Assessing Therapy Effects in Clinical Trials Using a Measure Theoretic Quanta Analysis

Lem Moyé, MD, PhD

Correspondence to:
Lem Moyé, MD, PhD
UTHealth School of Public Health
1200 Pressler St.
Houston, Texas 77030
Phone: (713) 500-9518
Email: Lemmoye@msn.com

Abstract

Statistical hypothesis testing is a fixture in clinical trials. However, its continued application has produced an artificially constrained clinical trial analysis paradigm that is tightly bound to type I error management. Not only is this difficult to helpfully apply in modern complex clinical trials, but it is not responsive to the *primum movens* of the investigators.

A measure theoretic approach is developed here that is based on the principals of health care research analysis and not those of statistical hypothesis testing theory. This new rubric permits all data collected by the trial that is responsive to a specific scientific question to quantitatively contribute to that question's answer. Thus, estimates for each of the following are obtainable: 1) the total available evidence in a clinical trial to answer the question, 2) the strength of that evidence, 3) the strength of evidence that supports benefit and the strength of evidence supporting harm, and 4) the magnitude of the beneficial effect and the magnitude of harm. The incorporation of sampling error in these estimates is achieved without formal hypothesis testing, obviating the need for type I error consideration with its attendant multiplicity corrections.

Keywords. Clinical trial, measure theoretic, benefit-risk ratio

Introduction

Ninety-two years have passed since the writings of Ronald Fisher introduced inference testing to the applied statistical community [1,2]. This theory of statistical hypothesis testing generated the *p*-value that subsequently garnered the support of US Food and Drug Administration (FDA) regulators, National Institutes of Health administrators, and medical journal editors in assessing clinical research [3]. With the advent of clinical trials, statistical hypothesis testing became a fixture among medical researchers and, despite the concerns voiced principally by epidemiologists [4,5,6,7,8], statistical hypothesis testing remains a fixture of clinical investigation today, including cardiology, the focus of this manuscript.

These statistical hypothesis testing requirement and its focus on *p*-values generated a collection of interpretative conundrums [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] for the cardiology research community. From this turmoil arose additional design and analysis requisites that a clinical trial must satisfy [23, 24]. These requirements of 1) differentiating prospectively declared analyses from *post hoc* (exploratory) endpoints and 2) conserving the overall type I error among a small number of prospectively declared endpoints (multiplicity corrections) instilled important and necessary discipline in conducting and interpreting clinical trial results. However, an unfortunate consequence of these tenets is that commonly only a fraction of the data that are collected in a clinical trial are actually used to directly answer the study question, a restriction that is required to control type I error propagation [25]. Thus, although many analyses are conducted, only a small subset of them (and commonly only one of them) is identified as a “primary”. These restrictions are applied not just in academic research but in clinical trials conducted by the private sector which follows the contemporary guidance of the federal Food and Drug Administration [26].

This analysis parsimony – a consequence of type I error control – is emblematic of the fundamental tension between biostatistics and clinical researchers; the inability of statistical hypothesis testing to directly address the primary, probing question that motivated investigators to execute their research.

The principal inquiry of interest to investigators in clinical trials is "Are participants better or worse off after exposure to the intervention when compared to the control group experience?" This is a global question that requires a comprehensive review of all analyses that bear on this query. However, in cardiology, the combination of 1) the daunting universe of possible assessments (e.g., heart function, renal function, the ability of the individual to exercise, how long the individual survives, number of hospitalizations, patient quality of life surveys) and 2) the different types of statistical estimators implemented to asses the exposures effects (e.g., Bayes procedures, non-parametric U statistics, regression analyses, imputation evaluations, survival analyses) have complicated all attempts to provide an answer.

The traditional approach of biostatistics is to require the investigator to select one or a small number endpoints, and then execute statistical hypothesis testing on each, converting every one of them into the familiar dichotomous decision framework (rejection or non-rejection of the null hypothesis); type I error is calculated and accumulated as each of these endpoints is assessed. Thus, to the investigator who has accepted the task of interpreting a complex study of a complicated disease, statisticians deliver 1) the results of a small number of analyses which the statisticians believe are dispositive, and 2) an accumulated type I error rate.

Although the investigators' point of view is quantitative, it is also contradistinctive. Investigators believe that an analysis finding (e.g., the mean difference in the change in exercise tolerance between the exposed and control group), because of sampling error and other sources

of imprecision, can support both a degree of benefit and a degree of harm simultaneously. It is this dualism – not hypothesis testing dichotomy – to which the investigators resonate, and it is this dualism – not type I error – that should be accumulated so that investigators can assimilate their research results. Unfortunately, the standard statistical analysis disappoints the investigators who find themselves left with 1) no quantitative answer to their principal question and 2) an accumulated type I error rate in which they have no direct interest. This is the disconnect between these two scientific disciplines.

This paper establishes a rubric which relies on measure theoretic tools to develop Lebesgue-Stieltjes functions that assess the evidence from all analyses in a clinical trial that are responsive to a global question, and that provide the degree to which that evidence supports benefit and harm. This approach provides a direct answer to the investigator's principal question with no reliance on statistical hypothesis testing.

Methods

This development assumes that there is one clinical trial that has been well designed and concordantly executed (i.e., carried out in accordance with the prospectively written protocol). It is also assumed that the investigators designed the study to answer one overall question *q,* e.g., "Does mesenchymal cell therapy improve the well-being of patients with heart failure?" This manuscript develops answers to the four following inquiries related to question *q:*

1. Within the scope of all of the clinical trial's analyses, what is the content of evidence that addresses specific question *q?*

2. Within the scope of all of the clinical trial's analyses, what is the strength of evidence that actually addresses the specific question q?
3. What is the strength of evidence supporting an affirmative answer to question q (i.e., a beneficial effect of cell therapy) and the strength of evidence in the trial suggesting the reverse (i.e., a harmful effect of cell therapy)?
4. From the evidence of all analyses, what is an estimator of that benefit? What is the estimate of harm?

