ELSEVIER

Seminars in
HEMATOLOGY

# Decision-Making When Data and Inferences Are Not Conclusive: Risk-Benefit and Acceptable Regret Approach

Iztok Hozo,[a] Michael J. Schell,[b] and Benjamin Djulbegovic[b]

**The absolute truth in research is unobtainable, as no evidence or research hypothesis is ever 100% conclusive. Therefore, all data and inferences can in principle be considered as "inconclusive." Scientific inference and decision-making need to take into account errors, which are unavoidable in the research enterprise. The errors can occur at the level of *conclusions* that aim to discern the truthfulness of research hypothesis based on the accuracy of *research evidence and hypothesis,* and *decisions,* the goal of which is to enable optimal decision-making under *present and specific* circumstances. To optimize the chance of both correct conclusions and correct decisions, the synthesis of all major statistical approaches to clinical research is needed. The integration of these approaches (frequentist, Bayesian, and decision-analytic) can be accomplished through formal risk: benefit (R:B) analysis. This chapter illustrates the rational choice of a research hypothesis using R:B analysis based on decision-theoretic expected utility theory framework and the concept of "acceptable regret" to calculate the threshold probability of the "truth" above which the benefit of accepting a research hypothesis outweighs its risks.**
**Semin Hematol 45:150-159** 

In 1960, the statistician John Tukey[1] warned about the need to clearly separate "conclusions" from "decisions." Tukey pointed out that scientific "conclusions" are concerned with establishment of "truth," while "decisions" deal with consequences of specific actions in specific circumstances. Conclusions are judged by the truthfulness under formalized inferential assumptions without regard to the consequences of specific actions under specific circumstances. Decisions, in contrast, are concerned with an assessment if, for example, it is rational "to act as if A > B (treatment A is superior to treatment B) in the present situation" while asserting no judgment as to "truth" or "certainty beyond a reasonable doubt."[1] Here we contrast drawing conclusions and decision-making in therapeutic research, focusing on how to rationally accept that treatment A is superior to treatment B given the *current* evidence and research hypothesis generated in a randomized controlled trial (RCT). We show that this end can be accomplished within a decision-analytic risk-benefit (R:B) framework, which requires integration of frequentist and Bayesian approaches (discussed elsewhere in this issue).

Four fundamental premises define our approach. The first is the philosophy of pragmatism.[2,3] It is recognized that clinical research is unavoidably associated both with research payoffs (benefits) and inadvertent consequences (harms, here termed *risks*). We follow the principles of classical decision theory founded on expected utility decision theory,[3] which maintains that rational decision-making is such that the research hypothesis should be preferred only when benefits of acting on it outweigh its risks, that is, our choice maximizes the value of consequences,[4-6] obtained by choosing the option with the higher expected utility[7-18] (see Glossary).

Second, knowledge of the absolute truth in research is impossible. Expressing this statement in the language of probability calculus, we can never conclude that research findings or research hypothesis are impossible ($P = 0$) or certain ($P = 1$). Therefore, all data and inferences can in principle be considered as "inconclusive."

Third, because of the second premise, a rational approach to decision-making must take into account errors with regard to evidence, inferences about the accuracy of a research hy-

[a]Department of Mathematics, Indiana University, Gary, IN.
[b]Department of Interdisciplinary Oncology, H. Lee Moffitt Cancer Center & Research Institute at the University of South Florida, Tampa, FL.
Address correspondence to Benjamin Djulbegovic, MD, PhD, Department of Interdisciplinary Oncology, H. Lee Moffitt Cancer Center & Research Institute at the University of South Florida, 12902 Magnolia Dr, Tampa, FL 33612. E-mail: Benjamin.Djulbegovic@moffitt.org

 0037-1963/08/$-see front matter 
doi:10.1053/j.seminhematol.2008.04.006

pothesis, and regret of mistakes with respect to ultimate choices.

Fourth, specification of errors (about evidence, research hypotheses, and decisions) will depend on the goals of research (if the trial is designed as *explanatory* or *pragmatic*).

The reader is referred to the Glossary for definition of terms used here.

## Frequentist Paradigm-Using Evidence From the Trial to Assess the Probability That Conclusions Are Wrong: $\alpha$, $\beta$, and $\gamma$ Errors

In a traditional frequentist approach to hypothesis testing, we typically postulate the following null hypothesis:

$H_o$ : Effect of treatment A = Effect of Treatment B and the alternative hypothesis:

$H_a$ : Effect of Treatment A $\neq$ Effect of treatment B.

For convenience, we shorten the expression the "effect of treatment A" as A, etc. The testing will result in one of three possible outcomes (O), which will be determined by the evidence from the trial:

$O : A = B$ ($O : H_o$ − the conclusion that effects of treatments A and B nearly are equal),

$O : A > B$ (the conclusion that the treatment A is superior),

or

$O : A < B$ (the conclusion that the treatment B is superior), while the reality (R) can be described as

$$R : A = B \ (R : H_o, \text{i.e. reality is } H_o)$$

[The reader should note that exact equality is impossible for continuously measured outcomes. Since classic frequentist inference favors the null hypothesis and focuses determination of the power on a simple alternative hypothesis, one may effectively consider R : A = B as meaning R : A − B $<$ $\Delta$ and B − A $<$ $\Delta$ for some $\Delta > 0$.]

$$R : A > B \text{ and}$$

$$R : A < B, \text{ respectively.}$$

[In this paper, when suitable, calculations are based on combining the last two outcomes into $O : A \neq B$ ($O : H_a$), and the last two realities into $R : A \neq B$ ($R{:}H_a$, ie, reality is $H_a$)].

In our attempt to distinguish accurately between findings under the null hypothesis and the alternative research hypothesis, we must take into account the possibility of drawing erroneous conclusions. Schwartz and Lellouch[19] describe three types of errors:

*Error of the first kind* (*false positive*)

The probability that we will deduce that $H_a$ is true, when in fact it is not:

$$\alpha = P\left(O : A \neq B | R : A = B\right)$$

*Error of the second kind* (*false negative*)

The probability that we will deduce that $H_o$ is true, when in fact it is not:

$$\beta = P\left(O : A = B | R : A \neq B\right)$$

*Error of the third kind*

The probability that we will deduce that treatment A is better (A $>$ B), when in fact the reverse is true.

$$\gamma = P\left(O : A > B | R : A < B\right)$$

It should be noted that the maximum $\gamma$ error is equal to $1/2\alpha$.

