by the single author (Beasley) of this response.

These transcriptions went through multiple checks but not by an independent reviewer, except for the data in the Ader, Kim, Catalano, et al. (2005) manuscript. It is a virtual certainty that some typographical errors exist in the transcribed data. However, conceptual summaries are faithful to the data. Finally, the tabular summary of findings from the 17 manuscripts that allows easy comparison in Table 17, Section 4, is consistent with the data.

References:

Ader M, Kim SP, Catalano KJ, Ionut V, Hucking K, Richey JM, Kbir M, Bergman RN. Metabolic dysregulation with atypical antipsychotics occurs in the absence of underlying disease. *Diabetes* 2005; 54:862-871.

Albaugh VL, Judson JG, She P, Lang CH, Maresca KP, Joyal JL, Lynch CJ. Olanzapine promotes fat accumulation in male rats by decreasing physical activity, repartitioning energy and increasing adipose tissue lipogenesis while impairing lipolysis. *Mol Psychiatr* 2015; 16:569-581.

American Diabetes Association, American Psychiatric Association, American Association of Clinical Endocrinologists, North American Association for the Study of Obesity. Consensus development conference on antipsychotic drugs and obesity and diabetes. *Diabetes Care* 2004; 27:596–601.

Beasley C, Tollefson G, Tran P, Satterlee W, Sanger T, Hamilton S. Olanzapine versus placebo and haloperidol: acute phase results of the North American double-blind olanzapine trial. *Neuropsychopharmacol* 1996a; 14:111-1123.

Beasley C, Sanger T, Satterlee W, Tollefson G, Tran P, Hamilton S. Olanzapine versus placebo: results of a double-blind, fixed-dose olanzapine trial. *Psychopharmacology* 1996b; 124:159-167.

Beasley C, Hamilton S, Crawford A, Dellva M, Tollefson G, Tran P, Blin O, Beuzen J. Olanzapine versus haloperidol: acute phase results of the international double-blind olanzapine trial. *Euro Neuropsychopharmacol* 1997; 7:125-137.

Beasley C, Berg P, Dananberg J, Kwong K, Taylor C, Breier A. Incidence and rate of treatment-emergent potential impaired glucose tolerance and potential diabetes with olanzapine compared to other antipsychotic agents and placebo. Annual Meeting of the American College of Neuropsychopharmacology. San Juan, PR, December 11, 2000

Beasley C, Sowell M, Cavazzoni P, Breier A, Steinberg H, Dananberg J
Assessment of insulin secretory responses using the hyperglycemic clamp in normal subjects treated with olanzapine, risperidone, or placebo**.** Annual Meeting of the American College of Neuropsychopharmacology. Kona, HI, December 10, 2001.

Beasley C, Sowell M, Henry R, Carlson C, Mukhopadhyay N, Dananberg J, Cavazzoni P, Breier A. Prospective evaluation of insulin sensitivity by the hyperinsulinemic, euglycemic clamp in healthy volunteers treated with olanzapine, risperidone, or placebo. Annual Meeting of the American College of Neuropsychopharmacology. San Juan, PR, December 11, 2002.

Bergman RN, Finegood DT, Ader M. Assessment of insulin sensitivity *in vivo. Endocr Rev* 1985; 6:45-86.

Boyda HN, Procyshyn RM, Pang CCY, Hawkes E, Wong D, Jin CH, Honer WG, Barr AM. Metabolic side-effects of the novel second-generation antipsychotic drugs asenapine and iloperidone: a comparison with olanzapine. *PLOS One* 2013; 8:e53459. doi: 10.1371/journal.pone.0053459.

Finegood DT, Bergman RN, Vranic M. Estimation of endogenous glucose production during hyperinsulinemic-euglycemic glucose clamps. *Diabetes* 1987; 36:914-924.

Girault EM, Alkemade A, Foppen E, Ackermans MT, Fliers E, Kalsbeek A. Acute peripheral but not central administration of olanzapine induces hyperglycemia associated with hepatic and extra-hepatic insulin resistance. *PLOS One* 2012; 7:e43244. doi: 10.1371/journal.pone.0043244.

Hahn MK, Chintoh A, Remington G, Teo C, Mann S, Arenovich T, Fletcher P, Lam L, Nobrega J, Guenette M, Chon T, Giacca A. Effects of intracerebroventricular (ICV) olanzapine on insulin sensitivity and secretion in vivo: an animal model. *Eur Neuropsychopharmacol* 2014; 24:448-458.

Hardy TA, Meyers AL, Yu J, Shankar SS, Steinberg HO, Porksen NK. Acute insulin response and β-cell compensation in normal subjects treated with olanzapine or risperidone for 2 weeks. *Diabetes Care* 2007; 30:157-158.

Hardy TA, Henry RR, Forrester TD, Kryzhanovskaya LA, Campbell GM, Marks DM, Mudaliar S. Impact of olanzapine or risperidone treatment on insulin sensitivity in schizophrenia or schizoaffective disorder. *Diabetes Obes Metab* 2011; 13:726-735.

Houseknecht KL, Robertson AS, Zavadoski W, Gibbs EM, Johnson DE, Rollema H. Acute effects of atypical antipsychotics on whole-body insulin resistance in rats: implications for adverse metabolic effects. *Neuropsychopharmacol* 2007; 32:289-297.

Kopf D, Gilles M, Paslakis G, Medlin F, Lederbogen F, Lehnert H, Deuschle M. Insulin secretion and sensitivity after single-dose amisulpride, olanzapine or placebo in young male subjects: cross-over glucose clamp studies. *Pharmacopsychiatry* 2012; 45:223-228.

Kowalchuk C, Teo C, Wilson V, Chintoh A, Lam L, Agrwal SM, Giacca A, Remington GJ, Hahn MK. In male rats, the ability of central insulin to suppress glucose production is impaired by olanzapine, whereas glucose uptake is left intact. *J Psychiatr Neurosci* 2017; 42:424-431.

Krentz AJ, Heinemann L, Hompesch M. Methods for quantifying insulin sensitivity and determining insulin time-action profiles. Chapter 1 In: Krentz AJ, et al., editors. *Translational Research Methods for Diabetes, Obesity and Cardiometabolic Drug Development: A Focus on Early Phase Clinical Studies*. 2015a; London: Springer-Verlag.

Krentz AJ, Heinemann L, Hompesch M. Assessment of β-cell function. Chapter 2 In: Krentz AJ, et al., editors. *Translational Research Methods for Diabetes, Obesity and Cardiometabolic Drug Development: A Focus on Early Phase Clinical Studies*. 2015b; London: Springer-Verlag.

Kumar R, Nandhini LP, Kamalanathan S, Sahoo J, Vivekanadan M. Evidence for current diagnostic criteria for diabetes mellitus. *World J Diabetes* 2016; 7:396-405.

Lipkovich I, Jacobson JG, Caldwell C, Hoffman VP, Kryzhanovskaya L, Beasley CM. Early predictors of weight gain risk during treatment with olanzapine: analysis of pooled data from 58 clinical trials. *Psychopharmacol Bull* 2009; 42:23-39.

Malik M and Camm AJ. Evaluation of drug-induced QT interval prolongation. *Drug Safety* 2001; 24:323-351.

Martins PJF, Hass M, Obici S. Central nervous system delivery of the antipsychotic olanzapine induces hepatic insulin resistance. *Diabetes* 2010; 59: 2418-2425.

Molina JM, Baron AD, Edelman SV, Brechtel G, Wallace P, Olefsky JM. Use of variable tracer infusion method to determine glucose turnover in humans. *Am J Physiol–Endoc M* 1990; 258:E16-E23.

Muniyappa R, Lee S, Chen H, Quon MJ. Current approaches for assessing insulin sensitivity and resistance in vivo: advantages, limitations, and appropriate usage. *Am J Physiol–Endoc M* 2008; 294:E15-E26.

Park S, Hong SM, Ahn IL, Kim DS, Kim SH. Estrogen replacement reverses olanzapine-induced weight gain and hepatic insulin resistance in ovariectomized diabetic rats. *Neuropsychobiology* 2010; 61:148-161.

Remington GJ, Teo C, Wilson V, Chintoh A, Guenette M, Ahsan Z, Giacca A, Hahn MK. Metformin attenuates olanzapine-induced hepatic, but not peripheral insulin resistance. *J Endocrinol* 2015; 227:71-81.

Sowell M, Mukhopadhyay N, Cavazzoni P, Shankar S, Steinberg H, Breier, Breier A, Beasley C, Dananberg J. Hyperglycemic clamp assessment of insulin secretory responses in normal subjects treated with olanzapine, risperidone, or placebo. *J Clin Endocrinol Metab* 2002; 87:2918-2923.

Sowell M, Mukhopadhyay N, Cavazzoni P, Carlson C, Mudaliar S, Chinnapongse S, Ray A Davis T, Breier A, Henry RR, Dananberg J. Evaluation of insulin sensitivity in healthy volunteers treated with olanzapine, risperidone, or placebo: a prospective, randomized study using the two-step hyperinsulinemic, euglycemic clamp. *J Clin Endocrinol Metab* 2003; 88:5875-5880.

Teff KL, Rickels MR, Grudzia J, Fuller C, Nguyen H-L, Rickels K. Antipsychotic-induced insulin resistance and postprandial hormonal dysregulation independent of weight gain or psychiatric disease. *Diabetes* 2013;62:3232-3240.

Tollefson G, Beasley C, Tran P, Street J, Kruger J, Tamura R, Graffeo K, Thieme M. Olanzapine versus haloperidol in the treatment of schizophrenia and schizoaffective and schizophreniform disorders: results of an international collaborative trial. *Am J Psychiatry* 1997; 154:457-465.

Vidarsdottir S, de Leeuw van Weenen JE, Frolich M, Roelfsema F, Romijn JA, Pijl H. Effects of olanzapine and haloperidol on the metabolic status of healthy men. *J Clin Endocrinol Metab* 2010; 95:118-125.

Wu C, Yuen J, Boyda HN, Procyshyn RM, Wang CK, Asiri YI, Pang CCY, Honer WG, Barr AM. An evaluation of the effects of the novel antipsychotic drug lurasidone on glucose tolerance and insulin resistance: a comparison with olanzapine. *PLOS One*. 2014; 9:e107116. doi: 10.1371/journal.pone.0107116.

Zyprexa® US Prescribing Information. Revised 2010.

July 4, 2019

c. Barry Blackwell's Comment on Beasley's Response

Background

When Tom Ban, at the suggestion of Ned Shorter, asked me to reply to the recent substantial INHN posting of Beasley and Tamura's statistical methodology for tackling the difficulty of identifying adverse drug reactions, initially in general but at Shorter's suggestion focusing on Olanzapine, (marketed by Eli Lilly as Zyprexa for schizophrenia), I entertained serious reservations.

It is 10 years since I retired, for the third time, and although I am a trained psychopharmacologist I am neither a statistician nor an endocrinologist although I held academic professorships in Psychiatry, Pharmacology and Medicine.

But I decided to accept in the context of perspectivism**;** OED: "The philosophical theory that knowledge of a subject is inevitably partial and limited by the individual perspective from which it is viewed."

This perspective includes an historical point of view derived from more than 10 years working with Tom Ban, first on the *Oral History of Neuropsychopharmacology: The First Fifty Years* (2011), and since 2013 as editor of Biographies and Controversies on the INHN website.

In that context I posted a critical essay on "Corporate Corruption in the Pharmaceutical Industry" in which Charles Beasley spent a distinguished career as the lead biostatistician for Eli Lilly. Charles was working on Olanzapine and its side effects until 2002 when higher authority ordered him to cease. His recent postings on INHN are the product of his work post retirement.

The Historical Context

Beasley's work on INHN includes 10 separate postings between November 29, 2018, and April 4, 2019 (Pages 1-31 of the Ebook), summarizing the approach he developed at Eli Lilly intended to identify adverse drug reactions. On April 25, 2019, Edward (Ned) Shorter made a brief comment congratulating the authors on their "commitment to the high ground of science" but challenging them to provide a specific demonstration of the method with regard to Olanzapine and the risk of diabetes. Charles Beasley responded with a 56-page posting (pages 4-60 of the Ebook) which is the topic of my comments.

Significant Historical Findings

For retired folks who no longer keep up with academic sources of wisdom Google is an appealing contemporary substitute. Type in Olanzapine and 3,000,000 postings are recorded. Limit the search to Olanzapine and Diabetes and the number falls to just under a million, 986,000. Looking for a cogent publication the most relevant and frequently cited (516 times) was a study published in the British Medical Journal titled "Assessment of the independent effect of Olanzapine and Risperidone on the risk of Diabetes among patients with schizophrenia; population based nested case control study" **(**Koro, Fedder, L'Italien et al., 2002).