The goal is to create a sample space Ω of clinical trial analyses, from which a standard σ-algebra Σ is formulated. With this as a foundation, a formal measure ψ is developed on which analysis-measurable functions operate and can be integrated with respect to (Ω, Σ). Their integrals produce answers to queries 1-4.

The full development is available (Appendix). To recapitulate, since a clinical trial's product is a collection of analyses, $\{\omega_i\}$, a sample space Ω containing all of the analyses is generated by the study. Each element $\omega_i \subset \Omega$ contains the constitutive components of the i^{th} analysis. One element of ω_i is the question that motivated the analysis. Denote this element of ω_i as q_i. Another collection of components of ω_i is the group of analysis characteristics denoted as $\delta_i(j), j = 1,,,$ where j indexed the design and operational features of the analysis. There are many of these characteristics of the analysis, e.g., planning of the analysis (prospective versus retrospective), and the type of analysis (e.g., survival analysis, mean different analysis, etc., subgroup analysis). The remaining components of ω_i are the participants used in the analysis, the variables used in the analysis (not their values, but their identities), and the analysis' estimate of effect size and its standard error.

From this perspective one can, for example, collect a set of analyses *A* containing all subgroup assessments evaluating the role of an antidiabetic medication on changes in micro albuminuria, or a set of analyses *B* containing all analyses conducted that examined the difference in the change over time in systolic blood pressure by therapy group. This

multicomponent structure of ω_i offers a wide latitude in the creation of sets of analyses. These analysis collections, or "regions of analysis" can then serve as the domain on which Lebesgue-Stieltjes integrals operate.

With this framework, define the content of the analysis ω_i, as $\psi(\omega_i)$, and write

$$\psi(\omega_i) = n_i v_i.$$

This defines the content of an analysis as the product of the number of participants whose data contribute to the evaluation multiplied by the number of variables that are required for the analysis. Denote the content of this intersection as $\psi(\omega_i \cap \omega_j)$ where ω_i and ω_j are not disjoint and define

$$\psi(\omega_i \cap \omega_j) = n_{ij} v_{ij}.$$

Here, n_{ij} is the number of participants and v_{ij} the number of variables common to both analyses. In general, the content of the intersection of *k* analyses $\omega_1, \omega_2, \omega_3, ...\omega_k$ is

$$\psi\left(\bigcap_{i=1}^{k} \omega_i\right) = n_{i...k} v_{i...k}$$

It has been demonstrated that $\psi(\omega_i)$ meets the formal definition of a measure (Appendix).

Since the measure of an analysis is simply based on the number of observation and variables it contains, it is easily anticipated that analyses are in general not pairwise disjoint, complicating the computation of $\psi\left(\bigcup_{k=1}^{n} \omega_k\right)$. To expedite this computation, define $\{B_k\}$ as the sequence of sets created from the increasing sequence of sets $C_k = \bigcup_{i=1}^{k} \omega_i$ where $B_k = C_k \cap C_{k-1}^c$. Then while it is true that $\bigcup_{k=1}^{n} \omega_i = \bigcup_{k=1}^{n} B_i$, it is also true that $\psi\left(\bigcup_{k=1}^{n} \omega_k\right) = \sum_{k=1}^{n} \psi(B_k)$ since $\{B_k\}$ consists of pairwise disjoint sets. (Figure 1.) This collection of sets $\{B_k\}$ represent the analysis fragments or quanta that make separate contributions to the measure of the union of all analyses responsive to a question *q*. Furthermore, the measure of any analysis quanta B_k can be computed as

$$\psi(B_k) = (nv)_k - \sum_{j_1=1}^{k-1} (nv)_{j_1 k} + \sum_{j_1=1}^{j_2-1} \sum_{j_2=2}^{k-1} (nv)_{j_1 j_2 k} - \sum_{j_1=1}^{j_2-1} \sum_{j_2=2}^{j_3-1} \sum_{j_3=3}^{k-1} (nv)_{j_1 j_2 j_3 k} + \ldots$$

Where $(nv)_{ij\ldots}$ is simplifying notation for $n_{ij\ldots} v_{ij\ldots}$ (Appendix).

Thus the measure or content of the collection of non-disjoint analyses, $A = \bigcup_{i=1}^{k} \omega_i$ can be assembled from the sum of measures of mutually disjoint combinations of analysis quanta $\{B_i\}, i = 1, 2, 3, \ldots, k$, thereby permitting the expression of $\psi(A) = \int_A d\psi = \int_{\bigcup_{i=1}^{k} B_i} d\psi = \sum_{i=1}^{k} \psi(B_i)$ where $\bigcup_{i=1}^{k} B_i = A$. An adaptation of this measure for the circumstance in which the variables both within and across analyses are correlated is available (Appendix).

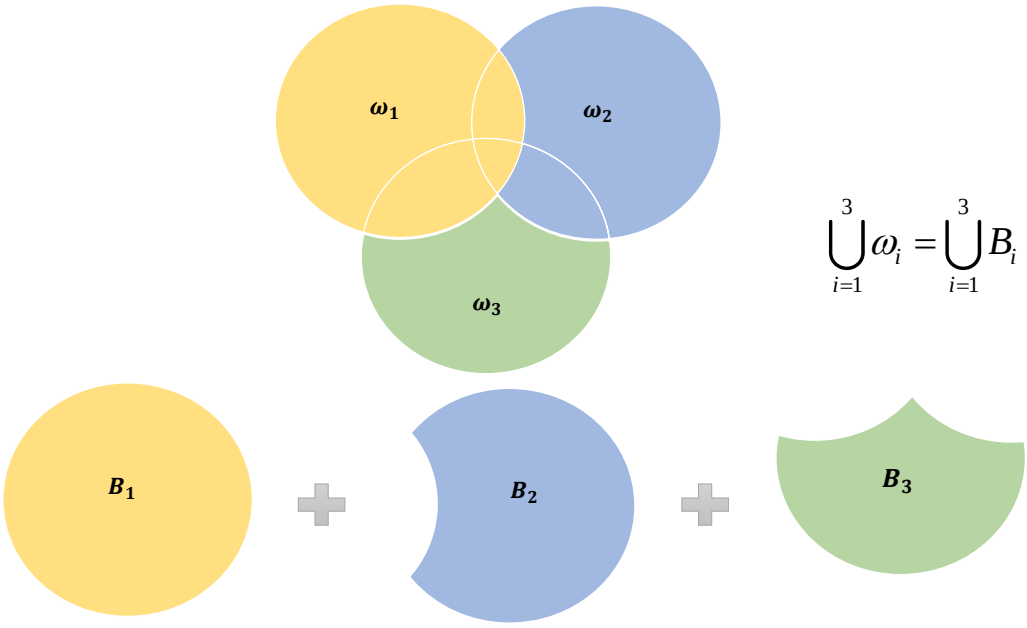

Figure 1. Decomposing the overlapping analyses $\{\omega_1, \omega_2, \omega_3\}$ into non-overlapping analysis components $\{B_1, B_2, B_3\}$

With this as background, $\psi -$ measureable functions will now be created to address the four inquiries.