## Bayesian Approach—Assessing the Probability That a Research Hypothesis Is False: 1 − PRHT

We are usually most interested in the assessment of the probability that the research hypothesis is true (PRHT) or false (1-PRHT).[20,21] We assume a "simple" research hypothesis with a specific effect difference (for example,15 units difference in outcomes) for which the power of the test statistics (1 − $\beta$) applies.[21,22] PRHT depends on the prior probability of it being true (before doing the study), the statistical power of the study (1 − $\beta$), and the level of statistical significance ($\alpha$ and $\gamma$ errors) and can be calculated via Bayes' theorem.[20-24] Figure 1 shows a Bayesian tree illustrating the relationship between true state of treatment effects ("reality") and observed research findings ("outcomes"). Table 1 shows the calculation of the conditional posterior probabilities for the various relationships shown in Figure 1.

## Assessing the Consequences of a Wrong Decision: Decision-Analytic and Acceptable Regret Approach

The PRHT quantifies the likelihood of the research hypothesis being correct, but it does not tell us how high this probability should be before we can accept it.[9,10,11,13,14,22] In other words, when should a result be considered sufficiently convincing to allow action under specific circumstances of evidence on benefits and risks? This question may refer to the situations when the evidence is emerging or when it is more mature, under the conduct of an explanatory or pragmatic trial paradigm, respectively (see below). As stated in the Introduction, according to normative decision theory, rational decision-making means that we should select the hypothesis with higher expected utility involving benefits and risks associated with our decision.[4] Applying this decision-theoretic

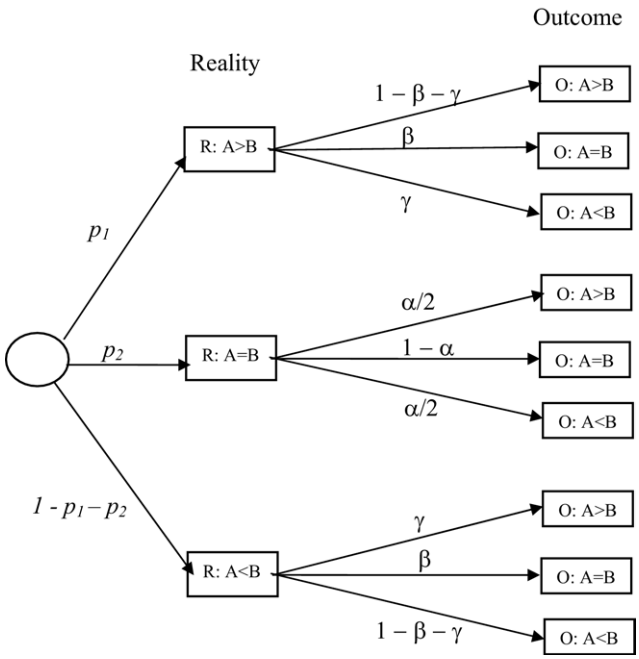


**Figure 1** Bayesian (decision-tree) illustrating the relationship between true state of treatment effects ("reality") and research findings ("outcomes"). A > B indicates that treatment A is better than B; A = B, treatments are equal; B > A, treatment B is superior to A; $p_1$, prior probability that A > B; $p_2$, prior probability that A = B. See text for definition of $\alpha$, $\beta$, and $\gamma$ error. The values on the arrows connecting realities with outcomes are likelihoods.

framework, there is some probability above which the results of the study will be sufficient for researchers to accept them[10,11]; a research hypothesis should be accepted when it is coherent with beliefs "upon which a man is prepared to act."[2] This will occur when the benefits of an action (of accepting the results of research hypothesis) outweigh its risks.[10] Mathematically this can be expressed as[9,10,11,13,14,25]

$$P_t = \frac{1}{1 + \left(\frac{B}{R}\right)} \quad \textbf{(1)}$$

where ($p_t$) is the threshold probability, B is *net benefit* and R is *net risk* (Figure 2). In a decision-analytic framework, the threshold probabilities are equated with actionable research findings. The threshold becomes the "working truth indicator"; it is this relationship between PRHT and $p_t$ that determines what represents our best decision "here and now" in present circumstances of the clinical trial.

If PRHT is above $p_t$ we can rationally accept the results of research findings. Similarly, if PRHT $< p_t$ we should not reject the null hypothesis. Note that research payoffs (benefits) and inadvertent consequences (risks) [equation (1)] can be expressed in a variety of units (like morbidity, mortality, life expectancy, etc).

However, acting according to the threshold model does not guarantee that we cannot make a mistake. Therefore, if we subsequently conclude that our initially positive research conclusions were in fact false, the alternative (the null hypothesis) would have been preferable.[11,26-30] When initially positive research findings lead to a wrong decision, this may bring a sense of loss or regret.[11,26-30] However, under certain conditions making a wrong decision will not be particularly burdensome to the decision-maker.[11,30] We have previously described the concept of *acceptable regret* ($R_0$), which formalizes conditions under which a wrong decision to accept hypothesis and act upon on it is tolerable.[11,18,30]

Formally, acceptable regret, $R_0$, is defined as the utility we find acceptable to lose when we are wrong. It can further be shown that we should be willing to accept results of potentially false research hypothesis as long as the probability (PRHT) of it being true is above the threshold probability, $p_r$

$$PRHT \geq p_r = 1 - \frac{R_0}{R} \quad \textbf{(2)}$$

This equation describes the effect of acceptable regret on the threshold probability [equation (1)] in such a way that if regret about a *wrong decision* is taken into account, the PRHT now also needs to be above the threshold defined in equation (2) for the research results to become acceptable. Providing for the possibility of making a wrong decision results in the requirement that the probability of the research hypothesis being true (PRHT) is higher than would have been estimated based solely on the expected utility theory [equation (1)]. Since optimal decision-making is typically connected to the evaluation of benefits and risks, we can further express acceptable regret, $R_0$, as either a fraction of net benefits lost or net risks incurred due to a wrong decision. The exact empirical relationship between $R_0$ and benefits and/or risks is not known, but herein we will assume a linear relationship. Thus, we choose to characterize our acceptable regret as either the percentage (denoted *r*) of the benefits (denoted B), or risks (denoted R) that we are willing to lose/incur in case our decision is the wrong one, depending upon the goal of the trial. In explanatory trials, we suggest employing a relationship $R_0 = r{\cdot}B$ because we are more concerned about missing benefits; in pragmatic trials using $R_0 = r{\cdot}R$ may be more appropriate (see also below). Equation (2) now becomes:

$$p \geq p_r = 1 - \frac{R_0}{R} = 1 - r \cdot \frac{B}{R} \quad \textbf{(3a)}$$

(to be used in explanatory trials)

or

$$p \geq 1 - \frac{R_0}{R} = 1 - r. \quad \textbf{(3b)}$$

(to be used in pragmatic trials)

It follows from these equations that if we cannot accept any error in our decision-making, we can operate only at the level of absolute truth in research hypothesis (ie, PRHT=1), clearly an unachievable goal.

