

# **DO SENSITIVITY ANALYSES REALLY CAPTURE PROBLEM SENSITIVITY?**

*An Empirical Analysis Based on Information Value*

James C. Felli

Naval Postgraduate School

Monterey, CA 93943

Gordon B. Hazen

Northwestern University

Evanston, IL 60208

March 1998

Please direct all queries and requests for reprints to James C. Felli, DRMI (Code 64 FL), Naval Postgraduate School, 1522 Cunningham Road, Monterey, CA, 93943-5201.

## ABSTRACT

The most common methods of sensitivity analysis (SA) in decision-analytic modeling are based either on proximity in parameter-space to decision thresholds or on the range of payoffs that accompany parameter variation. As an alternative, we propose the use of the expected value of perfect information (EVPI) as a sensitivity measure and argue from first principles that it is the proper measure of decision sensitivity. EVPI has significant advantages over conventional SA, especially in the multiparametric case, where graphical SA breaks down. In realistically sized problems, simple one- and two-way SAs may not fully capture parameter interactions, raising the disturbing possibility that many published decision analyses might be overconfident in their policy recommendations. To investigate the extent of this potential problem, we re-examined 25 decision analyses drawn from the published literature and calculated EVPI values for parameters on which sensitivity analyses had been performed, as well as the entire set of problem parameters. While we expected EVPI values to indicate greater problem sensitivity than conventional SA due to revealed parameter interaction, we in fact found the opposite: compared to EVPI, the one- and two-parameter SAs accompanying these problems dramatically *overestimated* problem sensitivity to input parameters. This phenomenon can be explained by invoking the flat maxima principle enunciated by von Winterfeldt and Edwards.

## INTRODUCTION

The traditional role of sensitivity analysis (SA), as applied *post-hoc* to decision models, has been to provide the decision maker (DM) with insight into the susceptibility of the optimal alternative to changes in the values of uncertain model parameters. Graphical 1-way and 2-way SA techniques are widely practiced, but difficulties often arise in N-way SA for  $N \geq 3$ . As a result, analysts typically perform only one-way SAs, supplemented by two-way (and occasionally three-way) SAs on parameters which are thought important. Because published decision analyses can contain 20 or more parameters, it is conceivable that such limited scale SA might overlook extensive parameter interactions and thereby significantly underestimate problem sensitivity. This raises the disturbing possibility that significant numbers of decision analyses are overconfident in their recommendations, and that many more “toss-ups” exist than are reported.

We believe that the expected value of perfect information (EVPI) is the proper measure of problem sensitivity to input parameters. In addition to its theoretical soundness, it has the practical advantage that computation can be done for parameter sets of arbitrary size. Therefore, if there exist higher-level parameter interactions which affect problem sensitivity, a multi-parameter EVPI calculation should detect them. Consequently, we sought to evaluate the adequacy of traditionally performed SAs by comparing them with the “gold standard” of multi-parameter EVPI. For a selection of 25 problems from the literature, we calculated EVPIs for parameter sets corresponding to the authors’ SAs and compared our EVPI-based sensitivity labels with the labels the authors’ assigned. We found the opposite of what we expected: instead of a widespread understatement of sensitivity, we discovered a pervasive tendency for overstatement. In many cases we found that entire sets of parameters to which analysts deemed a decision problem sensitive individually had little information value, even as a group. As we will show, our findings can be explained by invoking the flat maximum principle first enunciated by von Winterfeldt and Edwards (1973, 1986).