Inquiry 1. What is the content of evidence in the trial that addresses question q?

A clinical randomized trial designed to address the question "Does mesenchymal cell therapy delivered intravenously improve the cachexic state of advanced heart failure?" conducts many analyses to address this inquiry. However, although the analyses are germane, their different methodologies (imputation analyses, regression analyses, Bayes analyses, etc.) challenge any attempt to combine them. The ultimate goal is to use Lebesgue-Stieltjes integrals to accumulate these findings. As a preamble to these steps, from the set of analyses $\{\omega_i \,|\, q_i \subset q\}$ assemble the collection of analyses quanta $\{B_i\}$, and compute the measure of these quanta using $\psi-$ measure (Appendix).

Begin with the identification and collection of the subset of all analyses conducted that address question *q*. Denote this subset as $A_q = \{\omega_i \,/\, q_i = q\}$. The content of evidence that addresses question *q* is obtained by measuring or accumulating the content of analyses that contribute to $A_{q.}$ Thus, if the content of evidence that addresses question *q* is $\mathbf{\Gamma_q}$ then write $\mathbf{\Gamma_q} = \int_{A_q} d\psi = \psi(A_q)$ where the integral signifies Lebesgue-Stieltjes integration over the collection of analyses $\omega_i \subset A_q$.

Since $\psi(A_q)$ involves the measure of overlapping sets $\omega_i \subset A_q$ the inequality $\int_{A_q} d\psi = \psi(A_q) = \psi\left(\bigcup_{i=1}^{n} \omega_i 1_{\omega_i \subset A_q}\right) \le \sum_{i=1} \psi(\omega_i) 1_{\omega_i \subset A_q} = \sum_{i=1}^{n} n_i v_i 1_{\omega_i \subset A_q}$ (where there are *n* member analyses contained in A_q) is available but not sharp. However the creation of the collection of disjoint analyses quanta $\{B_i\}$ of the previous section permits $\int_{A_q} d\psi = \psi(A_q) = \psi\left(\bigcup_{i=1}^{n} \omega_i 1_{\omega_i \subset A_q}\right) = \sum_{i=1}^{n} \psi(B_i) 1_{\omega_i \subset A_q}$. Thus, the measure of evidence that addresses

question q is $\mathbf{\Gamma_q} = \int_{A_q} d\psi = \psi\left(A_q\right) = \sum_{i=1}^{n} \psi\left(B_i\right) 1_{\omega_i \subset A_q}$ with analysis ω_i is represented by its quantum B_i.

Inquiry 2. What is the strength of evidence for any analysis in the trial that addresses question q?

The measure ψ is based only on the number of participants and the number of variables that is contained in an analysis, representing only the data that is incorporated in a body of analyses $A_q = \{\omega_i / q_i = q\}$. Inquiry 2 addresses not just evidence brought to bear to address question q but the strength of that evidence provided by the set of analyses $A_q = \{\omega_i / q_i = q\}$. Begin by asserting that the strength of evidence contained in each analysis is determined by the research community and transmitted to $A_q = \{\omega_i / q_i = q\}$ through the decisions of the investigators. For example, in a clinical trial, prospectively declared analyses are commonly held to be of greater value than *post hoc* or exploratory analyses. As another example, measures of organism function (e.g., survival, amputation free survival, walking distance, quality of life), can be of greater value than changes in measures of organ function (e.g. left ventricular ejection fraction (LVEF)), which are themselves of greater value than isolated findings for biomarkers. Thus the strength of evidence contained in an analysis is determined *a priori* by the investigators. This in turn determines the importance of the contribution that the analysis makes to answering the scientific question *q*. The nomenclature commonly used to communicate this concept is the use of adjectives such as “prospective”, “primary”, “secondary”, etc.

Thus, the formal evaluation process in a clinical trial involves a priority ordering of the analyses’ contributions from the most influential and important to the least contributory. This

clinical trial methodology is incorporated here [Appendix]. The choice of a well defended analysis priority *a priori* is equivalent to creating a function *T* that oversees the reordering of the set of analyses $\{\omega_i\}$ from essentially a random sequence of analyses to a specifically ordered set, i.e., $T(\omega_1, \omega_2, \omega_3, ..., \omega_n) = \omega_{[1]}, \omega_{[2]}, \omega_{[3]}, ..., \omega_{[n]}$ where the subscript $[i]$ denotes the i^{th} analysis in the priority order from highest to lowest priority. Note that the function *T* also converts the sequence $\{B_i\}, i = 1, 2, 3, ...$ to $\{B_{[i]}\}, i = 1, 2, 3, ...$ the sequence of disjoint analyses quanta corresponding to the sequence of analyses ordered by priority. The reordering of $\{B_i\}$ is critical, because an implication of the collection of quanta $\{B_i\}$ being pairwise disjoint is that their contribution to $\psi\left(\bigcup_{i=1}^{n} \omega_i\right)$ depends on their location in the priority sequence. Thus the ordered sets $\{B_{[i]}\}, i = 1, 2, 3, ...$ manifests the *a priori* sense of the importance of the evidence (as reflected by the magnitude of $\psi(B_{[i]})$ to be provided by the analysis.