It can further easily be shown that our error in wrongly accepting the null hypothesis is largely a function of (a fraction of) forgone benefits, while our error in wrongly accept-

**Table 1** Calculation of the Conditional Posterior Probabilities

| Reality | Outcome (research findings) | *P* (Reality\|Outcome) |
|---|---|---|
| R: A > B | O: A > B | $\frac{(1-\beta-\gamma)p_1}{(1-\beta)p_1+\left(\frac{\alpha}{2}\right)p_2+(\gamma)(1-2p_1-p_2)}$ |
| | O: A = B | $\frac{(\beta)p_1}{(1-\alpha)p_2+(\beta)(1-p_2)}$ |
| | O: A < B | $\frac{(\gamma)p_1}{(\gamma)p_1+\left(\frac{\alpha}{2}\right)p_2+(1-\beta-\gamma)(1-p_1-p_2)}$ |
| R: A = B | O: A > B | $\frac{\left(\frac{\alpha}{2}\right)p_2}{(1-\beta-\gamma)p_1+\left(\frac{\alpha}{2}\right)p_2+\gamma(1-p_1-p_2)}$ |
| | O: A = B | $\frac{(1-\alpha)p_2}{(1-\alpha)p_2+\beta(1-p_2)}$ |
| | O: A < B | $\frac{\left(\frac{\alpha}{2}\right)p_2}{(\gamma)p_1+\left(\frac{\alpha}{2}\right)p_2+(1-\beta-\gamma)(1-p_1-p_2)}$ |
| R: A < B | O: A > B | $\frac{(\gamma)(1-p_1-p_2)}{(1-\beta-\gamma)p_1+\left(\frac{\alpha}{2}\right)p_2+(\gamma)(1-p_1-p_2)}$ |
| | O: A = B | $\frac{(\beta)(1-p_1-p_2)}{(\gamma)p_1+\left(\frac{\alpha}{2}\right)p_2+(\beta)(1-p_1-p_2)}$ |
| | O: A < B | $\frac{(1-\beta-\gamma)(1-p_1-p_2)}{(1-\beta)(1-p_1-p_2)+\left(\frac{\alpha}{2}\right)p_2+(\gamma)(2p_1+p_2-1)}$ |

NOTE. The first three rows represent the posterior probability that the research hypothesis is true (PRHT). These formulae are derived from Figure 1, where the prior probability and likelihood for a given reality-outcome pair are multiplied together to obtain the numerator, while the denominator is the sum of all pairs with the same outcome, with the top and bottom formulae re-arranged slightly.

ing the alternative (research) hypothesis is largely a function of the magnitude of risks we are willing to incur.[11,30]

It is important to note that our tolerance for making mistakes with respect to benefits may dramatically differ from our tolerance to mistakes made with respect to harms. That is, we act differently to the possibility of wrongly concluding that intervention is harmful (if in fact it is not) than falsely concluding that treatment is beneficial (if in fact it is not). Humans are cognitively more ready to wrongly accept the signal of potential benefits than one that carries the potential of harm. As a consequence, we need a higher PRHT to accept the alternative (research) hypothesis than not to reject the null hypothesis. (This discussion can be differently framed from the point of view of classic inferential statistics. Here, when it comes to benefits, we accept more false-negative [typically set at $1-\beta=0.2$] than false-positive signals [typically, $\alpha=0.05$], indicating that before we recommend treatment to everyone we require a higher level of certainty that the intervention is in fact beneficial. However, when it comes to harms, we accept more false positive than false negative evidentiary signals out of a desire to minimize the risk of missing a danger-

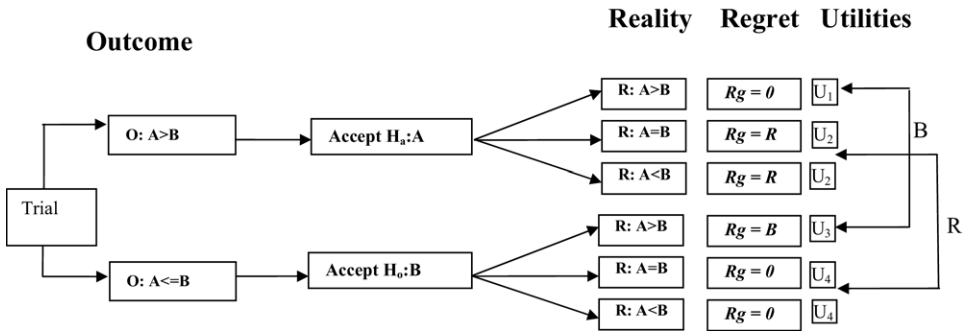


**Figure 2** Decision tree outlining the choice in a typical clinical research setting, with a one-sided research hypothesis. The choice is between null hypothesis: $H_0$: A ≤B, claiming that the experimental treatment A is not significantly better than treatment B which, if accepted, would lead to continuation of administering the treatment B – the proven and standard treatment. The alternative (research) hypothesis $H_a$: A > B means that treatment A is significantly better than treatment B which, if accepted, implies that we should administer treatment A. We define net benefit as the difference between the utilities (research payoffs) of the actions taken based on the outcomes when research hypothesis is true. Net risks are defined as the difference between the utilities of the actions taken based on the outcomes when the null hypothesis is true. That is, the net benefit and net risk are defined as: B-net treatment's benefit, $B = U_1 - U_3$, R – net treatment's risk, $R = U_4 - U_2$ (after Djulbegovic and Hozo I[11]). Rg = regret (defined as the difference between the utility of the outcome of the action taken and the utility of the outcome of best action we could have taken, in retrospect). See refs 11, 18, and 31 for details of calculations.

ous signal, which seems to be evolutionarily wired in human cognition.)

