

Do We Know How to Set Decision Thresholds for Diabetes?

Y. Ben-Haim, PhD,^{1,4} M. Zacksenhouse, PhD,² C. Keren, BSc² and C. C. Dacso, MD³

¹Yitzhak Moda'i Chair in Technology and Economics,
Faculty of Mechanical Engineering,
Technion, Israel Institute of Technology, Haifa, Israel.

²Faculty of Mechanical Engineering,
Technion - Israel Institute of Technology, Haifa, Israel.

³John S. Dunn Sr. Research Chair in General Internal Medicine,
The Methodist Hospital Research Institute, Houston, TX.

⁴Corresponding author: yakov@technion.ac.il
Telephone: +972-4-829-3262, Fax: +972-4-829-5711,

Abstract: The diagnosis of diabetes, based on measured fasting plasma glucose level, depends on choosing a threshold level for which the probability of failing to detect disease (missed diagnosis), as well as the probability of falsely diagnosing disease (false alarm), are both small. The Bayesian risk provides a tool for aggregating and evaluating the risks of missed diagnosis and false alarm. However, the underlying probability distributions are uncertain, which makes the choice of the decision threshold difficult. We discuss an hypothesis for choosing the threshold that can robustly achieve acceptable risk. Our analysis is based on info-gap decision theory, which is a non-probabilistic methodology for modelling and managing uncertainty. Our hypothesis is that the *non-probabilistic* method of info-gap robust decision making is able to select decision thresholds according to their *probability* of success. This hypothesis is motivated by the relationship between info-gap robustness and the probability of success, which has been observed in other disciplines (biology and economics). If true, it provides a valuable clinical tool, enabling the clinician to make reliable diagnostic decisions in the absence of extensive probabilistic information. Specifically, the hypothesis asserts that the physician is able to choose a diagnostic threshold that maximizes the probability of acceptably small Bayesian risk, without requiring accurate knowledge of the underlying probability distributions. The actual value of the Bayesian risk remains uncertain.

Keywords: Diabetes diagnosis, decision thresholds, Bayes risk, info-gap uncertainty.

1. INTRODUCTION

Diagnosis and treatment is a complex interaction of subjective information and impressions, objective data, patient needs and preferences, and resource constraints. The integration of quantitative information into clinical decisions often relies on thresholds or intervals (which are pairs of thresholds.) For example, "normal limits" generally refers to the interval in which a parameter value for a healthy population falls. Above or below these limits is the region of abnormality. The quantitative diagnostic question becomes: is the patient above or below a threshold? Prior to this question comes the query: what value of threshold should trigger a response? Decision thresholds are difficult to set because of diverse and complicated uncertainties. In this paper we explore this difficulty, and we propose an hypothesis based on the idea of satisficing as it is quantified in info-gap decision theory (1). The hypothesis, if true, provides a tool for the responsible and successful management of this uncertainty in choosing decision thresholds. We illustrate the hypothesis with the case of diagnosing diabetes.

Thresholds are often based on clinical trials with populations which may not reliably reflect the target population to which the individual patient belongs. Both patient and physician are sometimes quite uncertain about the relevance of the clinical trials to their specific case. Greenfield *et al* (2) note that randomized controlled trials, which underlie clinical guidelines and decision thresholds, typically enroll patients with less severe disease and exclude older patients, making the resulting thresholds of uncertain applicability to the excluded populations. Feinstein and Horwitz (3) warn against the prevalence of randomized clinical trials in which "the data do not include many types of treatments or patients seen in clinical practice". Morimoto *et al.* (4) note that clinical guidelines, developed in the U.S. for use of aspirin in primary prevention of cardiovascular events, need modification before application in Japan. McLaughlin (5) reports the conclusions of a roundtable discussion of implications of heterogeneity of treatment effects (HTE). He concludes that, due to HTE, and especially in the absence of "sound data", "care has to be individualized, using the clinician's best judgment regarding available treatment options." Glasziou *et al.* (6) explain that "There is generally a weak signal-noise ratio in cholesterol level monitoring. The signal of a small increase in cholesterol level will be difficult to detect against the background of a short-term variability". The problem is two-fold: first, the statistical variability itself, and second, the difficulty in characterizing this variability which can change over time and be quite idiosyncratic. This second problem is in the presumed probability distribution, which may be poorly known. The distribution may be estimated from a limited sample, so it may include substantial estimation errors. Furthermore, as detailed above, the target population may differ from the sample population.