Therefore, the strength of evidence offered by any analysis $\omega_i = \psi(B_{[i]})$ and the relative strength of evidence provided by the ω_i^{th} analysis to address question *q* is **RSE** where

$$\mathbf{RSE}[\omega_i] = \frac{\psi\left(B_{[i]} \mathbf{1}_{\omega_i \subset A_q}\right)}{\int_{A_q} d\psi} = \frac{\psi\left(B_{[i]} \mathbf{1}_{\omega_i \subset A_q}\right)}{\boldsymbol{\Gamma}_{\mathbf{q}}}.$$

Inquiry 3. What is the strength of evidence in the trial supporting an affirmative answer to question q ? What is the strength of evidence in the trial suggesting a negative answer?

The two parts of inquiry 3 will be addressed in turn. Assume that question *q* concerns the benefit or harm of an intervention in a clinical trial, e.g., "Does the provision of mesenchymal cells to patients with heart failure ameliorate their signs and symptoms when compared to the experience of controls?" The process to be followed to address this question is to first identify the statistical estimate of effect from each analysis $\omega_i \subset A_q = \{\omega_i / q_i = q\}$ and then for each estimator, 1) consider the distorting role of sampling error and imprecision on this estimate, 2) parse the resulting region into a region supporting benefit, 3) quantify this region, 4) norm this by the measure of its quanta, and 5) accumulate this evidence over all $\omega_i \subset A_q$.

One of the components of each ω_i is the effect size produced by the analysis, identified now as e_i. This quantity e_i can be the difference between therapy groups of the mean blood pressure change over time, or the relative risk of death associated with an intervention. However, due to sampling variability and the measurement's relative imprecision, this estimate of benefit cannot be relied upon in and of itself. The impact of these two distorting effects is to blur the exact position of the population measure of effect that could be deduced from the value of the statistical estimator from the sample; variability and imprecision each suggest that both larger values and smaller values of the estimator are admissible for consideration. This range of values will be termed the estimator's region of plausible effects. It is not just the estimator that provides a sense of the effect of the intervention; it is the estimator's region of plausible values that is most informative about the possible effect size that would be seen in the population.

The region of plausible effect will always provide values of the effect size that are larger than the statistical estimator, and others that are smaller. In many cases, the distorting effects of imprecision and sampling error can actually reverse the direction of effect, signifying that not benefit, but harm might be produced in the population at large.

The observation that the statistical estimator produces a plausible region of effect that together and simultaneously supports both larger benefit values and smaller ones (that sometimes includes harm) is here termed duality. Estimators refract the data on which they are based into both larger and smaller effect sizes including effect sizes that are indicative of harm. It is this duality that the functions developed in this section will first segregate and capture (a process termed analysis parsing) and then accumulate using $\psi -$ measure.

Define the upper e_i^+ and lower e_i^- bounds of an interval of plausible effect for the analysis as ω_i, computing

$$e_i^+ = e_i + a_i$$
$$e_i^- = e_i - b_i$$

where a_i and b_i are constants based on variability and imprecision. Note that this interval need not be symmetric around the actual estimator e_i. The region of plausible effect is signified as $\left[e_i^-, e_i^+\right]$.

This plausible effect interval can be parsed into two subintervals, one a region of benefit, the other of harm. In order to locate these sub-regions, knowledge of the value of the statistical estimator's effect that is neutral (i.e., denotes neither benefit nor harm) is required. Define this value of neutral effect as $e_i(0)$. Similarly, let $e_i(b)$ and $e_i(h)$ be the values of the greatest possible benefit and the greatest possible harm permitted by the estimator respectively. The introduction of $e_i(b)$ and $e_i(h)$ is necessary since values of harm need not always be less than values of benefit. For example, if the i^{th} analysis is a total mortality hazard function analysis, then $e_i = 1$ indicates no effect on the time to death , $e_i(h) = \infty,$ and $e_i(b) = 0.$ Alternatively, if ω_i is an evaluation of changes in mean differences where the greater differences are salubrious,

then the value of $e_i = 0$ reflects no mean effect, $e_i(h) = -\infty,$ and $e_i(b) = \infty.$ Using this notation, then the interval $\left[\min\left(e_i(h),\, e_i(b)\right), \max\left(e_i(h),\, e_i(b)\right)\right]$ is the range of possible values of the estimate.

Consider the case where $e_i(b) > e_i(h)$. We now define the plausible benefit interval $\chi_i^{(b)}$ as;

$$\chi_i^{(b)} = \left[b_i^-,\, b_i^+\right] = \left[e_i^-,\, e_i^+\right] \cap \left[\min\left(e_i(0), e_i(b)\right), \max\left(e_i(0), e_i(b)\right)\right] = 1_{\left[e_i^-, e_i^+\right]} 1_{\left[e_i(0),\, e_i(b)\right)} = 1_{\left[b_i^-, b_i^+\right]}$$

This is the portion of the plausible effect size region that supports benefit. For example, larger values of left ventricular ejection fraction are considered beneficial *ceteris paribus*; its increases are beneficial and its decreases are harmful. Thus, if the plausible effect region for a change in left ventricular ejection fraction is $[-1, 7]$ and the region of these changes that are beneficial is $\left(e_i(0), e_i(b)\right) = (0, \infty),$ then $\chi_k^{(b)} = [-1, 7] \cap (0, \infty) = (0, 7]$ is the plausible benefit region. The plausible region for harm is based on $\left(\min\left(e_i(h), e_i(0)\right), \max\left(e_i(h), e_i(0)\right)\right) = (-\infty, 0),$ and is

$$\chi_i^{(h)} = \left[h_i^-,\, h_i^+\right] = \left[e_i^-,\, e_i^+\right] \cap \left[e_i(h),\, e_i(0)\right] = 1_{\left[e_i^-,\, e_i^+\right]} 1_{\left[e_i(h),\, e_i(0)\right)} = 1_{\left[h_i^-,\, h_i^+\right]}$$

which in this example is $\chi_k^{(h)} = [-1, 7] \cap (-\infty, 0) = (-1, 0].$

Now define the contribution function

$$\mathbf{Y}\left(\chi_i^{(b)}\right) = \mathbf{Y}\left(1_{\left[b_i^-,\, b_i^+\right]}\right) = \frac{1}{\left(b_i^+ - b_i^-\right)}\left(\frac{b_i^+ + b_i^-}{2} + b_i^-\right)$$

as the unit-less benefit function that maps the interval of plausible benefit to an assessment of the level of that benefit. $\mathbf{Y}\left(\chi_i^{(b)}\right)$ penalizes the benefit estimate derived from ω_i for a wide interval,

315 while amplifying benefit if the minimum value of the plausible region is different than $e_i(0)$

316 (Figure 2).

