When are randomized trials unnecessary? A signal detection theory approach to approving new treatments based on non-randomized studies

Benjamin Djulbegovic[1], Marianne Razavi[2], and Iztok Hozo[3]

[1]City of Hope National Medical Center
[2]Affiliation not available
[3]Indiana University Northwest

June 17, 2020

Abstract

Rationale, aims and objectives New therapies are increasingly approved by regulatory agencies such as the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) based on testing in non-randomized clinical trials. These treatments have typically displayed "dramatic effects" (i.e., effects that are considered large enough to obviate the combined effects of bias and random error). The agencies, however, have not identified how large these effects should be to avoid the need for further testing in randomized controlled trials (RCTs). We investigated the effect size that would circumvent the need for further RCTs testing by the regulatory agencies. We hypothesized that the approval of therapeutic interventions by regulators is based on heuristic decision-making whose accuracy can be best characterized by the application of signal detection theory (SDT). Methods We merged the EMA and FDA database of approvals based on non-RCT comparisons. We excluded duplicate entries between the two databases. We included a total of 134 approvals of drugs and devices based on non-RCTs. We integrated Weber-Fechner law of psychophysics and recognition heuristics within SDT to provide descriptive explanations of the decisions made by the FDA and EMA to approve new treatments based on non-randomized studies without requiring further testing in RCTs. Results Our findings suggest that when the difference between novel treatments and the historical control is at least one logarithm (base 10) of magnitude, the veracity of testing in non-RCTs seems to be established. Conclusion Drug developers and practitioners alike can use the change in one logarithm of effect size as a benchmark to decide if further testing in RCTs should be pursued, or as a guide to interpreting the results reported in non-randomized studies. However, further research would be useful to better characterize the threshold of effect size above which testing in RCTs is not needed.

When are randomized trials unnecessary? A signal detection theory approach to approving new treatments based on non-randomized studies

Running head: When are randomized trials unnecessary?

Benjamin Djulbegovic[a,b] , Marianne Razavi[a,b] and Iztok Hozo[d]

[a] Department of Supportive Care Medicine,[b] Department of Hematology, [c]Evidence-based Analytics and Comparative Effectiveness Research, City of Hope, 1500 East Duarte Rd, Duarte, CA

[d] Department of Mathematics, Indiana University, Gary, IN

Word count:

Abstract: 299

Main text: 3,712

Corresponding author :

Benjamin Djulbegovic, MD, PhD

City of Hope

1500 East Duarte Rd.

Duarte, CA. 91010

+1 626 218-7502

E-mail: bdjulbegovic@coh.org

Key words: randomized trials - observational studies- large effects- decision making-signal detection theory-Weber-Fechner law – priority heuristic

Abstract

Rationale, aims and objectives

New therapies are increasingly approved by regulatory agencies such as the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) based on testing in non-randomized clinical trials. These treatments have typically displayed "dramatic effects" (i.e., effects that are considered large enough to obviate the combined effects of bias and random error). The agencies, however, have not identified how large these effects should be to avoid the need for further testing in randomized controlled trials (RCTs). We investigated the effect size that would circumvent the need for further RCTs testing by the regulatory agencies. We hypothesized that the approval of therapeutic interventions by regulators is based on heuristic decision-making whose accuracy can be best characterized by the application of signal detection theory (SDT).

Methods

We merged the EMA and FDA database of approvals based on non-RCT comparisons. We excluded duplicate entries between the two databases. We included a total of 134 approvals of drugs and devices based on non-RCTs. We integrated Weber-Fechner law of psychophysics and recognition heuristics within SDT to provide descriptive explanations of the decisions made by the FDA and EMA to approve new treatments based on non-randomized studies without requiring further testing in RCTs.

Results

Our findings suggest that when the difference between novel treatments and the historical control is at least one logarithm (base 10) of magnitude, the veracity of testing in non-RCTs seems to be established.

Conclusion

Drug developers and practitioners alike can use the change in one logarithm of effect size as a benchmark to decide if further testing in RCTs should be pursued, or as a guide to interpreting the results reported in non-randomized studies. However, further research would be useful to better characterize the threshold of effect size above which testing in RCTs is not needed.