This study was derived from the UK General Practice Research Data Base. It involved 19,637 patients diagnosed and treated for schizophrenia, compared to 2,696 controls. The conclusions were, "Olanzapine is associated with a clinically important increased risk of diabetes. Patients taking Olanzapine had a significantly higher risk of developing diabetes than both non-users of

antipsychotics and Risperdal (risperidone). Taking Risperdal had a non-significant risk compared to both non users and those taking conventional antipsychotics."

There is little doubt this was a seminal study in a refereed international medical journal that had an impact on regulatory and labelling practices. On October 17, 2003, the FDA issued a warning letter to manufacturers about the risk of diabetes followed the next year by practice guidelines from the American Diabetes Association and the American Psychiatric Association. These required Weight and BMI measures at every visit and a fasting blood sugar and blood lipid level at week 12 and annually. In March 2004 Eli Lilly added a warning statement to the labelling of Olanzapine describing the increased risk of hyperglycemia and diabetes.

The impact on research and regulatory procedures.

The timing of the BMJ study in 2002 coincides with top management at Eli Lilly requiring Charles Beasley to cease work on Olanzapine although that study is not cited in the 34 references which include 22 published during or after 2002. I speculate that the timing of that publication served to render moot any further need for costly in-house statistical analysis. In Section 5 of the statistical approach Charles outlines "A hypothetical study to resolve the uncertainties." But he concedes that search for co-morbid variables influencing risk of diabetes pale into insignificance now that the severity and prevalence of the disorder itself is known. He speculates that a manufacturer would be unwilling to devote profits to such a large and expensive enterprise.

Working for Industry; a personal perspective

In 1968 at the age of 34 after completing my psychiatric training at the Maudsley Hospital in London I migrated to America (still the land of opportunity) to accept the position as Director of Psychotropic Drug Development at the Merrell pharmaceutical company in Cincinnati. They had recently emerged from the fiasco of marketing thalidomide to pregnant women as a safe hypnotic. This triggered Congress to enact legislation to empower the FDA to develop regulatory procedures for the development of new drugs. Merrell, like other pharmaceutical companies hope to benefit from this new lucrative field of research.

Early on in the evolution of modern psychopharmacology I soon realized I was better suited to research than marketing so, after two years, I returned to academic life as a Professor of Psychiatry and Pharmacology at the University of Cincinnati.

I believe that Charles Beasley was a person of similar temperament and capability who in his mid to late career became snared in the 1980's industry transition from credible research on innovative compounds to the ingenious marketing of me-too compounds coupled with a steady erosion of ethical standards in top management focused on profit who established tight controls over talented scientists like Charles Beasley. I believe that the statistical methodology he describes had the capacity to fulfill his personal and ethical goals as well as his desire to see it posted on INHN post retirement which does credit to his distinguished career in industry.

References:

Beasley CM Jr, Tamura R. 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) – Full text. inhn.org.ebooks. November 21, 2019.

Koro CE, Fedder DO, L'Italien GJ, Weiss SS, Magder LS, Kreyenbuhl J, Revicki DA, Buchanan RW. Assessment of independent effect of olanzapine and risperidone on risk of diabetes among

patients with schizophrenia: population based nested case-control study. *BMJ* 2002; 325:243-247.

November 28, 2019

d. Charles Beasley's Response to Blackwell

First, I thank Barry Blackwell for accepting Ned Shorter and Tom Ban's suggestion to comment on my response to Edward (Ned) Shorter's comment. Tom and INHN have received fewer than expected comments on the e-book by Beasley and Tamura (Beasley and Tamura, 2019). Writing a comment such as Barry's requires significant effort, and, again, I want to thank Barry for expending the effort.

It might be helpful to those who read this response to review the evolution of interchanges that led to what I am writing here. Barry wrote his Corporate Corruption in the Psychopharmaceutical Industry (Blackwell, 2016) and a subsequent revised version (Blackwell, 2017). Tom Ban suggested I write a comment regarding Barry's work, which I did (Beasley, 2017). Barry and I continue to exchange comments and responses regarding his original posting concerning corporate corruption.

As part of this progressive interchange with Barry (Beasley, 2018), I included the sample sizes resulting from several sample size calculations based on the sample sizes required for a study with 80% power. The hypothetical study would definitively address the question as to whether an AE that is observed rarely to very infrequently is an ADR or a coincidental background AE not caused by the drug being studied[17]. These sample size calculations considered two situations, first, where the background incidence of the AE is virtually 0, and second, where the background incidence of an AE is relatively high compared to the AE incidence when it is an ADR.

Having written this short piece regarding sample sizes, after a discussion with Tom Ban, I decided to write a longer piece for publication as a book (with my statistical colleague, Roy Tamura). This longer work was intended to explain the limitations on definitive assessment of whether any given AE is or is not (two vastly different questions that cannot be appropriately addressed by the same statistical methods) an ADR. This work appeared as a series of nine postings (an outline plus eight separate Sections – all nine now chapters) on the INHN website

[17] This sentence as I wrote it is technically incorrect. The hypothetical study would be designed to "prove" that the AE is an ADR. The null hypothesis that would be tested with an inferential statistical test would be that the AE was not an ADR. If the inferential test rejected the null hypothesis, the AE would be considered an ADR. The study sample sizes for test drug and comparator were sufficient to give the study 80% power. 80% power is generally considered adequate power for a definitive test. However, a Type II error is a possibility, and the study could falsely fail to reject the null hypothesis. In this case, the study would falsely fail to "prove" that the actual ADR is an ADR. The sentence might be considered incorrect from a second perspective. I said that the study could "prove" that the AE was, or was not an ADR. Proving that the AE is not an ADR would require a non-inferiority inferential test with a comparison to placebo as was discussed in Chapter 6. The hypothetical study would require a statistical analysis plan that allowed testing for both an effect and a lack of effect. The sample sizes that were included in our first INHN posting discussing large and impractical sample sizes were not calculated for sequential testing for a difference from and then non-inferiority to placebo. Our sample sizes could be inadequate for robust tests of an AE being an ADR and not being an ADR.

with a collated version posted on November 21, 2019 (Beasley and Tamura, 2019). A postscript was posted (Beasley, 2019b), now Chapter 10 of the book. For me, there were two stimuli for the writing of this more extended work. First, for full transparent disclosure, I wanted to make it clear that the sample sizes based on 80% power in the posting 2018 (Beasley, 2018) responding to Barry were larger than would be necessary to achieve a statistically significant finding ('proving' a given AE to be an ADR) under certain circumstances. If the hypothetical study designer estimated correctly or underestimated the incidence of the AE in the experimental drug group and did not underestimate (*sic*, "overestimate" incorrectly substituted for "underestimate" in the original posting) this incidence in the control group, smaller sample sizes would result in a significant finding. With perfect estimates of the observed incidences, sample sizes that result in a study of only ~50-51% power lead to significant findings based on the conventional definition of significance ($p \leq 0.05$).

The second stimulus was that the matters discussed had been my interests during most of my career at Eli Lilly. Writing the series of postings allowed me to formalize my thinking and share my thoughts on concepts I believe to be essential for medical professionals and other relevant parties, i.e., legislators, plaintiffs' and defense attorneys involved in product liability litigation, and the public in general who receive a prescription medication.

The series of postings has generated several comments to which I have written responses, with some posted and some not yet posted. Ned offered a comment (Shorter, 2019), wondering if I might comment on Lilly's work to investigate and understand the relationship between olanzapine and diabetes mellitus. The exact contents of Ned's comment are important, and I quote them in entirety below:

"Readers of this website will look forward with special interest to the comments of Charles Beasley, in particular on the issue of side effects and their measurement, given that in his long tenure at Eli Lilly he often confronted these issues on an almost daily basis. In the late 1990s there was an intense in-house discussion about possible hyperglycaemia, weight gain and diabetes associated with olanzapine and much of this correspondence has, in connection with discovery in litigation, now become part of the public record. In these exchanges, Alan Breier and Dr. Beasley come across very much as the in-house investigators committed to the high road of science and one hopes that in the coming instalments (sic) of this thread, Dr. Beasley might illustrate his points with references to some of this material."

Ned was partially correct regarding internal Lilly correspondence and other documents relevant to olanzapine and glycemic dysregulation investigation. Documents obtained as part of discovery by the plaintiffs' attorneys had been posted on the web. However, the postings were illegal, and the documents were successfully removed (at least from websites on the conventional web accessible from US IP addresses and from the conventional web from IP addresses appearing to be in Eastern Europe). Ned might have preferred my response to his comment to discuss the material in the posted documents, but these were not available to me.

As our series of postings (Beasley and Tamura, 2019), now chapters, addressed what can be considered high-quality 'proof' that an observed AE is or is not an ADR, Ned's comment allowed me to briefly review the sequential analyses that had been performed with laboratory analyte data relevant to glycemic control. These were data collected in olanzapine Phase 3-4 clinical trials. Our response to Ned then turned to a detailed discussion of both hyperglycemic glucose clamp studies (evaluating pancreatic insulin production) and hyperinsulinemic-

euglycemic clamp studies (evaluating the effectiveness of insulin in disposing of glucose [causing glucose to be taken up by the body's cells]). For various reasons, glucose values observed in patients with schizophrenia during mainly outpatient clinical trials (with few patients treated for extended periods) are highly variable. My discussion of these data in my response to Ned (Beasley, 2019a) hopefully illustrated this point clearly. I reviewed placebo-controlled clamp studies (human and animal) in great detail because as I read the literature, the combination of both clamp studies in a reasonable sample size of human subjects stands the greatest potential likelihood of addressing the questions as to whether diabetes is likely to be an ADR associated with a drug.[18]

There are two confounders with this pair of studies that must be addressed. First, experts in diabetes and these studies must reach a consensus regarding study methods and analyses that address pancreatic function and insulin action efficiency for muscle, liver, and the entire body. My view from the studies reviewed is that such a consensus does not yet exist. Second, if the drug of interest is associated with weight gain, then methods that would likely be expensive and perhaps difficult to implement would be necessary to isolate any direct diabetogenic effect from an indirect effect due to weight gain. I am not convinced that statistical analysis methods that would adjust for weight gain would be sufficient to definitively separate a direct effect from a secondary effect in this domain.

Besides analyses of the laboratory data from well-controlled clinical trials and the conduct of three clamp studies, Lilly undertook other lines of research to address this important clinical question. I did not review these in response to Ned (Beasley, 2019a) as I do not consider them to be as definitive as analyses of clinical trial data and clamp studies in assessing diabetes as an ADR. However, as Barry, in his comment to Ned, mentions, an epidemiological study conducted using an extensive database found an excess risk of diabetes with olanzapine. I briefly describe the results of this study cited by Barry and two additional studies conducted during the same period using that same database, one study conducted by Lilly.

Two matters in Barry's comment require clarification. The first of these matters is my academic training and functional responsibilities while an employee of Lilly. The second of these matters are the circumstances surrounding my transitioning from working primarily with olanzapine to other primary responsibilities within Lilly in 2001 (not 2002).

In his comment on Ned's comment (Blackwell, 2019), Barry states: "In that context I posted a critical essay on 'Corporate Corruption in the Pharmaceutical Industry' in which Charles Beasley spent a distinguished career as the lead biostatistician for Eli Lilly. Charles was working on Olanzapine and its side effects until 2002 when higher authority ordered him to cease."

My work within Lilly was not as a biostatistician but as a Research Physician (my medical specialty training was as a psychiatrist, and I joined Lilly directly from completing my residency in 1987). I did have extensive computer science training and worked in artificial intelligence research and the development of a database and some analyses methods for an evoked potential

[18] In the original posting of this response to Blackwell, I should have qualified the reference to using the two clamp studies to determine if diabetes is an ADR for a drug to determining if there was a direct diabetogenic effect through either impairment of insulin sensitivity or impairment of pancreatic β-cell insulin production and release (as well as other physiological processes known to facilitate response to an appropriate disposal of glucose loads as well as mitigate endogenous production.

laboratory (two different work experiences) before beginning medical school in 1979. While I had some statistics and research design training and have maintained a keen interest in the interface among statistical methods, research design, and data analyses, I should not be considered a statistician. I performed all the initial programming and sample size computations for our posting 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) (Beasley and Tamura, 2019). My statistical colleague Roy Tamura (biostatistics faculty member at a US university medical center [University of South Florida]) reviewed and checked my work and suggested essential revisions for the final posting.

Barry is quite correct that I transitioned from a role that reported directly to the President of the Neurosciences Business Unit that involved, for the most part, the in-depth review of the safety of olanzapine to an alternative role. That alternative role was Medical Director for the Tadalafil (Cialis™) Product Development Team. The change in primary responsibilities occurred in July 2001. Barry's comment implies that Lilly's senior management were dissatisfied with my work on olanzapine and perhaps my suggestions for studies, analyses, and information dissemination. I cannot affirm that Barry's implication is incorrect.