317

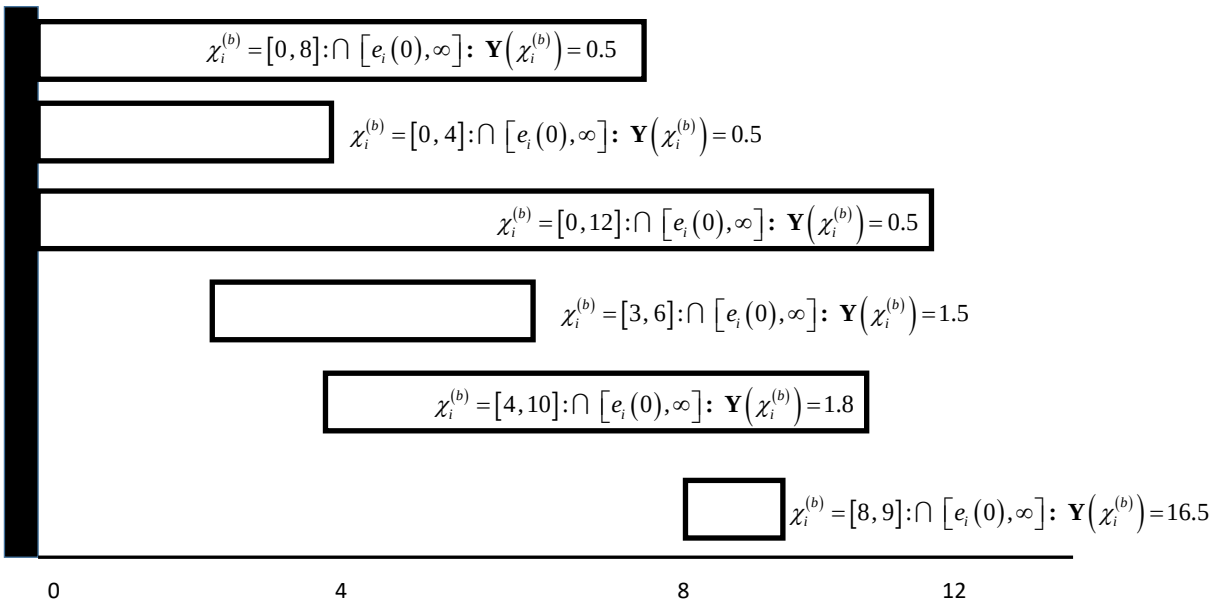

Figure 2. Operation of the benefit function for different levels of analyses effects.

318

319 From Figure 2, the circumstance where $\chi_i^{(b)} = \left[b_i^-, b_i^+\right] = [0,8]$, $\chi_i^{(b)} = \left[b_i^-, b_i^+\right] = [0,4]$ and

320 $\chi_i^{(b)} = \left[b_i^-, b_i^+\right] = [0,12]$ each generate a contribution function value of only

321 $\mathbf{Y}\left(\chi_i^{(b)}\right) = \mathbf{Y}\left(1_{\left[b_i^-, b_i^+\right]}\right) = 0.5$, reflecting some addition of benefit from this region, but penalizing it

322 because their lower bound includes $e_i(0)$, the value of no effect. The contribution's function

323 value is greater when $b_i^- > e_i^0$, as is the case of the remaining two examples in Figure 2.

324 Analogous contribution computations manage the harm concern.

325 With the benefit interval and contribution function in hand, it remains to compute the

326 benefit over all of the analysis $\omega_i \subset A_q$. The integral $\int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) d\psi$ is the assessment of the benefit

function on each set $\omega_i \subset A_q$ with respect to ψ–measure. A normed version over the measure of A_q can be written as $\mathbf{B}_q = \left[\psi\left(A_q\right)\right]^{-1} \int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) d\psi$. Thus $\mathbf{B}_q$ is the normed measure of benefit derived from all of the analyses responsive to question *q*.

A similar quantity can be computed to assess harm. With the plausible harm interval $\chi_k^{(h)}$ defined as above where $\chi_i^{(h)} = \left[h_i^-,\, h_i^+\right] = \left[e_i^-,\, e_i^+\right] \cap \left[-\infty, e_i\left(0\right)\right] = 1_{\left[e_i^-, e_i^+\right]} 1_{\left[-\infty, e_i(0)\right]} = 1_{\left[h_i^-, h_i^+\right]}$. define $\mathbf{Y}\left(\chi_i^{(h)}\right) = \left| \frac{1}{\left(h_i^+ - h_i^-\right)} \left(\frac{h_i^+ + h_i^-}{2} + h_i^+ \right) \right|$ and $\mathbf{H}_q = \left[\psi\left(A_q\right)\right]^{-1} \int_{A_q} \mathbf{Y}\left(\chi_i^{(h)}\right) d\psi$. With these quantities, write the benefit to harm ratio as

$$\frac{\mathbf{B}_q}{\mathbf{H}_q} = \frac{\left[\psi\left(A_q\right)\right]^{-1} \int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) d\psi}{\left[\psi\left(A_q\right)\right]^{-1} \int_{A_q} \mathbf{Y}\left(\chi_i^{(h)}\right) d\psi} = \frac{\int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) d\psi}{\int_{A_q} \mathbf{Y}\left(\chi_i^{(h)}\right) d\psi}$$

Inquiry 4. If there is a benefit, what is an estimator of that benefit? If there is harm, what is the estimate of that harm?