INTRODUCTION

A central aspect of scientific inference is distinguishing coincidence from cause and effect. David Hume – one of the greatest philosophers of all times – maintained that this is empirically impossible.[1] Methodological development during last 50-70 years have identified randomization as a necessary condition to establish the relationship between cause and effect[2-4], such as ascertaining treatment effects tested in randomized controlled trials (RCTs).[5] Since the early 1960's [3,6], regulators such as Food and Drug Administration (FDA) and later on, the European Medicines Agency (EMA) have typically required RCTs to approve new treatments for use in medical practice. However, developments in basic science and signals can be identified

without testing in RCTs. [7]An unintended consequence of these developments is that a formal mechanism for rigorous cause-effect inferences is removed and, by and large, cannot be reliably compensated by using techniques based on non-randomized comparisons.[8] Nevertheless, under some circumstances, the therapeutic signal observed in non-RCTs testing is accepted as "truthful".[9] This, for example, occurs when effects of treatments are so large ("dramatic") that they are believed to override the combined effects of biases and random errors that potentially affect the study's results.[9] The mechanism for accepting such results as "true" resides in the way our minds distinguish these signals as "true" from "false". A theoretical framework for such inferences can be postulated within the heuristic theory of decision-making[10,11] linked with signal detection theory (SDT).[12,13] Indeed, when formal methods such as RCTs are not available people resort to heuristics.[10] Here we argue that inference about dramatic effects is driven by two heuristics: first, the large effects must be recognized (recognition heuristic) and second, the magnitude of that effect to cross the decision threshold reflects the heuristic according to the Weber-Fechner law.

In this paper, we propose how SDT-related heuristics can be used to interpret the results of non-RCT comparisons of drug approvals by the FDA and EMA. We believe that wider familiarity with these principles has important educational values for trainees, practicing physicians, researchers, and policy-makers.

METHODS

The regulators have accepted non-randomized studies as the basis for licensing of new treatments

A notion that that large treatment effects can sometimes obviate the need for data from RCTs has also been accepted by the FDA and the EMA. The agencies established special pathways-the EMA's PRIME (Priority Medicines) and Adaptive Pathways programs and the FDA's "Breakthrough Therapy Designation" programs - designed to support approval of drugs that demonstrate substantial improvement over the existing therapies, which may not require further testing in RCTs. [14] We recently analyzed drug approvals by the EMA and FDA based on non-randomized drug comparisons and confirmed that larger effect sizes in non-randomized studies are associated with higher rates of licensing approval.[15,16] Overall, we found that between 7 to 10% of drug approvals are based on non-RCT comparisons and that between 2% and 4% of these approvals displayed "dramatic" effects[15,16], defined as relative risks (RR)>2[8], RR [?]5 [17], or RR[?]10.[9]

Although the probability of approval increased with larger effect sizes, we could not identify a specific threshold of effect size above which the regulatory agencies would always grant licensing approval.[15,16] Similarly, to date, the agencies have not formally quantified the effect size necessary to preclude the need for additional testing in RCTs before granting licensing approval. Thus, the treatment effect size that is dramatic enough to convince the FDA and EMA that the treatment differences observed are real and free of bias and random error and hence can be approved without RCTs remain unclear. The proposed definitions of dramatic effects- (RR)>2[8], RR [?]5 [17], or RR[?]10[9]- have not been theoretically or empirically justified; instead, they represent a decision rule based on*heuristics* - powerful, rule-of-thumb, decision-making strategies that are often more accurate than complex statistical models.[10,11,18]

The heuristic theory of decision-making has been linked to SDT[12,13] and the threshold model of decision-making[12] to show how seemingly unrelated theories in different disciplines can lead to discovery of new relationships and explanations. Here, we argue that when the direct, empirical evaluation of a treatment effect is not possible, an alternative approach is to employ SDT to define the circumstances under which the "signal" (e.g., treatment effect) is credible and reproducibly detected to allow approval of new drugs without further testing in RCTs.[19,20] [21]

We rely on the generalizability of SDT to account for two heuristics that we believe influence the FDA and the EMA's approval of new treatments: 1) the likelihood of approval without testing in RCTs will increase if the difference in treatment effect between the experimental and control arm is at least one logarithm of magnitude (i.e., reflecting heuristic based on the *Weber-Fechner law*) [2021] [22]; 2) the specific threshold of effect size above which the agencies will not require further RCTs will reflect heuristic known as *recognition heuristic* .[23] We focus on the effect size heuristics, but also discuss the heuristic related to the use of p-values. Recently, there has been an increasing attempt to modify a century-old inferential rule-of-thumb to reject

the null hypothesis at p[?]0.05[24]; some authors vehemently oppose a hard cut-off for p-values, [25] while others propose a new heuristic rule of p[?]0.005 [24] as an acceptable evidentiary standard against a null hypothesis.