## Errors Should Be Set as a Function of the Goals of Research: Explanatory and Pragmatic Trials

Forty years ago, in one of the most highly cited papers in medical research, Schwartz and Lellouch defined the foundational logic of clinical trials.[19] They distinguished between *explanatory* and *pragmatic* trials.[19] The goal of an explanatory trial is to provide a scientific answer to a research question. Such trials focus on the proof of a concept or mechanism, as for example whether an intervention works under ideal circumstances ("efficacy"). Consequently, the most important error to avoid is a "false-positive" ($\alpha$) error—the goal is to not conclude that a treatment works when in fact it does not. Hence, the $\alpha$ error is kept as small as possible.

The goal of pragmatic trials is more practical and aims at the question "Which treatment (of already proven efficacy) is better?" That is, which of the interventions will work better in a representative sample of patients to whom the results of the trial will likely be extrapolated ("effectiveness")? Here, the concern is about making the error of concluding that one intervention (say, treatment A) is better, when in reality B is superior to A. This is the $\gamma$ error—the often neglected error of the third kind.[19]

The articulation of the goal of research—explanatory versus pragmatic—will determine the types of errors that need to be accounted for in the trial. The frequentist and Bayesian approaches are only concerned with the effects on a single outcome, such as the benefit of treatments. This may be acceptable in explanatory trials, for which our goal is focused on testing biological mechanisms. However, using a conventional non-directional two-tailed test is inappropriate when our decision relates to calculation of the probability of concluding the direction wrong[31] as in pragmatic trials. In the latter case, the appropriate approach is to focus on the error of third kind—the $\gamma$ error.[31]

The main point is that the types of errors that drive the design to answer our research questions are fundamentally different. Similarly, R:B formulas will vary depending on whether we have formulated research questions under an explanatory or pragmatic framework. Schwartz and Lellouch lucidly noted that $\alpha$ error is irrelevant for pragmatic trials since, *if A = B, it does not matter which treatment we choose.*[19] Since the $\alpha$ error does not matter, we can let $\alpha = 1$. We further assume that there is some difference between two treatments. Therefore, we set $\beta$ error = 0. The $\gamma$ error now becomes crucial—our main concern is that we will deduce that A > B, when in fact the reverse is true.

## Examples

### Decision-Making Under the Paradigm of Explanatory Trials

On January 27, 2007, Amgen (Thousand Oaks, CA) released the following Drug Safety Alert for its blockbuster, multibillion dollar erythropoietin-stimulating agent (ESA), Aranesp (darbepoetin alfa), which was tested against placebo (iron) in the treatment of anemia in patients with active malignant disease.[32]

> *The final analysis of the initial 16-week treatment period did not show a statistically significant effect on the primary efficacy endpoint (hazard ratio 0.89; 95% confidence Interval:[0.65,1.22]), with an incidence of red blood cell transfusions of 24% in the placebo vs. 18% in the Aranesp group, P = .15. In the 16-week treatment phase of the study, more deaths were reported in the*

**Table 2** Risk:Benefit Analysis (explanatory trials: $\alpha = 0.05$; $\beta = 0.2$;$\gamma = 0$): The Effect of Erythropoietin-Stimulating Agents on Anemia of Cancer

1. Calculation of the probability that research hypothesis is true (PRHT): ESA is truly superior to iron (placebo) in reducing incidence red blood cell (RBC) transfusion
- Assuming the following prior probabilities:
  - ESA > placebo: 70%; ESA = placebo: 30%; placebo > ESA: 0%

we obtain that
- PRHT (posterior probability that ESA is truly superior to placebo after we observed equality of these two treatments; outcome (A = B) = 32.9%.
  - This also means that PRHT (posterior probability that ESA is truly equal to placebo under these conditions) = 67.1%.

2. Calculation of net benefits and risks

The calculations above cannot tell us how high the probability should be before the research hypothesis can be accepted. To arrive at the answer (step #3 below), *we first need to know what are the benefits and risks of the use of ESA in treatment of cancer-related anemia for each outcome of interest (ie, RBC transfusion, survival).*[11,46]

Primary outcome: incidence of RBC (red blood cell) transfusion
- Net benefits = RBC transfusion incidence in placebo arm - RBC transfusion incidence in ESA arm − harms (risks of thromboembolic events) = 24 − 18 − 2 = 4%
  - Net benefit/net risks = 4/2 = 2

3. Calculation of the threshold above which research hypothesis should be accepted
- B/R = 2 corresponds to $p_t$ = 33% (equation 1).
  - This means that we should use ESA only if PRHT that ESA is superior to placebo is ≥ 33%. Note that PRHT = 32.9% i.e. slightly below 33%.
- Therefore, ESA should be withheld in these patients.
  - Under these assumptions benefits of administering ESA still does not outweigh its risks.

4. Calculation of the regret threshold above which research hypothesis should be accepted

Outcome: RBC transfusion

How much regret we are willing to accept in case our decision not to recommend ESA was erroneous? Equation 3a) indicates that that our regret is mostly a function of forgone benefits.
- If we assume that we are willing to forgo 50% or more of benefits that we could expect from ESA, the threshold for accepting ESA as effective treatment in reduction of RBC transfusion drops to 0
  - Since 0 < 32.9% this means that under these circumstances we should not withhold ESA.

Outcome: overall mortality
- In terms of the effect on mortality, the trial indicates that risks > 0 > benefits (ie, the net benefit was <0).
  - Under these circumstances we cannot accept the research hypothesis as true until PRHT > 100% [equations (1) and (2)], which is, of course, impossible.[11]
- One of conclusions of R:B analysis is that treatment should never be given when the difference between net benefits and risks is negative. This also holds true under acceptable regret theory when regret are expressed in terms of benefits [equation (3a)].
  - When the acceptable regret is expressed in terms of risks alone [equation (3b)], we can accept research hypothesis whenever PRHT > 1− r [equation (3b)].