I address the reasons that I understand lead to the request that I make this transition. However, I did not ask senior management if, in addition to the reasons I understood for this request, other reasons were at play as well. If Barry is correct in his implication, senior management would have been unlikely to be honest with me if asked about other reasons. I was asked to make this transition shortly after tadalafil had been submitted for review for marketing approval to the FDA and the European Union (EMEA) review body. The Medical Director for that team had abruptly resigned from Lilly shortly after the submissions, which was a critical juncture in the development of a new drug. Also important, while Lilly was developing tadalafil, it was owned by another company at that time, Icos. At that point in its development, the loss of a Medical Director would not enhance the working relationship between partners in such a joint venture. I had extensive and recent successful experience in working with FDA, EMEA, and the Japanese regulatory authority on several complex matters and, at the risk of being immodest, agreed that I was the best physician within Lilly at the time to step into this role. Although the tadalafil role was intensive, I was consulted on several matters related to olanzapine through December 2002. I transitioned back to work in the Neuroscience Business Unit as a special consultant in January 2003 with work completed and tadalafil positioned for US and European approval, the expressed goals for my movement to the Tadalafil Team.

I now turn back to Lilly and other researchers' work to assess the relationship between olanzapine and glycemic dysregulation. In response to Ned's comment (Beasley, 2019a), I explained my rationale for the narrow focus on Lilly clinical trials data (illustrate many of the problems with such data when addressing an AE that is relatively infrequent to rare, has a high background incidence, and delayed onset), and glucose clamp studies (if conducted properly, probably the best way of addressing the question of a drug impact on glycemic dysregulation if it is an acute, direct effect that is not idiosyncratic). Lilly conducted other studies as well, including two epidemiological studies for which I advocated. One of these was US-based (Buse, Cavazzoni, Hornbuckle, et al., 2003), and the other was UK-based (Carlson, Hornbuckle, DeLisle, et al., 2006). The latter used the same database, the GPRD database, used by the study cited by Barry (Koro, Fedder, L'Italien, et al., 2002). These two Lilly epidemiological studies using large databases were planned to be performed before I transitioned to my work with

tadalafil in 2001, and I cannot address the timing of their publication as I was not involved with these publications as an author.

The Buse, Cavazzoni, Hornbuckle, et al. study (2003) found a greater hazard ratio (hazard ratio from a Cox proportion regression model adjusting for age, gender, and duration of exposure - relative to persons not treated with antipsychotics) for new diabetes diagnoses to be associated with: 1) all conventional antipsychotics combined; 2) two conventional antipsychotics that were analyzed separately (haloperidol, thioridazine); 3) all atypical antipsychotics combined; and 4) all atypical antipsychotics analyzed separately (clozapine, risperidone, olanzapine, quetiapine). The diagnosis of new cases with the four atypical agents was compared to the diagnoses of new cases with haloperidol. Significantly higher hazard ratios were found only with clozapine and risperidone. The hazard ratio for olanzapine was numerically higher (1.09). The hazard ratio for quetiapine was significantly lower (0.67).

The Carlson, Hornbuckle, DeLisle, et al. study (2006), using methods comparable to those in the Buse, Cavazzoni, Hornbuckle, et al. study (2003) but adjusting for obesity and not length of exposure, found greater hazard ratios with combined conventional antipsychotics and, separately, combined atypical antipsychotics, for thioridazine alone, both combined olanzapine, and risperidone and olanzapine alone (the only two atypical agents analyzed separately), but not fluopenthixol (*sic* - flupenthixol), trifluoperazine, chlorpromazine or haloperidol analyzed separately.

There are notable differences in both the findings and methods of the Carlson, Hornbuckle, DeLisle, et al. study (2006) and the Koro, Fedder, L'Italien, et al. study (2003) cited by Barry, although both were conducted using the GPRD database. The Carlson, Hornbuckle, DeLisle, et al. (2006) study used all patients treated with antipsychotics without a preexisting diagnosis of diabetes and all patients not treated with antipsychotics without a preexisting diagnosis of diabetes as controls. 59,089 patients treated with conventional antipsychotics and 9,059 patients treated with atypical antipsychotics (5,213 with risperidone and 2,374 with olanzapine) were included in the Carlson, Hornbuckle, DeLisle, et al. (2006) analyses. The Koro, Fedder, L'Italien, et al. (2003) study limited the investigational cohort to patients diagnosed with schizophrenia. Only 19,637 patients were treated with any antipsychotic, 1,683 risperidone-treated patients, and 970 olanzapine-treated patients were included in the Koro, Fedder, L'Italien, et al. (2003) study. The study was of a case-control design with six controls matched to each case. Controls were matched to cases by sex, age, length of follow-up, and date of being eligible as a control.

The model adjusted odds ratio was significantly greater for the use of conventional antipsychotics and olanzapine but not for the use of risperidone compared to no antipsychotic treatment. The model adjusted odds ratio was significantly greater for olanzapine but not for risperidone than for conventional antipsychotics.

While a case-control design restricted to patients with a diagnosis of schizophrenia could be expected to reduce the potential for unknown differences between groups being compared biasing the outcome of the analysis compared to the design of the Carlson, Hornbuckle, DeLisle, et al. (2006) study, the Koro, Fedder, L'Italien, et al. (2002) study included smaller comparative sample sizes. Nonetheless, the results of the two studies agreed, except for findings for risperidone. There were only seven new cases of diabetes in the risperidone group in the Koro, Fedder, L'Italien, et al. (2002) study.

Another case-control study (Kornegay, Vasilakis-Scaramozza, and Jick, 2002) was conducted using the GPRD database, and this study was conducted by FDA staff. This study found the adjusted odds ratio for ***current*** (emphasis added) use of conventional antipsychotics and atypical antipsychotics (separately) to be significantly higher than for non-use during the preceding year. The adjusted odds ratio (1.0) for ***recent*** (emphasis added) use of conventional antipsychotics compared to non-use within the preceding year was not significant. The adjusted odds ratio for ***recent*** (emphasis added) use of atypical antipsychotics compared to non-use in the preceding year could not be computed as no subject had this type of exposure as defined by the investigators (use within the seven to twelve months before the index date of diagnosis).

The Kornegay, Vasilakis-Scaramozza, and Jick (2002) study included patients with information recorded in the GPRD database on drug treatment between January 1994 and December 1998 (publication approximately four years after the data cut-off date). The Koro, Fedder, L'Italien, et al. (2002) study included patients diagnosed with schizophrenia and information on drug treatment recorded in the GPRD database between June 1987 and September 2000 (publication approximately three years after data cut-off date). The Carlson, Hornbuckle, DeLisle, et al. (2006) study included patients with information recorded in the GPRD database on drug treatment between January 1, 1994, and December 31, 2001 (publication approximately five years after data cut-off date). Based on the earliest to latest end dates for patient inclusion in these retrospective epidemiological analyses, they were conducted in the order in which they were described in the preceding three sentences. With each subsequent study, more data for atypical antipsychotics would come available. The differences in analysis results demonstrate the impact of increasing data (especially for the atypical antipsychotics) and methodological differences in the analyses. These methodological differences resulted from slightly different questions that the three research groups hoped to answer with their studies.

Not only did other groups publish the results of analyses strongly supporting the hypothesis that olanzapine has an association with a risk of incident cases of diabetes, but Lilly also conducted such research and published the results (Carlson, Hornbuckle, DeLisle, et al., 2006). However, Lilly's work found conventional antipsychotics as a group and atypical antipsychotics as a separate group associated with this risk. FDA analysis (Kornegay, Vasilakis-Scaramozza, and Jick, 2002) resulted in similar findings for both conventional and atypical antipsychotics as separate groups when considering the diagnosis of diabetes while being actively treated with an antipsychotic.

To me, the epidemiological study findings with olanzapine are not surprising. That olanzapine is associated with substantial weight gain has been well understood since the five registration trials for the drug were conducted and documented in the initial product labeling for the drug in the US and other regulatory venues. Any medical professional licensed to treat patients with pharmaceutical products (physicians, advanced nurse practitioners, physician assistants, clinical psychologists in some licensing venues) should have a clear understanding that weight gain, especially weight gain primarily in the form of visceral adipose tissue, is a major risk factor for the development of Type II diabetes mellitus.

What was surprising to me was that the results of the analyses designed by my statistical colleagues and me, executed by these statistical colleagues in the late 1990s and presented in multiple public scientific forums (e.g., Beasley, 2000), failed to demonstrate that glycemic dysregulation and diabetes was an ADR with olanzapine. These analyses demonstrated two important things. The analyses first demonstrated the magnitude of variability or 'noise' in what

are intended to be fasting glucose concentration values obtained from patients with schizophrenia (also mentioned above) and, therefore, the great difficulty in using such data to find evidence of impaired glucose regulation or diabetes in such a patient population. Second, these analyses demonstrated that there was no 'smoking gun' in the clinical trial data. Even with an additional four years of clinical trial data acquired since completing the registration studies, no statistically significant evidence demonstrated an association between olanzapine and glucose dysregulation or diabetes. 'Noisy' data collected from too few patients over a too short a time for cases of an ADR to develop where the time of onset might require months to years from treatment initiation to occurrence of the ADR is unlikely to offer 'proof that a drug is or is not associated with such an ADR. The difficulty with clinical trial data used for assessing such possible ADRs is exacerbated by a high background incidence of the medical event that might be an ADR, as is the case with diabetes. This matter was discussed at length, with examples in Beasley and Tamura (2019).

The interpretation of the results of our late 1990s (Beasley, 2000) studies should be highly limited. These analyses were conducted with the null hypothesis of olanzapine equivalent to placebo and equivalent to haloperidol (haloperidol being important for long-term comparisons), and these null hypotheses could not be rejected. As stated above, this was important because the analyses then demonstrated that there was no 'smoking-gun' within the existing clinical trial data demonstrating that olanzapine was causally related to diabetes. The results of these analyses underscored the need for additional research using alternative methods (e.g., glycemic clamp studies, epidemiological studies in 'big-data' databases). Lilly and others undertook and published such work.

It is important to underscore that while the analyses (Beasley, 2000) failed to 'prove' a diabetic effect, the results of the analyses most assuredly and unequivocally did not 'prove' the absence of a diabetic effect associated with olanzapine. As Paul Leber, Director of the Division of Neuro-Pharmacological Drug Products within the FDA from 1981 through 1999, said, "Absence of evidence is not evidence of absence." Stated alternatively, failure to reject the null hypothesis does not allow acceptance of the null hypothesis as correct. I would consider any use of the results of these analyses to suggest that they demonstrated that olanzapine was not associated with a diabetic effect to be a grossly inappropriate use of those results. If such use resulted from a lack of understanding of the principles of interpreting statistical analyses, such use could be attributed to ignorance. If such use did not result from a lack of understanding of these fundamental principles of interpretation, such use could be considered malignant.

I conclude this response to Barry with a return to the matter of weight gain and the question as to whether the diabetic ADR with olanzapine is a direct effect, an indirect effect mediated through visceral adipose tissue gain, or a combination of both. While others might substantially disagree, I do not believe this question has been adequately addressed (Beasley, 2019a). Furthermore, the question is fundamental in determining which antipsychotics to use with a given patient, assess individual patient risk-benefit, and the decision to switch antipsychotics if necessary. Weight gain is easily followed. For olanzapine, substantial long-term weight gain can be predicted with as few as 2-3 weeks of treatment (Lipkovich, Jacobson, Caldwell, et al., 2009). If this weight gain (likely most if not all adipose tissue) is the primary but indirect etiology of glycemic dysregulation, some ~50% or more of patients who experience substantial weight gain during longer-term olanzapine treatment (not all of whom would develop diabetes) can be easily identified. For these patients, alternative medication can be considered in terms of individual

risk-benefit assessment. In this case, olanzapine is easy to use with respect to this specific risk assessment

However, if there is a potentially abrupt and unpredictable direct diabetogenic effect of olanzapine and other antipsychotics, then the individual risk of glycemic dysregulation is unpredictable. In this case, olanzapine and other such antipsychotics become exceedingly difficult to use with respect to this individual risk assessment. Both robust response and compliance with frequent follow-up in patients judged to be at significant risk of consequent morbidity if glycemic dysregulation developed (i.e., diabetes due to a direct diabetogenic effect) would be required for me to consider olanzapine as the first choice for such an individual patient. Not all patients would be at significant and imminent risk of severe morbidity if new-onset diabetes were not detected very soon after the development of diabetes. However, for those at such risk (e.g., patients with preexisting atherosclerotic coronary artery disease, carotid artery disease, hypertension, dyslipidemia), I would personally view substantial efficacy and capacity on the part of the patient to participate in frequent follow-up examinations as a requirement to consider olanzapine (or any drug with a direct and unpredictable diabetogenic effect) as a first-choice treatment for such a patient.

Again, we thank Barry for his comments.

References:

Beasley C, Berg P, Dananberg J, Kwong K, Taylor C, Breier A. Incidence and rate of treatment-emergent potential impaired glucose tolerance and potential diabetes with olanzapine compared to other antipsychotic agents and placebo. Annual Meeting of the American College of Neuropsychopharmacology. San Juan, PR, December 11, 2000.