EMA and FDA databases of drug approvals based on non-RCT comparisons

To provide empirical support for our theoretical framework, for this analysis we merged the EMA and FDA database of approvals based on non-RCT comparisons[15,16], as described above. We merged both databases because of the high concordance rate (91-98%) in approval [26] between the two agencies. We excluded duplicate entries between the two databases, and reviewed all regulatory approval documents that addressed whether testing in subsequent RCTs is required. We included a total of 134 approvals of drugs and devices that were based on non-RCTs. For 35 of these treatments, the regulators required further evidence from RCTs, whereas for 99 treatments they did not. Although this is the best, contemporary dataset available[15],[16], it is important to note that the agencies often failed to provide explicit comparisons of these drugs and devices against comparators, arguing that in many cases "*efficacy has been assessed on the basis of [outcomes] in comparison to what would be expected by expert clinical evaluation and by comparison with previous experience in this type of patient*".[15] Indeed, the evaluation of treatment effects always depends on the comparison of experimental (direct or counterfactual) with a control intervention if one is to estimate the effect size. Therefore, when the agencies did not specifically provide comparison data, we imputed the control events either based on our interpretation of the agencies' judgments documented in the approval reports or the best available data available in the literature.[15,16] However, in our attempts to translate the FDA and EMA judgments into the effect sizes, we frequently imputed very low (often equal to zero) event rates such as response rate or survival in the control arm. As a result, we observed some empirically improbable high effect sizes. Nevertheless, our estimates seem to reflect what the agencies often believed – "*without new treatments, most patients would surely die* " [15] – implying that these effects are indeed considered self-evidently large, dramatic effects and hence confirming the role of heuristics in the decision-making process of treatment approvals.

Weber-Fechner law

In the early 19$^{\text{th}}$ century, psychologist Ernst Weber noted that in order for people to notice a given stimulus, the amount of the stimulus must increase (or decrease) by *a fraction* of its physical intensity to yield *"just noticeable differences" (jnd)* .[19,21] For example, Weber found that people could not discriminate between 20.5 and 20.0 g weights but could usually discriminate between 21 and 20 g.[20] When baseline weight was 40, 60, 80, and 100 g, the required increased in stimulus to represent *jnd* was 42, 63, 84, and 105 g, respectively.[20] That is, to appreciate the differences between weights *(jnd),* the weight (i.e., stimulus) should increase by at least 5% of the original weight.[20]Gustav Fechner, another 19$^{\text{th}}$ century psychologist, proposed that *"jnd"* can be conceptualized as units of psychological intensity instead of physical intensity.[19-21] Subsequently, the relationship between the intensity of a signal and how much more intense the signal needs to increase before a person can reliably tell that the signal has truthfully changed has become known as the *Weber-Fechner Law* of psychophysics. As expected for physiological systems, the law is valid within certain domains of stimulus-response ratios. In other words, it is approximately correct for a wide variety of sensory dimensions, although it may deviate at the extremes of the spectrum of stimuli.[20,21] Other authors have tried to improve upon the Weber-Fechner Law. Notably, Stevens[27] proposed a power law according to which a relationship between stimulus intensity and the magnitude of sensation can be plotted on a log-log axis as a straight line with a slope of the exponent.

Regardless of the exact mathematical description of the relationship between *"jnd"* and stimulus, the reproducible relationships between signal and perception was subsequently documented in other fields as well [28]: from influencing human behaviors by specific marketing stimuli[29] to the way people experience the value of money[19] to making risky choices [30] to the mental line for numbers.[22] In addition, psychological research[31] has demonstrated that people often use a simple heuristic in decision-making, based on *the prominent numbers* as powers of 10- defined as the powers of ten, their doubles, and their halves [e.g., 1,2, 5, 10, 20, 50, 100, 200...] that approximate the Weber-Fechner function. [31] [32]For example, when presented with monetary choices, people often make judgments according to the "1/10 aspiration level" (rounded to the closest

number) in such a way that if the gains, losses, or probabilities change by one order of magnitude (or more), they will stop further examination of the observed results and accept the findings.[31] Thus, the most common heuristics defined in the literature to categorize "dramatic" effects as RR>2 [8], RR [?]5[17], or RR[?]10[9] appear to be consistent with the Weber-Fechner law. However, a fundamental property of the Weber-Fechner law is that "*jnd*" occurs only when the increase or change in stimuli are a*constant* percentage of the stimulus itself[19,21]; this is also directly applicable to evaluation of treatment effects. Indeed, treatments effects are commonly assumed to remain constant over a range of predicted risks,[33]providing further justification for the application of Weber-Fechner law to assess the likelihood of approval of new therapeutics without testing them in RCTs. Appendix 1 demonstrates how a stimulus (effect size) and response [probability of not requiring further RCTs when$jnd = \log(OR)$] is derived from the Weber-Fechner law as:

$$\text{logit}\left(p\left(\text{no}n_{\text{RCT}}\right)\right) = A * \log\left(\text{OR}\right) + B$$

where OR is odds ratio, A and B are fitted constants, respectively.

