David Healy: Do Randomized Clinical Trials Add or Subtract from Clinical Knowledge

Charles M. Beasley, Jr's comment – Part 1

My Comment on Dr. Healy's Essay consists of two parts, with this being Part 1. This Part 1 consists of two sections:

1. Background Information and Introduction Concerning Clinical Studies and Randomized Clinical Trials: This section describes what my comment is about, defines terms that I use, and contains additional information pertinent to the second section.
2. Areas of Possible Disagreement: This section describes several areas in which I potentially disagree with things written by Dr. Healy and my bases for these possible disagreements.

Background Information and Introduction Concerning Clinical Studies and Randomized Clinical Trials

This Comment addresses Dr. David Healy's Essay, Do Randomized Clinical Trials Add or Subtract from Clinical Knowledge - The Fault lies in our Stars not in Ourselves: Randomized Controlled Trials and Clinical Knowledge? (Healy 2021), contained in his Reply to a Comment by Dr. Jean-François Dreyfus (2021) on Dr. Healy's Essay Do Randomized Controlled Trials Add to or Subtract from Clinical Knowledge? (Healy 2020). The Healy (2021) Essay was a revision of the Healy (2020) Essay. My Comment addresses the 2021 revised Essay, the most recently posted version.

There are two parts to my Comment. The first part describes areas where I possibly disagree with Dr. Healy and the basis for my possible disagreements. This INHN posting is Part 1. In Part 2, a future posting, I present in greater detail than in Part 1 my conceptual framework for characterizing the components of an RCT that progress from forming a hypothesis to be tested through the writing and publication of an academic manuscript or an in-depth regulatory report that might well extend to many thousands of pages. Both academic manuscripts and extensive regulatory reports include human cognitive interpretation of the results of the RCT. In Part 2, I

also offer suggestions on how some RCTs' components and their reporting to the medical community and regulators might be improved. These suggestions are based on my firsthand experiences in designing and reporting RCTs.

I find myself in "possibly disagreement" rather than "disagreement" with points in Dr. Healy's Essay for three reasons. First, after multiple readings of Dr. Healy's revised Essay (Healy 2021), I agree with many of his points. Second, after I read past what I consider Dr. Healy's rather vitriolic descriptions of his concerns with RCTs compared to other methods of assessing medical treatments, I find myself in greater agreement with some of his other points. Third, I am interpreting Dr. Healy's written text. If my interpretations (i.e., my beliefs, influenced by my life experiences, about what he intended to convey) are not in agreement with what he intended to convey, I might agree with what he intended to convey.

Words/terms (i.e., accurate) and their linguistic derivatives (i.e., accuracy) that I use in this Comment that have specific meaning as I use them are placed in single quotations ('WORD') and defined within the text. While I have endeavored to assure that all instances of these words are enclosed in their single quotation marks, there might be rare instances where the quotation marks are inadvertently missing.

The areas of possible disagreement are at both a macro- and a micro-level

There is only one over-arching, macro-level possible disagreement, although there are three components to this possible disagreement. This primary possible disagreement concerns RCTs' appropriate characterization regarding their nature, conduct, interpretation, and relative importance compared to other methods of acquiring knowledge about medical treatments. When I refer to medical treatments, I include those approved for commercial sales by regulatory authorities and those under study as potential treatments. Most frequently, these treatments are drugs, but treatments include devices, instruments, psychotherapies, among others. Diagnostic and laboratory measurement devices can be included as well. Health authorities (e.g., the Food and Drug Administration [FDA] in the US) do not regulate all treatments, with psychotherapies being one example. However, most treatments and diagnostic methods are regulated and require approval by a health authority. Research that increases our valuable knowledge of regulated treatments is acquired both before and after their commercial use approval.

The first component of this macro-level possible disagreement is about RCTs' potential to add clinically relevant information about treatments that improve their use and improve patients' outcomes when treatments are used. The second component is a possible disagreement concerning whether an RCT can be considered a scientific experiment. The third component is a possible disagreement that concerns the extent to which an open-label challenge-dechallenge-rechallenge ("CDR") evaluation of an adverse event (AE) can demonstrate that the AE is an adverse treatment (drug) reaction (ADR).

As I discuss AEs and ADRs, I must provide definitions for these terms. An AE is any untoward and harmful, or potentially harmful medical event (e.g., specific disease, disorder, syndrome, sign, symptom) experienced while receiving medical treatment, often a drug. An ADR is an AE with a reasonable probability of being directly or indirectly caused by or facilitated by the treatment (drug). In common terms, an ADR is a side-effect.

While I only discuss two examples of micro-level possible disagreements, I have more of these possible disagreements. The first of these is a possible disagreement about the completeness and objectivity of Dr. Healy's characterization of Sir Anthony Bradford Hill's beliefs in the relative strengths of RCTs in advancing clinical information that can improve patients' outcomes.

The second micro-level possible disagreement concerns the completeness and objectivity of Dr. Healy's summary of information available in 1959, in which he describes a causal association between imipramine treatment of depression and an increase in suicidality (the spectrum of thoughts and behaviors that progress from an excess/inappropriate preoccupation with death to a suicide attempt that results in death).

In Part 2 of this Comment, I offer my thoughts on how some of the components of the RCT's evolution from conceptual hypothesis formulation to public disclosure of the interpretation of the RCT's results can be improved. RCTs should add truthful and helpful information about treatments and not subtract from quality information. For an RCT to meet this purpose, it should be what I term a 'proper' RCT.

I must define what I mean by a 'proper' RCT. The simplest definition of my concept of a 'proper' RCT is that it maximizes the probability that it increases knowledge of a treatment that is truthful/'accurate'/correct. I consider an RCT in terms usually applied to measurements that are

made in RCTs. As I understand the concept, statistically, 'accuracy' is the extent to which a measurement of thing X approaches the true measurement. In some instances, a measurement can be a combination of measurements or descriptive statistical summaries of measurements. For example, an individual subject's PANSS total score at a visit is a measurement. As I am using the term "measurement," the mean change for a group of subjects is also a measurement (although actually a descriptive statistic). Combining the mean change within a treatment group with the proportion of patients responding to the treatment within that group, as a two-component endpoint, to better characterize a treatment's effect is also a "measurement."

I apply the terms "Type I error" and "Type II error" that have specific meanings for the formal null hypothesis and alternative hypothesis for the result of an inferential statistical analysis to the ultimate cognitive interpretation of the results of an RCT in defining an 'accurate' and therefore 'proper' RCT. An RCT that is 'accurate' does not result in either a Type I error or a Type II error in the RCT's cognitive interpretation in any regulatory or public dissemination of the RCT's results. With a Type I error, the inferential analysis result leads to the incorrect rejection of the null hypothesis. With a Type II error, the inferential analysis result leads to the incorrect failure to reject the null hypothesis.

An RCT studies a conceptual hypothesis. For example, the conceptual hypothesis is that treatment X reduces the symptoms of schizophrenia. This conceptual hypothesis is formalized in terms of a difference between treatment X and a control treatment (e.g., placebo) in one or more measures. The conceptual hypothesis is the alternative statistical hypothesis. The statistical null hypothesis is the absence of effect of treatment X expressed as a lack of difference between treatment group X and the control group based on the selected measures.

If a 'proper' RCT rejects the null hypothesis and results in an interpretation that treatment X reduces symptoms of schizophrenia, treatment X indeed and truthfully causes a reduction in the symptoms of schizophrenia *in the entire population for which the treatment might be used*.

Additionally, if the RCT fails to reject the null hypothesis, the RCT is 'accurate' if the treatment indeed and truthfully does not cause the potential effect being studied, at least *within the study's system, the specific experimental population* (no Type II error).

A ‘proper’ RCT’s interpretation must be ‘accurate’ from both perspectives of the lack of an interpretive Type I error and an interpretive Type II error.

I distinguished the populations relevant to the absence of a Type I error and a Type II error because the interpretation of a ‘proper’ RCT that rejects its null hypothesis should be generalizable to the entire potential population that might receive treatment X. However, failure to reject the null hypothesis should not be generalized to the interpretation that treatment X has been indeed and truthfully demonstrated *not* to improve the symptoms of schizophrenia in *any* population with schizophrenia. This matter is discussed immediately below.

The crucial point is that if the RCT fails to reject the null hypothesis when that null hypothesis was of no effect, the most complete and objective interpretation of the RCT is that it fails to support the belief that the treatment causes the effect being studied. It would be an inappropriate interpretation to conclude that the RCT supports the belief that the treatment does not cause the effect being studied.

As Roy Tamura and I have explained (Beasley and Tamura 2019), RCTs (non-inferiority [usually to placebo] studies) can be designed and analyzed to demonstrate the absence of an effect, but their design is quite specific. In the analyses of RCTs supporting a belief in the lack of effect, a slight effect must be considered clinically equivalent to the lack of an effect. An analysis supporting an absolute absence of effect would require infinite experimental treatment and control group sample sizes. However, this magnitude of ‘difference equivalent to no difference’ can be set quite small.

This slight acceptable difference resulted in observing less effect in the treatment group than in the control group when the RCT supported the lack of effect in many of these studies. Such observed numerical superiority for the treatment studied can be seen for the mean change in a continuous numerical variable (i.e., heart-rate corrected QT interval) or the proportion of subjects experiencing a binary outcome (i.e., an AE being studied such as Major Adverse Cardiovascular Events that are routinely performed with non-insulin anti-diabetic drugs).

The US Food and Drug Administration (FDA) explicitly requires several such studies to address safety questions for drugs to receive regulatory approval in the United States, as Tamura

and I (Beasley and Tamura 2019) have described, along with specific examples. Such studies are also required in other regulatory venues.

I am not providing specific references. However, I have read lay press coverage of such failed studies where the conceptual/alternative hypothesis was that the drug or treatment had some beneficial effect that incorrectly interpreted the "absence of evidence as evidence of absence," claiming that the study demonstrated that the drug or treatment did not have the hoped-for effect. I believe I first heard this succinct characterization from Dr. Paul Leber, former Director of the then Division of Neuropharmacological Drug Products within the FDA.

Again, the critical point here is the appropriate interpretation of an RCT where the null hypothesis is one of no effect, and the study's analysis fails to reject that null hypothesis. That appropriate interpretation is simply that the study failed. In the interpretive step of the RCT process, interpreting the result of such an RCT as supporting a lack of effect would make for bad science. Such an RCT would not be 'proper' based on a flawed interpretation.

I also suggest that RCTs must also be 'precise' to be 'proper.' As I understand 'precision' from a statistical perspective, it is the extent to which replicate measures of thing X are equivalent to one another. 'Precision' can be considered on multiple levels. 'Precision' can be adversely affected by occult changes in single thing X, the measurement instrument itself, and the measurement instrument's application to the single thing X. Every time I place a piece of a substance on a scale, some mass is eroded. With sufficient repetitions of measurement, the mass of the single exemplar of thing X has changed from previous measurement results.

Every measurement instrument and assay is associated with some magnitude of intra-assay and between assay variability. Clinical laboratories (and/or the assay developers) routinely evaluate their assays for the assays' magnitudes of assay variability with sets of different samples spiked with identical quantities (identical past the magnitude of discrimination of the assay). With no actual change in a patient's mental status, a single interviewer could administer an interview for the PANSS and score the PANSS in two rapidly sequential interviews and arrive at slightly different scores. Two separate interviewers/raters could also arrive at different scores based on hearing responses to the same interview or conducting separate but rapidly sequential interviews. Many factors could explain the variance for the single interviewer/rater, and additional factors could contribute to the variance between interviewers/raters.

Assessing the 'precision' of an RCT requires one or more replications of the study of the potential causation of the effect studied in the first or index RCT. Results of multiple RCTs of treatment A and potential effect B where the RCTs' results consistently support the interpretation that treatment A causes effect B reflect the RCTs' precision. However, multiple bad design, execution, and analysis factors can result in a failed study when treatment A does cause effect B, and the RCT was intended to demonstrate this causal relationship. So, as explained above, a failed study does not demonstrate the absence of an effect.