Beasley CM Jr. Comment on Barry Blackwell's Corporate corruption in the psychopharmaceutical industry. inhn.org.controvesies. March 23, 2017.

Beasley CM Jr. Response to Barry Blackwell's response to Charles Beasley's comment on Barry Blackwell's Corporate corruption in the psychopharmaceutical industry. inhn.org.controversies. January 12, 2018.

Beasley CM Jr. Response (Olanzapine and diabetes mellitus, evolution of data – illustrating the difficulties in identification of ADRs) to Edward Shorter's comment on Charles Beasley's and Roy Tamura's 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 (Chapter 1). inhn.org.ebooks. July 4, 2019 (2019a).

Beasley CM Jr. 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) – Postscript (Chapter 10). inhn.org.ebooks. October 24, 2019 (2019b).

Beasley CM Jr, Tamura R. 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) – Full text. inhn.org.ebooks. November 21, 2019.

Blackwell B. Corporate corruption in the psychopharmaceutical industry. inhn.org.controversies. September 1, 2016.

Blackwell B. Corporate corruption in the psychopharmaceutical industry (revised). inhn.org.controversies. March 16, 2017.

Blackwell B. Comment on Charles Beasley's response (Olanzapine and diabetes mellitus, evolution of data – illustrating the difficulties in identification of adverse drug reactions) to Edward Shorter's comment on Charles Beasley's and Roy Tamura's 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 (Chapter 1). inhn.org.ebooks. November 28, 2019.

Buse JB, Cavazzoni P, Hornbuckle K, Hitchins D, Breier A, Jovanovic L. A retrospective cohort study of diabetes mellitus and antipsychotic treatment in the United States. *J Clin Epidemiol* 2003; 56:164-170.

Carlson C, Hornbuckle K, DeLisle F, Kryzhanovskaya L, Breier A, Cavazzoni P. Diabetes mellitus and antipsychotic treatment in the United Kingdom. *Eur Neuropsychopharmacol* 2006; 16:366-375.

Kornegay CJ, Vasilakis-Scaramozza C, Jick H. Incident diabetes associated with antipsychotic use in the United Kingdom General Practice Research Database. *J Clin Psychiatry* 2002; 63:758-762.

Koro CE, Fedder DO, L'Italien GJ, Weiss SS, Magder LS, Kreyenbuhl J, Revicki DA, Buchanan RW. Assessment of independent effect of olanzapine and risperidone among patients with schizophrenia: population based nested case-control study. *BMJ* 2002; 325:243-247.

Lipkovich I, Jacobson JG, Caldwell C, Hoffman VP, Kryzhanovskaya L, Beasley CM. Early predictors of weight gain risk during treatment with olanzapine: analysis of pooled data from 58 clinical trials. *Psychopharmacol Bull* 2009; 42:23-39.

Shorter E. Comment on Charles Beasley's and Roy Tamura's 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 (Chapter 1). inhn.org.ebooks. April 25, 2019.

May 14, 2020

2. Edward Shorter's Comment on the Definition of Terms (Chapter 2), Followed by Beasley's Response: The Use of Dechallenge-Rechallenge Methods in the Assessment of a Potential ADR

a. Edward Shorter's Comment

Well, it would be hard to offer critical comments on these introductory thoughts. The difference between adverse effects not related to the drug and adverse drug reactions seems elementary. I would offer only the thought that there are other techniques for establishing the relationship between a drug and putative side effects, such as challenge-dechallenge-rechallenge. This, too, will answer the question without a lot of statistical complexities. We are hung up on RCTs simply because the US Congress appears to have mandated them in the Kefauver-Harris legislation of 1962 (the legislation used the phrase "well controlled" rather than "RCTs"). RCTs are not the word of God.

One further thought: in the world of drug litigation, retrospective data from RCTs must be analyzed rather than prospective data (as in challenge-dechallenge-rechallenge). It would be unfortunate if our discussion of these matters were entirely guided by the exigencies of litigation.

May 9, 2019

b. Charles Beasley's Response

We thank Prof. Shorter for raising the topic of dechallenge-rechallenge approaches to determine which AEs experienced by patients while taking medication are ADRs. The basic approach is one of discontinuing medication after observing an AE (dechallenge) occurrence. If the AE resolves in close temporal association to the discontinuation and clearance of the medication, there is a suggestion that the AE might be an ADR. The patient is then treated again with the medication (rechallenge), and if the AE recurs shortly after restarting the medication, that observation is considered substantial evidence that the AE is an ADR, at least for the specific patient.

Our focus was on randomized clinical trials (RCTs) and the extent to which RCTs can (or cannot) establish with the same robustness of scientific evidence required for a demonstration of efficacy for a medication to gain regulatory approval that an AE is (or is not) an ADR. In the posting of what is now Chapter 9, we briefly suggested that all parties should be aware of the limitations on the robustness of evidence for infrequent and rare AEs regarding whether they are ADRs. We firmly believe that the gold standard for such evidence is the placebo-controlled RCT (or a set of such studies). Acceptance of the uncertainty about whether an infrequent or rare AE is or is not an ADR is a practical necessity if we are to continue developing new medications. We also briefly alluded to the need for work directed at developing alternative methods to RCTs for substantially robust ascertainment of what are (and are not) the ADRs associated with a medication. The goal is to identify all ADRs, even those that are relatively rare, without false identification. Our brief discussion focused on evolving epidemiological methods using large, collaborative databases.

We were remiss in not discussing dechallenge-rechallenge approaches in Chapter 9. Properly conducted, a dechallenge-rechallenge study can provide as strong or stronger (compared to an RCT) scientific/statistical evidence as to whether an AE is or is not an ADR for an individual patient. If the dechallenge-rechallenge study 'proves' that the AE is an ADR for one patient, then the general case (the AE is an ADR) is 'proven', but unique patient characteristics (e.g.,

uncommon genetic susceptibility, treatment with a combination of other specific drugs) might need to be present for the occurrence of the AE as an ADR. However, if the dechallenge-rechallenge study either fails to 'prove' that the AE is an ADR or 'proves' that the AE is not an ADR for the patient, the AE could still be an ADR for other patients.

From our perspective, to be robust, the dechallenge-rechallenge study should be conducted as a blinded 'N-of-1' experiment with inferential statistical analyses (Kravitz and Dunn 2014). There are two limitations on the nature of AEs that can be studied with dechallenge-rechallenge methods. First, the method can be applied to AEs with a relatively rapid progression to maximal severity and, more importantly, resolve virtually completely on discontinuation of the causal medication. When such a trial design is used to study efficacy, statistical methods exist to consider continuous and ordinal assessments of the treated disorder. However, for assessing whether an AE is an ADR, we believe that the dependent (outcome) variable should be binary, the presence of the AE of any severity or its complete absence. If an AE only changes in severity with dechallenge and rechallenge, but it remains present on dechallenge, it would be difficult to interpret such results concerning the test medication as an etiological contributor to the AE. An anaphylactic reaction would be an example of an AE that would be a suitable candidate for assessment through an 'N-of-1' trial. Cases of agranulocytosis and aplastic anemia might be good candidates for study in an 'N-of-1' trial. Once they occur, some ADRs do not resolve, although the medication's contribution to the ADR's pathophysiology would resolve following the medication's discontinuation. An example of such an AE would be myocardial infarction, where the medication accelerated coronary artery atherosclerosis as the pathophysiological process. Although the resolution and stabilization of atherosclerotic plaques can occur, lowering the probability of a subsequent myocardial infarction, these are slow processes. Such a progression and resolution process would not lend itself well to assessment with an 'N-of-1' trial. Some other important AEs (malignancies, diabetes mellitus to some extent) would have the same difficulties in using an 'N-of-1' trial approach, slow development, and no resolution or delayed resolution following medication discontinuation.

The second limitation when using an 'N-of-1' trial design is the ethics of potentially causing an SAE in a patient. Closely related to the abstract matter of such a study's ethics is the practical matter of whether patients would consent to participate in such a study where the AE is clinically serious and might be associated with permanent harm. Those AEs of greatest need for improved assessment methods, infrequent and rare events, tend to be clinically serious events.

We are aware of instances of simple (unblinded, single sequence) dechallenge-rechallenge evaluation of AEs temporally associated with medications on which we worked while employed by Lilly. These simple dechallenge-rechallenge studies added valuable information to the assessment of individual cases. This valuable information went well beyond the quality of the information provided by merely observing an AE and discontinuing the medication. However, this information's quality still falls short of that provided by formal inferential comparison in a controlled, parallel RCT or a formal crossover 'N-of-1' trial.

Medical disorders can be episodic in their course. Disorders with such a course underscore formal 'N-of-1' trials' relative superiority over simple dechallenge observation and even dechallenge-rechallenge observation. Beasley recalls a personally critical example case that shaped his thinking about the quality of evidence when attributing an AE to treatment and declaring it an ADR based on dechallenge-rechallenge data. A subject developed significant neutropenia during a Phase 3 development study of a new molecular entity (NME), approaching

neutrophil indices values consistent with agranulocytosis but without infection symptoms. The trial medication was discontinued, and the condition resolved. On internal unblinding of the study medication for regulatory reporting purposes, it was found that the patient was being treated with a placebo. With additional follow-up information, it was ultimately concluded that this patient probably suffered from cyclic neutropenia, a rare condition (Dale, Bolyard, Aprikyan, 2002; Dale, Bolyard, Marrero, et al., 2012; Dale, Bolyard, Leung, et al., 2017). According to Dale and colleagues (2012), this disorder was first described in 1910, but descriptions of its genetic etiology were not described until 1999 (Dale, Bolyard, Aprikyan, 2002), several years after this case occurred. Beasley did not know about this disorder before learning this subject's neutrophil indices and reviewing the potentially pertinent literature.

If the patient had been rechallenged in a single rechallenge episode, recurrence would likely have been observed. If treatment assignment had been to the NME, such a sequence, the resolution on dechallenge and recurrence on rechallenge, could have been interpreted as compelling evidence of treatment causation. As odds were only one out of four of assignment to placebo in the trial in which the patient was participating, it was likely that this case would have occurred on the NME, and without an 'N-of-1' study design, it could have been easily concluded that there was substantial evidence for the neutropenia to be a treatment effect of the NME. This interpretation could have been reinforced by the experience with other drugs known to cause neutropenia and agranulocytosis if the NME was considered to have a comparable molecular structure to clozapine. Except for an unlikely treatment assignment, and especially if a simple dechallenge-rechallenge evaluation had been conducted, this one case could have had a profound negative impact on the development of the NME.

Better quality data is always preferable to lesser quality data. We agree that for a specific domain of AEs, dechallenge-rechallenge studies, even those without multiple, controlled, sequential dechallenge-rechallenge periods, can add useful data in separating AEs from ADRs. However, for patients' long-term good, any study method's limitations must be kept clearly in mind when interpreting study results and drawing conclusions about treatment effects.

References:

Dale DC, Bolyard AA, Aprikyan A. Cyclic Neutropenia. *Semin Hematol* 2002; 39:89-94.

Dale DC, Bolyard AA, Marrero TM, Bonilla MA, Link DC, Newburger PE, Shimamura A. Boxer LA, for The Abstract 2141 - Severe Chronic Neutropenia International Registry and Repository. The natural history of cyclic neutropenia: long-term prospective observations and current perspectives. 2012; *Blood* 120:2141.

Dale D, Bolyard AA, Leung J, Tran E, Marrero TM, Newburger PE. Cyclic neutropenia, congenital and idiopathic neutropenia. *Blood* 2017; 130:(Suppl 1):2275. https://doi.org/10.1182/blood.V130.Suppl_1.2275.2275

Kravitz RL, Duan N, editors, and the DEcIDE Methods Center N-of-1 Guidance Panel (Duan N, Eslick I, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S). Design and Implementation of N-of-1 Trials: A User's Guide. AHRQ Publication No. 13(14)-EHC122-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2014.

August 15, 2019

3. Hector Warnes' Comment on the Postscript (Chapter 10), Followed by Warne's Additional Comment on the Postscript (Chapter 10), Followed by Beasley's Response, Followed by Warnes' Response to Beasley's Response: The Potential for False Positive and False Negative Attribution of ADR Status to an AE in Product Labeling

a. Hector Warnes' Comment on the Postscript (Chapter 10)

I read your excellent study three times. I was impressed by the premarketing risk study (CDER) published in 2005 (p. 49) and by two key findings in your own study: a reduced peripheral insulin sensitivity based on lower Rd. and an increased corticosterone concentration from baseline. Were the patients with these findings more likely to experience metabolic dysregulation on follow up?

I also came across a basic science study by Li H, Peng S, Li S et al. published in Nature in 2019, "Chronic olanzapine administration causes metabolic syndrome through inflammatory cytokines in rodent models of insulin resistance," which has settled my doubts about the mechanism of adverse side effects of olanzapine. I wonder if you would agree with their findings.