It is important to note that the response – i.e., not requiring further testing in RCTs – is not linearly related to the size of treatment effect (i.e., OR), but rather to the *logarithm* of the effect size [log(OR)].

A signal detection analysis of "dramatic effects" based on recognition heuristics

Recognition is a natural mechanism for making inferences and solving problems – if the object or phenomenon is not recognized, further ascertainment and reasoning processes cannot proceed.[34] One strategy that relies on using recognition to make inferences is called the *recognition heuristic.*[35] Recognition heuristic has been demonstrated to provide accurate answers in a wide range of circumstances, particularly when information is limited and uncertain. It is considered to have a special status in our cognitive capabilities because, as explained, if the object is not recognized then it becomes impossible to draw any inferences.[35] Therefore, if the regulators do not recognize effect size as a criterion for making decisions on whether to request further RCTs, then effect size would not play a role in their decision-making. However, as discussed, we have demonstrated ecological validity between effect size and decisions whether to require further testing in RCTs: larger effect sizes were associated with a greater likelihood of approval based on nonrandomized data. [15,16] Thus, the magnitude of effect size serves as a recognition heuristic related to the decision to approve drugs based on non-randomized studies. Nonetheless, the specific decision will depend on beliefs (stemming from familiarity) that the effect size exceeding a certain threshold (T) (e.g., RR>2, 5, 10) is sufficient to obviate testing in further RCTs.

Use of recognition heuristics, like any other decision rule, may result in correct and incorrect inferences. [35] In turn, analyzing the proportion of correct inferences based on recognition heuristics lends itself to inquiry within a framework of SDT.[12,36] SDT resides on the notion that the two possible events (*signal* , e.g. treatment effect is "true"), and *noise* , e.g. treatment effect is not "true") have overlapping distributions on a given observation scale. Each of these distributions is further divided into two possible outcomes, which are determined by setting a decision criterion. The criterion divides the signal distribution into true positives (hits, or sensitivity) and false negatives (misses). The noise distribution is composed of true negatives (correct rejections, or specificity) and false positives, respectively.[37] To assess the accuracy of recognition heuristic of a continuous variable such as effect size, we assume that judges have a criterion set at one point along the possible values of their prior beliefs, which in our case corresponds to treatment effect size, $TE = (OR)$. If the TE exceeds the given threshold (T) consistent with the judges prior beliefs, the rule to activate recognition heuristic can formally be stated as:[36]

$$If\ TE = \ OR \geq T\ ,\ then\ "Approve\ without\ further\ testing\ in\ RCTs"$$

$$If\ TE = \ OR < T,\ then\ "\ Approve\ with\ request\ for\ further\ testing\ in\ RCTs"$$

Using the previously defined frameworks for integrating heuristic decision-making with SDT[12,36], for each possible cutoff value of $TE = (OR)$, we can calculate standard SDT statistics[36,37] including sensitivity, specificity, overall accuracy, positive predictive value (PPV) (i.e.*recognition validity*) and *d'* (discriminability). In turn, we define the optimal *recognition heuristic* as the maximum TE threshold for the largest *d'* value. [Note that *d'*(discriminability) represents the standardized distance between the signal i.e., no further RCTs needed and noise (further RCTs required) distributions and is defined as:

$$d' = zHit \quad zFA,$$

where z Hit and z FA are the z -scores of the true positive and the false alarm rate, respectively].

Results

Weber-Fechner law

When the Weber-Fechner equation is applied to the EMA-FDA data, there is a highly significant association between the logarithm of treatment effects and the likelihood that the regulators will not require further testing in RCTs (Table 1). To illustrate the significance of this finding, it is instructive to calculate differences in the change of treatment effects (stimuli):

$$\text{logit}\,(p) = 0.831 \cdot TE_{10} + 0.32$$

which means that an increase of $TE_{10} =$ ORby 1 leads to an increase in the *probability* by$\frac{e^{0.831}}{1+e^{0.831}} \approx 0.69$

Consistent with Weber-Fechner law, we found that if OR increases by one, the probability of not requiring further RCTs increases by 69%, which, we contend, is very large probability rarely observed in decision-making literature. Thus, our analysis suggests that the difference between new treatments and historical controls should be at least one logarithm of magnitude larger to omit subsequent requests for RCTs. (Figure 1). Strictly speaking, some would argue, evidentiary support in favor of this hypothesis is moderately strong at p=0.007, but reaches the recently recommended heuristic cut-off at p=0.005 [24] when 5 observations of treatment effects with empirically implausible large values (OR>250) are removed from the analysis (not shown).