What is the most objective interpretation of, say, 10 studies where eight failed to support causation of the effect being studied and two supported causation of the effect? That is a question that I cannot answer satisfactorily for myself. The larger the ratio of failed RCTs to RCTs supporting effect, the greater my concern about, at the very least, the size of effect if there even is an effect. However, I also consider the appropriate interpretation of a failed study in attempting to formulate an answer to this question satisfactory to me.

Here is a situation where a pair, or more, of RCTs with the conceptual hypothesis of 'equivalence to placebo' would be helpful. These RCTs would be designed and analyzed to potentially support the absence of causation of the effect of interest. However, if this pair, or more, of RCTs failed to reject their null hypotheses of a difference and therefore failed to support the lack of causation of the effect of interest, the interpretive waters regarding the treatment and what it causes or does not cause would be exceedingly murky.

For an investigational drug intended to treat a non-life-threatening, non-rare condition, two studies demonstrating efficacy are generally required by the FDA for the investigational drug to be granted regulatory approval, and failed studies are generally not considered (FDA CBER and CDER 2019).

However, even 'proper' RCTs have their limitations. RCTs cannot answer some crucial questions about a treatment for several reasons, in addition to the impractical size of the RCT that would be required to address some of those questions. Also, it must always be kept in mind that a single RCT designed to demonstrate a treatment effect can result in a Type I error. Assume that a single effect is tested in an RCT with a single experimental treatment compared to a single control treatment, based on a single parameter, and the statistical criterion for rejecting the null hypothesis of no difference is set at the conventional p-value of ≤ 0.05. Then, based on setting the p-value for

rejection of the null hypothesis at ≤0.05, 5% of a vast number of such RCT replicates (*all* factors that influenced the observed results in the index RCT replicated exactly in that vast number of RCTs) might be expected to result in a Type I error and support the belief that the treatment has the effect when in truth, the treatment does not have the effect. As noted above, repeated, independent RCTs supporting a specific effect provide confidence in the truth of the specific effect of interest.

Multiple reasons can lead to a Type II error in an RCT. It is difficult to quantitate the probability of a Type II error.

Intuitively, the larger the total sample size included in the RCT and the greater extent to which the total sample population and each treatment arm population matches the target population for all factors that influence demonstration of the effect of interest, the lower the probability of a Type II error (and Type I error). Statistics provide the power to extrapolate from a sample to an entire population. If the entire target population could be studied, a definitive answer would be obtained, and there would be no need for statistics. Of course, the legitimacy of that extrapolation depends on a host of correct assumptions about all RCT components. Additionally, as new persons are added to the target treatment population, the results of that RCT entering what was previously the entire treatment population would become potentially outdated.

The matter of factors that might influence the demonstration of an effect is also relevant to reducing the probability of either Type of error. Unfortunately, there is the potential for occult, unknown factors to be relevant to whether a treatment has the effect under study in specific individuals. Those things that we do not know that we do not know and those things that we think we know but what we know but are incorrect can contribute to 'inaccurate' results for an RCT. These occult factors could vary between treatment groups in a single RCT and could vary among multiple RCTs believe to be identical. Unfortunately, because these factors are occult, such within-study and among-study variability influencing study results and interpretation would not be known.

Areas of Possible Disagreement

First, Dr. Healy is to be congratulated for raising an important, fundamental question: Do RCTs increase useful clinical information, or do they sometimes, or more frequently, provide

inaccurate information that can ultimately degrade patient outcomes? The titles of both the original posted Essay (Healy 2020) and its revised version (Healy 2021) referred to "clinical knowledge," and he addresses his beliefs on the relative merits of RCTs as one potential method of increasing this "clinical knowledge."

My Comment focuses more specifically on my beliefs about the methods that provide the clinician with the best information (defined below) to maximize the probability of improving an individual patient's clinical status to the most significant overall extent possible when administering a treatment (e.g., a medication, a surgical intervention, a psychotherapy, an assistive medical device such as CPAP). The maximal overall improvement is an important concept. The treatment might cause some harm. This harm might be slight in the way of a nuisance effect of the treatment. However, the harm could be significant, including death. The ratio of improvement to harm in the individual patient is the relevant net outcome. The clinician needs the best information to select a treatment (or no treatment) to maximize the probability of a maximally positive outcome.

Again, I might not entirely agree with my interpretation of what Dr. Healy has written in several areas. However, as I said above, my possible disagreements are with my interpretations of Dr. Healy's written words, and my interpretations might not be concordant with his intended meanings when those words were written. Additionally, I agree with much of what Dr. Healy has said about RCTs' confounders and limitations in providing the best information that allows a clinician to maximize the probability of improving a patient's overall or net outcome.

My academic statistical colleague, Roy Tamura, Ph.D., and I (Beasley and Tamura 2019; Beasley 2019) have, I believe, demonstrated that for some critical questions about the safety of a treatment, it is virtually impossible for an RCT to contribute information relevant to the question because the required size of the RCT would make it impossible to conduct the RCT. Information obtained by other methods must be used along with information from RCTs to optimize individual patient outcomes. These alternative methods include individual observations by experts. I would characterize these alternative methods as of lesser quality for providing correct information than RCTs. Their potential deficiency in the correctness of the information they provide must be recognized when using their information to maximize a net positive patient outcome.

My global views on relevant data sources that might maximize the potential for best net patient outcomes are reflected in a publication by the Institute of Medicine (2001) that provides its position on sources of information with a supportive and explanatory quotation as follows:

> "Early definitions of evidence-based medicine or practice emphasized the 'conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients' (Sackett et al., 1996). In response to concerns that this definition failed to recognize the importance of other factors in making clinical decisions, more recent definitions explicitly incorporate clinical expertise and patient values into the decision-making process (Lohr et al., 1998). Contemporary definitions also clarify that 'evidence' is intended to refer not only to randomized controlled trials, the 'gold standard,' but also to other types of systematically acquired information.

For purposes of this report, the following definition of evidence-based practice, adapted from Sackett et al. (2000), is used:

> Evidence-based practice is the integration of best research evidence with clinical expertise and patient values. *Best research evidence* refers to clinically relevant research, often from the basic health and medical sciences, but especially from patient-centered clinical research into the accuracy and precision of diagnostic tests (including the clinical examination); the power of prognostic markers; and the efficacy and safety of therapeutic, rehabilitative, and preventive regimens. *Clinical expertise* means the ability to use clinical skills and past experience to rapidly identify each patient's unique health state and diagnosis, individual risks and benefits of potential interventions, and personal values and expectations. *Patient values* refers to the unique preferences, concerns, and expectations that each patient brings to a clinical encounter and that must be integrated into clinical decisions if they are to serve the patient."

I believe Dr. Healy and I agree that the best information helpful in maximizing patients' net outcome is based on multiple sources of information that includes clinical expertise as defined above (likely analogous to Dr. Healey's "clinical knowledge") and

patient values as defined above. The clinician obtains information about patient values in real-time during the evaluation and treatment determination processes and monitoring the treatment.

Patient values are essential because they must be considered when assessing whether a specific treatment (or no treatment) has optimized an individual patient's outcome. Several factors contribute to this assessment, such as:

- whether the patient achieved a more prolonged survival to death than would have been expected without treatment or with other treatments;
- the number of objectively assessable AEs and subjective complaints (whether the events and complaints are adverse reactions to the treatment) relative to what might be expected with alternative treatments or no treatment; or
- the amount of symptom improvement with treatment, such as the percentage of total skin affected by plaque psoriasis after some treatment period.

However, the patient's subjective satisfaction (or lack of satisfaction) with the treatment for reasons that cannot be objectively ascertained or measured is a critical component in assessing whether a treatment has optimized a patient's outcome relative to an alternative treatment or no treatment. Therefore, these patient values are essential to the assessment.

Dr. Dreyfus (2021) has already addressed my possible disagreement with Dr. Healy if I correctly understand Dr. Healey's primary thesis. My quotation of three paragraphs written by Dr. Dreyfus, beginning in the third paragraph below, expresses this potential disagreement succinctly.

I understand the thesis of Dr. Healy is that RCT results and the interpretation of these results are often inferior to the information obtained from alternative methods, including individual case observation by expert and objective clinicians. Dr. Healy appears to suggest that RCTs and the interpretation of their results can "subtract from clinical knowledge" as per the latter portion of the title of his Essay. He might be restricting this thesis to RCTs conducted by and published in the academic literature by or on behalf of pharmaceutical industry sponsors. However, such a possible distinction between RCTs developed, conducted, and published by or for industry, as opposed to academic researchers, and their

possible impact on "clinical knowledge" is not explicit in his Essay. In the Executive Summary to the revision of his Essay, Dr. Healy states: "Pharmaceutical company use of RCTs gives rise to another set of problems distinct from the ones outlined here." Therefore, I understand his reservations about RCTs' quality of information, enumerated in his Essay, that impact patient care to not be restricted to those conducted and published by pharmaceutical companies.

I do not believe that Dr. Healy completely rejects the utility of the information provided by RCTs. In his Executive Summary, he states:

> "Elements of RCTs, such as randomization, confidence intervals and primary endpoints, can help in treatment evaluation but their indiscriminate combination can cause problems."

As such, my possible disagreement with his thesis is relative and not absolute.

At the risk of presenting important components of Dr. Dreyfus' Comment out of context, I quote three important passages from his Comment (Dreyfus 2021) extracted from paragraphs three, four, and 14 of his Comment:

> "I fully agree that we should not dismiss individual data as second-rate but it is my contention it should be used as the basis for new hypotheses to be tested."
>
> "According to my doxa, using relevant groups and randomizing the group to which a participant belongs is the only way to equalize all the known and unknown factors that could confound the results."
>
> "To conclude, it is my contention that RCTs remain one of the pillars of knowledge, if only because they prevent us from jumping to conclusions that are not warranted by data and building a coherent system out of false premises. Not only can a coherent system be built on false premises but it has been proven that a system that is more coherent is not by essence closer to truth than a less coherent one (Bovens and Hartmann 2003)."

I understand Dr. Dreyfus' reference to "individual data" in the first sentence quoted above to be comparable to Dr. Healy's views on expert clinicians' observations. I agree with

Dr. Dreyfus that such "individual data" should lead to a hypothesis. In this sentence, "hypothesis" is an idea that some treatment does or does not result in some effect; "hypothesis" is not referring to a formal, statistical null hypothesis or its alternative hypothesis, but this conceptual hypothesis is then expressed as the formal, statistical alternative hypothesis. I might disagree with Dr. Dreyfus in a relative sense because not all hypotheses based on "individual data" can be assessed in an RCT for multiple reasons, including the size required of some RCTs. As stated, hopefully, other study methods, more robust than additional "individual data" collected through more expert clinicians' observations, can be employed in the case of such hypotheses. Tamura and I (Beasley and Tamura 2019; Beasley 2019, 2020) have briefly mentioned several such methods, including CDR, formal N-of-1 experimental paradigms, epidemiological studies, and others.

Where Dr. Healy and I likely disagree in relative terms, but the extent of the relative disagreement might be substantial, is in our beliefs in the extent to which RCTs that are appropriately planned (including the statistical planning), conducted-executed, analyzed, interpreted, and objectively reported (i.e., 'proper' RCTs) should be considered the most robust experimental method. I believe that 'proper' RCTs constitute the most robust experimental paradigm, especially when an experiment is intended to support the belief that a treatment causes one or more specific effects or that a treatment does not cause one or more specific effects. I hasten to emphasize that the RCT must be 'proper.' It must be 'properly' designed, executed, interpreted, and presented to an audience. I agree that experts must process descriptive and inferential results cognitively to determine the best interpretation of the RCT's results in its manuscript and/or its regulatory report discussion.