▪Assuming the following prior probabilities (ESA > placebo: 95%, placebo = ESA: 3%,placebo > ESA:2%), we get that PRHT (ESA > placebo) given the outcome in the Amgen ESA trial (ESA < placebo, see text) is 0%. This means that we could never accept the results that ESA is better than placebo even if we are willing to tolerate almost 100% of death risk incurred with ESA [equation (3b)]. Note: Under explanatory paradigm $\gamma$ is assumed to be zero, which is the reason why the posterior probability PRHT = 0 when the outcome is reverse from reality (ie, when B > A instead of A > B).

> *Aranesp treatment group (26% [136/515]) than the placebo group (20% [94/470]). With median survival follow-up of 4.3 months the absolute number of deaths was greater in the Aranesp treatment group (250/515 = 49%) than in the placebo group (216/470 = 46%) (absolute risk difference in death = 3%, hazard ratio 1.25; 95% confidence interval: 1.04, 1.51; P = .018). The trial also indicated 2% difference in increased risk of thromboembolic event (9.7% in Aranesp v 7.7% in control arm).*

In response to these unexpected results, Amgen further stated "Aranesp is not approved for use in this population. Aranesp is approved for the treatment of patients with anemia, which is caused by chemotherapy treatment of their malignant disease, rather than the underlying malignant disease itself." However, the drug has been widely used for this indication and recommended by all major guideline panels.[33,34] The results from this study prompted the Food and Drug Administration (FDA) black box safety warning, an FDA Oncology Drug Advisory Committee meeting that reassessed safety and efficacy of ESA, a massive change in Medicare coverage reimbursement policy, Congressional hearings as to the appropriateness of the FDA ruling and Medicare decision, and major modifications in the guidelines issued by the leading professional societies.[33-37] None of the discussions that preceded changes in recommendations was in-

**Table 3** Risk:Benefit Analysis (pragmatic trials: $\alpha = 1$; $\beta = 0$; $\gamma = 0$ to 0.5): The Effects of Allogeneic Versus Autologous Stem Cell Transplant/Chemotherapy in Acute Myelogenous Leukemia

1. Calculation of the probability that research hypothesis is true (PRHT):
alloSCT (allogeneic stem cell transplant) is truly superior to autologous SCT/chemotherapy in improving disease-free survival (DFS).
- Assuming the following prior probabilities:

AlloSCT > autoSCT/chemoRx: 50%
- PRHT can be calculated (see Table 1, first row) as PRHT = $1 - \gamma$.
  - Assuming a large variation in $\gamma$ error (0 to 0.5), we obtain
- PRHT = 50% ($\gamma = 0.5$) to 100% ($\gamma = 0$).

2. Calculation of net benefits and risks
(at 4 yrs, all patients)
- Net benefits = $DFS_{alloSCT} - DFS_{autoSCT/chemoRx} - (TRM_{alloSCT} - TRM_{autoSCT}) = 48 - 37 - (21 - 4) = -6\%$

where TRM is treatment-related mortality.
- Net risks = $TRM_{allpSCT} - TRM_{autoSCT} = 17\%$
- Benefit/risk ratio <0 (research hypothesis that alloSCT is superior treatment cannot be accepted.

Subgroup analysis
(at 4 yrs, good risk patients)
- Net benefits = $DFS_{alloSCT} - DFS_{autoSCT/chemoRx} - (TRM_{alloSCT} - TRM_{autoSCT}) = 72 - 64 - (6 - 4) = 6\%$
- Net risks = $TRM_{alloSCT} - TRM_{autoSCT} = 2\%$
- Benefit/risk ratio = 3.

3. Calculation of the threshold above which research hypothesis should be accepted
- Pt = 1/(1 + B/R) (eq 1) = 25%
- PRHT (that alloSCT is truly superior to autoSCT/chemoRx) is 50-100% > 25%.
  - Hence, we should accept the "truthfulness" of this hypothesis and administer alloSCT to this group of patients.

4. Calculation of the regret threshold above which research hypothesis should be accepted
The reader should note that the authors from whose paper data for calculation of benefits and risks were taken do not advocate use of alloSCT in acute myelogenous leukemia (AML) with good risk features.[41] This is because DFS between two groups was not statistically significant in their analysis. Our analysis based on normative decision model indicates that alloSCT should be used for these patients. However, we could have made a mistake in our recommendations.

- How much regret are we willing to accept in case our decision to recommend alloSCT was erroneous?
- The key ingredient related to decision to recommend alloSCT is how much treatment-related harms we are willing to tolerate (see text for details).
- Cornelisson et al[42] recommend alloSCT in intermediate risk of patients in whom the difference in TRM between two treatments is 16% (Net risks = $TRM_{alloSCT} - TRM_{autoSCT} = 19\% - 3\% = 16\%$), while difference between DFS was = $53\% - 41\% = 12\%$.
- We assume that the authors are willing to tolerate 4% of extra harms in order to realize perceived benefits in DFS (given that their recommendation was based on negative net benefits). Therefore, we will assume that our acceptable regret (see text) is $R_0 = 4\%$.
- According to equation (2): $p_r = 1 - 0.04/0.16 = 0.75 = 75\%$, ie, we should accept research hypothesis that alloSCT is truly superior to autoSCT/chemoRx only if $PRHT_{alloSCT} > 75\%$. If we assume a typical 5% for false positive rate, ie, $\gamma = .05$, then PRHT = 95%, which is greater than 75%, and hence alloSCT should be given.

formed by quantitative analysis. Table 2 shows the result of the R:B analysis, which may provide additional insights regarding the effects of ESA in anemia of cancer. Table 2 does appear to lend support to the recommendation given by the FDA and others. Note that this decision does not mean that it is scientifically affirmed ("concluded") that ESA increases mortality in cancer patients, but only that it is most rational "to act as if placebo (iron) ≥ESA in the present situation (see Table 2)" asserting no judgment as to "truth" or "certainty beyond a reasonable doubt."[1]