For a query q, quantitative assessments of the evidence for benefit $\mathbf{B}_q$ and harm $\mathbf{H}_q$ were produced. Here an actual estimate of the level of benefit (and of harm) will be developed.

Recall that the plausible benefit interval $\chi_i^{(b)}$ is defined as $1_{\left[b_i^-, b_i^+\right]}$. There are several functions that provide service in assessing the effect of therapy based on that interval. Let $\mathbf{I}$ be the condition where an increase in e_i reflects benefit and $\mathbf{D}$ reflect the circumstance where a decrease reflects benefit. Then one such function is $\mathbf{L}_{\max}\left(\chi_k^{(b)}\right) = \mathbf{L}_{\inf}\left(\chi_k^{(b)}\right) 1_{\mathbf{D}} + \mathbf{L}_{\sup}\left(\chi_k^{(b)}\right) 1_{\mathbf{I}}$. This represents the assessment of greatest benefit from the plausible interval. Alternative, one could

conservatively estimate benefit as $\mathbf{L}_{\min}\left(\chi_k^{(b)}\right) = \mathbf{L}_{\sup}\left(\chi_k^{(b)}\right)\mathbf{1}_{\mathbf{D}} + \mathbf{L}_{\inf}\left(\chi_k^{(b)}\right)\mathbf{1}_{\mathbf{I}}$. This represents the least effect value for benefit. Choosing the latter for this development, define the estimate of benefit from all of the analyses addressing question q as $\Lambda_{q\mathbf{B}}$

$$\Lambda_{q\mathbf{B}}(\min) = \left[\int_{\mathbf{A}_q} d\psi\right]^{-1} \int_{\mathbf{A}_q} \mathbf{L}_{\min}\left(\chi_i^b\right) d\psi.$$

This is the accumulation of unit less benefit with respect to the content of each analysis, normed by the accumulated content of all analyses.

A similar result is obtained for an estimator of harm produced by all analyses $\left\{\omega_i \,|\, \omega_i \subset A_q\right\}$.

$$\Lambda_{q\mathbf{H}}(\max) = \left[\int_{\mathbf{A}_q} d\psi\right]^{-1} \int_{\mathbf{A}_q} \mathbf{L}_{\max}\left(\chi_i^h\right) d\psi.$$

Where $\mathbf{L}_{\max}\left(\chi_k^{(h)}\right) = \mathbf{L}_{\sup}\left(\chi_k^{(h)}\right)\mathbf{1}_{\mathbf{D}} + \mathbf{L}_{\inf}\left(\chi_k^{(h)}\right)\mathbf{1}_{\mathbf{I}}$ is the worst case estimate of harm obtained from $\chi_k^{(h)}$ obtained from the plausible regions of harm. As with the benefit function, alternative views of harm are also available.

Discussion

This manuscript provides an alternative approach to clinical trial analysis that is based on both the principles of measure theory and clinical trial methodology. Its solutions provide answers to four inquiries of critical interest to clinical trialists using statistical estimation theory that are commonly not quantitatively addressed, while not relying on statistical hypothesis testing.

The current clinical trial analysis procedure requires the discrimination of prospective from exploratory analyses, and control of the familywise type I error among the former. This commonly restricts the study's conclusive analyses to a small number of evaluations addressing precise "primary endpoints" that are pre-designated to represent the principal findings of the well-designed, concordantly executed clinical trial.

Unfortunately, this standard approach is a symptom of the detachment of the goals of the clinical trial investigators from the work product of contemporary biostatistics. A clinical trial's primary analyses address important questions, but their answers are only contributory to the more global question of "Has the health, well-being, and sense of well-being of participants improved after exposure to the new therapy when compared to the experience of the control group?" This is the question that is of greatest interest to research investigators, participants, health care providers, formulary committees, and the regulatory community. The answer to this larger question requires a broad and integral appraisal of all responsive analyses.

This is not how biostatistics is applied. Its standard approach is to evaluate a small number of the many components (e.g., survival, or peak walking time, or improvement in LVEF)of this omnibus question one at a time, converting the question of, for example "What is the effect of the intervention on survival time?" into a dichotomous question "Is survival changed by therapy or not?" This bifurcation is modulated by consideration of the confidence interval, but in the end, it is then assessed using statistical hypothesis testing, whose product is a "yes-no" answer and a type I error measurement. Thus, in the end, the classic statistical analysis procedures proffer the combination of 1) a small collection of dichotomous responses to the primary endpoints and 2) an overall type I error expenditure as dispositive.

However, clinical investigators have an abiding interest in neither. Physicians and researchers understand that due to the role of measurement imprecision and sampling error, an effect size that is provided by an endpoint in fact stands for not just one value but for a range of effect sizes. Some of these effect sizes are supportive of benefit, while others – in a different part of the range – are less supportive and may even be consistent with harm. Thus, an effect size range can simultaneously contribute to an argument supporting benefit and also a contention for harm. It is this analysis-generated dualism that researchers require be drawn together into an ensemble of effects from different analyses that would be responsive to the global question. From the investigator's perspective, it is not type I error but the benefit/harm assessment that requires accumulation across analyses. Biostatisticians provide the former and investigators need the latter. This is the disconnect that the measure theoretic approach attempts to repair.

In this manuscript, the broad concepts of set and measure theory have been contoured to address clinical trial evaluations. Specifically a measure, and a collection of measureable functions have been developed with the single goal of incorporating salient features of clinical trial methodology into the realm of formal mathematical analysis and measure theory. The result is a system that is mathematically rigorous, flexible, and can be practically applied.

The computations involved in this system are straightforward. The steps are as follows:

1. Identify the set of analyses $\left\{\omega_i \,/\, \omega_i \subset A_q\right\}$ that address the question *q*.
2. Pre-specify all of the analyses to be conducted and their priority of importance in addressing question *q*.
3. For the set $\left\{\omega_i \,/\, \omega_i \subset A_q\right\}$ using compute $\psi\left(\omega_i\right)$.
4. Compute the set of quanta $\left\{B_{[i]}\right\}$ and for each member, compute $\psi\left(B_{[i]}\right)$.