Dr. Healy is substantially more familiar than me with the history of medicine and the developments intended to advance our knowledge, leading to improved patient outcomes. His lengthy Curriculum Vitae in medical history confirms this fact. His Essay (Healy 2021) provides an excellent summary of RCTs' origins in relatively modern medical history. However, the description of the concept of clinical trials as a method to systematically assess the potential utility of a specific treatment has a considerable history that pre-dates the first RCT described by Dr. Healy (Healy 2021). Perhaps the first clinical trial described (Bhatt, 2010) was an open-label, without randomization, with only a quasi-control group (a

convenience sample), and without statistical analysis. But the trial assessed a possible therapeutic intervention and resulted in the application of a new treatment. This trial is described in the *Holy Bible*, Book of Daniel (Versus 11-16) of the Old Testament. Bhatt (2010) suggested that the description of this trial was written about 500 B.C.

> "11. The king's official had put a guard in charge of Daniel and his three friends. So, Daniel said to the guard,
>
> 12. For the next ten days, let us have only vegetables and water at mealtime,
>
> 13. When the ten days are up, compare how we look with the other young men, and decide what to do with us.
>
> 14. The guard agreed to do what Daniel had asked.
>
> 15. Ten days later, Daniel and his friends looked healthier and better than the young men who had been served food from the royal palace.
>
> 16. After this, the guard let them eat vegetables instead of rich food and wine."

Dr. Healy (Healy 2021) attributes substantial credit to Sir Austin Bradford Hill for the design and conduct of what might well be the first multicenter RCT of medical treatment, streptomycin, for treating tuberculosis that resulted in multiple, sequential publications (Medical Research Council [MRC] 1948, 1955; Daniels and Hill 1952; Fox and Sutherland 1955). Dr. Healy states (Healy 2021):

> "A Medical Research Council (MRC) trial of streptomycin in 1947 demonstrated the feasibility of randomization as a control of the subtle biases involved in evaluating a medicine. Tony Hill, the MRC trial lead, got the idea of randomization from a horticultural thought experiment about fertilizers outlined two decades previously in which Ronald Fisher proposed that randomization could control for unknown confounders."

Bhatt (2010) describes other medical intervention clinical trials before the streptomycin trial but agrees with Dr. Healy that the streptomycin trial introduced randomization in the design.

My understanding of the MRC study evolution, based on a review by Crofton (2006), is that the MRC established an advisory committee to design a study or studies. Sir Geoffrey Marshall chaired this committee, Philip D'Arcy Hart served as the committee's secretary, and Sir Hill was a committee member. Sir Hill's contributing expertise was in the field of medical statistics (Hill 1937). Based on the list of authors contributing to what I believe was the first published manuscript (MRC 1948) for the first study, the committee included 15 members. This study (MRC 1948) was conducted in patients with miliary tuberculosis or tuberculosis meningitis. These conditions were considered universally fatal. According to Crofton (2006), Sir Hill was primarily responsible for this patient selection based on ethical reasons. The standard of care treatment at the time was only bed rest. Supplies of streptomycin were limited. Therefore, the most critically infected patients received a new treatment, and the control group received the standard of care as the control treatment. Sequential studies and manuscripts reporting their results followed this initial study with example manuscripts cited above. In reviewing the MRC streptomycin studies, I discuss the possible disagreement with Dr. Healy about his representation of Sir Hill's thoughts.

Dr. Healy (2021) states (entire paragraph included):

> "Before 1962, RCTs were not seen as offering gold standard knowledge about what drugs do. As Tony Hill put it in a 1965 lecture, RCTs have a place in the study of therapeutic efficacy, but they are only one way of doing so and any belief to the contrary is mad (Hill 1965). Hill's lecture ties RCTs to the investigation of one effect and places the information they yield within the framework of clinical judgement."

I now address my possible disagreement with Dr. Healy on his characterization of Sir Hill's beliefs about RCTs. From the text, I would understand that in the citation provided by Dr. Healy, Sir Hill explicitly used the word "mad" to describe the belief that RCTs are the only way of studying therapeutic efficacy. Although Dr. Healy, in his text, provides the citation "(Hill 1965)," there is no corresponding reference in his Reference List. However, there is a reference in the Reference List to "Hill AB. Reflections on the Controlled Trial. *Annals Rheum Disease* 1966; 25:107-113." This reference is Dr. Healy's sixth of 10 references. The reference is to the Heberden Oration of 1965, delivered on November 12,

1965. In reading this manuscript, I do not find Sir Hill stating: ". . . RCTs have a place in the study of therapeutic efficacy, but they are only one way of doing so and any belief to the contrary is mad." Additionally, a search for the word "mad" in the PDF and converted Word copies of the manuscript failed to find the word "mad" (several instances of the word "made" were found).

In reading the text (Hill 1966), what I did find on page 108 is the following paragraph:

> "Any belief that the controlled trial is the only way would mean not that the pendulum had swung too far but that it had come right off its hook. We need not argue, therefore, over the semantics of observation and experiment. What we can more profitably reflect upon is whether the modern controlled trial is a useful adjunct to therapeutics, whether it asks the right question or questions, whether there is any way in to-day's more sophisticated and computerized setting by which it could be appreciably improved?"

My summary interpretation of the lecture's published text is that it offers a balanced presentation of potential ways in which RCTs can support beliefs that, rather than resulting in improved patient outcomes, can degrade patient outcomes if these beliefs impact patient treatment. Sir Hill offers suggestions for improvement in RCTs such that their interpretations can improve patient outcomes. I question some of his suggestions, such as what I believe to be the extent to which he views unblinded RCTs as more valuable than blinded RCTs (although I would agree that unblinded studies have their place in hypothesis generation). Sir Hill indeed emphasizes the need for sound cognitive processing of the descriptive and inferential numerical results of an RCT that lead to its best interpretation.

Dr. Healy may have been present at the 1965 lecture, and Sir Hill might have spoken the words Dr. Healy wrote above. If such is the case, I would not disagree with Dr. Healy. However, if Sir Hill did not speak these words in 1965, I have an actual disagreement with Dr. Healy concerning his characterization of Sir Hill's thoughts about RCTs.

Dr. Healy did not reference or cite what is arguably Sir Hill's most notable contribution to the sciences of the investigation and evaluation of investigational and approved medical treatments, *Principles of Medical Statistics* (Hill 1937), in which he

described his views of RCTs. This work appeared in 12 editions, with the last coauthored by Sir Hill and I.D. Hill. The book could be considered one of the most influential and important books in the evolution of the scientific method's application in assessing medical treatments. The 1961 seventh edition was the last published edition before Sir Hill's Heberden Oration of 1965. The 1967 eighth edition was the first edition published after Sir Hill's Heberden Oration of 1965. It is possible that the 1967 eighth edition was written shortly before the 1965 Heberden Oration. The writing of the 1971 ninth edition would almost certainly have been completed after that lecture. Reference to these three book editions (all in my library) provides what I believe to be a more in-depth and complete view of Sir's Hills beliefs about RCTs' merits in furthering valuable clinical knowledge than Dr. Healy's summary of his understanding of the contents of the Heberden Oration of 1965.

The title of the 1977 tenth edition was changed to *A Short Textbook of Medical Statistics*, but the twelfth edition adopted the original title, *Principles of Medical Statistics* (Armitage 1991).

The disagreement with Dr. Healy's summary of Sir Hill's published text is perhaps not so much a disagreement about fact or belief as it is a disagreement about the optimal words used by Dr. Healy to summarize his understanding of Sir Hill's text. I consider Dr. Healy's use of the phrase containing the word "mad" in the light of Hanlon's Rasor generalization (Bloom 2021). I did not hear Sir Hill's lecture, and all that I had available was the cited publication, also cited by Dr. Healy (I believe cited by Dr. Healy but with the incorrect year in the text of his Essay but the correct year in his list of Reference). There are many viable alternative explanations for Dr. Healy's summary characterization and use of the word "mad" without an assumption of any negative intent.

I quote in full the last two paragraphs of Sir Hill's 1967 eighth edition of his book that were an expansion of his last two paragraphs of his 1961 seventh edition:

> "**Common Sense and Figures**
>
> This interpretation of statistical data turns, it should be seen, not so much on technical methods of analysis but on the application of common sense to figures and on elementary rules of logic. The common errors discussed in

Chapters XXI to XXIII are not due to an absence of knowledge of specialised [*sic*] statistical methods or of mathematical training, but usually to the tendency of workers to accept figures at their face value without considering closely the various factors influencing them - without asking themselves at every turn 'what is at the back of these figures? what factors may be responsible for this value? in what possible ways could these differences have arisen?' That is constantly the crux of the matter. Group A is compared with Group B and a difference in some characteristic is observed. It is known that Group A differed from Group B in one particular way - *e.g.* in treatment. It is therefore concluded too readily that the difference observed is the result of the treatment. To reject that conclusion in the absence of a full discussion of the data is *not* merely an example of armchair criticism or of the unbounded scepticism [*sic*] of the statistician. Where, as in all statistical work, our results may be due to more than one influence, there can be no excuse for ignoring that fact. And it has been said with truth that the more anxious we are to prove that a difference between groups is the result of some particular action we have taken or observed, the more exhaustive should be our search for an alternative and equally reasonable explanation of how that difference has arisen.

"It is also clearly necessary to avoid the reaction to statistics which leads an author to give only the flimsiest statement of his figures on the grounds that they are dull matters to be passed over as rapidly as possible. They may be dull - often the fault lies in the author rather than in his data - but if they are cogent to the thesis that is being argued they must inevitably be discussed fully by the author and considered carefully by the reader. If they are not cogent, then there is no case for producing them at all. In both clinical and preventive medicine, and in much laboratory work, we cannot escape from the conclusion that they are frequently cogent, that many of the problems we wish to solve *are* statistical and that there is no way of dealing with them except by the statistical method."

I understand Sir Hill to say that human cognitive processing of the statistical inferential results of an RCT (or any scientific study method to which inferential statistical analyses have been applied) is necessary where those inferential analyses support causation of an effect by treatment. Feasible alternative explanations to the observed differences between or among the treatment groups other than causation by the treatment must be carefully considered and sought. However, after such careful consideration, the appropriate inference is often that the treatment did cause the effect of interest.

In the 1971 ninth edition of his book, Sir Hill added seven paragraphs to the two paragraphs in the eighth edition quoted above. In these added paragraphs, I believe he emphasizes the need for careful and critical thinking about RCTs' inferential test results and descriptive results. He warns specifically about describing the inferential test results without providing the descriptive data compared by the inferential tests. I could not agree more with this caution. My agreement is particularly the case with the development of modern inferential test methods that adjust for baseline differences and/or missing data values at the individual subject and visit level when these methods also generate descriptive statistics. I say more about this matter in Part 2 of this Comment.

Knowing that Sir Hill continued to evolve and publish his book through twelve editions over 54 years, and reading his two paragraphs above, I find it difficult to believe that Sir Hill was not a staunch proponent of RCTs and other experimental methods based on inferential statistical analyses in 1965 at the time of his Heberden Oration. He, of course, believed that the researchers needed to remain critically skeptical and think about their results before arriving at a final, published interpretation.

In addition to Sir Hill's book, multiple authors/editors have published many volumes providing advice on the process of conceiving, planning, conducting, analyzing, interpreting, and reporting an RCT, as well as other study methods. Several important examples are listed below, along with brief notes. The first four books and the Gutkin (2019) are in my library, along with the Spilker (1991) book. The first focuses specifically on studies assessing psychiatric treatments:

- Prien and Robinson (1994): Donald S. Robinson, M.D., transitioned from an academic career to a distinguished and successful pharmaceutical industry career. He

was associated with both the New York State Psychiatric Institute and Bristol Myers Squibb. Robert F. Prien, Ph.D., was Chief, Clinical Treatment Branch Division of Clinical and Treatment Research of the National Institute of Mental Health. The book grew from a collaborative effort of the American College of Neuropsychopharmacology and the National Institute of Mental Health.