## Decision-Making Under the Paradigm of Pragmatic Trials

Conventional chemotherapy (ChemoRx), allogeneic (AlloSCT), and autologous (autoSCT) stem cell transplant (SCT) are all well-established treatments for acute myelogenous leukemia (AML).[38] Typically, it is believed that SCT is most effective when it is administered as a consolidation therapy in patients who are in complete remission.[39] Many studies suggested that there is a trade-off between treatment effects of allogeneic SCT (alloSCT) versus autoSCT/chemoRx: alloSCT may result in improvement in leukemia-free survival (LFS) but at the expense of the increased risk for treatment-related mortality (TRM).[38,40] Most of these studies, however, have suffered from bias in design, conduct and analysis, and, therefore, the issue of superiority of alloSCT versus autoSCT/chemoRx as a consolidation treatment has remained controversial. Recently, Cornelissen et el reported credible results comparing alloSCT versus autoSCT/chemoRx.[41] They employed a genetic randomization to compare treatment effects on an intention-to-treat analysis based on the availability of an allogeneic donor. Table 3 shows the results of the quantitative R:B analysis under a pragmatic trial paradigm attempt-

ing to answer the question of whether alloSCT is superior to autoSCT/ChemoRx or not. Table 3 also shows that if the regret threshold is calculated only based on harms we are willing to tolerate, the decision threshold for action becomes identical to the 1 - $\gamma$ error.

## Discussion

In this paper, we treat the acceptance of a research hypothesis as a decision problem. Since knowledge of absolute "truth" is impossible to achieve, we must account for errors that are unavoidable in research. These errors can occur at the level of *evidence and research hypothesis (conclusions)* as well as *decision-making*. The classical frequentist approach to clinical research typically deals with *evidentiary errors* by specifying false positive ($\alpha$ or $\gamma$) and false negative ($\beta$) error rates to draw inferences about the correctness of a research hypothesis. As discussed in detail elsewhere in this issue, the classical approach does not have formal tools to assess the probability of research hypothesis and a method to decide if a particular research hypothesis should be accepted and acted upon. Frequentists use an informal, intuitive decision-making process in which rejection of the null hypothesis is equated with acceptance of the alternative (research) hypothesis. This is, of course, logically incorrect since evidence against the null hypothesis is not equivalent to evidence to support the alternative (research) hypothesis.[42]

This problem with classical statistical theory dates at least to the1960s. For example, Howard Raiffa, one of the pioneers of decision analysis, stated in 1961: "We believe, however, that without decision analysis formalization, decisions under uncertainty have been and will remain essentially arbitrary, as evidenced by the fact that, in most statistical practice, consequences and performance characteristics receive mere lip service while decisions are actually made by treating the numbers 0.05 and 0.95 with the same superstitious awe that is usually reserved for the number 13."[43] He reiterated this point in his recent personal account of a relationship between classical statistical and decision theory stating that stress "on tests of hypotheses, confidence intervals and unbiased estimation was either wrong or not central" and that "little attention was paid to integration of inference and decision."[44] The situation, astonishingly, has not improved much to date.

To make this link between evidence and decisions, we first need an apparatus to help us calculate the probability of a research hypothesis being true, conditioned on the likelihood provided by the evidence. This goal can be accomplished using Bayes' theorem. The resultant calculation can produce assessment of the *probability of research hypothesis being true (PRHT) or false (1 - PRHT)*[21] (see Table 1). The key question, of course, is how high the PRHT needs be before it is accepted? Calculation of the likelihood of the hypothesis being correct does not itself inform which research result is optimal and whether it should be accepted.[9,10,11,13,14,22] As we argue here, a research hypothesis can only be accepted within a formal decision-analytic framework (Figure 2): the research hypothesis should be accepted only when the probability of research hypothesis being "true" (PRHT) exceeds the decision-threshold probability, $p_t$ [see equation (1)].[10,11] Acceptance occurs when benefits of action (of accepting the results of research hypothesis) outweigh its risks.[10]

Next, we must realize that although the decision-analytic threshold approach outlined here can optimize our chances of selecting the correct research hypothesis, it cannot protect us against making erroneous decisions. We therefore need to take into account the possibility that our decision may be

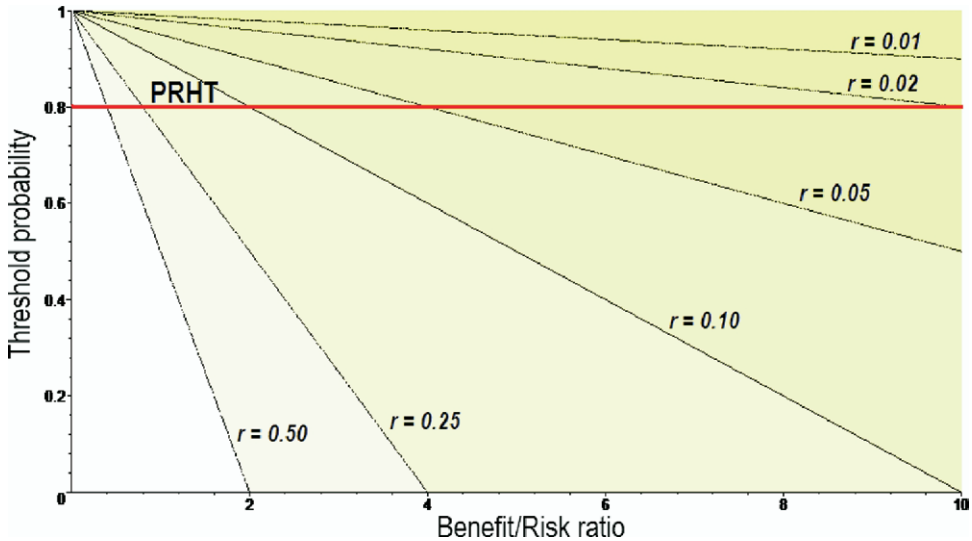


**Figure 3** Acceptable regret threshold (lines in the graph) as a function of the fraction (r) of benefits we are willing to forgo in case of a wrong decision and benefit/risk ratio. The less benefits we are willing to forgo, the higher threshold for accepting research hypothesis for the given benefit/risk ratio. The horizontal line illustrates an example of the actual probability that the research hypothesis is true (PRHT = .80) in relation to the threshold probability. This means that as long as the PRHT is above the acceptable regret threshold, the research findings could be accepted with tolerable amount of regret in case the research hypothesis proves to be wrong.

wrong, which can be best accomplished by explicitly taking into account regret of our mistakes.[45] In particular, specifying *acceptable regret*—regret of a wrong decision that we can tolerate—can help us calculate a more tolerable threshold above which a research hypothesis can be accepted [equations (2) and (3)]. When the PRHT is above the threshold, we will not regret if we accept research hypothesis even if our decision turns out to be wrong (Figure 3). We propose the use of acceptable regret as the ultimate decision criterion with respect to acceptability of research hypotheses.