Reference:

Li H, Peng S, Li S, Liu S, Lv Y, Yang N, Yu L, Deng YH, Zhang Z, Fang M, Huo Y, Chen Y, Sun T, Li W. Chronic olanzapine administration causes metabolic syndrome through inflammatory cytokines in rodent models of insulin resistance. *Sci Rep* 2019; 9:1582. doi: 10.1038/s41598-018-36930-y.

March 5, 2020

b. Hector Warnes' Additional Comment on the Postscript (Chapter 10)

Charles Beasley and Roy Tamura wrote an incisive elaboration of "the sample sizes required to infer with reasonable certainty that some adverse medical event is caused by a drug." By statistical method they "illustrated the sample sizes required to infer with reasonable medical certainty that some adverse medical events while possible observed during the administration of a drug is not caused by the drug" being tested. They further pointed out the temporal pattern of occurrence as a key factor in identifying an adverse medical event and undoubtedly the adverse effect might be etiologically related to the drug being tested or have other etiology. I would dare to say that one third to one half of the hundreds of adverse side effects attributed to the tested drug which are printed in the complete prospectus may not present "reasonable evidence."

Another confounding finding in drug research is the fact that often the conclusions are published based on positive outcome and a high percentage of research findings which have negative outcome are not published. Positive outcome would imply that the p-value of < 0.05 is statistically significant difference of the probability of occurrence of the given event while $P>0.05$ is not significant.

We are all aware that there is a consensus that an adverse side effect is considered very common when it occurs in more than 10% of the patient population receiving the drug; common or frequent when it occurs in less than 1%; uncommon or infrequent when occurs in 1/1000; rare when it occurs in 1/10,000; and very rare when it occurs in more than 1/10,000.

The authors, using the Fisher Exact Test, reached the conclusion that the treatment group would have a 51% power should the events be estimated at 0.08 events with the control group and 1.67 events with the test drug. It would have 80% power if it were found 0.17 events with the control group and 3.33 with the experimental drug. It would require 7.905 patients for the treatment group to validate the results.

It is widely known that post-marketing drug prescription may detect a higher incidence of adverse side effects not previously detected during the research studies of up to 3,000 patients with a control group or using a double-blind-cross over design. At times, the post-marketing side effect is not reported. It is possible that the adverse effect is due to drug interaction because rarely is a patient only taking one drug. It is considered that the doctor should weight the benefits versus the risks. Apparently one out of 5,000-10,000 compounds that enter preclinical testing are approved. Every year some drugs are withdrawn from the market because of frequent side effects which may cause harm to the patient.

I recognize my limitations of the statistical methods and, like the authors, came to the conclusion that it is not an exact science that would drive us to the latest trend of a personalized medicine (tailoring pharmacotherapy to individual phenotypes).

I found through Google a synthesis of the limitations of the conventional statistical methods written by Pooja Mehta, MD, entitled "8 Main limitations of Statistics – explained," posted on the "Pooja Mehta Economics Discussion website". According to Mehta:

1. The statistical methods do not study the nature of phenomenon which cannot be expressed in quantitative terms. They need a conversion of qualitative data into quantitative data.
2. They do not deal with individual items. They consist of aggregates of facts or items placed in relation to each other.
3. They do not depict the entire story of phenomenon; when phenomena do happen there may be several causes involved that cannot be expressed in terms of data.
4. The data may have been collected by inexperienced persons or they may have been dishonest or biased.
5. Laws are not exact, e.g., the law of inertia of large numbers and the law of statistical regularity are usually approximations.
6. Results are true only on average. Statistics largely deal with averages and these averages may be made up of individual items radically different from each other.
7. When several statistical methods are used the results vary with each method used. Although we use many laws and formulae in statistics, the results achieved are not final and conclusive.
8. Statistical results are not always beyond doubt. They deal only with measurable aspects of things and, therefore, can seldom provide the complete solution to the problem. They provide a basis for a judgement but not the whole judgement.

Of course, we all agree that the efficacy of the compound being tested and its potential harm to the patient is of the utmost significance. Adverse side effects to the point of lethality are frequently seen in hospital wards (Ernst and Grizzle, 2001).

Xiaodong Feng and Hong-Guang Xie published an excellent survey, Applying Pharmacogenomics in Therapeutics, in which Cong Liu, Weiguo Chen and Wei Zhang pointed out: "The majority of the Adverse Drug Reactions are due to genetic polymorphism of the

enzymes which metabolize the drugs and may be as well due to polymorphism of the transporter of the psychotropic compound: e.g., SERT for serotonin, HLA-B 1502 for carbamazepine; CYP2D6 for venlafaxine, CYP2C9- 19 y D6, etc. Drug toxicity and positive drug response are often related to the Cytochrome P450." Human error, age, ethnicity, weight, co-morbidity, diet, gender and polypharmacy must also be taken into account" (Feng and Xie, 2016; Liu, Chen and Zhuang, 2016).

The question of responders and non-responders to psychotropic drugs has also raised controversies.

References:

Ernst FR, Grizzle AJ. Drug-related morbidity and mortality: Updating the cost-of illness model. *J Am Pharm Assoc* 2001; 41: 192-199.

Feng X and Xie HG (editors). Applying *Pharmacogenomics in Therapeutics*. Boca Raton: CRC Press (Taylor and Francis Group), 2016.

Liu C, Chen W, Zhang W. Essential pharmacogenomic biomarkers in clinical practice. In: Feng X, and Xie HG (editors). *Applying Pharmacogenomics in Therapeutics.* Boca Raton: CRC Press (Taylor and Francis Group), 2016.

July 11, 2019

c. Charles Beasley's Response to Warnes

We thank Dr. Warnes for his comments about our recent set of postings. Dr. Warnes has clearly understood one of our most important points. Likely, some proportion of what are listed as ADRs in product labeling lack the same level of 'proof' that they are ADRs for the drug as the level of 'proof' required to approve an efficacy claim for treating a medical disorder. He goes on to "dare to say" that perhaps one-third to one-half of listed ADRs lack 'proof' as being ADRs equivalent to 'proof' of efficacy.

Our experience over our 28 years in the pharmaceutical industry suggests the United States Food and Drug Administration's expectations for labeling AEs as ADRs have moved toward a standard of at least some credible evidence that an AE is an ADR. For fluoxetine, the first drug that reached approval shortly after I joined Eli Lilly and Company in 1987, with which I worked directly, virtually all AEs reported in the clinical development trials were listed as ADRs. For olanzapine, approved in 1996, AEs that were highly non-specific or with substantial reason to believe they were not ADRs were not listed as ADRs. For tadalafil, approved in 2003, only those AEs with reasonable evidence of being ADRs or AEs of such major clinical significance that medical prudence suggested the need to include them were listed. I cannot speak to the specifics of ADR listing standards in other regulatory venues. However, a widely held concept is that being over-inclusive of AEs that are unlikely to be ADRs dilutes product labeling's clinical utility.

We did not offer any estimation about the proportion of listed ADRs lacking robust 'proof' of status. However, we implicitly suggest that for many drugs, the AEs labeled as ADRs would need to occur with an incidence >2-3% with a substantially lower incidence in the appropriate control group to have robust 'proof' of being ADRs in what is now Chapter 8. This required incidence varies depending on the investigational drug's studies' sample sizes and the proper control group included in the drug development program's set of studies. Among a few other classes, cardiovascular disorder drugs, and anti-diabetic drug classes often have much larger sample sizes useful for proper comparisons in their development databases than other classes such as drugs for psychiatric disorders. The larger the useful comparative sample sizes, the greater the sensitivity to smaller differences in incidences or ratios of incidences between groups.

Lack of robust 'proof' that an AE is an ADR for a given drug does not, in our opinion, imply that AEs with lesser evidence of being ADRs should not be listed as ADRs. Consistent with the first principle of first do no harm, it is reasonable to expect a lower standard of 'proof' than required for efficacy to list an AE as an ADR. We believe that what is most fundamentally important is for any individual who uses these lists of ADRs for any purpose to recognize the potential for

false-positive inclusion of an AE in the list of ADRs. Most persons probably recognize that if a medical condition (AE) is not listed as an ADR, this is not strong evidence that the medical condition is not an ADR with a very low incidence. This matter was addressed in our discussion of the 'Rule-of-3' in what is now Chapter 9.

As Dr. Warnes pointed out and was illustrated by our work, rare ADRs are almost always identified after initial drug approval. This identification begins with the observation and reporting of AEs. We briefly discussed the need for better (higher quality, more robust 'proof') and more rapid means of determining whether such events are or are not ADRs.

About the eight points of Dr. Mehta cited by Dr. Warnes, we believe numbers seven and eight are particularly relevant to our work. Number seven addresses alternative statistical methods. We showed that alternative inferential analytical methods for the same outcome data could require different sample sizes. Expert statistical consultation can optimize the analytical methods for both planned, *a priori* analysis of a specific data type collected in an experimental design, and *post-hoc* analysis of such data.

Point number eight is critical to the process of performing the best assessment possible in the development of a list of ADRs for a given drug. Inferential statistical results (a p-value and/or a confidence interval of some magnitude) provide what is essentially a probability estimate for the 'proof of the truth' when the null hypothesis is rejected. If the null hypothesis is rejected based on a p-value of 0.05, there is at least a 95% probability that rejection of the null hypothesis was the correct thing to do and that the alternative hypothesis is the 'truth'. 95% is not 100%, and there remains a probability approaching 5% that the alternative hypothesis is not the truth and a Type I statistical error has occurred. Best decisions about what are or are not ADRs result from complex cognitive processes involving multiple levels and data types. These data types can range from information about the drug's pharmacological actions and well-accepted consequences of those actions, through information about the drug's kinetics and metabolism, individual case reports, and finally to formal studies with or without randomization with or without proper control. However, statistically significant evidence from studies with randomization and proper control comparison is one important and robust type of data.

August 15, 2019

d. Hector Warnes' Response to Beasley

I am grateful for Charles Beasley's reply. Their excellent scientific study has brought to light the issue of adverse drug-reaction versus adverse drug event with utmost clarity.

From a clinical point of view the variables are staggering pharmacogenomics, ethnicity, gender, co-morbidity, drug-interaction, age, weight, stage at which the illness is treated (acute, chronic or cyclical), the intensity and or the clustering of symptoms the type and severity of the adverse drug reaction and many epigenetic and environmental factors including diet that may impact on the metabolism of the drug.

I would agree with Xiaodong Feng and Hong-Guang Xie (2016) that eventually we shall be able to tailor pharmacotherapy to individual phenotypes.

Reference:

Feng X and Xie HG (editors). *Applying Pharmacogenomics in Therapeutics*. Boca Raton: CRC Press (Taylor and Francis Group) 2016.

December 12, 2019

e. Hector Warnes' Additional Response to Beasley

I am most impressed and overwhelmed by Charles Beasley's research paper. He has provided us with all the basic necessary, sufficient and contingent variables to be considered in a good practice by the clinical pharmacologist.

Usually all the possible side effects listed and read before prescribing any drug are not enough in order to take precautionary measures given the multiple variables to be considered, including the facts of polypharmacy, individual response specificity and co-morbidity. I congratulate Professor Beasley for his outstanding contribution.

December 3, 2020

4. Barry Blackwell's Comment on Beasley's and Tamura's Book, Followed by Beasley's Response, Followed by Blackwell's Final Comment, Followed by Beasley's Final Response, Followed by Jay Amsterdam's Comment on Blackwell's Final: Additional Factors Potentially Influencing the Inclusion of an AE as an ADR in Product Labeling

a. Barry Blackwell's Comment

Charles Beasley and Roy Tamura have produced 30 pages of statistical wizardry to demonstrate that the pharmaceutical industry, if it was so motivated, might be able to demonstrate whether or not a new drug does or does not cause a specific side effect or Adverse Drug Reaction (ADR), distinguished from an Adverse Event (AE), that is coincidental with a study but not due to the drug.

Their purpose is to illustrate the large sample sizes and elaborate statistical techniques to accomplish this task. Over the course of five months (November 2018 to March 2019) they published an outline of their thesis in seven different episodes including an introduction, comments, potential sampling errors, proof of or absence of presence of ADR, the real incidence of AE and the requirements to undertake massive and costly measures that might be imposed by regulatory agencies.

Lacking the statistical weaponry to digest, critique or refute such an impressive body of work, now an e-book, I would normally not attempt to do so were it not for the fact that authors view their enterprise as a reposte to comments made in and subsequent to my essay on corporate corruption in the pharmaceutical industry (Blackwell, 2016).

My rebuttal rests more on logic rather than statistical wisdom. To begin with their "definition of terms" states that whether or not an AE can be distinguished from an ADR depends on whether there is "reasonable evidence". But they note: "To the best of our knowledge this has never been operationally defined or even qualified by any regulatory entity or drug safety organization." They cite two international organizations and the American FDA.