We now consider the difficulty of choosing a threshold value for a single variable, which may be a specific measured quantity or a combination (e.g. regression) of several measurements.

If we confidently knew the values of the measured variable under normal and abnormal conditions, and their statistical variations, then we could reliably choose a threshold value for triggering an intervention. Specifically, we could seek a threshold

for which the Bayesian risk – which combines the probabilities of false alarm and missed detection – is minimal. However, as discussed earlier, we usually do not have accurate knowledge of the variable used for diagnosis. Furthermore, past values only partially indicate future values because the patient's condition evolves.

We face a gap between what we *do know* about the patient and what we *need to know* in order to reliably choose a threshold. This *information gap* motivates the choice of a threshold which yields adequate Bayesian risk even if our best understanding errs substantially. Instead of minimizing our best estimate of the Bayesian risk, we seek a threshold for which the Bayesian risk is acceptable under the widest possible range of error in our knowledge about the patient. This strategy is called *robust-satisficing*: to be robust to information-gaps while also satisfying critical requirements (which is Simon's (7) definition of 'satisficing').

The *hypothesis* of this paper is that this robust-satisficing strategy can be used to choose a decision threshold for reliably achieving acceptably small Bayesian risk. Info-gap robust-satisficing is based on non-probabilistic models of the uncertainties underlying the evaluation of the Bayesian risk. Nonetheless, robust-satisficing acts as a proxy for the unknown probabilities and thus is able to select the threshold with higher probability of success. The actual value of the Bayesian risk remains uncertain. If this hypothesis is true, then the physician can make reliable diagnostic decisions in the absence of extensive probabilistic information.

In section 2 illustrate our hypothesis and the resulting strategy for diagnosis of diabetes. In section 3 we formalize and discuss the hypothesis.

2. DIAGNOSING DIABETES

The robust-satisficing choice of a decision threshold hinges on evaluating the robustness to uncertainty. The robustness is a quantitative answer to the question: how wrong can our models and data be, and the threshold will still yield acceptable outcomes. It is our hypothesis that acceptable outcomes are most reliably obtained by using the info-gap robust-satisficing strategy (1).

We formulate this hypothesis in the context of an example: choosing the decision threshold for diagnosing diabetes based on measuring the fasting plasma glucose (FPG) concentration. This will facilitate a precise formulation of our hypothesis in section 3.

2.1 Threshold Decision

Let x denote the measured FPG level. We will diagnose the patient as diabetic if and only if x exceeds a threshold, θ :

$$x \geq \theta \tag{1}$$

Estimated means and standard deviations for FPG are given by de Vegt *et al* (1998), based on American Diabetic Association criteria, as 5.4 ± 0.5 (mmol/l) for the healthy individual, and 9.6 ± 3.3 for the diabetic. 5% of the population are diabetic, so the prior probabilities of health and disease are taken as $\pi_h = 0.95$ and $\pi_d = 0.05$. The relevance of these population estimates for any individual patient, are subject to substantial uncertainty as discussed earlier: the individual, due to idiosyncrasies of personal history, present medical condition, and genetics, may belong to a sub-

population which is not well represented by the population which was used to estimate these means and variances, as discussed in section 1.

2.2 Probabilities and Bayes Risk

We will consider normal distributions of the measured indicator, FPG, with uncertain mean and variance. Let $Z(x, \mu, \sigma)$ denote the normal cumulative distribution function with mean μ and variance σ^2 .

Let μ_h and σ_h denote mean and standard deviation for the healthy state, with estimated values $\tilde{\mu}_h$ and $\tilde{\sigma}_h$. Let ε_{μ_h} and ε_{σ_h} denote errors of these estimates. Analogous quantities for the diseased state are denoted with the subscript d instead of h , i.e., μ_d , σ_d , $\tilde{\mu}_d$, $\tilde{\sigma}_d$, ε_{μ_d} and ε_{σ_d} . In our numerical example we have $\tilde{\mu}_h < \tilde{\mu}_d$.

A *false alarm* occurs when a healthy patient is mis-diagnosed as being diseased. A *missed detection* occurs when a diseased patient is diagnosed as healthy.