Our argument should be understood as a didactic attempt to articulate a very complex area of scientific testing, inference and decision-making. For this reason, we have not extended our approach to include uncertainties about our estimates. The methods outlined here are, therefore, purposefully and necessary simplified. The analysis is always based on the *current state of knowledge*, focuses on a single hypothesis testing involving two outcomes (benefit and harm), although the method can easily integrate multiple harms.[46] Acceptable regret must be expressed in the same units as benefits and risks. Important also is the judgment as to which data should be employed in R:B analysis—the trial data or the totality of data external to the trial. Which data to select for R:B analysis will remain a critical problem (as in other area of scientific inference), but the issue is not mathematically solvable[47]; the best we can do is to outline our reasoning in a transparent and explicit way.[47] All of these limitations notwithstanding, we believe that the approach discussed here represents an improvement over current statistical practice, since it focuses on two dimensions (benefits and risks) instead of the usual single dimension (primary outcome).

## Glossary

**Acceptable regret** (see also **Regret**)**:** a loss in utility when undertaking a wrong decision that will not be particularly burdensome to the decision-maker.

**Bayesian statistics:** a branch of statistics that employs Bayes' theorem for calculation of conditional probabilities to update one's prior beliefs or knowledge with research evidence to derive the estimates about the accuracy of research hypothesis.

**Errors in drawing conclusions** (see also **Frequentist statistics**)**:** based on frequentists' definition of probability, we typically distinguish 3 types of inferential error:

—$\alpha$ error (*error of the first kind; false positive error*) = the probability that we will deduce that ***research hypothesis*** ($H_a$) postulating that the effects of treatment A differs from the effects of treatment B is true, when in fact it is not.

—$\beta$ error (*error of the second kind; false negative error*) = the probability that we will deduce that ***the null hypothesis*** ($H_o$) of no difference in treatment effects is true, when in fact it is not.

—$\gamma$ error (*error of the third kind;* the maximum $\gamma$ error is equal to $1/2\alpha$) = the probability that we will deduce that treatment A is better (A > B), when in fact the reverse is true.

**Explanatory clinical trials** (see also **Pragmatic clinical trials**)**:** clinical trials whose main aim is to test the proof of a concept or mechanism, ie, whether an intervention works under ideal circumstances ("efficacy").

**Frequentist statistics** (see also **Errors in drawing conclusions**)**:** a branch of statistics that defines the probability of an event's occurring in a particular trial as the *frequency* with which it occurs in a long sequence of similar trials. It employs inferences based on calculations of the errors of drawing wrong conclusions.

**Decision analysis:** originally conceived as applied decision theory. It is a logical structure for the balancing of the factors that influence a decision. At its basic level, decision-analysis involves the distinction between the actions (choices), probabilities of events and their relative values (payoffs, outcomes, consequences).

**Disutility** (see also ***Utility***)**:** undesirability, or strength of preferences that individuals or societies have against a particular outcome such as loss of length of life, morbidity or mortality rates, presence of pain, cost, etc, a complement of utility, typically expressed as 1 − utility.

**Expected utility** (see also ***Utility***)**:** the average of all possible results weighted by their corresponding probabilities. It is normative criterion of rationality according to which a rational decision-maker should select the alternative (eg, research *v* null hypothesis) with the highest expected utility value.

**Net benefit:** benefit minus risks when research hypothesis is true. More specifically, net benefit is defined as the difference between the utilities (research payoffs) of the actions taken under the research and the null hypothesis when in fact research hypothesis is true.

**Net risks:** differences in risks (harms) when the null hypothesis is true. More specifically net risks is defined as the difference between the utilities (research payoffs) of the actions taken when the null hypothesis is true.

**Pragmatic clinical trials:** clinical trials whose main goal is to answer the question "Which treatment (of already proven efficacy) is better?" That is, which of the interventions will work better in a representative sample of patients to whom the results of the trial will likely be extrapolated ("effectiveness")?

**Regret:** an example of counterfactual thinking that anticipates our reactions (emotions) to comparisons of outcomes under scenarios of what has happened versus *what might have happened* had we chosen differently. It a psychological reaction to making a *wrong decision*, when wrong is determined on the basis of actual outcomes rather than on the information available at the time of the decision. In decision-analytic language, regret can mathematically defined as the difference between the utility of the outcomes of the action taken and the utility of the outcomes of another action, which, in retrospect, we should have taken.

**Utility:** desirability, or strength of preferences that individuals or societies have for a particular outcome (research pay-

off) such as length of life, morbidity or mortality rates, absence of pain, cost, etc.