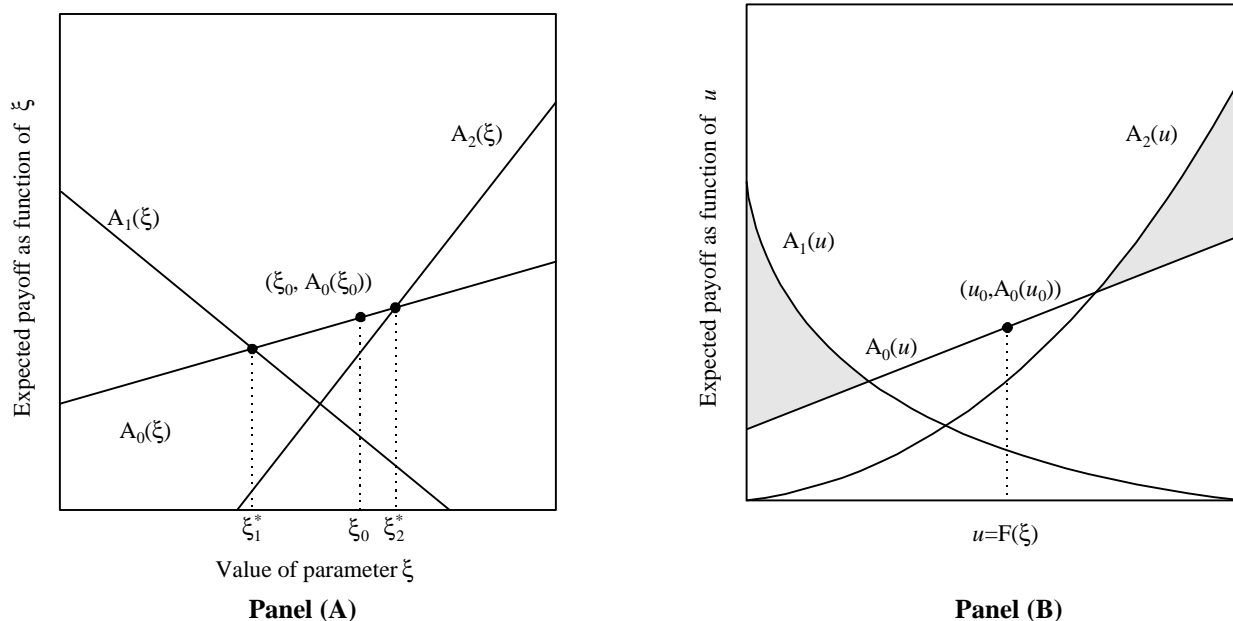
## SENSITIVITY ANALYSIS AND INFORMATION VALUE

The application of SA with which we are concerned is the *post-hoc* appraisal of the stability of a base optimal alternative (BOA) for a probabilistic decision model. In the following sections, we briefly discuss parameter-space threshold-proximity SA and value-range SA, then motivate the use of EVPI as a sensitivity measure.

### Parameter-space threshold proximity methods

*Parameter-space threshold proximity methods* (PSTP) use nearness in parameter-space to a decision threshold as a proxy for sensitivity. Figure 1A depicts a graphical threshold proximity SA for a parameter  $\xi$  with base value  $\xi_0$ . The curves  $a_0(\xi)$  and  $a_i(\xi)$  for  $i \neq 0$  denote the expected payoff curves of the BOA and competing alternatives, respectively, as functions of  $\xi$ .

In Figure 1A, the parameter space is partitioned into three intervals,  $\xi \leq \xi_1^*$ ,  $\xi_1^* < \xi < \xi_2^*$ , and  $\xi \geq \xi_2^*$ , where a different alternative is optimal. The point boundaries shared by two intervals are referred to as *thresholds* and designate points of indifference between adjacent alternatives.  $\xi_0$  is the base value of  $\xi$  and identifies  $a_0$  to be BOA. By examining the proximity of  $\xi_0$  to its neighboring thresholds and contrasting it with her beliefs about the likely values of  $\xi$ , the DM can get a feel for how the optimal alternative is likely to change with variation in  $\xi$ .



**Figure 1.** The payoff curves for three competing alternatives, plotted (A) as functions of a parameter  $\xi$  and (B) as functions of the fractiles  $u = F(\xi)$ .

PSTP approaches direct the DM's attention to the amount of parametric perturbation required to rescind the optimality of the BOA. They are quickly constructed and easily interpreted in the one- and two-parameter case. Still, sensitivity conclusions are critically tied to the DM's beliefs about the behavior of the parameter(s) under examination. In Figure 1A, for example, two DMs, one fairly certain that  $\xi_1^* < \xi < \xi_2^*$  and one only agreeable that  $\xi > \xi_1^*$ , could draw different conclusions regarding the sensitivity of the problem to  $\xi$ .