As noted in my essay this is hardly surprising. The FDA derives 50% of its budget from industry payments for approval of a new drug application (NDA), a requirement imposed by the Republican Reagan administration (1980-1988). This creates a massive conflict of interest, encouraging the FDA to turn a blind eye to imposing costly sample sizes and elaborate statistical techniques that might and has discouraged industry innovation in the contemporary me-too era

So, where is the evidence that any pharmaceutical company has attempted to accomplish a costly endeavor it is not obligated to initiate? There are no contemporary examples cited. It is less expensive to cynically build the cost of potential class action lawsuits into the price of a drug before a missed ADR makes its presence known and then have court settlements impose silence on the victims.

Charles Beasley has spent a distinguished and blameless career in industry. This body of work embellishes his statistical ingenuity and if his former employers were to use it they might salvage a glimpse of the integrity their greed has consumed.

Reference:

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

July 25, 2019

b. Charles Beasley's Response

We thank Barry for taking the time to carefully read our work and provide his responsive comments posted on July 25, 2019. We want to elaborate on and clarify several points in Barry's response. Barry states: "The FDA derives 50% of its budget from industry payments for approval of a new drug application (NDA), a requirement imposed by the Republican Regan Administration (1980-1988)." These "industry payments" are termed 'user fees', and it is quite essential to understand that they are use taxes. The industry's user fees are paid for the review of potential commercial products requiring regulatory review and approval before the potential products can be commercially marketed. The payment of these user fees is required independent of the review results and whether the FDA does or does not approve the potential product for commercial marketing. Therefore, the payment of these user fees perhaps creates less "conflict of interest" than is suggested by Barry's comments.

The approved operating budget for the FDA for the fiscal year 2019 was $5.725B, as provided in Table 1 (pp. 5-6) of the Congressional Research Service document, The Food and Drug Administration (FDA) Budget Fact Sheet. Of this total, $2.575B was paid for through user fees. This budget was spread across 14 program areas. Most of these program areas had some bearing on human drugs (divided between pharmaceuticals and biologics [a protein or other substance derived from a biological source, including rDNA and monoclonal antibody products, and vaccines]). Extracting the program areas of foods, animal drugs, and feeds, tobacco products, and export–color certification program areas from the budget figures above because these are the 4 of 14 program areas not related to or minimally related to the diagnosis or treatment of human disease, the following budget figures result: total budget - $3.747B; the amount paid by user fees - $1.836B (49%). The 10 program areas include various administrative and infrastructure costs. If the program areas are restricted to only human drugs, biologics and devices, and radiological health, then the total budget is $2.859M, with $1.570M (55%) paid for by user fees. Of course, if these user fees were not being paid, then the entirety of FDA funding would come from general federal revenues (primarily income taxes). Is it more appropriate for companies potentially deriving benefit from the FDA's decisions to pay for these reviews when the payments for reviews are not contingent on the outcome of the reviews? Or, is it more appropriate for the American taxpayer, including individuals and corporations not involved in human health care / not regulated by the FDA, to pay for these reviews?

Barry states: "This creates a massive conflict of interest, encouraging the FDA to turn a blind eye to imposing costly sample sizes and elaborate statistical techniques that might and has discouraged industry innovation in the contemporary me-too era."

To suggest that the FDA is biased in favor of industry regarding establishing policy and standards for adequate drug development and decisions regarding individual product approvals or rejections based on FDA funding being derived in part from user fees is only speculation. This speculation ignores the relevance of the facts that not all funding is derived from user fees

and, more importantly, these user fees are paid regardless of whether a regulatory decision is to the economic benefit of a company or deals a severe economic blow to the company.

Our speculation, based on interactions with one Reviewing Officer, four Review Division Directors, and several more senior FDA staff, is that, for the most part, they are excellent scientists attempting to serve the American public. About safety assessments, they are appropriately conservative. Favoritism for industry does not influence them. Perhaps on occasion, they can be influenced by congresspersons and senators (and presidents) without a good understanding of science and with significant interests in their political agendas.

The requirements for the development programs for new pharmaceutical agents used to treat non-life-threatening diseases and used on a longer-term basis promulgated by FDA are in line with those established by committees of regulatory authorities, as we have previously discussed. These other regulatory authorities are funded by alternative methods to those that fund the FDA. Are all these regulatory agencies conspiring to benefit the human health industry? Alternatively, are these regulatory agencies designing development requirements that reasonably protect human safety while still allowing new pharmaceutical agents to be developed within a reasonable period that often extends well past five years from the first human dose?

The development program requirements for an individual potential new drug are sometimes modified by regulatory agencies from the general guidance that further extends the development timeline due to both pre-clinical and clinical observations during development. Such development program requirement adjustment is, as it should be, in the service of protecting public health from both the consequences of medical disorders needing treatment and the minor harm and significant harm that can be done by the treatments for those disorders.

Reference:

Congressional Research Service. The Food and Drug Administration (FDA) Budget: Fact Sheet. 2019.
https://www.everycrsreport.com/files/20190508_R44576_2584438761cef9cafc63c1bdf44ecca14f9b9335.pdf

December 5, 2019

c. Barry Blackwell's Final Comment

This pertinent and fascinating topic has been the subject of polite and civilized debate between the authors, Ned Shorter, Hector Warnes, Carlos Morra and me, now 138 pages long, already appearing as an e-book and perhaps a potential volume in the forthcoming annual INHN series.

My comments have been careful to stress a lack of statistical competence and a bias towards negative connotations based on political influences and corporate corruption that overtook the industry towards the tail end of the authors' unblemished and distinguished careers.

For this reason I confined my comments to historical and personal issues and later to logical concerns.

This type of controversy is exactly the kind that INHN is designed to handle, particularly now that traditional medical and scientific journals have become mired in controversy about their blemished publication practices. We owe a debt of gratitude to Charles Beasley and Roy Tamura for their integrity and courage.

April 9, 2020

d. Charles Beasley's Final Response

The response below is likely to be the final response to Barry Blackwell connected with our work (Beasley and Tamura 2019; Beasley, 2019). The work intended primarily to:

1. detail the magnitude of certainty about whether adverse events (AEs) observed in temporal association with the administration of a drug that are included in the product labeling for that drug are adverse reactions (ADRs) to that drug; and
2. remind readers that for many AEs included in product labeling, as ADRs, this magnitude of certainty is much less than the magnitude of certainty for the drug's efficacy for its approved indications.

We agree with Barry that the interchange with all parties involved has been polite and civilized. We would characterize our original work, comments, and replies as an in-depth review and exchange of information. The series of comments and replies have not felt like a debate to us, and for that, we are grateful.

With his commentary on Corporate Corruption in the Psychoparmaceutical Industry (Blackwell, 2016), Barry provided the stimulus for our work (Beasley and Tamura 2019; Beasley, 2019), and we thank him again for the stimulus. Our work's contents were of great interest to Beasley during his last 15 years or so working within the industry, but if it had not been for Barry, the work would not have been written. Beasley, at the encouragement of Tom Ban, wrote a comment (Beasley, 2017) on that commentary by Barry (Blackwell, 2016) in which Beasley provided some sample sizes for robust assessment of whether an AE observed during a randomized clinical trial (RCT) or set of RCTs was an ADR. The sample sizes described in that response were for 80% power and a single inferential analysis type. After posting that response, Beasley became concerned that his response was inappropriately simplistic regarding the sample sizes included in the response. Depending on the outcome observed in a prospective RCT and the inferential statistical method applied to the RCT results, a smaller sample size could result in a robust demonstration that an AE was an ADR. The more extensive work was written with Roy Tamura's assistance (Beasley and Tamura, 2019) along with a post-script (Beasley, 2019), and these works were intended to provide objective and full disclosure in this somewhat complicated domain.

The more extensive work (Beasley and Tamura 2019; Beasley, 2019), the various comments and questions about the more extensive work, and the responses and replies to the comments and questions have intertwined Barry's commentary (Blackwell, 2016), Beasley's comment on that commentary (Beasley, 2017), and also Ned Shorter's commentary on Mellaril and QT prolongation (Shorter, 2013). QT prolongation became an intertwined topic because Beasley's and Tamura's work (2019) addressed not only the sample size requirements to determine that an AE is an ADR, but several other topics, including studies intended to demonstrate that a potential ADR is unlikely to be an ADR for a given drug. QT prolongation, potentially leading to Torsades de Pointes (TdP), is one potential ADR for which an RCT design exists for a robust demonstration of the absence of a clinically meaningful effect predictive of TdP.

Besides thanking Barry, we thank all others who have provided comments/questions, further stimulating our thinking about these matters.

References:

Beasley CM Jr. Comment on Barry Blackwell's Corporate corruption in the psychopharmaceutical industry. inhn.org.contoversies. March 23, 2017.

Beasley CM Jr, Tamura R. 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) – Full text. inhn.org.ebooks. November 21, 2019 (2019a).

Beasley CM Jr. 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) – Postscript (Chapter 10). inhn.org.ebooks. October 24, 2019.

Blackwell B. Corporate corruption in the psychopharmaceutical industry. inhn.org.controversies. September 1, 2016.

Shorter E. The Q-T interval and the Mellaril story: a cautionary tale. inhn.org.controversies. July 18, 2013.

June 11, 2020

e. Jay Amsterdam's Comment on Blackwell's Final Comment

I just read Barry Blackwell's terrific commentary on your excellent series of articles describing "What We Know and Do Not Know…" about psychotropic drug side effects, in this week's INHN posting (April 9, 2020). I always knew that there was something different about your approach to clinical psychopharmacology – and that you were, of sorts, different from other researchers based in Pharma.

Back in the early 90s, when I first worked with you as the lead scientist on the Lilly HCEX long-term fluoxetine trial, I suspected that you were a fish swimming upstream, out of familiar waters, against the current of academia. In fact, there were times when I wondered what the reward was at Lilly for an academic-minded individual like yourself (beyond mere financial reward). Despite some minor professional ups and downs between us, during the highly successful HCEX project, I always admired your scholarly intellect and concise approach to problem solving. I do not believe that I could have happily survived in industry that regulated what I could think or say. I probably would not have lasted very long in that environment. I did interview for several positions at a few Pharma companies in the early to mid 1990s – always with you in mind, as a role model. However, at the time, I feared that I could not survive in an atmosphere where the epistemological approach to clinical trials was: "Dr. Amsterdam, we value your opinion; and, when we want it, we'll give it to you"! Of course, this simplistic view of industry writ small was a reflection of my own intellectual and personal short-coming; and, thankfully, I recognized its presence and did not bail from academia, as I seemed to require the confusion and intellectual disorganization of academic research to survive. (Note – Back in those Halcyon Days of pharmaceutical industry research, I had absolutely no knowledge or insight into the academic corruption and intellectual dishonesty that was brewing all around us within the field of psychopharmacology research. At the time, I was a hopelessly romantic, young researcher in search of fact. I didn't know what a KOL was until February 2011).

Despite all of the seismic changes that have occurred in our field over the past 35 years, it has been a privilege for me to have played in the same intellectual sand box with a scholar like yourself. Who knows, perhaps there may even be a future project on which we may collaborate, once again.

Reference:

Blackwell B. Final comment on Charles Beasley's and Roy Tamura's 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). inhn.org.ebooks. April 9, 2020.

June 25, 2020

5. Donald Kline's Comment on Beasley's and Tamura's Book, Followed by Beasley's Response: The use of large databases and Machine Learning / Artificial Intelligence in the Identification of ADRs

a. Donald Kline's Comment

Charles Beasley and Roy Tamura's treatise shows that the statistically minded clinician knows that rare events are trouble When it comes to the minuscule - 1/1000, therapeutic or toxic effects - enormous, impractical samples would be necessary for RCTs. We have given up hope of concluding from RCTs if It is there or not. Beasley and Tamura have assiduously put numbers on this generalization but have not contradicted it. In keeping with the substantial literature on the detection of rare events, they show that big, impractical sample sizes are required for the lucky RCT experimenter to have a fair chance of coming to an accurate conclusion. They use a variety of modern statistical approaches and show that different sample size estimates occur, but the conclusions do not break out of impractical sample space. I believe that the rapidly developing area of machine learning has not been applied but am not optimistic. My current opinion is that the RCT is too impractical to fulfill this goal.

There are multiple confounded approaches to naturalistic data. The great benefit of randomization, which balances out the variables you don't know about, is not available. I suspect reasonable, if shaky, conclusions will require similar large sample sizes that may be available from Scandinavian archives.

August 8, 2019

b. Charles Beasley's Response to Kline

We thank Don Klein for his comments concerning What We Know and What We Do Not Know by Conventional Statistical Standards About Whether a Drug Does or Does Not Cause a Specific Side Effect (ADR). Dr. Klein has underscored what we believe to be an important proposition in our work. Sample sizes can be minimized by selecting the appropriate experimental design combined with the optimal inferential statistical method best suited to the experimental design and the experiment's anticipated observations. Expert statistical consultation can be highly beneficial. However, robust 'proof' of effect or lack of effect remains an impractical goal for highly infrequent or rare events.