## References


1. Tukey J: Conclusions vs. decisions. Technometrics 2:423-433, 1960
2. deWaal C: On Pragmatism. Belmont, CA, Wadsworth, 2005
3. Edwards W, Miles R Jr, vonWinterfeld D: Advances in Decision Analysis. From Foundations to Applications. New York, NY, Cambridge University Press, 2007
4. Bell DE, Raiffa H, Tversky A: Decision Making. Descriptive, Normative, and Prescriptive Interactions. Cambridge, UK, Cambridge University Press, 1988
5. Ciampi A, Till JE: Null results in clinical trials: The need for a decision-theory approach. Br J Cancer 41:618-629, 1980
6. Hastie R, Dawes RM: Rational Choice in an Uncertain World. London, UK, Sage, 2001
7. Browner WS, Newman TB: Are all significant *p* values created equal. The analogy between diagnostic tests and clinical research. JAMA 257: 2459-2463, 1987
8. Dagli A, Morse RM, Dalton C, Owen JA, Hayden GF: Formulating clinical questions during community preceptorships: A first step in utilizing evidence-based medicine. Fam Med 35:619-621, 2003
9. Djulbegovic B, Desoky AH: Equation and nomogram for calculation of testing and treatment thresholds. Med Decis Making 16:198-199, 1996
10. Djulbegovic B, Hozo I: At what degree of belief in a research hypothesis is a trial in humans justified? J Eval Clin Practice 8:269-276, 2002
11. Djulbegovic B, Hozo I: When should potentially false research findings be considered acceptable? PLoS Med 4:e26, 2007
12. Pater JL, Willan AR: Clinical trials as diagnostic tests. Controlled Clin Trials 5:107-113, 1984
13. Pauker S, Kassirer J: Therapeutic decision making: A cost benefit analysis. N Engl J Med 293:229-234, 1975
14. Pauker SG, Kassirer J: The threshold approach to clinical decision making. N Engl J Med 302:1109-1117, 1980
15. Djulbegovic B, Hozo I, Schwartz A, McMasters K: Acceptable regret in medical decision making. Med Hypoth 53:253-259, 1999
16. Djulbegovic B, Hozo I, Lyman GH: Linking evidence-based medicine therapeutic summary measures to clinical decision analysis. Med Gen Med 2:E6, 2000
17. Hozo I, Djulbegovic B: Using Internet to calculate clinical action thresholds. Comp Biomed Res 32:168-185, 1999
18. Hozo I, Djulbegovic B: When is diagnostic testing inappropriate or irrational? Acceptable regret approach. Med Decis Making (in press)
19. Schwarz L, Lellouch J: Explanatory and pragmatic attitudes in therapeutic trials. J Chron Dis 20:637-648, 1967
20. Ioannidis JP: Why most published research findings are false. PLoS Med 2:e124, 2005
21. Wacholoder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N: Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 96: 432-442, 2004
22. Goodman SN: Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med 130:1005-1013, 1999
23. Gabbay J, le May A: Evidence based guidelines or collectively constructed "mindlines?" Ethnographic study of knowledge management in primary care. BMJ 329:1013-1010, 2004
24. Goodman SN: Toward evidence-based medical statitistics. 1: The *p* value fallacy. Ann Intern Med 130:995-1004, 1999
25. Sinclair JC, Cook RJ, Guyatt GH, Pauker SG, Cook DJ: When should an effective treatment be used? Derivation of the threshold number needed to treat and the minimum event rate for treatment. J Clin Epidemiol 54:253-262, 2001
26. Bell DE: Regret in decision making under uncertainty. Operations Res 30:961-981, 1982
27. Loomes G, Sugden R: Regret theory: An alternative theory of rational choice. Economic J 92:805-824, 1982
28. Zeelenberg M, Pieters R: A theory of regret regulation 1.0. J Consumer Psychol 17:3-18, 2007
29. Zeelenberg M, Pieters R: A theory of regret regulation 1.1. J Consumer Psychol 17:29-35, 2007
30. Djulbegovic B, Hozo I, Schwartz A, McMasters KM: Acceptable regret in medical decision making. Med Hypoth 53:253-259, 1999
31. Leventhal L, Huynh C-L: Directional decisions for two-tailed tests: Power, error rates, and sample size. Psychol Methods 3:278-292, 1996
32. FDA Safety Information and Adverse Event Reporting Program. http://www.fda.gov/medwatch/safety/2007/safety07.htm Aranesp; accessed April 2008.
33. Rizzo D, Lichtin AE, Woolf SH, Seidenfeld J, Bennett CL, Cella D, et al: Use of epoetin in patients with cancer: evidence-based clinical practice guidelines of the American Society of Clinical Oncology and at American Society of Hematology. J Clin Oncol 20:4083-4107, 2002
34. Rodgers GM 3rd, Becker SP, Bennett CL, Cella D, Chanan-Khan A, Chesney C, et al: Cancer- and treatment-related anemia. http://www.nccn.org/professionals/physician_gls/PDF/anemia.pdf; accessed April 2008
35. Rizzo JD, Somerfield MR, Hagerty KL, Seidenfeld J, Bohlius J, Bennett CL, et al: Use of epoetin and darbepoetin in patients with cancer: 2007 American Society of Clinical Oncology/American Society of Hematology Clinical Practice Guideline Update. J Clin Oncol 26:132-149, 2008
36. https://www.cms.hhs.gov/scripts/ctredirector.dll/.pdf?@_CPR0a0a043a07d1.YE_Qa3N_cvb
37. Fromer M, Hogan M: ESAs: Further labeling restrictions endorsed, questions about safety remain. Oncol Times 29:23-25, 2007
38. Cassileth PA, Harrington DP, Appelbaum FR, Lazarus HM, Rowe JM, Paietta E, et al: Chemotherapy compared with autologous or allogeneic bone marrow transplantation in the management of acute myeloid leukemia in first remission. N Engl J Med 339:1649-1656, 1998
39. Lowenberg B: Post-remission treatment of acute myelogenous leukemia. N Engl J Med 332:260-262, 1995
40. Zittoun RA, Mandelli F, Willemze R, de Witte T, Labar B, Resegotti L, et al: Autologous or allogeneic bone marrow transplantation compared with intensive chemotherapy in acute myelogenous leukemia. European Organization for Research and Treatment of Cancer (EORTC) and the Gruppo Italiano Malattie Ematologiche Maligne dell'Adulto (GIMEMA) Leukemia Cooperative Groups. N Engl J Med 332:217-223, 1995
41. Cornelissen JJ, van Putten WL, Verdonck LF, Theobald M, Jacky E, Daenen SM, et al: Results of a HOVON/SAKK donor versus no-donor analysis of myeloablative HLA-identical sibling stem cell transplantation in first remission acute myeloid leukemia in young and middle-aged adults: Benefits for whom? Blood 109:3658-3666, 2007
42. Senn SJ: Falsificationism and clinical trials. Stat Med 10:1679-1692, 1991
43. Miles R Jr: The Emergence of Decision Analysis. New York, NY, Cambridge University Press, 2007
44. Raiffa H: Decision Analysis: A Personal Account of How It Got Started and Evolved. New York, NY, Cambridge University Press, 2007.
45. Kahneman D: Maps of bounded rationality: Psychology for behavioral economics. Am Econ Rev 93:1449-1475, 2003
46. Djulbegovic B, Hozo I, Lyman G: Linking evidence-based medicine therapeutic summary measures to clinical decision analysis. Med Gen Med http://www.medscape.com/Medscape/GeneralMedicine/journal/2000/v02.n01/mgm0113.djul/mgm0113.djul-01.html, January 13, 2000
47. Djulbegovic B: Articulating and responding to uncertainties in clinical research. J Med Philosophy 32:79-98, 2007