Recognition heuristics

How accurate is the recognition heuristic? Figure 2a shows the relationship between recognition validity (i.e., PPV) and discriminability (i.e., *d'*). As expected, the greater *d'* , the greater the separation between signal and noise. However, the degree of separation is relatively modest, with maximum*d'* ˜1.0. Fig 2b displays the relationship between*d'* and *TE* value. The optimal cut-off is approximately*log(OR)* of 1, which is equivalent to *jnd* as per the Weber-Fechner law. Similarly, the largest evidentiary support is found for *TE* of about 1(=1.17) at moderate p=0.033. P-values for larger TE, which theoretically should have higher recognition validity, has not reached traditional statistical evidence at p<0.05 (Fig 2c), which probably explains why larger treatment effects are not invariably associated with drug approval without testing in RCTs. In fact, only 11 observations were associated with effect size exceeding OR > 2.

Discussion

We sought to address one of the most important clinical question in contemporary clinical research: how large of an effect size is large enough to allow approval of treatments without further testing in RCTs? To address this question, we illustrate the application of the signal detection and heuristic theory of decision-making to interpret the effect size that regulatory agencies may use to approve treatment without further testing in RCTs. We propose that two heuristics can explain the agencies' decision-making: first, signal is*recognized* as large, and second, the *magnitude of that signal* is assessed via the Weber-Fechner law. Our findings suggest that when the difference between novel treatments and historical controls is at least one

logarithm of magnitude, the veracity of testing in non-RCTs seems to be established.[38] These findings based on the convergence of the Weber-Fechner and recognition heuristics agree with the heuristic rule suggested by Glasziou et al[9]: further RCTs may not be necessary when RR of experimental treatment is [?] 10 in comparison with control.

Theories of decision making are divided into those that deal with 'large-' or 'small'- world phenomena.[39] In a small world, time constraint is not an issue, decision-makers have access to the best available evidence – ideally from well-designed and powered RCTs – regarding all competing management alternatives, consequences and probabilities. Signal detection theory is a prototype of the small world theories and is a normative theory that provides a framework for how people *"should" or "ought to"* make their decisions and draw inferences. (This is also known as the theory of *"ought"*). [40-42]

In contrast, in a 'large' or real-world context, decision-makers are typically under time constraints, with limited knowledge about the complete set of alternatives, consequences, and probabilities. This means that making rational inferences requires adaptation to environment/context (*adaptive or ecological rationality*) and respecting epistemological, environmental and computational *constraints* of human brains.[11,40] Because finding the optimum solution to a given problem can be resource and computationally intensive, adaptive behaviors typically rely on *satisficing* (finding a good enough solution), rather than striving to find a "perfect" solution (via optimising/maximizing procedures). [40,43] The principle behind satisficing is that there must exist *a point (threshold)* at which obtaining more information or engaging in more computation becomes overly costly and thereby detrimental. Identifying this threshold, at which a decision-maker should stop searching for more information, is often accomplished by using "heuristics"[11] for implementation of bounded rationality.[44] The heuristic theory of decision-making is a descriptive theory, which helps explain how people *actually* make their decisions (also known as theory of *"is"*). [40-42] Surprisingly, simple heuristic-based inferential and decision-making strategies are often more accurate than more complex statistical models (the phenomenon known as "less-is-more").[11]

Recently, we[12] and others[13,36]integrated small-world SDT with heuristics decision-making to show how connecting apparently unrelated theories in different disciplines likely leads to discovery of new relationships. In this paper, we extend the theory integration program[45] to the application of the Weber-Fechner law and recognition heuristics in order to provide descriptive explanations of the decisions made by the FDA and EMA to approve new treatments based on non-RCT studies without further testing in RCTs. By integrating heuristic reasoning with SDT, it is sometimes possible to derive *"ought"* rule from *"is"* observations.[40-42,46] That is, if had observed high discriminability of Weber-Fechner, or recognition heuristic, we may then argue that these empirically derived observations may, in turn, be normatively used by drug developers and practitioners alike: one log effect size magnitude could serve as a benchmark to decide if further testing in RCTs should be pursued, or as a guide in interpretation of the results reported in non-RCT studies.

Throughout this study, we found some support for "one logarithm of treatment magnitude" rule, but we should acknowledge the study limitations. First, the strength of evidence supporting high accuracy related to the decision to pursue further RCTs based on the one log effect size is moderate. Second, as discussed in the papers leading to this one[15,16], in addition to effect size, other factors play a role in the decision to grant licensing approval; these seem to include issues such as approval for rare diseases where few effective treatment exists, risk tolerance in the attempt to strike a balance between failing to approve effective drugs and approving ineffective or dangerous drugs [47], political pressures like conflict of interest, feasibility of undertaking of RCTs, small sample sizes and bias in the assessment of control event rates, as outlined above.