The total probabilities of false alarm and missed detection are

$$P_{fa}(\theta, \mu_h, \sigma_h) = [1 - Z(\theta, \mu_h, \sigma_h)]\pi_h \quad (2)$$

$$P_{md}(\theta, \mu_d, \sigma_d) = Z(\theta, \mu_d, \sigma_d)\pi_d \quad (3)$$

The Bayesian risk $R(\theta)$ is a weighted average of these two probabilities of false diagnosis:

$$R(\theta) = \lambda P_{fa}(\theta, \mu_h, \sigma_h) + (1 - \lambda) P_{md}(\theta, \mu_d, \sigma_d) \quad (4)$$

where $0 \leq \lambda \leq 1$ is chosen by the analyst to express the relative importance of missed detection and false alarm. The diagnostic procedure – which depends on the decision threshold θ and on the population moments – is successful if the Bayesian risk is small.

The choice of the value of λ entails a difficult value judgment of the relative importance of false alarm and missed detection. On the one hand, false alarm can result in unneeded and perhaps costly or harmful medical intervention. On the other hand, missed detection can result in severe medical consequences due to lack of treatment. It is well known that these two considerations trade-off, one against the other. Furthermore, examination of eqs.(2)-(4) shows that it is important to consider the quantities $\lambda\pi_h$ and $(1 - \lambda)\pi_d$, since in many situations (like the diagnosis of diabetes), π_h and π_d differ greatly in magnitude. In our subsequent example we use $\lambda = 0.025$, which results in a factor-of-two emphasis on missed detection over false alarm: $[(1 - \lambda)\pi_d] / [\lambda\pi_h] = 2.05$.

2.3 Info-Gap Models of Uncertainty

We know estimated means and standard deviations for healthy and diseased states, $\tilde{\mu}_h$, $\tilde{\sigma}_h$, $\tilde{\mu}_d$, and $\tilde{\sigma}_d$. However, we are uncertain of the pertinence of these data to the sub-population to which a specific individual patient belongs. More precisely, the fractional-error between each estimated value and the true value for the sub-population of that patient is unknown. We will use the following non-probabilistic fractional-error info-gap models to represent this uncertainty in the means and standard deviations of the estimated distributions:

$$U_h(\alpha) = \left\{ \mu_h, \sigma_h : \left| \mu_h - \tilde{\mu}_h \right| \leq \varepsilon_{\mu_h} \alpha, \quad \left| \sigma_h - \tilde{\sigma}_h \right| \leq \varepsilon_{\sigma_h} \alpha, \quad \sigma_h \geq 0 \right\}, \quad \alpha \geq 0 \quad (5)$$

$$U_d(\alpha) = \left\{ \mu_d, \sigma_d : \left| \mu_d - \tilde{\mu}_d \right| \leq \varepsilon_{\mu_d} \alpha, \quad \left| \sigma_d - \tilde{\sigma}_d \right| \leq \varepsilon_{\sigma_d} \alpha, \quad \sigma_d \geq 0 \right\}, \quad \alpha \geq 0 \quad (6)$$

$U_h(\alpha)$ is an unbounded family of nested sets of μ_h and σ_h values. In the absence of uncertainty, i.e., $\alpha = 0$, the set contains only the estimated values. As the horizon of uncertainty, α , increases, the set becomes more inclusive. $U_h(\alpha)$ is a non-probabilistic representation of the uncertainty in the moments of the healthy distribution. Analogous statements are true for $U_d(\alpha)$ as well, regarding the diseased distribution.

2.4 Info-Gap Robustness

The robustness of threshold value θ is the greatest horizon of uncertainty, α , at which the Bayesian risk, $R(\theta)$, does not exceed a specified critical level, R_c . The robustness function is a quantitative answer to this question: by how much can the estimated moments of the probability distributions err, and the Bayesian risk will not exceed R_c ? The robustness is denoted by $\hat{\alpha}(\theta, R_c)$, and is defined in eq.(10) of the appendix.

The robustness depends on the estimated moments of the measurement x , and on the decision threshold θ . Since more robustness is better than less robustness, we can use the robustness function to choose between different values of the threshold, θ .

2.5 Numerical Example

Fig. 1 shows robustness curves for three different choices of the decision threshold θ , evaluated as described in the appendix. We will discuss three central features of these curves.

Trade-off between robustness and risk. The positive slope of the curves shows that increased robustness against uncertainty in the estimated moments can be obtained only by accepting greater Bayesian risk. Requiring lower risk entails accepting lower immunity against uncertainty. This trade-off is an unavoidable mathematical property of all robustness curves.