5. Compute the measure of the body of evidence that addresses question q as by using the quanta as $\mathbf{E}_q = \int_{A_q} d\psi = \psi\left(A_q\right) = \sum_{i=1}^{n} \psi\left(B_i\right) \mathbf{1}_{\omega_i \subset A_q}$ and the strength of evidence for proffered for each analysis as $\frac{\psi\left(B_{[i]} \mathbf{1}_{\omega_i \subset A_q}\right)}{\mathbf{\Gamma_q}}$

6. Compute the evidence for benefit and the evidence for harm $\mathbf{B}_q = \left[\psi\left(A_q\right)\right]^{-1} \int_{A_q} \mathbf{Y}\left(\chi_i^{(b)}\right) d\psi$, $\mathbf{H}_q = \left[\psi\left(A_q\right)\right]^{-1} \int_{A_q} \mathbf{Y}\left(\chi_i^{(h)}\right) d\psi$. and the benefit to harm ratio $\frac{\mathbf{B}_q}{\mathbf{H}_q}$

7. Compute the estimates of benefit $\Lambda_{q\mathbf{B}}\left(\min\right)$ and harm $\Lambda_{q\mathbf{H}}\left(\max\right)$.

The research community can choose the functional form for assessing the intervals of benefit and harm. $\mathbf{Y}\left(\chi_i^{(b)}\right)$ was specifically chosen here in order to reduce the impact of the plausible region of benefit by its length, and increase its impact if its lower boundary was different from that value delineating no effect. There are other choices available though. Triangle functions and scaled beta functions require attention. However, it is best to keep in mind that this system is designed to be relatively easy to use; more complicated forms of $\mathbf{Y}\left(\chi_i^{(b)}\right)$ increase the complexity of the evaluations.

One challenge in the application of this measure theoretic approach is the interpretation of the measure of benefit $\Lambda_{q\mathbf{B}}\left(\min\right)$ and measure of harm $\Lambda_{q\mathbf{H}}\left(\max\right)$. These are derived as unit-less quantities, but the research community has no experience with their interpretation. It is therefore proposed that for the immediate future both the traditional analysis and this measure

theoretic approach be conducted in the evaluation of clinical trials. This would provide calibration for the research community as it works to interpret these new values $\Lambda_{q\mathbf{B}}(\min)$ and $\Lambda_{q\mathbf{H}}(\max)$.

Flexibility of analyses is an advantage of this approach. There is no need to focus on a particular type of estimator. Standardly used estimators, e.g., mean differences, relative risks, Bayes procedures, regression estimates, imputation generated effects, can each be incorporated.

No prior example of the assignment of a formal measure to a clinical analysis (separate and apart from the equivalence of the Lebesgue and Riemann integral when the Riemann integral exists) has been identified. While it seems clear that the "amount" of data on which an analysis relies is an appropriate contributor to the measure of that analysis, defining the measure of $\psi(\omega_i) = n_i v_i$ is not the only definition available, and there are clearly alternative measures that one could apply to the (Ω, Σ) collection of analyses. However, the framework developed here is simple, reasonable, and produces tractable computations.

A requirement of the approach of this manuscript is to ensure that the structure of the measure theoretic framework be permeable to clinical trial design requirements. The importance of priority of analysis is critical in research methodology and therefore is incorporated in the proposed analysis rubric; the size of the independent contribution of analysis ω_i depends on where it lies in the sequence of evaluations. In fact, this measure theoretic approach provides a mathematical justification for the long established practice of selecting high priority analyses in clinical trials; these are the analyses which makes the greatest contribution to the $\psi\left(\bigcup_{i=1}^{n}\omega_i\right)$. By matching the sequence of quanta $B_{[i]}$ to the priority of evaluations chosen by the investigators,

emphases on analyses results $\mathbf{B}_q, \mathbf{H}_q, \mathbf{\Lambda}_{q\mathbf{B}}(\min)$, and $\mathbf{\Lambda}_{q\mathbf{H}}(\max)$ are placed precisely where the investigators have *a priori* stipulated. It is recommended to investigators that this sequence be chosen based on the importance of the analysis in contributing to the understanding of the effect of the exposure.

The investigator determination of analysis priority viewed from a measure theoretic perspective provides new approaches to challenging problems in trial design. For example, it is beyond question that safety evaluations in clinical trials are paramount. However, safety evaluations are not typically part of the type I error control structure in traditionally analyzed clinical trials; for example, type I error is commonly not first accrued for safety, with the remainder being distributed across primary endpoints. The safety analysis lies awkwardly outside the alpha accumulation structure in the traditional paradigm. However, the measure theoretic structure presented in this manuscript permits the safety evaluation to be prioritized first, followed by efficacy evaluations. The impact of the efficacy endpoints would be reduced, but this is wholly consistent with a *primum non nocere* philosophy. In addition any reduced measure that is seen in the primary efficacy evaluations because of the first consideration of safety is partially offset by the accumulation of benefit using $\mathbf{Y}\left(\chi_{t(\omega_i)}^{(b)}\right)$ for the safety evaluation. This approach avoids the analytic disconnect; the safety evaluations are incorporated mathematically and smoothly into the scope of the analyses.

In addition, when viewed from a measure theoretic perspective, the door is open for the investigators to examine different scenarios to optimize the size of $\psi\left(B_{\omega_i}\right)$ for the analyses of most interest. This optimization requires not just concern for sequencing, but for maximizing or minimizing the measure of the intersections between the analyses.

This paper is not an argument for the abandonment of rigor. The discipline that epidemiologists and biostatisticians have helped to instill in investigators is laudable; it is not argued that the stringent execution of a protocol be dismissed. All analyses to be incorporated should be identified prospectively and thoroughly vetted before the research endeavor commences. Endpoint measures should be obtained from state of the art equipment known for their satisfactory precision. If possible, evaluations that would support or refute the purported mechanism of action should be incorporated. While the need for statistical hypothesis testing may be removed when one implements these procedures, the rules of epidemiology and the need for discipline still apply.