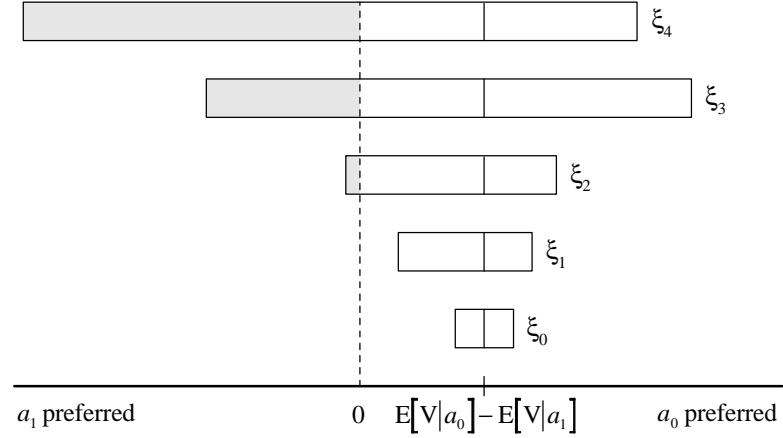
The kind of graphical support in Figure 1 is unfortunately unavailable for N-way SA when  $N > 3$ . This is regrettable as a decision may be insensitive to the variation of some sets of parameters individually, but sensitive to their simultaneous variation (O'Brien et al. 1994, Felli and Hazen 1998). Some researchers have recommended the calculation of the distance (in parameter space) from the vector of base parameter values to the nearest threshold as a numerical proxy for the likelihood of a decision change for multiparametric problems (Isaacs 1963, Starr 1966, Fishburn et al. 1968, Evans 1984, Buckley 1988).

The choice of an appropriate distance metric then becomes a concern, and difficulties related to non-commensurable units of measure across parameters, or equivalent but different methods of jointly defining parameters, can abrogate the validity of the final measure (Schneller and Sphicas 1985, Rios Insua and French 1991). Regardless of dimensionality, the lack of established guidelines for discriminating between sensitivity and insensitivity can result in essentially arbitrary interpretation of analytic results (O'Brien et al. 1994, Sonnenberg et al. 1994).

### **Value-range measures of decision sensitivity**

Tornado diagrams and scenario analyses comprise the most common *value-range methods* (VR) of SA encountered in the literature. Tornado diagrams have been employed primarily for single parameter SA whereas scenario analyses have been reserved for multiparametric sensitivity investigations. For a detailed discussion of tornado diagrams, we refer the reader to Clemen (1991); for illustrative examples of scenario analyses, see Velanovich and Tapper (1993) and Zwetsloot-Schonk and Leer (1993).

Tornado diagrams are usually employed to compare the BOA  $a_0$  to another alternative  $a_1$  and employ horizontal bars to represent ranges of payoffs when specified parameters are allowed to vary. An example is given in figure 2, where the bars represent the payoff difference  $E[V|a_0, \xi] - E[V|a_1, \xi]$  as each parameter  $\xi$  varies. Bars are arranged from widest to narrowest, giving a “tornado” shape around  $E[V|a_0] - E[V|a_1]$ , which occurs when all parameters are at their base values. The portion of the bar beyond which payoff differences become negative is shaded, and corresponds to values of  $\xi$  for which alternative  $a_1$  becomes optimal. The size of the shaded portions of each bar are used to make sensitivity judgments. For example, Figure 2 would indicate substantial sensitivity to parameters  $\xi_4$  and  $\xi_3$ , and possible sensitivity concerns about  $\xi_2$ .



**Figure 2.** Hypothetical tornado diagram for a dichotomous, five parameter decision problem with BOA  $a_0$ .

In a scenario analysis, certain problem parameters are given (often extreme) values to bias an analysis in favor of or against a specific alternative while remaining parameters are held fixed at base values. The problem is considered sensitive to the parameter set under investigation if the scenario causes a change in the BOA. Results of scenario analyses are most often expressed numerically.