Zero robustness of the estimated risk. It is a mathematical theorem that each robustness curve reaches the horizontal axis – where the robustness equal zero – precisely at the estimated value of the Bayesian risk. When we use our estimates of the healthy and diseased moments to evaluate the Bayesian risk in eq.(4), we obtain a particular value for each value of θ ; call this value of the estimated risk $\tilde{R}(\theta)$. If we adopt this value of risk as the critical value, $R_c = \tilde{R}(\theta)$, then we find the robustness precisely equals zero: $\hat{\alpha}(\theta, R_c) = 0$. This means that we cannot rely on attaining the estimated value of the Bayesian risk: the robustness of the estimate of that risk is zero.

In light of the trade-off between risk and robustness, we conclude that we can reliably attain only those values of risk which exceed the estimated risk. The robustness curve provides an assessment of the enhanced confidence – robustness to uncertainty – obtained in exchange for greater Bayesian risk. This holds for *any* decision threshold θ . Changing the threshold may reduce the estimated Bayesian risk, $\tilde{R}(\theta)$, but the robustness for achieving that reduced risk will still be zero. In short, in evaluating the performance of any proposed decision threshold, θ , we must "migrate up" the robustness curve for that threshold, seeking a point whose robustness is adequately large and whose risk is adequately small.

Robustness curves can cross one another. We see in fig. 1 that the robustness curves cross each other. In particular, the curves the $\theta = 6$ and $\theta = 7$ cross one another quite near the horizontal axis. The estimated robustnesses for these thresholds

are $\tilde{R}(6) = 0.0094$ and $\tilde{R}(7) = 0.0105$. That is, the average Bayesian risk is 94 in 10,000 for the smaller threshold, and 105 in 10,000 for the larger threshold. Based on these estimated risks one would prefer the smaller threshold, $\theta = 6$.

We know, however, that the robustness to uncertainty for achieving either of these low risks is zero; only larger risks have positive robustness. As we "migrate up" the robustness curve for $\theta = 6$ we cross below the robustness curve for $\theta = 7$. This intersection occurs at risk $R_x = 0.0130$ and at robustness $\hat{\alpha}_x = 0.23$. For any Bayesian risk in excess of 130 in 10,000, the larger threshold ($\theta = 7$) is more robust to uncertainty than the smaller threshold ($\theta = 6$). Furthermore, the robustness at the intersection is quite low in light of how the info-gap model has been implemented (details in the caption to fig. 1). Robustness of 0.23 implies that the diseased mean or standard deviation can err by as much as $0.23 \times 0.2 \tilde{\mu}_d$, and $0.23 \times 0.2 \tilde{\sigma}_d$ but no more, in order to assure that the Bayesian risk does not exceed 0.0130. A tolerable error of 4.6% of the mean or standard deviation is quite small. The healthy mean or standard deviation can err by no more than $0.23 \times 0.1 \tilde{\mu}_h$, and $0.23 \times 0.1 \tilde{\sigma}_h$; tolerable error of only 2.3% in each moment. If the analyst judges that the moments could reasonably err more than these margins, then one would prefer $\theta = 7$ over $\theta = 6$.

It is useful also to consider the level of Bayesian risk which can be guaranteed at large robustness. For instance, at $\hat{\alpha} = 1.5$, we see from fig. 1 that threshold $\theta = 6$ can guarantee a Bayesian risk of 0.0375; that is, an average probability of false diagnosis of 375 per 10,000. In contrast, threshold $\theta = 7$ can guarantee a Bayesian risk of 0.0287; substantially better than the lower threshold.

Finally, let us note that the crossing of robustness curves for different thresholds implies that the robust-satisficing decision may differ from the decision which is optimal based on the estimated risk. Specifically, these decisions *will* differ if the analyst needs robustness greater than the value at which the curves cross, $\hat{\alpha}_x$, and can accept risk greater than the crossing value, R_x .

3 THE ROBUST-SATISFICING HYPOTHESIS

3.1 What is Hypothesized?

The Bayesian risk is the average probability of false diagnosis, either missed detection or false alarm, defined in eq.(4). The physician and patient aspire to achieve small Bayesian risk, but this is difficult since the moments of the statistical distributions are uncertain. Our hypothesis is that the robustness function (which can be calculated without probabilistic information about the uncertain entities) can be used to reliably choose the decision threshold to assure acceptably small Bayesian risk. That is, a non-probabilistic strategy is able to select between threshold values according to their probability of success. We will not know the actual value of the Bayesian risk, but we will be able to maximize the probability that the Bayesian risk is acceptably small.