Dr. Klein has suggested the utility of large databases in the early and accurate identification of ADRs and specifically mentioned the Scandinavian countries as repositories of such databases. As briefly mentioned in what is now Chapter 9 of our work, we concur with this opinion. The utility of such databases is highly dependent on the accuracy and completeness of the information they contain. Unfortunately, even hospital-based medical records (e.g., discharge summaries) may contain inaccuracies on occasion. There is a good reason that when a clinical trial or development program's primary objective involves identifying an AE that might be an ADR, it is customary for the trial or development program's sponsor to rely on an Event Assessment Committee. The Committee would review, blinded to treatment, all clinical information made available regarding an AE reported clinically as the AE of interest and decide whether the reported AE is the AE of interest. For example, such a committee would generally be used in Major Adverse Cardiovascular Events (MACE) studies to develop new anti-diabetic medications.

In the US, the Food and Drug Administration (FDA) has recognized the importance of such an extensive database as a significant advantage over simply receiving reports of possible ADRs. FDA has developed a program to create such a database, and more information on this effort can be reviewed at https://www.fda.gov/safety/fdas-sentinel-initiative.

Dr. Klein has wondered about the potential utility of applying machine learning technology, and by implication, other artificial intelligence processes, to such a database resulting in a further advancement in speed and accuracy of identification of ADRs. He is not optimistic about the utility of such computational methods. We are cautiously optimistic. However, to be maximally helpful artificial intelligence technology will likely need to progress to the point that the programs can recognize novel patterns and associations against an extensive background of 'noise'. The programs will need to achieve something close to human creativity and imagination.

November 14, 2019

6. Daniel Kanofsky's Comment Followed by Beasley's Response: The Use of Patient Registries in the Assessment of Potential ADRs for Drugs Used Infrequently

a. Daniel Kanofsky's Comment

The authors state: "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."

I want to elaborate a bit on a mentioned complication of this quest. Drug-drug interactions have been reported in cases of adverse drug reactions as possible contributing actors (Kanofsky, Woesner, Harris, et al., 2011). Focusing only on clozapine, valproate has been implicated as a possible contributing factor in "clozapine induced acute renal failure" (CIARF) and clozapine induced myocarditis (Kanofsky, Woesner, Harris, et al., 2011; Woesner and Kanofsky, 2015; Nielsen, Manu, and Kane, 2015; Ronaldson, Fitzgerald, and McNeil, 2015; Kanofsky and Woesner, 2017). Our group used a Fisher's Exact test on the set of all reported CIARF cases which at that time was only eight. The analysis suggested co-treatment with antibiotics may exacerbate CIARF (Kanofsky, Woesner, Harris et al., 2012). A definitive statement would require many more reported cases.

In keeping with these thoughts, these serious but rare inflammatory responses to clozapine which include clozapine induced pancreatitis, colitis and pericarditis should lead to termination of clozapine but since clozapine can be a very effective antipsychotic drug when no other antipsychotic is effective the highly clinically relevant question emerges: is clozapine rechallenge safe and meaningful (Nielsen, Manu, Kane, and Correll, 2015)? Few cases of rechallenge have been reported. This lack of statistical power makes overarching conclusions impossible. Under these circumstances what can guide clinical decision making? Nielsen et al. respond: "As these low numbers illustrate, it is highly important that any patient who experienced a serious/potentially life-threatening ADR with clozapine who is later rechallenged is reflected in the literature, so that we can learn more about under which circumstances clozapine rechallenge is or is not safe."

Our group has expanded on this recommendation. We believe there is a need for a clozapine rechallenge case file or special registries. A case file or registry could encourage a greater and more accessible flow of information and expedite learning under what conditions a clozapine rechallenge can be safely conducted (Kanofsky and Woesner, 2017). Clozapine is the only psychiatric medication in this country that is currently dispensed using a national registry - the Risk Evaluation and Mitigation Strategy (REMS). The REMS program has the potential to become an ideal resource to locate American-based clozapine rechallenge cases. We hope this opportunity will be realized.

References:

Kanofsky JD, Woesner ME. Clozapine-valproate adverse drug reactions and the need for a clozapine rechallenge case file. *Prim Care Companion CNS Disord* 2017; 19(1):16. doi: 10.4088/PCC.16l01968.

Kanofsky JD, Woesner ME, Harris AZ, Kelleher JP, Gittens K, Jerschow E. A case of acute renal failure in a patient recently treated with clozapine and a review of previously reported cases. *Prim Care Companion CNS Disord* 2011; 13(3):e1-e5. PCC.10br01091. doi: 10.4088/PCC.10br01091.

Kanofsky JD, Woesner ME, Harris AZ, Kelleher JP, Gittens K, Jerschow E. Antibiotic treatment may exacerbate clozapine induced renal failure. *Intern Med J* 2012; 42:1272.

Nielsen J, Manu P, Kane JM, Correll CU. Dr. Nielsen and colleagues reply. *J Clin Psychiatry* 2015; 76:1694-1695.

Ronaldson KJ, Fitzgerald PB, McNeil JJ. Clozapine-induced myocarditis, a widely overlooked adverse reaction. *Acta Psychiatr Scand* 2015; 132:231-240.

Woesner ME, Kanofsky JD. Revisiting the discussion: termination of clozapine treatment due to renal failure. *J Clin Psychiatry* 2015; 76:1694.

April 16, 2020

b. Charles Beasley's Response to Kanofsky

We appreciate Dr. Kanofsky's Comment (Kanofsky, 2020) on our work (Beasley and Tamura, 2019) and are gratified to see that it has stimulated a call for data that should be readily available and would be of some aid to physicians in making critical risk-benefit decisions regarding patient care.

We consider schizophrenia to be a critical illness for which clozapine can be a life-saving treatment. However, as pointed out by Dr. Kanofsky, highly infrequent (approaching only one case in 1,000 treated patients) to rare (< one case in 1,000 treated patients) and very rare (< one case in 10,000 treated patients) cases of acute renal failure and inflammatory disorders (colitis, pancreatitis, pericarditis, myocarditis, among others) have been reported during treatment with clozapine. Given the infrequent to very rare occurrences (observations and reports) of these disorders in temporal association with clozapine treatment, it cannot be said with conventional statistical standards that such cases are ADRs to clozapine or only coincidental cases not due to or influenced by clozapine, and therefore, simply AEs observed during clozapine treatment. To complicate clinical care further, as pointed out again by Dr. Kanofsky, the occurrence of these disorders might be facilitated by clozapine, but clozapine treatment by itself does not directly cause the disorders. Concomitant treatment with clozapine and other drugs might be required to induce what still might be an infrequent to very rare ADR in patients treated with the relevant combinations.

Readers will have their own opinions about whether these cases are ADRs to clozapine, ADRs to clozapine plus another drug, or AEs not related to clozapine or clozapine combined with an additional, single medication or multiple medications.

Consistent with the medical dictum of "first, do no harm", discontinuing clozapine in the face of one of these reactions and not reintroducing clozapine as a treatment is good medical practice. However, there is uncertainty about whether these serious and potentially fatal events are ADRs to clozapine (or a clozapine combination) or AEs with no relationship to clozapine other than temporal co-occurrence of treatment with clozapine and the observation of the medical disorder. Given the gravity of schizophrenia, especially in some specific patients with the disorder, careful consideration of the individual patient's risk-benefit ratio in receiving subsequent treatment with clozapine favors re-treatment with clozapine, and some patients have received such treatment.

Such patients constitute 'rechallenge cases'. We have discussed the use of the rechallenge paradigm and the more formal 'N-of-1' experimental design (Beasley, 2020) in response to Shorter (2019) that had not appeared in INHN at the time Dr. Kanofsky provided his comment

(Kanofsky, 2020). As we pointed out, there are limitations to how the outcome of a single rechallenge can be interpreted. However, even with interpretive limitations and the potential for supporting incorrect beliefs about a rechallenge's safety, such data are likely more helpful than the absence of such data when making critical treatment decisions.

At the very least, with a sufficient number of rechallenge cases with their outcomes reported for a specific medical disorder (disorder recurred or did not recur on rechallenge), the binomial probability of the ratio of non-recurrence to recurrence could be computed. The following example illustrates the potential utility of only a relatively small number of cases.

Assume that the probability of non-recurrence of myocarditis on rechallenges with clozapine is 0.50 (the outcome is binary [recurrence or non-recurrence], and non-recurrence is an entirely random event with a probability similar to that of getting a 'head' on a single coin toss). With 20 reported rechallenge cases, non-recurrence is observed in 14 cases. Then, the probability of this outcome is 0.037 (https://www.thecalculator.co/math/Binomial-Calculator-741.html). Believing that the observed AE is not an ADR, assigning a higher probability to non-recurrence with each rechallenge case would be reasonable. Assigning a probability of 0.80 to non-recurrence with the same 14-to-6 non-recurrence to recurrence ratio, the probability of that outcome is 0.11. With the belief that the outcome is an ADR and assigning a probability of 0.20 to non-recurrence, the probability of observing the same 14-to-6 non-recurrence to recurrence ratio is 2.0×10^{-6}. More sophisticated analytical models could likely be employed to consider potential confounding factors and mediating variables if the data were available for the cases. Without a comparative group that would include non-treatment periods in 'N-of-1' designs, the assumptions about the probability of observing non-recurrence in an individual rechallenge in a series of such cases are quite an important consideration in the application of numerical methods to the assessment and then the interpretation of such case series data.

Such information is far from rigorously definitive based on what we have previously described as definitive by conventional standards. Furthermore, there are the caveats regarding open rechallenge compared to formal 'N-of-1' experiments. However, we believe such data could be beneficial in the real-world clinical treatment of schizophrenia with what might well be the most powerful treatment option currently available. Notably, myocarditis is not a symptom where a variety of factors might influence its reporting. While its diagnosis might be missed or diagnosed when not present, these possibilities are much less likely than an inaccurate Positive and Negative Syndrome Scale (PANSS) Total score (Kay, Fiszbein and Opler, 1987) due to an unsatisfactory rating interview without good patient-rater rapport. The utility of such single rechallenge case report data is highly dependent on the extent and quality of the available data. Nevertheless, our example demonstrates that even a small number of cases can provide beneficial guidance if there is a substantial preponderance of recurrences or non-recurrences.

Dr. Kanofsky has pointed out that all patients treated with clozapine in the US are registered in a Risk Evaluation and Mitigation Strategy (REMS) related database. Physicians who treat with clozapine and do rechallenge patients with the medication following serious medical disorders should be strongly encouraged to report these results in the medical literature. However, it should be kept in mind that such work would be another 'unfunded activity' in the life of many very busy, and in the view of some, under-compensated physicians, perhaps under the pressure of 'performance quotas'. With clozapine available generically, no single manufacturer of the medication is highly vested in extending the knowledge of both its positive and negative clinical potentials.

References:

Beasley CM Jr. Response to Edward Shorter's Comment on Charles Beasley's and Roy Tamura's 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) – Definition of terms used in this book (Chapter 2). inhn.org.ebooks. August 15, 2019.

Beasley CM Jr, Tamura R. 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) – Full text. inhn.org.ebooks. November 21, 2019.

Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull* 1987; 13(2): 261-76.

Shorter E. Comment on 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) – Definition of terms used in this book (Chapter 2). inhn.org.ebooks. May 9, 2019.

June 4, 2020

7. Carlos Morra's Question to Charles Beasley Followed by Beasley's Reply: The Use of Small Studies with Intense Monitoring to Assess Potential ADRs

a. Carlos Morra's Question to Charles Beasley

Are there any infrequent-rare Adverse Drug Reactions, such as myocardial infarction, aplastic anemia and Stevens-Johnson syndrome that you mentioned in your Commentary, that can be effectively evaluated with intensive monitoring in Phase 1 studies, with or without the use of biomarkers?

October 3, 2019

b. Charles Beasley's Reply

We thank Carlos Morra for his question. His question provides us an opportunity to expand on a matter discussed in our book. This matter is the conduct of studies intended to 'prove' the absence of an ADR analyzed using non-inferiority inferential methods, presented in three separate postings by Beasley to INHN. These three submissions dealt with QTc prolongation and the so-called Thorough QT Study (TQT Study) (Beasley, 2015, 2016, 2018). Carlos's question about whether any highly infrequent or rare ADRs in the large, clinically-treated population can be 'addressed' in a small, Phase 1 human clinical trial is a qualified yes. By 'addressed', I mean determined or predicted (with a reasonable probability of accuracy) to be an ADR (not merely an AE) in the population treated with the drug being studied and therefore requiring discussion in the drug's product labeling.

Our answer to this question is perhaps surprising given that we are discussing highly infrequent or rare ADRs that have little chance of being observed even once in sizeable clinical trial populations (e.g., populations between 1,000 – 5,000). The qualified yes is only for ADRs for which there is a biomarker or predictive surrogate for the ADR observable in small populations if a drug does have a risk liability for an ADR of interest.