An advantage of developing a foundation based on clinical trial methodology for the mathematical interpretation of the trial's results has the advantage of extensibility. The role of exploratory analyses has been problematic for the traditional analysis rubric, in which exploratory analyses are not incorporated into the trial's endpoint analysis. In this measure-theoretic structure, the exploratory analysis ω_i can be incorporated into the final result, but its role in affecting the final result depends on where in the sequence $[i]$ lies. In addition, the integration of analyses that appear from the same trial in separate manuscripts can be accomplished as well, providing an overall picture of the benefit and harm risks posed by the exposure being studied. Nonrandomized, observational studies in health care are amenable to the application of this measure theoretic approach as well, although at this stage of development, neither $\psi-$ measure nor the integrals $\mathbf{B}_q, \mathbf{H}_q, \mathbf{\Lambda}_{q\mathbf{B}}(\min)$, and $\mathbf{\Lambda}_{q\mathbf{H}}(\max)$ explicitly take into account the universe of biases that can vitiate the results of the observational study. Finally, there may be meta-analytic implications of this work.

Statistical hypothesis testing has played an important role in clinical trials. However, the connection between its standard application and the goal of clinical trials is broken. The rubric described here provides a measure theoretic mathematical structure developed specifically for clinical trials, allowing the investigator access to the global result they require and for which they designed the study.

References

1. Fisher, R A. (1925) *Statistical methods for research workers.* Edinburg. Oliver and Boyd.
2. Fisher RA. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*. September 503 - 513.
3. Goodman, S.N. (1999). Toward Evidence–Based Medical Statistics. 1: The *p*-value fallacy. *Annals of Internal Medicine* 130:995–1004.
4. Walker A.M. (1986). Significance tests represent consensus a and standard practice (Letter) *American Journal of Public Health.*76:1033. (See also *Journal erratum* 76:1087.
5. Fleiss J.L. (1986). Significance tests have a role in epidemiologic research; reactions to A.M. Walker. (Different Views) *American Journal of Public Health* 76:559–560.
6. Fleiss J.L. (1986). Confidence intervals versus significance tests: quantitative interpretation. (Letter) *American Journal of Public Health* 76:587.
7. Fleiss J.L. Dr. Fleiss response (Letter) (1986). *American Journal of Public Health* 76:1033–1034.
8. Walker A.M. (1986). Reporting the results of epidemiologic studies. *American Journal of Public Health* 76:556–558.
9 . Multiple risk factor intervention trial. Risk factor changes and mortality results. Multiple Risk Factor Intervention Trial Research Group. JAMA. 1982 Sep 24;248(12):1465-77.
10. Feldman AM, Bristow MR, Parmley WW et.al (1993) Effects of vesnarinone on morbidity and mortality in patients with heart failure. *N Engl J Med* 329:149-55
11. Cohn J, Goldstein SC, Feenheed S et.al. (1998) A dose dependent increase in mortality seen with vesnarinone among patients with severe heart failure. *N Eng J Med* 339:1810-16
12. Pitt B, Segal R, Martinez FA. et al. on behalf of the ELITE Study Investigators (1997). Randomized trial of losartan versus captopril in patients over 65 with heart failure. *Lancet* 349:747–52.
13. Pitt B, Poole-Wilson PA., Segal R, et. al (2000). Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure randomized trial–The losartan heart failure survival study. ELITE II. *Lancet.*355:1582–87.

14. Packer, M., O'Connor, C.M., Ghali, J.K., et al for the Prospective Randomized Amlodipine Survival Evaluation Study Group (1996). Effect of amlodipine on morbidity and mortality in severe chronic heart failure. *New England Journal of Medicine*.335:1107–14.
15. Packer, M. (2000). Presentation of the results of the Prospective Randomized Amlodipine Survival Evaluation-2 Trial (PRAISE-2) at the American College of Cardiology Scientific Sessions, Anaheim, CA, March 15, 2000.
16. Packer M., Bristow MR. Cohn JN. et al. (1996). The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. *New England Journal of Medicine.* 334:1349-55.
17. Moyé LA, Abernethy D. (1996) Carvedilol in Patients with Chronic Heart Failure (Letter). *New England Journal of Medicine.* 335: 1318-1319.
18. Packer M, Cohn JN., Ccolucci WS. Response to Moyé and Abernethy. *New England Journal of Medicine* 335:1318-1319.
19. Fisher L. (1999). Carvedilol and the FDA approval process: the FDA paradigm and reflections upon hypothses testing. *Controlled Clinical Trials* 20:16-39.
20. Fisher LD, Moyé LA. (1999) Carvedilol and the Food and Drug Administration Approval Process: An Introduction. *Controlled Clinical Trials*. 20:1-15.
21. Moyé L.A. (1999) *P* Value Interpretation in Clinical Trials. The Case for Discipline. *Controlled Clinical Trials* 20:40-49.
22. Fisher LD. Carvedilol and the Food and Drug Administration-Approval Process: A Brief Response to Professor Moyé's article. *Controlled Clinical Trials*. 20:50-51.
23. Pfeffer M. A Second Prospective Randomized Amlodipine Survival Evaluation (PRAISE-2). Cardiology Update. te.
24. Pfeffer MA, Skali H. PRAISE (prospective randomized amlodipine survival evaluation) and criticism .JACC Heart Fail. 2013 Aug;1(4):315-7. doi: 10.1016/j.jchf.2013.05.005. Epub 2013 Aug 5. No abstract available. PMID: 24621934

25 . Hare JM, Bolli R, Cooke JP, Gordon DJ, Henry TD, Perin EC, March KL, Michael P. Murphy MP, Pepine CJ, Simari RD, Skarlatos SI, Szady A, Taylor DA, Traverse J, Willerson JT,Vojvodic RW, Yang PC, Moyé L for the Cardiovascular Cell Therapy Research Network (CCTRN). Phase II Clinical Research Design in CardiologyLearning the Right Lessons Too Well: Observations and Recommendations from the Cardiovascular Cell Therapy Research Network (CCTRN). *Circulation* 2013;127:1630-35.

26 Guidance for industry. Cellular Therapy for Cardiac Disease. http://www.fda.gov/BiologicsBloodVaccines/ GuidanceComplianceRegulatoryInformation/ Guida nces/default.htm.