We have explained in section 2.5 how the robustness curves are used to select between alternative values of the decision threshold, by assessing the enhanced confidence – robustness to uncertainty – obtained in exchange for greater Bayesian risk. But robustness is not necessarily the same as likelihood of success. Nonetheless our hypothesis asserts that robustness is indeed a proxy for probability:

Greater robustness corresponds to greater probability of not exceeding the specified value of Bayesian risk.

Let us make this hypothesis perfectly clear, in order to understand why its truth is not simply a matter of definition.

Consider two different decision thresholds, θ_1 and θ_2 . Suppose that θ_1 is more robust to uncertainty for guaranteeing Bayesian risk no larger than R_c :

$$\hat{\alpha}(\theta_1, R_c) > \hat{\alpha}(\theta_2, R_c) \quad (7)$$

What is the *probability* that the Bayesian risk will in fact be no greater than R_c , if we choose threshold θ_1 or θ_2 ? That is, what is the probability that:

$$R(\theta_i) < R_c \quad (8)$$

We have argued that eq.(7) implies that we should prefer threshold θ_1 over θ_2 because θ_1 is more robust to uncertainty. But is θ_1 actually more likely than θ_2 to achieve Bayesian risk below R_c ? We are unable to answer this question, since the uncertainty we are concerned with is uncertainty in the statistical moments of the distribution of FPG in health and disease. ($R(\theta_i)$ depends on these moments; see eq.(4).) We do not know the probability distribution of these moments, so we cannot calculate the probability that the Bayesian risk will not exceed the value R_c .

Our hypothesis is that robustness is a *proxy for probability*. Namely, we hypothesize that eq.(7) implies that θ_2 is more likely than θ_1 to result in Bayesian risk less than R_c :

$$\text{Prob}[R(\theta_1) \leq R_c] > \text{Prob}[R(\theta_2) \leq R_c] \quad (9)$$

That is, we hypothesize that eq.(7) implies eq.(9). This is a risky hypothesis because the robustness function, $\hat{\alpha}(\theta, R_c)$ has no probabilistic information behind it. The info-gap models on which it is based are much less informative than probability models.

3.2 Testing the Hypothesized

One can test the hypothesis by repeatedly estimating the Bayesian risk based on observed rates of missed detection and false alarm. To do this, consider two thresholds, θ_1 and θ_2 , whose robustnesses are ranked as in eq.(7). Use these thresholds for diagnosing diabetes in separate populations, and calculate the Bayesian risk for each population based on the observed rates of false alarm and missed detection. The hypothesis is rejected if and only if the observed frequency with which θ_1 satisfies eq.(8) is less than the observed frequency with which θ_2 satisfies eq.(8).

The hypothesis could be falsified for either or both of two reasons. First, the info-gap models, eqs.(5) and (6), could be wrong. For instance, we have assumed that the distributions are in fact normal and only the moments are info-gap-uncertain; the distributions may be non-normal on their tails. Error in the info-gap models would result in erroneous robustness curves which could cause erroneous ordering in eq.(7).

Second, the hypothesis could fail if robustness is not a proxy for probability of success. That is, failure could result if eq.(7) does not imply eq.(9).

3.3 What is the Hypothesized Plausible?

The info-gap models of uncertainty which underlie our example, eqs.(5) and (6), assume two things: (i) that the cumulative distribution functions (cdf's) are normal (Gaussian) and (ii) that the estimated moments deviate fractionally by an unknown amount. The assumption of normality is reasonable if the population to which the

patient belongs is statistically homogeneous and large enough for the central limit theorem to assure normality.

The assumption of unbounded fractional-error assumes that we know estimated values of the moments, and errors of these estimates. However, these errors do not constitute worst cases and we don't know the extent to which the measured population actually represents the population to which our patient belongs. This info-gap model imposes a very weak type of order on the uncertainty. It treats the estimated moments as "best guesses" around which increasingly different possible values evolve as the horizon of uncertainty rises. The actual horizon of uncertainty is unknown, though there is an assumption of gradualness or continuity in the uncertain deviation of our patient's population from the measured population. The continuity between cause (health or disease) and effect (the corresponding cdf) underlies inference and decision in info-gap theory as well as in probability theory (9).