Nevertheless, it is clear that the larger the effect size, the higher the probability that treatments will be approved without further testing in RCTs. [15,16] When integration of multiple factors are difficult people resort to heuristics, which are often defining characteristics of psychology of decision-making. [10]However, one of the reasons that we could not provide more definitive evidence related to the specific effect size above which drugs should be approved based solely on non-RCT data is that our database, even most comprehensive to date, is relatively small (n=134). Mere exposure to "dramatic effects" does not account for the mechanism of recognition heuristic.[35] Rather, repeated experience and internalization of the rule is required for the ease of

retrieval to rely on recognition for making inferences from memory about the phenomenon of interest.[35] We suspect that as databases- and experience- with approval of drugs based on non-RCTs increase, regulators and practicing physicians will encounter many more instances that will help improve the quality of recognition memory and the use of the methods described here will be more applicable.

Conclusions

We sought to address one of the most crucial clinical question in contemporary clinical research: how large of an effect size is large enough to circumvent the need for further RCTs testing by the regulatory agencies? Our results suggest that drug developers can use the change in one logarithm of effect size as a benchmark to decide if further testing in RCTs should be pursued, or as a guide to interpreting the results reported in non-randomized studies. Further research would be helpful to better characterize the threshold of effect size above which testing in RCTs is not needed.

Acknowledgements, Funding and Conflict of Interest Disclosures:

Authors would like to report that their study was supported by grant :

616210515 from GlaxoSmithKlines (BD and MR)

Author statement

The conception and design of the study were completed by B.D. and I.H. Acquisition of data was done by B.D. and M.R. Analysis of the data was done by B.D., I.H and M.R. Interpretation of data was handled by B.D., I.H and M.R. The manuscript was completed by B.D. with critical revisions done by all authors. All authors gave final approval of the completed manuscript.

Figure 1 : The likelihood that testing in further RCTs will not be required by the regulatory agencies reflects the Weber-Fechner law: the higher the signal [effect size, here expressed as logarithm base 10 of odds ratio, [log(OR)], the greater the probability of the treatment approval without requirement for conducting confirmatory RCT.

Figure 2: 2a) a relation between *recognition validity*(expressed as positive predictive value, PPV) and *d'* [i.e., the standardized distance between distributions of the signal (i.e., no further RCTs needed and noise (further RCTs required)]; 2b) a relationship between *d'* and *"recognition heuristic"*effect size *threshold value* . Optimal cut-off is about log(OR) of 1; 2c) the largest evidentiary support is found for effect size of 1.11 at moderate evidentiary support p=0.033. p values for larger thresholds have not reached traditional statistical evidence at p<0.05. (Fig 2c).

Appendix

Derivation of a relation between stimulus (effect) size and response [probability of not requiring further RCTs whenjnd $= \log(\mathbf{OR})$**] based on analogy to Weber-Fechner law**

Let s be the magnitude of a measurable stimulus and Δs the increase in stimulus just required to discriminate between stimuli as:[21]

$$r = \frac{\Delta s}{s} = constant$$

This means that the noticeable difference in sensation occur only when the increase, or change in stimuli (such as change in the magnitude of effect due to the experimental treatment compared with control treatment) are a **constant** percentage of the stimulus (s) itself. This is Weber's law.

Fechner proposed a method of scaling that takes Weber's law into account; let s_o be a fixed value of s to

allow us to calculate the nearest noticeably higher stimulus as[21]:

$$r = \frac{s_1 - s_o}{s_o} \text{ or } r \cdot s_o = s_1 - s_o$$

$$s_1 = s_o + r \cdot s_o = s_o \cdot (1 + r) = s_o \cdot q \text{ where } q = 1 + r$$

So the "next" stimulus is q times the previous level of "noticeable" stimulus. Similarly

$$s_2 = s_1 \cdot q = s_o \cdot q \cdot q = s_o \cdot q^2 \quad \text{and} \quad s_3 = s_2 \cdot q = s_1 \cdot q^2 = s_o \cdot q^3$$

which leads to general equation

$$s_n = s_o \cdot q^n$$

Which, after taking a natural logarithm (or logarithm base 10) is analogous to

$$n = A \cdot \ln s + B$$

where $A = \frac{1}{\ln q}$ and $B = -\frac{\ln s_o}{\ln q}$.