- Hamilton (1961): While Max Hamilton, M.D., was a psychiatrist and probably most recognized for his Hamilton Depression Scale, his book applies to assessing treatments in multiple therapeutic areas.
- Bailey (2008): This book sits at the interface of study design and appropriate descriptive and inferential statistical analyses and presentation. It is applicable across a broad domain of potential causation studies by various interventions, not just medical treatments. This book is one of the two intensively statistical books among those I am listing. I find Chapter 1 understandable, and it explains important concepts for which both the clinical subject matter expert and the expert statistical consultant need to share a mutual understanding in the development of an RCT. Chapter 7 is the chapter most relevant to assessing investigational treatments of human diseases.
- Montgomery (2020): As with Bailey (2008) above, this book deals with the design and appropriate analyses for a range of study designs. It is also intensively statistical. It describes experiments in systems theory terms. Two treatments in an RCT are inputs in a system. Chapter 1, primarily Section 1.4, provides an excellent description of the experimental process.
- Spilker (1984, 1986, 1987, 1990, 1991, 2008) and Spilker and Schoenfelder (1990, 1991): Bert Spilker, Ph.D., M.D., held joint appointments as Professor in the departments of pharmacology and pharmacy at the University of North Carolina Schools of Medicine and Pharmacy and an appointment as Professor at the University of Minnesota School of Pharmacy at the time he authored/edited (one with his colleague) the referenced works. Three of these works were on my desk when I arrived at my office on my first day at Eli Lilly, July 7, 1987. These books included an earlier edition of Spilker 1991). The 1,078 pages of the Spilker 1991 text can be considered the definitive reference work for the design of ‘proper’ RCTs.

- Gutkin (2019): While the book's title emphasizes its focus on manuscript and report writing, it covers aspects of study design and analysis. I have included this book because, as the cover text indicates, one purpose is to provide advice on how to ". . . Disclose More Transparently" in written reports and manuscripts.

I have not read all the works above in my library but have skimmed them in their entirety and read specific sections.

Notably, the authors of these works above are highly esteemed scientists and/or deeply experienced in RCTs. I suggest that the volume of information published on research methods for determining if treatment effects occur and how these methods should proceed from developing a conceptual hypothesis to a report or manuscript underscores the belief that the RCT is a bedrock for progress in clinical medicine. These books aim to instruct on the development of a 'proper' RCT.

I have not listed any works that deal exclusively with inferential statistical analyses of data collected from medical treatment studies and, more specifically, RCTs. Many such books exist, and while I own several, statistics was not a domain upper-level graduate study for me, and the field continues to advance rapidly. Therefore, I am ill-equipped to suggest a good representative set of such references.

I hope that it is clear that I believe that while RCTs have some limitations, a 'proper' RCT is the most robust method for advancing sound clinical knowledge. I believe Sir Hill's writings support this position along with the work of many other scholars.

Should an RCT be considered an experiment? My answer to this question is the second component of macro-level potential disagreement with Dr. Healy because this question is at the heart of the credibility we give to RCTs regarding whether they increase valuable clinical knowledge. The question is not one that I had ever asked myself, and if I had ever been asked the question, I would have immediately replied, "yes." However, after reading the following text (complete paragraph) by Dr. Healy (2021), I considered the question further:

> "In recent years, there has been sophisticated consideration of the statistical techniques employed in epidemiological studies, including RCTs (Greenland, Senn, Rothman et al. 2017), and of the merits of RCTs applied to complex

> situations in the social sciences (Deaton and Cartwright 2019). Both considerations have stressed the role of judgement in deciding what populations and experimental design are appropriate and how results should be interpreted. Both view RCTs, and related designs using statistics, as assay systems yielding results specific to the system, rather than experiments that generate the 'knowledge from nowhere' that means we don't have to worry whether the laws of gravity will apply to the next patient."

Dr. Healy is citing the opinions of two groups of authors. I agree with Dr. Healy's summary interpretation of both groups of authors he cites and my interpretation of Dr. Healy's text about several specific matters. I have often said, somewhat in jest, to my statistician collaborators that besides assisting in designing an RCT and optimizing its descriptive and inferential analyses, their primary job is to keep me and our work honest. That work cannot be over-interpreted or incorrectly interpreted.

Judgment and careful thinking by both experts in the medical disorder for which an investigational treatment is studied and statistical experts with knowledge of specific optimal inferential analyses methods for RCTs is critical in optimizing an RCT. This collaboration designs the RCT and its analyses. A component of the design is the definition of the experimental population such that the RCT's interpretation can be generalized to the clinical population for whom an investigational treatment is targeted. An RCT's results' generalizability is a critical topic that involves both ethical considerations with considerations of the best design to allow for generalization. I briefly address this topic in Part 2 of this Comment.

I do not know if I correctly understand part of Dr. Healy's summary of his interpretation of the two groups of authors that he provides in his last sentence of his paragraph that I quoted above: "Both view RCTs, and related designs using statistics, as assay systems yielding results specific to the system, rather than experiments that generate the 'knowledge from nowhere' that means we don't have to worry whether the laws of gravity will apply to the next patient."

I believe Dr. Healy might be offering caution about ever being able to generalize from an RCT to the population that might be treated with the medication studied. If the

interpretation of an RCT cannot be generalized, can that RCT be considered an experiment? Additionally, he might be suggesting that actual experiments ". . . generate 'knowledge from nowhere.'" Perhaps Dr. Healy suggests that only studies demonstrating constant and consistent causation of effect across all members of a class of objects, as with gravity on two or more objects with mass, are actual experiments.

Indeed, medical treatments that effectively treat a disease with a single and well-defined etiology do not have a constant and consistent effect across all patients with that effectively treated, specific and well-defined disease with a single etiology. This lack of consistency is observed within an RCT, clinical practice, and most broadly across the entire target treatment population. Infectious diseases offer perhaps the best examples of diseases with specific, well-defined etiologies. Pneumonia is a syndrome, while pneumococcal pneumonia is a disease if a disease is distinguished from a syndrome based on a disease having a single and well-defined etiology. The treatment of choice for community-acquired pneumococcal pneumonia in the US as of January 28, 2021, is a parenteral β-lactam antibiotic, most specifically the third-generation cephalosporin ceftriaxone, 1 gm IV every 24 hours (Musher and Tuomanen 2021). Across the US, 98% of pneumococcal isolates respond to this treatment (Musher and Tuomanen 2021), but the proportions of resistant strains are higher in the Pacific Northwest and Southeast regions of the US (Musher and Tuomanen 2021). Therefore, in a large RCT evaluating the efficacy of ceftriaxone in treating pneumococcal pneumonia conducted in the US, it would be reasonably expected that at least 2% of cases would fail to respond to this treatment.

Does such expected and almost always observed variability in any RCT exclude RCTs that study medical treatments from being experiments?

The online Oxford English Dictionary (2021) gives six definitions, with several additional variants, for "experiment." The definition relevant to science is: "An action or operation undertaken in order to discover something unknown, to test a hypothesis, or establish or illustrate some known truth."

Wikipedia (2021) offers an extensive definition of "experiment" and the concept's history in advancing human knowledge. The basic definition that is offered is in paragraphs 1 and 3 of the articles as follows:

> "An experiment is a procedure carried out to support, refute, or validate a hypothesis. Experiments provide insight into cause-and-effect by demonstrating what outcome occurs when a particular factor is manipulated. Experiments vary greatly in goal and scale, but always rely on repeatable procedure and logical analysis of the results. There also exists natural experimental studies *[RCTs are not natural experiments]*."
>
> "Experiments typically include controls, which are designed to minimize the effects of variables other than the single independent variable. This increases the reliability of the results, often through a comparison between control measurements and the other measurements. Scientific controls are a part of the scientific method. Ideally, all variables in an experiment are controlled (accounted for by the control measurements) and none are uncontrolled. In such an experiment, if all controls work as expected, it is possible to conclude that the experiment works as intended, and that results are due to the effect of the tested variables."

I believe that RCTs are intended to be consistent with the definitions above and that a 'proper' RCT meets these definitions. Generalizability of the RCT's interpretation to the broad target population of treatment is a component of my definition of a 'proper' RCT. However, I do not believe that variability in the magnitude of the effect observed in persons treated with the investigational treatment within an RCT that supports a causal effect precludes the RCT from being considered an experiment. As pointed out above, the response of pneumococcal pneumonia to a third-generation cephalosporin, its treatment of choice, is variable.

Such variability can be observed in the measurement of physical forces that can then be explained by mediating influences. Gravity, to which Dr. Healy alludes in his concern about RCTs' generalizability, provides another example of variability in observed physical effects. If gravity's effect (acceleration of an object [e.g., a ball] of small mass moving toward a much more massive object [e.g., the earth]) is measured with extreme accuracy for the same object at distinct positions on the earth's surface, the measured effect is variable, as described below.

> "The variation in apparent gravitational acceleration (g) at different locations on Earth is caused by two things (as you implied). First, the Earth is not a perfect sphere—it's slightly flattened at the poles and bulges out near the equator, so points near the equator are farther from the center of mass. The distance between the centers of mass of two objects affects the gravitational force between them, so the force of gravity on an object is smaller at the equator compared to the poles. This effect alone causes the gravitational acceleration to be about 0.18% less at the equator than at the poles.
>
> "Second, the rotation of the Earth causes an apparent centrifugal force which points away from the axis of rotation, and this force can reduce the apparent gravitational force (although it doesn't actually affect the attraction between two masses). The centrifugal force points directly opposite the gravitational force at the equator, and is zero at the poles. Together, the centrifugal effect and the center of mass distance reduce g by about 0.53% at the equator compared to the poles." (Department of Physics, University of Illinois at Urbana-Champaign 2021)

Also, the measurements of that acceleration at the top of Mount Everest would be different from the measurement made at a point on the earth's surface below sea level. I agree that some RCTs fail at being 'proper' RCTs. In such cases, they are not good experiments. However, variance observed in the effect within an RCT potentially raises questions that further experiments might address. In my opinion, such observed variance does not exclude an RCT from being an experiment.

Each reader will have their own opinion about RCTs as a general class of human activity and specific RCTs concerning what they add or subtract from generalizable and valuable clinical knowledge.

The following is an examination of my potential disagreement regarding the robustness of a CDR paradigm study of an AE. This discussion of the CDR paradigm is mixed with the second and last example of a possible micro-level disagreement with Dr. Healy. This possible disagreement illustrates the importance of subtle variations in oral and written linguistic productions, possibly leading to misunderstandings between the person

producing the language (spoken or written) and the person interpreting the language. Dr. Healy (2021) wrote the following two contiguous paragraphs but without a citation.

> "By 1959, clinicians praising imipramine's benefits also noted it could cause agitation and suicidality in some patients that cleared when the drug was stopped and reappeared when restarted. This Challenge-Dechallenge-Rechallenge (CDR) evidence, especially as it was replicated by several clinicians with different patients, offers close to Fisherian [*sic*] expert like certainty that imipramine causes suicide in certain individuals.
>
> "Despite being able to cause suicide, in an RCT of melancholic patients, imipramine seems likely to protect against suicide on average by reducing the risk from melancholia to a greater extent than placebo. In contrast, in the RCTs that brought SSRIs to the market, these drugs doubled the rate of suicidal acts. This was because, weaker than imipramine, SSRIs had to be tested in people with mild depression at little risk of suicide. The low placebo suicidal act rate revealed the risk from the SSRI – as it does for imipramine when put into trials of mild depression. RCTs can, in other words, mislead as regards cause and effect – potentially getting results all the way along a spectrum from 'causes,' to possible risk, likely protective and 'cannot cause.'"