In addition to the plausibility of the very weak assumption of unbounded fractional error, we will cite two examples in which this assumption has proven useful. These two examples also suggest that the hypothesis that robustness is a proxy for probability may be true.

Carmel and Ben-Haim (10) study foraging behavior of animals from a very wide range of taxa. They show that a robust-satisficing foraging strategy, based on a fractional-error info-gap model, seems to be more consistent with field and laboratory observation than strategies based on maximizing the energy intake. This makes sense since, though an animal needs energy in order to survive, maximal intake is not necessary; a sub-maximal critical amount of energy will suffice. If robust-satisficing is prevalent in nature, as suggested by the study, then robust-satisficing should have a survival advantage over other strategies. This survival advantage is that robust satisficing is more likely to achieve at least the critical energy intake than other strategies. This is contingent on the realism of the fractional-error info-gap model, and on robustness being a proxy for the probability of success, as suggested by that study.

Our second example concerns the equity premium puzzle in financial economics. Risky assets like stocks have higher average returns than risk-free assets like government bonds. This "equity premium" of risky assets is generally understood as necessary to attract investors. The puzzle is that standard economic models do not explain the size of the equity premium (11, 12). These standard models assume that investors attempt to maximize their returns. However, an investment can be justified if its return exceeds that of alternatives, even if not maximal. Like foraging animals, investors do not need to maximize profits in order to survive; it is sufficient to beat the competition. In fact, a robust-satisficing model of investment behavior, based on a fractional-error info-gap model, seems to explain the equity premium puzzle (1, section 11.4). The apparent prevalence of robust-satisficing among investors suggests that robust-satisficing has survival value in economic competition: robust-satisficing is more likely than other strategies to yield competitive returns. This is contingent on the validity of the underlying fractional-error info-gap model, and on the robustness being a proxy for probability of success.

3.4 Why is the Hypothesized Medically Important?

The hypothesis of this paper, if true, has an important practical implication for clinical medical practice, as we now explain.

A physician making a differential diagnosis and a patient working to understand the risks and benefits of a therapeutic plan, are both dealing with severe uncertainty. At present their option is to understand the population based literature and try to

understand how their problem fits in the context of a trial or observation. The only thing that they know for certain about the model that they choose as a reference is that the model is incorrect. It is incorrect because the person about whom the decision is to be made was not a part of that study or trial. Thus, he or she is compelled to make a decision by estimating how much he or she is actually like the population for which data are available.

Conventionally, physicians and patients alike seek to optimize. They ask the question, "what can I do for myself (or my patient) that will assure the most favorable outcome?" This formulation, as we have shown, although attractive on the surface, is highly info-gap uncertain. This means that the person making the decision may have little confidence that the chosen outcome will actually occur. In the context of false positives and false negatives, a satisficing approach does not attempt to optimize the set of diagnostic inclusions, because doing so requires exhaustively exploiting data which are suspect of error. Rather, the robust-satisficing hypothesis, if true, provides maximal confidence that the diagnostic threshold satisfactorily includes illness and excludes those without illness.

It is important to note that there are many cases, perhaps even most cases, where the robust satisficing approach and the optimizing approach yield the same decision threshold. In this case, the user can have enhanced confidence that the threshold values chosen are indeed correct. In the cases where the optimizing and robust satisficing approach differ, there is a graphic and useful display of the decision trade-offs that must be made to increase confidence. If the premium to pay for increasing confidence is too high, the user may still elect the optimizing approach. However this is done with eyes wide open, knowing that confidence in the outcome will decrease.

Robust satisficing in the design of diagnostic thresholds adds a valuable dimension to disease classification and acknowledges the enormous variation in disease expression, the heterogeneity of the population, and the uncertainties underlying every medical interaction.