In our case, we actually only want to distinguish whether the difference in stimulus is noticeable, i.e., we are not interested in the absolute size of the stimulus, but rather whether the ratio $r$is large enough.

We express the ratio as effect size measured in terms of odds ratio (OR), or proportional reduction of OR (1-OR, in terms of reducing bad events); or, increase in terms of improving good outcomes (OR-1). These effects are commonly assumed to remain constant over the range of predicted risks[33], providing further justification for application of Weber-Fechner law:

$$r = \frac{\Delta s}{s} = \frac{ODDSexp - ODDSctrl}{\text{ODDctrl}} = \frac{\text{ODDSexp}}{\text{ODDSctrl}} - 1 = OR - 1$$

and

$$q = (1 + r) = (1 + OR - 1) = OR$$

where ODDSexp and ODDSctrl, are odds of events related to the experimental and control treatment, respectively.

Rather than using the size of the stimulus (n), we are using the size of the ratio (r or q) as the independent variable. Since this is a ratio, the use of logarithmic scale is more appropriate, and rather than modeling the scale or intensity of the stimulus (n) as in the Weber's law, we are modeling whether the size of this ratio will indicate a distinguishable difference :

$$\text{diff (yes or no)} = A \cdot \ln(q) + B = A \cdot \ln(\text{OR}) + B$$

Since in this case, this difference can only take two values (0=no=further RCTs are required or 1=yes=further RCTs are not required), we employ logistic regression[48] to obtain probability of not requiring further RCT:

$$\text{logit}\left(P\left(\text{no}n_{\text{RCT}}\right)\right) = A \cdot \ln(\text{OR}) + B$$

Similar expression can be derived using relative risks instead of OR.

References

1. Hume D. *Philosophical Essays Concerning Human Understanding*London:Millar; 1748.

2. Steel D. Causal Inference and Medical Experiments. In: Gifford. F, ed. *Handbook of the Philosophy of Science. Volume 16: Philosophy of Medicine. (Gabbay DM, Thagard P, Woods J, eds).* Vol 16. London: Elsevier BV.; 2010.

3. Djulbegovic B, Guyatt GH, Ashcroft RE. Epistemologic inquiries in evidence-based medicine. *Cancer Control.* 2009;16(2):158-168.

4. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet.* 2017;390(10092):415-423.

5. Hill AB. The clinical trial. *N Engl J Med.* 1952(247):113-119.

6. Matthews JR. *Quantification and quest for medical certainty.*Princeton, NY: Princeton University Press; 1995.

7. Micheel CM, Nass SJ, Omenn GS, eds., eds. *Evolution of Translational Omics: Lessons Learned and the Path Forward.*Washington,DC: The National Academies Press; 2012.

8. Collins R, Bowman L, Landray M, Peto R. The Magic of Randomization versus the Myth of Real-World Evidence. *New England Journal of Medicine.* 2020;382(7):674-678.

9. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ.*2007;334(7589):349-351.

10. Gigerenzer G, Brighton H. Homo heuristicus: why biased minds make better inferences. *Top Cogn Sci.* 2009;1(1):107-143.

11. Gigerenzer G, Hertwig R, Pachur T, eds. *Heuristics. The foundation of adaptive behavior.* New York: Oxford University Press; 2011.

12. Hozo I, Djulbegovic B, Luan S, Tsalatsanis A, Gigerenzer G. Towards theory integration: Threshold model as a link between signal detection theory, fast-and-frugal trees and evidence accumulation theory. *J Eval Clin Pract.* 2017;23(1):49-65.

13. Luan S, Schooler LJ, Gigerenzer G. A signal-detection analysis of fast-and-frugal trees. *Psychol Rev.* 2011;118(2):316-338.

14. Sherman RE LJ, Shapley S, Robb M, Woodcock J: Expediting. Expediting Drug Development — The FDA's New "Breakthrough Therapy" Designation. . *New England Journal of Medicine.* 2013(369(20):1877-1880).

15. Djulbegovic B, Glasziou P, Klocksieben FA, et al. Larger effect sizes in nonrandomized studies are associated with higher rates of EMA licensing approval. *Journal of Clinical Epidemiology.*2018;98:24-32.

16. Razavi M, Glasziou P, Klocksieben FA, Ioannidis JPA, Chalmers I, Djulbegovic B. US Food and Drug Administration Approvals of Drugs and Devices Based on Nonrandomized Clinical Trials: A Systematic Review and Meta-analysis. *JAMA Network Open.* 2019;2(9):e1911111-e1911111.

17. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence-publication bias. *J Clin Epidemiol.* 2011.