In the first sentence above, he notes that ". . . clinicians noted that imipramine could cause agitation and suicide." As I read and interpret the sentence, I believe Dr. Healy suggests that these two alterations in behavior and thinking are independent and that he is not suggesting that agitation is necessary before observing suicidality. This interpretation on my part is reinforced by the contents of the last sentence in the first paragraph where he suggests that ". . . offers close to Fisherian [*sic*] expert like certainty that imipramine causes suicide in certain individuals." Also, Dr. Healy has shifted from the term "suicidality" (as I understand the term, a continuum of ideation and behaviors, with the most clinically severe a completed suicide) in the first sentence above to "suicide" in the last sentence of the first paragraph above as caused by imipramine. The first sentence of the second paragraph also reinforces my interpretation that Dr. Healy states that by 1959 there was some degree of expert clinical belief that imipramine directly caused an increase in suicidality and suicide without intervening or explanatory/mediating mental status or

psychomotor status changes. Readers must carefully judge whether my interpretation of the words written by Dr. Healy is likely concordant with the meaning Dr. Healy intended to convey. Indeed, my interpretation of his meaning could be other than what he intended to convey.

I present below a detailed history of early descriptions of the relationship between imipramine and suicidality. First, I offer my understanding of the first report of an antidepressant explicitly causing suicidality without first causing an adverse increase in psychomotor activity status (a phenomenon described with a term such as agitation or anxiety or resolution of psychomotor retardation) (Damluji and Ferguson 1988). These authors reported a case series suggesting a paradoxical, *de novo* appearance of suicidality with worsening of depression but without an excessive increase in psychomotor activation in at least two of the four reported patients (one of the four patients did experience agitation and one experienced insomnia) during antidepressant treatment with desipramine. Patients 1, 2, and 3 were described as melancholic.

Patient 1 experienced worsening depression and new-onset suicidal ideation. Agitation accompanied these symptoms. Amoxapine was the first medication used after desipramine was discontinued. The response to amoxapine was like the response to desipramine. After amoxapine was discontinued, the patient showed partial improvement. A course of fluoxetine was then initiated that was associated with complete remission of the depression.

Patient 2 experienced worsened depression, new suicidal ideation, and decreased energy with no mention of agitation, anxiety, or other terms reflecting excess psychomotor activation. Suicidal ideation resolved within 24 hours of discontinuing desipramine. A course of amoxapine resulted in similar symptomatology. The patient had an excellent response to tranylcypromine.

Patient 3 experienced worsening of depression and new-onset suicidal ideation and insomnia that can be considered an adverse increase in psychomotor activity status. Her symptoms quickly improved partially on discontinuation of desipramine. Trazadone treatment was initiated. Following some initial improvement, the clinical status deteriorated, and there was a suicide attempt by a trazodone overdose. The patient finally responded well to ECT.

Patient 4 experienced worsening of depression and new-onset suicidal ideation after beginning desipramine. After the patient discontinued the desipramine, the depression returned to

pre-desipramine levels. A trial of nortriptyline resulted in a similar profile of symptom worsening. Fluoxetine led to remission of the depression.

Damluji and Ferguson (1988) cite other authors' work describing a case of a "paradoxical response" to amitriptyline in patients with Borderline Personality Disorder and an emergent dysphoric and depressed mood in normal volunteers treated with imipramine. The authors cited this adverse mood change as occurring during the first two weeks of imipramine administration and resolving by the third week of administration.

After two preliminary paragraphs immediately below, I discuss three aspects of the last two paragraphs written by Dr. Healy (2021) quoted above. These three aspects of Dr. Healy's two paragraphs are discussed in the following order:

- the extent to which the CDR methodology can offer definitive support for causation of an adverse reaction by treatment in his first paragraph above;
- how 'proper' RCTs can 'accurately' assess the potential paradoxical causation or worsening of a sign or symptom of a disorder by treatment when that treatment effectively treats a disorder that includes the sign or symptom that has paradoxically worsened and effectively treats that sign or symptom in most patients; and
- precisely what was said in 1957-1959 about imipramine and suicidality as briefly summarized by Dr. Healy in the first sentence of his first paragraph above.

I am not reviewing my beliefs and the data influencing those beliefs about whether one antidepressant, a mechanistic class of antidepressants (e.g., selective serotonin reuptake inhibitors, selective norepinephrine reuptake inhibitors), or antidepressants in general do or do not induce suicidality without induction of the mediating phenomenon of increasing psychomotor activity. To adequately discuss my beliefs about this matter and the reasons for those beliefs, the length of the discussion would constitute a small book. More importantly, I would need to disclose information that I obtained during an academic manuscript review for a prominent US psychiatric journal in 1994. As such reviews are confidential, my disclosure of the information would be a significant breach of my ethical obligation to maintain confidentiality. For the same reasons, I do not discuss any data reported to support the conceptual hypothesis that any class of antidepressants increases suicidality risk.

What I can say is that I agree with Dr. Healy that the study of the conceptual hypothesis that treatment A causes event B in a small subset of patients when treatment A is an effective treatment for a disorder where B is one sign or symptom of the disorder and B improves along with other signs and symptoms in a substantial proportion of patients is exceedingly difficult. Studies of some conceptual hypotheses in this class are probably impossible. However, once this conceptual hypothesis has been put forward, RCTs and their analyses must address the hypothesis if possible.

I cannot entirely agree with Dr. Healy about the assumed 'accuracy' of the CDR method in establishing the causation of an ADR. In a Response to Edward Shorter (2019), I described a patient who had been assigned to placebo in a study with a 3-to-1 assignment to an experimental medication (not including an active control that was another treatment arm in the study) who experienced marked neutropenia that was ultimately diagnosed as a rare genetic disorder, cyclic neutropenia (Beasley 2020). This neutropenia was clearly not pharmacologically related to the investigational drug because the patient was assigned to a placebo. The cyclic periodicity of cyclic neutropenia is such that neutropenia would have quite likely recurred on rechallenge with the CDR method, and the CDR method would have 'inaccurately' supported causation of the neutropenia by treatment. If the subject were assigned to the experimental drug and the CDR method had been applied with recurrence on rechallenge, an 'inaccurate' attribution of causation would have occurred.

N-of-1 studies build on the CDR paradigm concept, adding the necessary elements of the blinded method, a control group, and appropriate statistical analyses. I mention references to relevant literature in my Response to Professor Shorter (Beasley 2020). An N-of-1 RCT can offer robust support for the causation or lack of causation of a phenomenon for an individual subject. However, it is more difficult to generalize from an N-of-1 RCT to a broad population of interest, either concerning the possible causation of an AE or demonstration of an efficacy effect.

An extension of an N-of-1 design is a parallel study design that uses an enriched population with both rechallenge and parallel control. I explain and outline an example of such a study focusing on assessing whether a specific mechanistic class of antidepressants (e.g., selective norepinephrine reuptake inhibitors) cause paradoxical worsening of depression, specifically, suicidality, without induction of excessive psychomotor activation. Patients who ostensibly experienced the phenomenon based on their treating clinicians' global impression while treated

with a selective norepinephrine reuptake inhibitor would be recruited from a large, diverse set of practices.

Informed consent would be thorough, and the ethical considerations involved in CDR, N-of-1, and enriched rechallenge with parallel control studies are extraordinarily complex. In all three of these designs, a patient is possibly at risk of experiencing adversity that could range from dysphoric to fatal without appropriate safeguards.

These patients would be assigned by random allocation to double-blind treatment with the suspect class of antidepressant or an alternative class of antidepressant (such as an SSRI). If the suspect treatment were all antidepressants, this design study would be virtually impossible to conduct. One would need an effective antidepressant but of a different mechanism than the suspect treatment. ECT would be the only option for the control treatment, except perhaps for esketamine. Many eligible patients would likely not consent to the possibility of receiving ECT. It might be possible to blind ECT by administering the rapid-acting general anesthesia to all patients, but the degree of complexities should be apparent. The blinding of esketamine treatment would be extremely difficult, although infusion of a low dose of rapid-acting general anesthesia might suffice.

To maintain blinding, if the suspect treatment produced a distinctive pattern of ADRs, it would be necessary to administer an exclusively peripherally acting agent that would produce the experience of those ADRs of the suspect treatment.

The treatment environment would be comparable to the environment in which the ostensible qualifying event occurred. This treatment place becomes another complex matter because it should afford patient protection while replicating the index occurrence environment to the greatest extent ethically possible. As most cases would likely have occurred in an out-patient setting, replicating the initial event environment while adequately protecting patient safety might be challenging.

The comparative treatment period would extend to the most extended period between initiation of treatment and the ostensible emergence or worsening of the AE of interest.

Sensitive and specific measurement instruments for any specific characteristics of the exacerbation of depression and/or suicidality thought to be caused by the drug class being studied

would be needed. Assume it was conceptually hypothesized that the norepinephrine reuptake inhibitor caused a severe, quantum increase in suicidal ideation that was dysphoric and ego-dystonic with no intent to act and was obsessional. Then a measurement instrument would be needed that was sensitive and specific for small to moderate changes in the global severity of suicidal ideation and these specific characteristics. Additionally, with one component of the hypothesis being that suicidality emerged without any increase in psychomotor activity status, a validated and reliable measurement instrument of sufficient sensitivity would be required to assess this status. The instrument's sensitivity assessing suicidality severity would protect patient safety and reduce the probability of Type I and II errors. Specificity for the suicidality assessment would be necessary because of the need to minimize the potential for false-positive identification of the hypothesized phenomenon that could contribute to either a Type I or II error.

Quade and colleagues (1980) noted that for events with an incidence <50%, higher specificity (reduced false positives) has a greater beneficial effect on the accuracy of incidence estimation and power to detect true differences between groups than higher sensitivity (reduced false negatives). This need for specificity that results in identifying fewer total cases (importantly less false-positive cases in comparative groups that can dilute a true difference) of an event or is often not easily recognized, even by experienced clinicians working in pharmacovigilance.

In addition to sensitivity and specificity, the instrument would need to be valid and reliable from a psychometric perspective.

During my professional career in pharmaceutical medicine that spans my time as an Eli Lilly and Company employee and my time as an independent consultant to multiple pharmaceutical and biotechnology companies, I began collaborating with other consultants to design one enriched population rechallenge and parallel control study. The study's purpose was to evaluate whether a drug that effectively treated disorder X and worsened disorder X in a small subset of patients treated with the drug. As the question the study was intended to address was adequately addressed through an alternative process, this study was not conducted.

I now return from the discussion about the relative merits of a CDR paradigm and alternative designs to Dr. Healy's summary history of the idea of a relationship between imipramine treatment and suicidality in patients with depression. What follows is a detailed presentation of material from published academic manuscripts of 1957 through 1959 and one book that reports and

recounts the material presented at a 1959 Cambridge symposium devoted to treating depression to allow the reader to carefully consider Dr. Healy's statements in the last two paragraphs that I fully quoted earlier regarding imipramine and suicidality.

Dr. Healy provided no citations for his text in these two paragraphs. I performed literature searches using both PubMed and Google Scholar, restricted initially to the year 1959. The text string for both searches was: ("imipramine" AND "suicide"). These searches resulted in 11 published manuscripts relevant to a possible relationship between imipramine and suicidality. As I knew that Ronald Kuhn, the developer of imipramine, had published a description of its clinical effects in 1957, I performed an additional literature search for 1957-1958 using Google Scholar with the text string ("imipramine"). This search added two manuscripts of relevance to the potential basis of Dr. Healy's statement of the relationship between some instances of suicidality and imipramine. These manuscripts are summarized below. I cannot warrant that my searches found all relevant manuscripts, as additional, proprietary search engines exist.