Acknowledgement:

This work was supported in part by a grant from the Abramson Center for the Future of Health, Houston, TX

References

1. Ben-Haim Y, 2006, *Info-Gap Decision Theory: Decisions Under Severe Uncertainty*, 2nd edition, Academic Press, London.
2. Greenfield S, Kravitz R, Duan N, Kaplan SH. Heterogeneity of treatment effects: Implications for guidelines, payment, and quality assessment. *Am. J. Med.* 2007; **120**:S3-S9.
3. Feinstein AR, Horwitz RI. Problems in the "evidence" of "evidence-based medicine". *Am. J. Med.* 1997; **103**:529-535.
4. Morimoto T, Fukui T, Lee TH, Matsui K. Application of U.S. guidelines in other countries: Aspirin for the primary prevention of cardiovascular events in Japan. *Am. J. Med.* 2004; **117**:459-468.
5. McLaughlin MJ, for the members of the HTE Policy Roundtable Panel. Healthcare policy implications of heterogeneity of treatment effects. *Am. J. Med.* 2007; **120**:S32-S35.
6. Glasziou PP, Irwig L, Heritier S, Simes RJ, and Tonkin A, for the LIPID Study Investigators, Monitoring cholesterol levels: Measurement error or true change? *Annals of Internal Medicine*, 2008; **148**:656-661.
7. Simon, HA, A Behavioral Model of Rational Choice, *Quarterly Journal of Economics*, 1955; **69**:174-183.
8. de Vegt F, Dekker JM, Stehouwer CDA, Nijpels G, Bouter LM and Heine RJ, The 1997 American Diabetes Association criteria versus the 1985 World Health Organization criteria for the diagnosis of abnormal glucose tolerance, *Diabetes Care*, 1998; **21**:1686-1690.
9. Ben-Haim, Y, Set-models of information-gap uncertainty: Axioms and an inference scheme, *Journal of the Franklin Institute*, 1999; **336**:1093-1117.
10. Carmel Y. and Ben-Haim, Y, Info-gap robust-satisficing model of foraging behavior: Do foragers optimize or satisfice?, *American Naturalist*, 2005; **166**: 633-641.
11. Mehra, R. and Prescott, E.C., The equity premium: A puzzle, *Journal of Monetary Economics*, 1985; **15**: 145-161.
12. Kocherlakota, N.R., The equity premium: It's still a puzzle, *Journal of Economic Literature*, 1996; **34**: 42-71.

Appendix A: Evaluation of the Robustness

The robustness function is defined as:

$$\hat{\alpha}(\theta, R_c) = \max \left\{ \alpha : \left(\max_{\mu_d, \sigma_d \in U_d(\alpha); \mu_h, \sigma_h \in U_h(\alpha)} R(\theta) \right) \leq R_c \right\} \quad (10)$$

We will evaluate the robustness function by evaluating its inverse.

Let us denote the inner maximum in eq.(10) by $m(\alpha)$. The robustness, $\hat{\alpha}(\theta, R_c)$ is the greatest value of α at which $m(\alpha) \leq R_c$. Note that $m(\alpha)$ increases monotonically as α increases due to the nesting of the info-gap model. Thus the robustness is the greatest α satisfying $m(\alpha) = R_c$. In other words, $m(\alpha)$ is the inverse of $\hat{\alpha}(\theta, R_c)$.

To evaluate $m(\alpha)$ let us define two new functions:

$$m_d(\alpha) = \max_{\mu_d, \sigma_d \in U_d(\alpha)} P_{md}(\theta, \mu_d, \sigma_d) \quad (11)$$

$$m_h(\alpha) = \max_{\mu_h, \sigma_h \in U_h(\alpha)} P_{fu}(\theta, \mu_h, \sigma_h) \quad (12)$$

From eq.(4) we see that:

$$m(\alpha) = \lambda m_h(\alpha) + (1 - \lambda) m_d(\alpha) \quad (13)$$

One can readily find expressions for $m_h(\alpha)$ and $m_d(\alpha)$. One finds:

$$m_d(\alpha) = \begin{cases} Z(\theta, \tilde{\mu}_d - \varepsilon_{\mu_d} \alpha, \tilde{\sigma}_d + \varepsilon_{\sigma_d} \alpha) \pi_d & \theta \leq \tilde{\mu}_d - \varepsilon_{\mu_d} \alpha \\ Z(\theta, \tilde{\mu}_d - \varepsilon_{\mu_d} \alpha, r[\tilde{\sigma}_d - \varepsilon_{\sigma_d} \alpha]) \pi_d & \theta > \tilde{\mu}_d - \varepsilon_{\mu_d} \alpha \end{cases} \quad (14)$$

where we have defined the ramp function $r(x) = x$ when $x \geq 0$, and $r(x) = 0$ otherwise. Note that $m_d(\alpha) = 1/2$ when $\theta = \tilde{\mu}_d - \varepsilon_{\mu_d} \alpha$. Thus, $m_d(\alpha)$ is continuous but not necessarily smooth as eq.(14) moves from the top to the bottom line.