18. Gigerenzer G, Todd P, ABC-Research-Group. Simple heuristics that make us smart. 1999.

19. Hastie R, Dawes RM. *Rational choice in an uncertain world.2nd edition.* Los Angeles: Sage Publications, Inc.; 2010.

20. Lanzara RG. Weber's law modeled by the mathematical description of a beam balance. *Math Biosci.* 1994;122(1):89-94.

21. Batschelet E. *Introduction to Mathematics for Life Scientists.* New York: Springer-Verlag; 1979.

22. Dehaene S. The neural basis of the Weber-Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences.*2003;7(4):145-147.

23. Pleskac TJ. A signal detection analysis of the recognition heuristic. *Psychon Bull Rev.* 2007;14(3):379-391.

24. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician.*2016;70(2):129-133.

25. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature.* 2019;567(7748):305-307.

26. Kashoki M, Hanaizi Z, Yordanova S, et al. A Comparison of EMA and FDA Decisions for New Drug Marketing Applications 2014-2016: Concordance, Discordance, and Why. *Clin Pharmacol Ther.*2020;107(1):195-202.

27. Stevens SS. Neural events and the psychophysical law.*Science.* 1970;170(962):1043-1050.

28. Rescher N. *Epistemetrics.*

. New York: Cambridge University Press; 2006.

29. Thaller RH. *The winner's Curse. Paradoxes and Anomalies of Economic Life.* Princeton, New Jersey: Princeton University Press; 1992.

30. Kacelnik A, Brito e Abreu F. Risky Choice and Weber's Law.*Journal of Theoretical Biology.* 1998;194(2):289-298.

31. Brandstatter E, Gigerenzer G. The Priority Heuristic: Making Choices Without Trade-Offs. *Psychological Review* 2006;113:409-432.

32. Converse BA, Dennis PJ. The role of "Prominent Numbers" in open numerical judgment: Strained decision makers choose from a limited set of accessible numbers. *Organizational Behavior and Human Decision Processes.* 2018;147:94-107.

33. Kent DM, Hayward RA. Limitations of Applying Summary Results of Clinical Trials to Individual Patients: The Need for Risk Stratification. *JAMA.* 2007;298(10):1209-1212.

34. Simon HA. Invariants of human behavior. . *Annual Review of Psychology.* 1990;41:1-19.

35. Pachur T, Todd P, Gigerenzer G, Schooler L, Goldstein D. The Recognition Heuristic: A Review of Theory and Tests. *Frontiers in Psychology.* 2011;2(147).

36. Pleskac TJ. A signal detection analysis of the recognition heuristic. . *Psychon Bull Rev* 2007;14(3):379-379.

37. Stanislaw H, Todorov N. Calculation of signal detection theory measures. *Behav Res Methods Instrum Comput.* 1999;31(1):137-149.

38. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ.*2007;334(7589):349-351.

39. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz L, Woloshin S. Helping doctors and patients make sense of health statistics.*Psychol Sci Public Interest.* 2007;8(2):53 - 96.

40. Djulbegovic B, Elqayam S. Many faces of rationality: Implications of the great rationality debate for clinical decision-making. *J Eval Clin Pract.* 2017;23(5):915-922.

41. Elqayam S. Grounded rationality:descriptivism in epistemic context.*Synthese.* 2012;189:39-49.

42. Elqayam S, Evans JSBT. Subtracting 'ought' from 'is': Descriptivism versus normativism in the study of human thinking. . *Behavioral and Brain Sciences.* 2011;34:233-248.

43. Simon HA. *Models of man: social and rational. Mathematical essays on rational human behavior* New York: Wiley; 1957.

44. Katsikopoulos KV, Gigerenzer G. One-reason decision-making: Modeling violations of expected utility theory. *Journal of Risk and Uncertainty.* 2008;37(1):35.

45. Gigerenzer G. A Theory Integration Program. *Decision.*2017;81(3):133–145.

46. Elqayam S, Thompson VA, Wilkinson MR, Evans JS, Over DE. Deontic introduction: A theory of inference from is to ought. *J Exp Psychol Learn Mem Cogn.* 2015;41(5):1516-1532.

47. Djulbegovic B, Hozo I. When Should Potentially False Research Findings Be Considered Acceptable? *PLoS Medicine.* 2007;4(2):e26.

48. Wileyto EP, Audrain-McGovern J, Epstein LH, Lerman C. Using logistic regression to estimate delay-discounting functions. *Behavior Research Methods, Instruments, & Computers.* 2004;36(1):41-51.

Hosted file

`Table 1.docx` available at https://authorea.com/users/334179/articles/460226-when-are-randomized-trials-unnecessary-a-signal-detection-theory-approach-to-approving-new-treatments-based-on-non-randomized-studies