Kuhn (1957) offered a hypothetical caution regarding the possibility for an effective antidepressant drug, specifically imipramine, to indirectly contribute to a range of inappropriate ideations and behaviors by induction of an alteration psychomotor state, more specifically, conversion to mania. He said:

> "In this connection, a question of great general importance presents itself, namely, whether, and to what extent, imipramine hydrochloride influences healthy impulses of conscience, consciousness of guilt and resistance against criminal or immoral actions. The seriousness of the moral and social implications involved in this question cannot be ignored. It is essential that investigation of this matter should be undertaken on a wide scale. For the time being perhaps the following can be stated: Undoubtedly the possibility exists of influencing people's ethical and moral behavior by administration of particular substances. The best known example is alcohol. It is therefore certainly within the realm of possibility that newly discovered medicaments may exert a similar effect. Furthermore, it should be remembered that in certain individuals suffering from depression, and in whom manic phases occur, their whole moral structure may altogether deviate from the normal.

> "This does not mean that such factors belong to the characteristics of the psychosis. But it is known that manic states particularly may give rise to criminal actions and to the absence of inhibition or moral scruples. If then a medicament such as imipramine hydrochloride exerts an effect on the mood and possibly provokes manic-like reactions, then it must be expected that in certain persons their moral structure may be imperiled. The inhibition against committing suicide must also be seen in this light; just as in the spontaneous course of a depression, phases occur in which resistance against suicide is lessened, during the course of imipramine hydrochloride treatment there may be an increased risk of suicide. It is essential to take this into account, and in spite of the possibilities which imipramine hydrochloride offers, to commit to an institution patients who are endangered in this way.
>
> "In the course of the treatment we have carried out to date, we have not seen any particularly striking signs of interference with ethical or moral standards. In one case we treated a patient with depression who was about to appear on trial in court for homosexual offences against a youth. The depression practically disappeared. The self-reproaches, hysterical collapse with crying and moaning disappeared completely within 3 days. But there was absolutely no evidence that the patient's own moral condemnation of his actions had suffered. The depressive phase gave place to a perfectly adequate sense of his own moral failure and to a natural reaction of repentance. The homosexual desires became strikingly less prominent during treatment."

Kuhn focuses on switching from depression to mania as a mediating phenomenon that could facilitate a range of inappropriate ideations and behaviors, including increases in suicidality. He noted that such increases in suicidality occur in the natural course of a depressive illness, probably referring specifically to a depressive episode in the longitudinal course of bipolar disorder. Although it discusses "ethical or moral standards," the second paragraph quoted above implies that Kuhn had not observed any inappropriate change in ideations during treatment with imipramine, even with improvement in depressive symptoms. The adult patient treated with imipramine, who was about to go on criminal trial for sexual activity with a minor, demonstrated

improvement in his depressive symptoms. However, he maintained his sense of "moral failure" regarding his inappropriate behavior with a minor. While this example does not appear to involve a manic switch, but only the resolution of depression, and illustrates the absence of a change in ideation rather than the absence of suicidality, it suggests the absence of the potential phenomenon in the "over 500 psychiatric patients" treated with imipramine about whom Kuhn was reporting.

In the decades that have followed, Kuhn's hypothetical adverse reaction has morphed into several variants. The concern with mania has expanded to include any state of excessive psychomotor activity, particularly those that are subjectively dysphoric to the patient. Examples include such a state described by many terms such as agitation, anxiety, nervousness, hyperactivity, and others. Akathisia, which is very subjectively distressing, can be included as a manifestation of excess psychomotor activity. Additionally, the concept has come to include improvement/resolution of psychomotor retardation before improvement in the other subjectively painful components of depression.

Lehmann, Chan, and de Verteuil (1958) described the treatment of 84 patients with imipramine. Increased agitation occurred in three, and a shift to "hypomanic excitement" occurred in two. The only mention of suicidality is as follows: "It should be noted, however, that the effects of the drug are much less spectacular than the therapeutic action of electroconvulsive treatment as regards both immediacy and intensity of its results. A deeply depressed patient who is suicidal might still require electroconvulsive therapy in order to control the situation rapidly, particularly if the patient is not hospitalized."

Freyhan (1959) described the treatment of 58 patients with a depressive episode, including patients with bipolar disorder. One patient ". . . developed extreme agitation with hallucinations necessitating immediate termination [of imipramine treatment]." No patient was described as experiencing an increase in suicidality.

Azima (1959) and Azima and Vispo (1959) reported on an expanding group of patients (N=145 in the second publication [Azima and Vispo 1959]) experiencing various depressive episodes treated with imipramine. Five cases likely experienced an increase in agitation, and two, a manic switch (Azima and Vispo 1959). All received chlorpromazine. There were no cases described as experiencing an increase in suicidality. In Azima (1959), 20% of the cohort of 100

patients were reported to experience "tremor and agitation" as "side effects," but all "side effects" were characterized as ". . . of low intensity and did not require the cessation of medication."

Lancaster and Foster (1959) describe a 29 y/o female with a recurrent depressive episode previously treated with electroconvulsive therapy (ECT), taking imipramine 25 mg three times daily for 16 days. In the evening of Day 16, she took 1,500 mg of imipramine with suicidal intent. She slept through the night and the following day was agitated, experienced gross involuntary movements, and a possible seizure. She was found by a relative who determined she had taken an overdose and was taken to a hospital. Although she experienced multiple AEs consistent with those expected with a large overdose of imipramine, she survived with the resolution of the adverse effects after three days. Subsequently, she received another course of ECT. The authors did not suggest that imipramine contributed to or caused her suicide attempt with imipramine. Notably, she survived an overdose with imipramine of 1,500 mg, a dose that might have been reasonably expected to result in death.

Foster and Lancaster (1959), in a separate manuscript, also describe a series of AEs of a motor function nature. These AEs included, for example, multiple falls with injuries, gross tremor, dysarthria, and ataxia. No of these events were described with terms suggestive of an increase in psychomotor activity.

Ball and Kiloh (1959) reviewed nine earlier reports (1957-1959) of various trials with imipramine and their trial in out-patients with depression. Cases of both "endogenous" (N=55) and "reactive" (N= 42) depression were included in their out-patient trial. Patients were excluded because the study was of outpatients if: 1) the patients were "showing gross retardation or extreme agitation"; and 2) "those in whom the risk of suicide was considerable." Within the two diagnoses, patients were assigned on a 1-to-1 basis to either imipramine or placebo. One patient receiving a placebo who refused hospitalization committed suicide. His diagnosis was not specified, and the means of suicide was not described. The authors offer the following advice: "Nevertheless, adequate observation and supervision is necessary because of the toxic effects of these drugs as well as the continuing risk of suicide in some cases and the possibility that other forms of treatment may be required." This advice is consistent with the concept that the risk of suicide for a patient experiencing a depressive episode continues to be elevated without substantial improvement in all symptoms and signs of depression. There is no further mention of suicide or suicidality.

Bram (1959) described his clinical practice experience treating with imipramine 65 cases of "endogenous depression," most of them within the category of "agitated depression." He stated: "At first all the cases treated were started on imipramine, but it was soon evident that they could not tolerate imipramine on account of increased anxiety, restlessness, or agitation." He was clearly suggesting that the imipramine was inducing excessive psychomotor activation. He adopted the practice of co-treatment of patients with agitated depression with imipramine and either chlorpromazine or trifluoperazine with the antipsychotic initiated "a few days" before initiation of imipramine. He summarized his clinical results as follows: "As to the indications for imipramine, I fully agree with Drs. Ball and Kiloh [1959 – summarized in the paragraph immediately above] that by using this drug, and, in my series, by using it in combination with trifluoperazine or chlorpromazine, the patients could continue with their work or could return to their work much sooner than with the previous forms of treatment." He goes on to caution: "Imipramine does not replace electroconvulsive therapy (E.C.T.) because, as was agreed at the symposium in Cambridge [this symposium and its presentations are described below], in the more serious forms of depression there is no response to imipramine. The risk of suicide being very great, it would be dangerous to start treatment with imipramine alone without prior E.C.T." The only reference to suicidality or suicide is in the sentence just quoted. While Bram believed that imipramine could exacerbate agitation in patients with agitated depression and that agitated depression is associated with a particularly increased risk of suicide and that ECT is the first treatment of choice for such patients, he did not suggest that imipramine directly induces suicidality without increasing psychomotor activation.

Mann, Catterson, and Macpherson (1959) reported a 21 y/o female treated as an in-patient for depression with psychotherapy and unspecified "tranquillizing agents, but no imipramine. Before her hospitalization, she stole a bottle of 100 25 mg imipramine tablets from a physician's office and had the tablets with her in the hospital. One evening, at about 7:45 PM, she ingested six tablets and about one hour later ingested all the remaining tablets (total ingestion of 2,500 mg). A "few moments later," she informed the nursing staff of her actions. She was closely observed without treatment until she experienced a grand-mal seizure at 11:10 PM. The seizure resolved after 10 seconds. She was then given sodium amytal 500 mg by IM injection and then catheterized. About 200 ccs of urine were obtained. The day following the overdose, she received one grain of sodium amytal three times (route of administration was not specified). She was described as

recovered from all adverse effects 36 hours after the ingestion. As the patient was not treated with imipramine, there could be no suggestion that this patient's suicidal behavior was induced by imipramine. The subsequent treatment of her depression was not described. The authors describe five more cases of medically serious AEs of a non-psychiatric nature observed in temporal association with the administration of imipramine. As with the survival of a patient ingesting 1,500 mg of imipramine, the survival is notable with no attempt to limit the absorption of the imipramine.

Garrett (1959) summarizes 21 cases treated with imipramine for depression on an in-patient basis. One patient with bipolar disorder was described as experiencing the AE of "highly agitated" and "became actively suicidal" that resulted in treatment discontinuation. Garrett opined that the suicidality ". . . was probably due to the fact that Tofranil cleared the psychomotor retardation." Garrett links the increase in suicidality with a partial or complete reversal of excessive psychomotor retardation, or as in this case, a shift to a state of excessive psychomotor activation, while still under the influence of depressive affect and other components of depression (implicit in this case) such as guilt. This idea that any form of potentially successful treatment of a depressive episode, including ECT, where psychomotor retardation can improve before other components of depression create risk for increased suicidality, including suicide attempts, recurs in additional literature of 1959 and onward in time. Two patients described as having "reactive depression" experienced the AE of "restlessness."

Kline (1959) proposed an extensive nomenclature for chemical entities and plant extracts with psychoactive effects, described the on-target effects (desired/therapeutic) and the off-target effects (undesired, adverse reactions). He proposed the term "ataraxic" for the antipsychotics available at the time, including reserpine. He proposed the term "psycho-stimulant" for imipramine (the only drug in this group) and the term "psychic energizer" for the MAOI antidepressants based on his belief that imipramine and the MAOIs resulted in somewhat different changes in subjective and observable psychobehavioral phenomena. He described suicidality as an AE but only when using an ataraxic agent in the treatment of depression:

> "*Suicide.* In our own very extensive series of several thousand patients treated by ataraxics, we have not found substantial evidence that the rate of suicide is appreciably increased. Care should be taken, as in the treatment of any

> depressed patient, and special caution should be observed at the period of transition, i.e. when the previously "immobilized" patient begins showing signs of increased activity but is not yet completely relieved of the depression. He has then sufficient energy to put suicidal thoughts into effect. A number of recent papers have stressed the fact that the overwhelming majority of suicidal patients either volunteer information as to their intention, or readily admit it if questioned. The patient who is responding to treatment but who goes through a phase of threatening suicide should be taken seriously and watched with extreme care, and there should be no hesitation in using hospitalization if it appears to be the only way protection can be given through this phase."

Kline espouses the same concern about the risks in an early improvement from a state of psychomotor retardation without improvement in other components of depression when using an antidepressant for treatment. He does not extend this concern to imipramine in this manuscript published in 1959.