Similarly, $m_h(\alpha)$ is given by:

$$m_h(\alpha) = \begin{cases} [1 - Z(\theta, \tilde{\mu}_h + \varepsilon_{\mu_h} \alpha, r[\tilde{\sigma}_h - \varepsilon_{\sigma_h} \alpha])] \pi_h & \theta \leq \tilde{\mu}_h + \varepsilon_{\mu_h} \alpha \\ [1 - Z(\theta, \tilde{\mu}_h + \varepsilon_{\mu_h} \alpha, \tilde{\sigma}_h + \varepsilon_{\sigma_h} \alpha)] \pi_h & \theta > \tilde{\mu}_h + \varepsilon_{\mu_h} \alpha \end{cases} \quad (15)$$

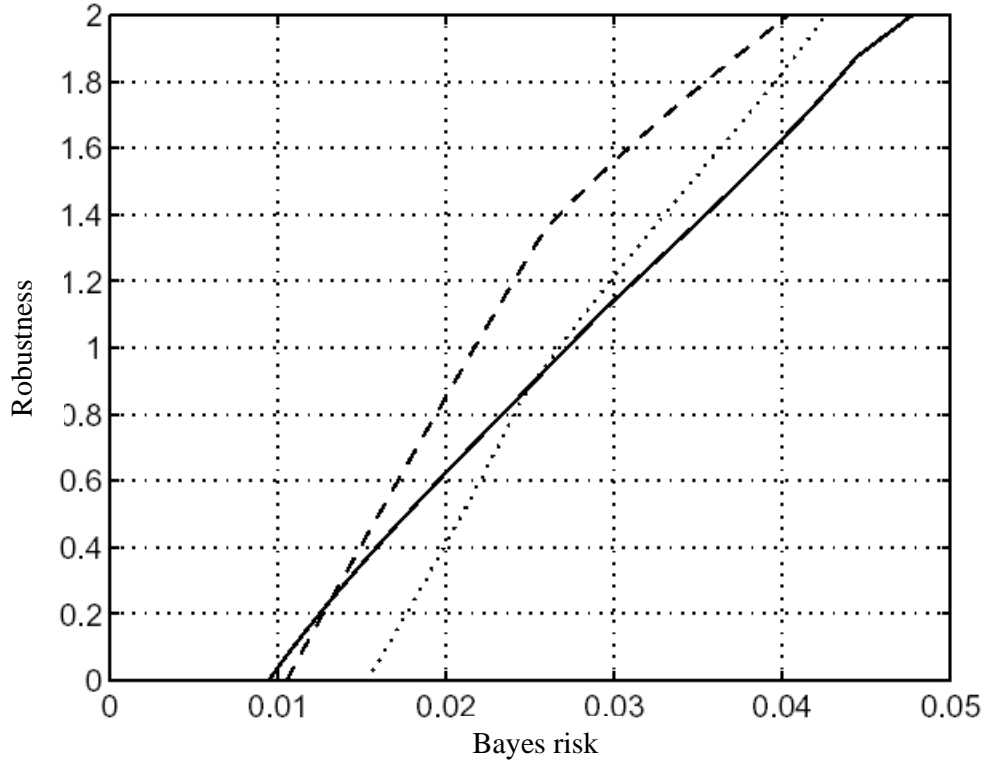


Figure 1: Robustness, $\hat{\alpha}(\theta, R_c)$, vs. critical level R_c for three thresholds θ : $\theta = 6$ (solid), $\theta = 7$ (dash) and $\theta = 8$ (dot). Parameters for the normal distribution for the healthy state: $\tilde{\mu}_h = 5.4$, $\tilde{\sigma}_h = 0.5$, $\varepsilon(\mu_h) = 0.1\tilde{\mu}_h$, $\varepsilon(\sigma_h) = 0.1\tilde{\sigma}_h$; Parameters for the normal distribution for the diseased state: $\tilde{\mu}_d = 9.6$, $\tilde{\sigma}_d = 3.3$, $\varepsilon_{\mu_d} = 0.2\tilde{\mu}_d$, $\varepsilon_{\sigma_d} = 0.2\tilde{\sigma}_d$; The relative importance of missed detection and false alarm is determined by $\lambda = 0.025$.