This biomarker must first be extremely sensitive (have a high negative predictive value). It is essential that when the biomarker is absent in the experimental population, the probability of observing the ADR of interest in the population treated in clinical practice approaches 0. With clinically severe, potentially fatal ADRs, reliance on biomarkers that would not be positive (not differentiate the drug from control) in a small experimental population, but in the real-world clinical use of the drug in large populations, even a few cases of the ADR would occur, would not be in the best interest of public health.

Ideally, the biomarker should also be highly specific (have a relatively high positive predictive value). It is desirable that when the biomarker is present in the experimental population, the probability of observing the ADR of interest in the population treated in clinical practice is high approaches 100%). From a public health perspective, it is not desirable for Phase 1 studies that result in the false-positive prediction of serious ADRs to keep potentially helpful medications from reaching patients.

In the August 20, 2015, INHN Announcement, a Comment written by me (Beasley, 2015), commented on Edward (Ned) Shorter's commentary (Shorter, 2013), The QT Interval and the Mellaril Story: A Cautionary Tale. This initial comment did not directly address Ned's information regarding Mellaril but rather provided background information on the epidemiology of drug-induced Torsade de Pointes (TdP) and my views on the appropriateness of FDA's

restriction on the maximum approved dose of citalopram based on a TQT Study that failed to reject the null hypothesis of a difference between citalopram and placebo in a non-inferiority analysis (the standard inferential analysis for a TQT Study). The study failed to show that citalopram was not different from placebo.

On February 25, 2016, the INHN Announcement posted a brief Further Comment on Ned Shorter's essay (Shorter, 2013) The QT Interval and the Mellaril Story: A Cautionary Tale authored by me (Beasley, 2016). This piece described a recently published article regarding changes in QTc with SSRIs describing prolongation with citalopram relative to other SSRIs that I thought relevant to judgment regarding the FDA's product labeling change's appropriateness for citalopram.

On January 25, 2018, a Final Comment on Ned Shorter's essay (Shorter, 2013), The QT Interval and the Mellaril Story: A Cautionary Tale, authored by me (Beasley, 2018) posted on INHN. This piece did not elaborate further on the FDA's action's appropriateness regarding the approved dose of citalopram. This piece described my thoughts on the evolving understanding and use of the heart-rated corrected QT interval (QTc) and the pathophysiology of TdP and other malignant ventricular tachydysrhythmias. This 2018 work pointed out several matters of importance to Carlos' question with the TQT Study as an example of where a small Phase 1 study might provide a better answer to a critical safety question than randomized clinical trials that would be of such enormous size and so lengthy that they would be impossible to conduct.

This material is reproduced from that Final Comment below.

While QTc prolongation is a biomarker for the risk of TdP and other ventricular tachydysrhythmias, even substantial QTc prolongation does not invariably lead to TdP. Multiple drugs that prolong QTc are not associated with TdP, including amiodarone, carvedilol, ebastine, loratadine, phenobarbital, ranolazine, salbutamol, tamoxifen, and tolterodine (Hondeghem, 2008a). Additionally, drugs that prolong QTc can be antiarrhythmic, e.g., amiodarone.

Hondeghem and colleagues (Hondeghem, 2001; 2008a,b; Shah, 2005) have proposed a set of four drug-induced changes (or characteristics of the changes) in cardiac electrophysiology that appear to be necessary to result in either TdP that can spontaneously revert to normal sinus rhythm (~80% of occurrences) or degrade into ventricular fibrillation (Vfib) or result directly in Vfib. These cardiac electrophysiologic changes are best assessed through cardiac action potential studies in tissue preparations, but some have biomarkers evaluable on the surface ECG.

This set of changes is referred to by the acronym of TRIaD. The first of these changes is triangulation (T), the lengthening of ventricular action potential (AP) duration specifically by prolonging Phase 3 of the AP. Triangulation lengthens QTc that reflects the AP if Phase 2 of the AP (plateau phase) does not shorten. However, triangulation does not extend QTc (or the AP's total duration) if Phase 2 is shortened. Prolongation of Phase 3 repolarization is specifically defined as an increase in AP30-90 duration in action potential studies (Shah 2005). The ECG manifestation of triangulation is a widening and flattening of the T-wave (Shah, 2005). Such widening and flattening could be quantitated by measuring the onset to the end of the T-wave, the T-wave amplitude, ratios of these two parameters, and the absolute values of these two parameters. Phase 3 repolarization is strongly contributed to by potassium influx resulting in the I_{Kr} current, and blockade of that current can result in triangulation.

The second factor, a characteristic of change, is reverse use dependence (R) of the triangulation/prolongation of Phase 3 repolarization – a more significant effect at slower heart rates (Shah, 2005). A negative correlation between QTc length and heart rate would reflect reverse use dependence, but this cannot be assessed on a standard 10-second ECG, although it might be evaluated on an extended recording (Holter) if the recording interval captured a sufficient range of different, sustained heart rates.

The third alteration is temporal variability in the action potential duration on a cycle-to-cycle basis, referred to as instability (Ia) (Shah, 2005). The ECG manifestation of instability is T-wave alternans (Shah, 2005) that is a beat-to-beat change in the T-wave morphology, including its amplitude, sometimes so large to result in the alternating polarity of the T-wave. Variations in width (including width from onset to peak vs. peak to end reflecting symmetry) and amplitude of the T-wave might quantitate such morphological change.

The fourth change is transmural dispersion (D) of ventricular repolarization (Shah, 2005). There is an ordered progression of repolarization across the ventricular wall. Normal repolarization begins with epicardial repolarization, followed by endocardial repolarization and, finally, M-myocyte (mid-myocyte, deep subendocardial) repolarization. Disruption and desynchronization of this sequence, particularly with M-myocytes, is dispersion. The ECG manifestation of dispersion is the lengthening of the time between the peak and the end of the T-wave, referred to as Tpe. This length is sometimes corrected for QT (Tpe/QT). The terminology is confusing across the relevant literature because some authors refer to the absolute length as Tpe, and some authors refer to that length corrected for QT as Tpe rather than Tpe/QT.

TRIaD predisposes to the development of TdP that might or might not progress to Vfib and the development of Vfib without preceding TdP. Other aspects of cardiac electrophysiology that drugs can influence due to blockade of other cardiac ion channels and currents (besides I_{Kr}) and alterations in the autonomic tone, among other influences, predispose to Vfib's occurrence in the presence of TRIaD. λ is the product of the Effective Refractory Period (ERP) and Conduction Velocity (CV) (λ = ERP * CV). The ERP is the time from myocyte depolarization initiation through partial repolarization (Phase 3) when stimulation does not result in a propagated AP (a second AP). The CV is the speed of transmission of depolarization. As λ decreases, there is an increased risk of Vfib (abrupt onset or evolution from TdP), and as λ increases, there is a higher likelihood of spontaneously terminating TdP (Shah, 2005).

In general, most non-cardiac drugs that lengthen QTc, do so by blocking the I_{Kr} current, and drugs that block I_{Kr} will often, but not always, be associated with all components of TRIaD. Therefore, while not perfect, QTc prolongation can be used with some caution as a biomarker for the risk of TdP. One notable exception to this general association between I_{Kr} blockade and TRIaD and risk of TdP is when the drug that blocks I_{Kr} also blocks Na and/or Ca currents as these pharmacological actions can offset the effect of I_{Kr} blockade (fluoxetine is one example of such a drug).

Based on the information briefly reviewed above, except for drugs intended for cardiac conditions that alter the activity of multiple cardiac ion currents other than I_{Kr} and non-cardiac drugs that block I_{Kr} but also possess compensatory pharmacological activity, the TQT Study is a reasonable method of risk prediction. It might well result in more false-positive signals than missing drugs with the potential for causing TdP. The information above suggests that a set of pre-clinical studies might be superior to a Phase 1 human study for risk prediction in this area.

The TQT Study has wide, international regulatory acceptance as a way of predicting a potential risk of TdP (more technically correct, given the non-inferiority analysis relative to placebo, predicting the lack of potential risk of TdP) in that this is an arrhythmia that occurs in a small proportion of persons with an inappropriately prolonged QTc. As noted above, about 80% of cases of TdP revert to normal sinus rhythm. However, 20% progress to fatal (without proper medical management) Vfib, and with what would have been a brief loss of consciousness but occurring in the wrong circumstances (e.g., while swimming), additional fatalities might occur with TdP.

We are aware of one other ADR for which there is regulatory acceptance for using a specialized Phase 1 study for risk prediction (again, the absence of risk is analyzed with a TQT study). This ADR is Substance Abuse Disorder (of the drug under evaluation). If the drug has pharmacological activity similar to that of other drugs of abuse or results in patients' subjective experience as similar to that of persons that abuse other drugs/substances, then a Phase 1 study (Human Abuse Potential [HAP] Study), using a population enriched for being prone to non-medical use of drugs similar to the one under evaluation can be employed to address this potential ADR.

As a final note regarding the two Phase 1 studies discussed above, both the TQT and HAP Studies are conducted with positive controls to confirm the studies' assay sensitivity. Furthermore, the inferential analysis for the drug-placebo comparison is a non-inferiority analysis, as noted above. The null hypothesis that must be rejected for the study to be a success (from the investigator/sponsor perspective) is that the drug and placebo are different. If the experiment is a success, it 'proves' (within an a priori magnitude of acceptable observed difference) that the drug is not inferior to placebo in causing more cases (or greater mean change) of the biomarker/predictor of the ADR than placebo. Failure to reject the null hypothesis cannot be correctly interpreted as proving that the drug does have a risk of causing the ADR.

There is an additional potential ADR, Type II diabetes mellitus, where we believe the risk for this ADR can be potentially adequately assessed with a set of two Phase 1 studies. These two studies are a hyperglycemic glucose clamp study (evaluates pancreatic β-cells' capacity to produce and release an appropriate amount of insulin in response to an increase in systemic glucose) and a hyperinsulinemic-euglycemic glucose clamp study (evaluates hepatic cells' and cells of other tissues [primarily muscle and adipose tissues] capacity to respond to insulin and dispose of glucose [transport glucose into the cells]). I have previously discussed the details of these studies (Beasley, 2019). Unfortunately, based on a review of virtually all placebo-controlled clamp studies conducted with olanzapine, there appears to be a lack of consensus on how these studies should be conducted and analyzed. Without robust consensus among experts on both the adverse medical event of interest that might be an ADR for a drug of interest and the optimal conduct and analyses of such studies, the use of the studies for ruling-in or ruling-out risk is limited.

The pair of glucose clamp studies differs from the TQT Study and HAP Study because regulators do not require these studies for approval to market a drug. As such, this pair of Phase 1 studies does not have implicit regulatory acceptance as a means of excluding the risk of the medical event of diabetes mellitus as associated with a drug. However, this pair of studies has demonstrated adequate sensitivity in demonstrating the risk of diabetes mellitus (or hyperglycemia) with a range of drug classes such as corticosteroids and β-blockers.

The following summarizes our response to Carlos. If the cascading elements of pathophysiology that lead to a clinically significant adverse medical event are well understood, and a biomarker or risk predictor for these elements of pathophysiology can be found that would manifest itself in a substantial proportion of a small population treated with a drug of interest, then small Phase 1 studies might be able to determine risk (or lack thereof) for the adverse medical event as an ADR. There must be a robust consensus on conducting and analyzing a study that uses the biomarker / risk predictor as a dependent variable in an experiment for such a Phase 1 study to be truly useful.

References:

Beasley CM Jr. Comment on Edward Shorter's The Q-T interval and the Mellaril story: a cautionary tale. inhn.org.controversies. August 20, 2015.

Beasley CM Jr. Further Comment on Edward Shorter's The Q-T interval and the Mellaril story: a cautionary tale. inhn.org.controversies. February 25, 2016.

Beasley CM Jr. Final Comment on Edward Shorter's The Q-T interval and the Mellaril story: a cautionary tale. inhn.org.controversies. January 25, 2018.

Beasley CM Jr. (Olanzapine and diabetes mellitus, evolution of data – illustrating the difficulties in identification of adverse drug reactions) Response to Edward Shorter's comment on Charles Beasley's and Roy Tamura's 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 (Chapter 1). inhn.org.ebooks. July 4, 2019.

Hondeghem LM, Carlsson L, Duker G. Instability and triangulation of the action potential predict serious proarrhythmia, but action potential duration prolongation is antiarrhythmic. *Circulation* 2001; 103:2004-2013.

Hondeghem LM. QT prolongation is an unreliable predictor of ventricular arrhythmia. *Heart Rhythm* 2008a; 5:1210-1212.

Hondeghem LM. Use and abuse of QT and TRIaD in cardiac safety research: importance of study design and conduct. *Eur J Pharmacol* 2008b; 584:1-9.

Shah RR, Hondeghem LM. Refining detection of drug-induced proarrhythmia: QT interval and TRIaD. *Heart Rhythm* 2005; 2:758-882.

Shorter E. The Q-T interval and the Mellaril story: a cautionary tale. inhn.org.controversies. July 18, 2013.

September 2, 2021