Sloane, Habib, and Batt (1959) reported using imipramine to treat depression in an open, general medical ward. Eighteen patients were treated with open-label imipramine, and 12 patients were assigned in a 1-to-1 ratio to placebo or imipramine. The total sample of 30 patients included 14 with bipolar disorder. The depressive episode presentations ranged from "melancholic stupor" to "agitated depression." The depression's severity ". . . would have justified the use of electro-shocks before the introduction of imipramine" in all 30 cases. The dose of imipramine ranged from 200 mg/day (initial and most common continued dose) to 600 mg/day, generally administered in four divided doses. The only references (four separate references) to suicidality are as follows:

> "In three patients severe restless agitation and suicidal impulses were controlled by slow intravenous injection of 50-100 mg. dissolved in 20 c.c. of sterile water. Notable was the relatively slow onset of the calming and euphorizing [*sic*] effect of the drug over the 10-30 minutes after such injections. In three patients the imipramine was combined with a phenothiazine compound - two received chlorpromazine 300 mg. and 200 mg. daily for 10 and 24 days respectively, with good effect, and the third TP 21 (Mellaril, made by Sandoz) 600 mg. daily with similar improvement.

> "After the first eight patients had been given the active drug to familiarize us with its action, a placebo was introduced. Patients under the care of other physicians were excluded for this reason, but only one patient was excluded because of the extreme suicidal severity of his illness.
>
> "One patient, a severely retarded and suicidally depressed man, had two ECT treatments at the beginning of his illness and three more in the seventh and eighth week. From then onwards he made an uninterrupted recovery. Before this treatment, however, his rating scales had in fact shown a progressive improvement to the level of 50% reduction, with, in the last week before the second series of ECT was given, a slight worsening, but still a 20% improvement.
>
> "Thus, we know that depressive illness even at its greatest severity is not a killer, provided that nutrition is maintained and suicide guarded against."

This 1959 manuscript did not suggest that imipramine exacerbated or induced suicidality and reported three cases where IV imipramine led to rapid improvement in agitation and suicidal "impulses." However, these three patients might have been the three patients that also received a phenothiazine. Unfortunately, the manuscript does not clarify this point.

In addition to the 13 manuscripts reviewed above, the symposium mentioned by Bram (1959) is relevant to this topic. As I collect books of historical interest to early psychopharmacology development, I have a copy of the book containing the text from this symposium (Davis 1964). I quote in full all relevant sections from the book below. I cannot confirm that the text published in 1964 was identical to the spoken lectures and discussions delivered in 1959.

The topic index of the book contained the following entries and pages:

- Imipramine and suicide: 82, 230, 255, 295, 320
- Imipramine and anxiety: 58, 82
- Suicide after imipramine: 82, 230, 255, 295, 320
- Agitated depression and imipramine: 84, 256

There are eight unique page references for these topics. Neither "agitation and imipramine" nor "imipramine and agitation" were listed in the index as topics.

Below, I quote the complete passages in sequential pages: 58, 82, 84, 230, 255, 256, 295, 320.

> Page 58: Roth (begins on page 57 – partial paragraph): "All of us have seen cases of endogenous depression which appeared to be reactive to external stresses. But the appearance of reactivity surely proves to be spurious in the light of everyday clinical observation. Let us take the example of an old man who has had a urinary tract infection after prostatectomy, or one who has had a recent myocardial infarction and who develops a depressive illness with classical endogenous features. Most of us have had the experience in such cases of seeing the depressive symptoms subside rapidly after physical treatment, such as electroplexy or imipramine, despite some continuing disability owing to angina of effort or residual kidney damage. Many intelligent patients in this situation remain aware of the fact their physical health will now be permanently impaired, but the hopelessness and despair this knowledge formerly evoked have been banished by a simple physical procedure."

The word "anxiety" is not on page 58. The text that begins on page 57 and extends to page 59 discusses endogenous and reactive depression.

> Page 82: D. H. Clark: "To go back to chemotherapy versus electroplexy, I am wondering whether there is greater risk of losing patients on chemotherapy from suicide."

> Page 82: Garai: "I think one of the problems of imipramine treatment is the intense anxiety which patients may sometimes show afterwards."

> Page 82: D. H. Clark: "There is the danger of relief of retardation before the relief of depression which we used to know in the old days before electroplexy."

> Page 82: Kristiansen: "There is a point in connexion [*sic*] with these anxiety states. I think that during treatment with imipramine there are two different situations of importance. First there are depressive states combined with open anxiety, and next there is something quite different. During the treatment I have seen now and then delirious periods of three to four days' duration, where patients complained of dreams of aggressive content, dealing with relatives or other persons important to them. This state is accompanied by great anxiety, and sometimes also by suicidal impulses. The symptoms disappear when imipramine is withdrawn or dosage is reduced, and often the depression disappears simultaneously."

> Page 82: Garai: "If one is treating the more severely depressed patients with imipramine one does not feel very happy about leaving them as out-patients. The suicide risk is still an unknown quantity."

> Pages 82-83: Russell Barton: "I think imipramine may produce anxiety. I have had good results with it, but undoubtedly some patients have become more anxious or agitated – it is difficult to know which – and they do not make a good response."

This material on pages 82-83 reiterates the concern over psychomotor retardation relieved before other components of depression and the development of anxiety and agitation with imipramine as facilitating suicidality.

> Page 84: Jensen: "With regard to drug treatment, it has been suggested that imipramine might be better in agitated depression than in retarded depression. This may well be because of its chemical relationship with chlorpromazine; in fact in using imipramine I have noticed that agitation may increase some days before the appearance of the first beneficial effects. This may appear with somatic malaise or a sense of oppression."

Jensen reports that he has observed an increase in agitation, and it is unclear whether this includes the *de novo* appearance of agitation in patients who would have been characterized as non-agitated before imipramine treatment.

Page 230: Deniker and Lemperier:

“CHEMOTHERAPY IN THE TREATMENT OF DEPRESSION

The action of imipramine, and to a lesser extent of iproniazid, is not merely sedative and symptomatic, like that of the neuroleptics, but is curative, which undoubtedly contributes something new to psychiatric therapeutics. It is difficult at present to define the part they should play in the treatment of melancholia in relation to E.C.T. which is still sometimes indicated. Severe melancholic stupors, or those states with great anxiety and agitation, or in which active ideas of suicide are present, present a therapeutic emergency which often demands E.C.T. Electroplexy in combination with chemotherapy is often advisable as drugs appear to hasten and enhance its effects.

“In melancholia of moderate intensity, and in simple depressions, it is possible to dispense with E.C.T. Doctors have often been confronted with opposition by patients to treatment of which they have unpleasant memories, or with the refusal of their families, who feel that such ‘brutal’ treatment is unjustified in an apparently harmless condition, even with the imminent risk of suicide. Chemotherapy is readily acceptable by contrast, which enables patients to be treated quickly. It is perhaps in protracted involutional melancholia that imipramine gives results really superior to those of E.C.T. Up to now patients have had repeated treatments by E.C.T., which has resulted in distressing confusion and persistent failure of memory. It is now possible to provide maintenance treatment with imipramine for months or even years, and at the same time make possible rapid social and professional rehabilitation. We have always kept the strict rule in the treatment of melancholia of sending the patient to hospital whenever there is risk of suicide. This danger, inherent in melancholia as such, may be further increased by treatment, for some patients are temporarily protected from suicide by their inhibition. In such cases improvement with imipramine would be dangerous in the same ratio as the loss of will-power confronted by the basic pessimism of the disease. Chemotherapy might then, as with E.C.T., enable the patient to carry out his wish to end his

> life. In some cases which are refractory to chemotherapy the increase in anxiety and insomnia might bring to the surface an imperative suicidal urge. When we consider how difficult it is to assess the risk of suicide, and we realize how frequent are the side-effects of drug treatment, we feel that hospitalization in such psychiatric conditions is completely justified during the initial treatment of melancholia by imipramine."

Deniker and Lemperier appear to offer a hypothetical caution, as did Kuhn. Also, they include ECT within this hypothetical warning regarding biological treatments that can result in early relief of an inability to act preceding improvements in other components of depression that drive and facilitate suicide.

> Page 255: Fogarty (speaking of clinical experience with 33 outpatients treated with imipramine): "One suicidal attempt was made by a patient who was improving; this was a woman of about 50 with an agitated depression, not particularly retarded, who suddenly felt extreme despair, and what she called 'a dreadful quiet feeling'." [*No further history provided for this patient.*]
>
> Page 256: Tauber (speaking of treatment with over 400 cases treated with imipramine or iproniazid in Switzerland): "Intuitively we have generally used imipramine for agitated depressives and iproniazid where apathy and retardation predominate."
>
> Pages 295-296: Cazzullo, De Martis, and Terranova: "We agree with previous writers about two therapeutic risks: first that of suicide during the transition from the acute phase to that of stabilization. This situation is equally dangerous in treatment by electroplexy. Second, that of the possible production of manic manifestations, undoubtedly alarming in cases of manic-depressive psychoses, in which the manic phases are usually severe, long-lasting and not easily controlled by therapy. At this moment, to our knowledge, studies on this subject are still incomplete."
>
> Page 320: Davies: "We have learnt that electroplexy is apt to relieve retardation more than depression in the early stages of treatment, in

> consequence of which some patients show a paradoxical suicidal urge after treatment has begun. It looks as though imipramine may have a similar effect."

The bullets below are a low-level summary of my interpretation of the points made in the written text of the Cambridge symposium (Davies 1964):

- Imipramine treatment has been observed to be temporally associated with increases in a spectrum of increased, excessive psychomotor activation described with the terms "anxiety" and "agitation," among others. I am uncertain whether these observations were of exacerbations of preexisting, excessive psychomotor activation or the *de novo* occurrence of the excessive psychomotor activation in patients whose presentation was retarded or otherwise without excessive psychomotor retardation.
- Patients who demonstrate new or worsened excess psychomotor activation might not show a good global response to imipramine.
- Patients treated with imipramine have been observed to experience dysphoric confusional states accompanied by dreams of violence directed at family members or acquaintances. These new signs and symptoms resolve along with the depression on discontinuation of imipramine. This phenomenon was not described in association with ECT.
- Despite the possibility of new or worsened excess psychomotor activation, some thought agitated depression might be more responsive to imipramine than iproniazid.
- Imipramine treatment and ECT have been observed to be temporally associated with switches from depression to mania.
    - It is unclear as to whether new or worsened excess psychomotor activation and manic switches were viewed as manifestations of a single process or as distinct processes.
- As Kuhn had warned hypothetically, there was concern that both imipramine and ECT might result in improvement in patients presenting with retarded depression before other components of depression improved and thereby increase the risk of suicidality.
- Hospitalization and ECT were the treatments of choice for patients at significant risk of suicidal behaviors. However, the risk of suicidal behaviors was known to be challenging to predict with accuracy.

- One case of an exacerbation of what might be considered a core depressive symptom, of a possibly psychotic degree (". . . felt extreme despair, and what she called 'a dreadful quiet feeling.'") was associated with a suicide attempt.

My higher-level interpretation of the relevant passages of the written text of the Cambridge symposium (Davies 1964) is that it suggests that:

1. Imipramine can be associated with excess psychomotor activation ("anxiety," "agitation"), and this is not described for ECT unless manic switches, seen with both imipramine and ECT, are considered the same phenomenon.
2. Excess psychomotor activation was not directly linked to an increased risk of suicidality, except as described in #3 immediately below.
3. Both imipramine and ECT can result in improvement from a retarded state to a more normal/improved psychomotor state, and this can increase the risk of suicidal behaviors when the components of depression that facilitate suicidality (e.g., guilt, hopelessness, feelings of lack of self-worth) have not correspondingly improved.

Those who read this Comment will have their own opinions about the objectivity of the material I have presented and RCTs' relative merits.

To conclude Part 1 of this Comment, it should be evident that I believe that a 'proper' RCT is the gold standard for increasing useful and 'accurate' knowledge of treatments that maximizes the probability that those treatments can be used to optimize patient outcomes. At the same time, many important questions are difficult or impossible to address with RCTs. We cannot avoid seeking the most 'accurate' answers that we can obtain for such questions. I believe we must employ methods that might (or might not) result in less 'accurate' answers than an RCT if it were possible to conduct. We must also strive to develop new scientific research methods that improve the probability of obtaining 'accurate' answers to these tough questions.

RCTs' results and the collective interpretation from multiple RCTs might be influenced by many factors between a sponsor conceiving a development plan made up of the multiple RCTs and the decision for or against the authorization of the commercial marketing of the potential new drug and/or publication of a manuscript by an academic study group. In Part 2, I offer some thoughts on how sponsors, commercial or academic, can bias RCTs to favor

or disfavor a study treatment at the individual RCT level. I describe the creation of these biases and how some might be avoided, made evident to the scientific community, and potentially prevented. I examine the possible creation of these biases organized by the various steps between RCT planning and submission of a regulatory report and/or academic manuscript for an RCT. The academic community, journal editors, and regulators are likely most interested in biases that might favor a treatment to prevent these biases from adversely impacting clinical care. Sponsors are most likely interested in biases that might disfavor a drug to prevent costly failed RCTs. I describe several steps from concept to publication and/or regulatory report that might not be well recognized as study steps leading to a final RCT product – the report and manuscript.

I have a final caveat. This Comment contains extensive and lengthy quotations. Most quotations have been copied by hand from their source documents into the Comment. For longer quotations, the source documents were scanned and converted to MS Word documents using Kofax OmniPage Ultimate 19.2. The scanning and MS Word conversion process was imperfect and required corrections for these quotations. Because neither process for quotation transcription was completely automated, there is the possibility that some quoted material is not an exact copy of the source material. Although I have checked the quotations' accuracy with another person's assistance, I still cannot guarantee the reader of perfect transcription. Any errors that readers might find if they compare the cited source documents to my text are entirely my responsibility. However, I believe that the quotations are accurate transcriptions and that I have exercised due diligence in reproducing the quoted material.

References:

Armitage P. Obituary: Sir Austin Bradford Hill, 1897-1991. *J R Stat Soc* 1991; 154:482-485.

Azima H. Imipramine (Tofranil): a new drug for the depressed. *Can Med Assoc J* 1959; 80:535-540.

Azima H, Vispo RH. Effects of Imipramine (Tofranil) on depressive states: a clinical and psychodynamic study. *AMA Arch Neurol Psy* 1959; 81:658-664.

Bailey RA. *Cambridge Series in Statistical and Probabilistic Mathematics: Design of Comparative Experiments*. Cambridge: Cambridge University Press, 2008.

Ball JRB, Kiloh LG. A controlled trial of imipramine in treatment of depressed states. *Brit Med J* 1959; 2:1052 (November 21, 1959).

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. inhn.org.ebooks. October 24, 2019.

Beasley CM Jr. Additional Reply to Edward Shorter's comments on his "introductory comments." inhn.org.ebooks. May 21, 2020.

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.

Bhatt A. Evolution of clinical research: a history before and beyond James Lind. *Perspect Clin Res* 2010; 1:6-10.

Bloom S. https://twitter.com/sahilbloom/status/1372173620592603142. March 17, 2021. Accessed March 21, 2021.

Bovens L, Hartmann S. Solving the riddle of coherence. *Mind* 2003; 112:601-633.

Bram G. Imipramine in depression (letter). *Brit Med J* 1959; 2:1486 (December 26, 1959).

Crofton J. The MRC randomized trial of streptomycin and its legacy: a view from the clinical front line. *J Roy Soc Med* 2006; 99:531-534.

Damluji NF, Ferguson JM. Paradoxical worsening of depressive symptomatology caused by antidepressants. *J Clin Psychopharmacol* 1988; 8:347-349.

Daniels M, Hill AB. Chemotherapy of pulmonary tuberculosis in young adults; an analysis of the combined results of three Medical Research Council Trials. *Brit Med J* 1952; 1:1162 (May 31, 1952).

Davies EB (Ed.). *Depression: A Cambridge Postgraduate Medical Course, Proceedings of the Symposium Held at Cambridge 22 to 26 September 1959.* London: Cambridge University Press, 1964.

Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. *Soc Sci Med* 2018; 210:2-21.

Department of Physics, University of Illinois at Urbana-Champaign. https://van.physics.illinois.edu/qa/listing.php?id=64061&t=how-gravitational-force-varies-at-different-locations-on-earth. Accessed April 7, 2021.

Dreyfus J-F. Comment on David Healy's Essay: Do Randomized Clinical Trials Add or Subtract from Clinical Knowledge? inhn.org.controversies. January 28, 2021.

Foster AR, Lancaster NP. Disturbance of motor function during treatment with imipramine. *Brit Med J* 1959; 2:1452 (December 26, 1959).

Fox W, Sutherland I. The clinical significance of positive cultures and of isoniazid-resistant tubercle bacilli during the treatment of pulmonary tuberculosis: report to the tuberculosis chemotherapy trials committee of the Medical Research Council. *Thorax* 1955; 10:85-98.

Freyhan FA. Clinical Effectiveness of Tofranil in the treatment of depressive psychosis. *Can J Psychiat* 1959; 4:(1:suppl):S86-S99.

Garrett GM. The use of Tofranil in mental hospital practice. *S Afr Med J* 1959; 33:993-994.

Greenland S, Senn SJ, Rothman K, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 216: 31:337-350.

Gutkin SW. *Writing High-Quality Medical Publications: A User's Manual*. Boca Raton, Fl: CRC Press, 2019.

Hamilton M. *Methodology of Clinical Research*. Edinburgh: E. & S. Livingstone Ltd., 1961.

Healy D. Do Randomized Controlled Trials Add to or Subtract from Clinical Knowledge? inhn.org.controversies. December 3, 2020.

Healy D. David Healy's Reply to Jean-François Dreyfus' Comment. inhn.org.controversies. February 18, 2021.

Hill AB. *Principals of Medical Statistics.* First Edition. London: Lancet, 1937.

Hill AB. *Principals of Medical Statistics.* Seventh Edition. London: Lancet, 1961.

Hill AB. Reflections on the Controlled Trial. Annals Rheum Disease 1966; 25:107-113.

Hill AB. *Principals of Medical Statistics.* Eighth Edition. London: Lancet, 1967.

Hill AB. *Principals of Medical Statistics.* Ninth Edition. London: Lancet, 1971.

Holy Bible. Book of Daniel (Versus 11-16) of the Old Testament (Contemporary English Version [https://www.bible.com/bible/392/DAN.1.CEV]). Accessed March 17, 2021.

Institute of Medicine 2001. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Eleventh Printing. Washington, DC: The National Academies Press, 2014. P 147.

Kline NS. Psychopharmaceuticals: effects and side effects. *B World Health Organ* 1959; 21:397-410.

Kuhn R. The treatment of depressive states with G 221355 (imipramine hydrochloride). *Schweiz Med Wochenschr* 1957; 87:1135-1140. English translation reprint: https://depts.washington.edu/psychres/wordpress/wp-content/uploads/2017/07/100-Papers-in-Clinical-Psychiatry-Depressive-Disorders-The-treatment-of-depressive-states-with-G-22355-imipramine-hydrochloride.pdf. Accessed March 21, 2021.

Lancaster NP, Foster AR. Suicidal attempt by imipramine overdosage. *Brit Med J* 1959; 2:1458 (December 26, 1959).

Lehmann HE, Chan CH, de Verteuil RL. The treatment of depressive conditions with imipramine (G 22355). *Can J Psychiat* 1958; 3:155-164.

Lohr, KN, Eleazer K, Mauskopf J. Health Policy Issues and Applications for Evidence-Based Medicine and Clinical Practice Guidelines. *Health Policy* 1998; 46:1-19.

Mann AM, Catterson AG, Macpherson AS. Toxicity of imipramine: report of serious side effects and massive overdosage. *Can Med Assoc J* 1959; 81:23-28.

Montgomery DC. *Design and Analysis of Experiments*. Hoboken, NJ: John Wiley & Sons, Inc., 2020.

MRC. Streptomycin treatment of tuberculous meningitis. *Lancet* 1948; 251:582-596.

MRC. Various combinations of isoniazid with streptomycin or with PAS in the treatment of pulmonary tuberculosis. *Brit Med J* 1955; 1:435. (February 19, 1955).

Musher DM, Tuomanen EI. Pneumococcal pneumonia in patients requiring hospitalization. *UpToDate*. Article updated January 28, 2021. https://www.uptodate.com/contents/pneumococcal-pneumonia-in-patients-requiring-hospitalization?search=pneumococcal%20pneumonia%20treatment&source=search_result&selectedTitle=1~88&usage_type=default&display_rank=1. Accessed April 12, 2021.

Oxford English Dictionary. (Online edition). experiment, n. : Oxford English Dictionary (oed.com) https://www.oed.com/view/Entry/66530?rskey=anihbN&result=1&isAdvanced=false#footerWrapper. Accessed March 22, 2021.

Prien RF, Robinson DS. (Eds.) *Clinical Evaluation of Psychotropic Drugs: Principals and Guidelines*. New York: Raven Press, 1994.

Quade D., Lachenbruch PA, Whaley FS, McClish DK, Haley RW. (1980). Effects of misclassifications on statistical inferences in epidemiology. *American Journal of Epidemiology* 111:503-515.

Sackett DL, Rosenberg WMC, Muir Gray JA, Haynes RB, Richardson WS. Evidence Based Medicine: What It Is and What It Isn't. *Brit Med J* 1996; 312:71-72.

Sackett, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice & Teach EBM*. Second Edition. London: Churchill Livingstone, 2000.

Shorter E. Comments on Introductory Comments [about Beasley and Tamura's second installment: 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) – Introductory Comments. inhn.org.ebooks. December 13, 2018]. inhn.org.ebooks. May 9, 2019.

Sloane RB, Habib A, Batt UE. The use of imipramine (Tofranil) for depressive states in open ward settings of a general hospital: a preliminary report. *Can Med Assoc J* 1959; 80:540-546.

Spilker B. *Guide to Clinical Studies and Developing Protocols*. New York: Raven Press, 1984.

Spilker B. *Guide to Clinical Interpretation of Data*. New York: Raven Press, 1986.

Spilker B. *Guide to Planning and Managing Multiple Clinical Studies*. New York: Raven Press, 1987.

Spilker B. *Presentation of Clinical Data*. Philadelphia: Lippincott Williams & Wilkins, 1990.

Spilker B. (Ed.) *Guide to Clinical Trials*. Philadelphia: Lippincott-Raven Publishers, 1991.

Spilker B. Guide to Drug Development: A Comprehensive Review and Assessment. Philadelphia: Lippincott Williams & Wilkins, 2008.

Spilker B, Schoenfelder J. *Presentation of Clinical Data*. New York: Raven Press, 1990.

Spilker B, Schoenfelder J. *Data Collection Forms in Clinical Trials*. New York: Raven Press, 1991.

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.

Wikipedia. https://en.wikipedia.org/wiki/Experiment. Accessed March 22, 2021.

Note on reference format: Journal name abbreviations are those used by Web of Science (https://images.webofknowledge.com/images/help/WOS/C_abrvjt.html. Accessed March 29, 2021) or PubMed if not included in Web of Science. For manuscripts published in the *Brit Med J* before 1994, the citation is provided for the publication year, the indicator 1 (for the months of January-June) or 2 (for the months July-December), and page number for the first page of the manuscript. The indicator of 1 or 2 and the publication date are required to quickly locate the manuscript in the Brit Med J archives. These dates are provided for *Brit Med J* manuscripts published before 1994.

July 8, 2021