### EVPI as a sensitivity measure

EVPI( $\xi$ ) is defined as the difference between the optimal expected payoff with knowledge of  $\xi$  at the time of decision and the optimal expected payoff without such knowledge (Howard 1966, 1967). A well-known equivalent definition casts EVPI( $\xi$ ) as the expected foregone payoff (or expected opportunity loss) suffered for not utilizing perfect information about  $\xi$  (Raiffa and Schlaifer 1961). If  $a^*$  maximizes  $E[V|a]$  and  $a^*(\xi)$  maximizes  $E[V|a, \xi]$ , then the expected foregone payoff can be expressed as  $EVPI(\xi) = E_{\xi} [E[V|a^*(\xi), \xi] - E[V|a^*, \xi]]$ . We believe that this representation of EVPI as foregone payoff is particularly relevant from a sensitivity analysis perspective: the DM's concern that  $\xi$  might cross a decision threshold should be *value-based*. That  $a^*$  might be *post-hoc* suboptimal should matter only to

the degree that the corresponding foregone payoff suffered by not selecting  $a^*(\xi)$  is large. Because  $EVPI(\xi)$  is expected foregone payoff, it provides a convenient measure of sensitivity.

EVPI shares an intrinsic kinship with traditional SA that lends further credibility to its use as a sensitivity measure.  $EVPI(\xi)$  can be expressed mathematically as the product of the probability of a change in the optimal alternative due to variation in  $\xi$  and the average foregone payoff given such a change, encapsulating likelihood and value issues in a single quantity. In fact, modifying a one-way SA graph to plot expected payoff against the *fractiles* of  $\xi$  presents  $EVPI(\xi)$  as the area above the expected payoff curve for  $a^*=a_0$  and below the expected payoff curve for some other alternative (Eschenbach and Gimpel 1990). A plot of this type is portrayed in Figure 1B. It follows that if a parameter  $\xi$  is uniformly distributed across its range, then the analogous area in a typical one-way SA graph (e.g., figure 1A) is proportional to  $EVPI(\xi)$ .

Under certain simplifying assumptions, tornado diagrams are usefully related to EVPI as well. The shaded region of a tornado bar for a parameter  $\xi$  corresponds to the values of  $\xi$  producing a decision change. Suppose that  $\xi$  is uniformly distributed across its range, and that the payoff function  $E[V|a_0, \xi]$  is linear in  $\xi$ , then (1) the *proportion* of the bar for  $\xi$  which is shaded is the probability of a decision change, and (2) the *length* of the shaded portion of the bar for  $\xi$  is proportional to the expected foregone payoff given a decision change. Therefore the *product* of the proportion (1) and the length (2) is proportional to  $EVPI(\xi)$ . We will have more to say about the relationship between tornado diagrams and EVPI in a forthcoming paper.

Above and beyond the kinship between EVPI and SA, we believe there are four fundamental reasons why EVPI should be the preferred measure of problem sensitivity. Philosophically, subjective probability and subjective expected utility have well-established axiomatic foundations, yet these foundations are mute on the necessity for SA as conventionally practiced. Because it can be cast as the value of a (perhaps hypothetical) information acquisition, there is no such difficulty in employing EVPI a sensitivity measure.

Second, because EVPI can be expressed as a product of the probability of decision change and the expected payoff improvement given that change, it serves as a combined measure of both decision and payoff sensitivity. Third, EVPI handles N-way SA at no extra conceptual cost since  $EVPI(\xi)$  can be readily calculated for vector  $\xi$ . Regardless of dimensionality, the units of EVPI remain the same as problem payoffs (e.g., dollars, quality adjusted life duration), units which are presumably familiar to the DM. Last, as a consequence of its applicability to N-way analysis, EVPI can be used to expose multiple-parameter interactions which might be missed by conventional one- or two-way SA.

The use of EVPI does, however, require the analyst to specify probability distributions for all problem parameters. Even so, in conventional SA, a base value, lower bound and upper bound for each parameter  $\xi$  must be specified; it is a small additional step to treat these values as the median, 5<sup>th</sup> percentile, and 95<sup>th</sup> percentile of the parameter, and apply, for example, the Pearson-Tukey approximation (Keefer and Bodily 1983) to obtain an approximate distribution. A second burden is the actual calculation of EVPI. Typically, this cannot be done in closed form, and some numerical approximation is required. We have found that Monte Carlo simulation works well for this purpose, and may be easily implemented in a spreadsheet environment. Perhaps most difficult is determining what levels of EVPI will indicate sensitivity. We believe that the DM's determination of whether a decision problem is sensitive to a group of parameters with a given EVPI should depend on whether she believes that EVPI to represent a significant forgone payoff. This latter consideration is clearly subjective, but no more so than would be her assessment of whether or not a base point is "sufficiently close" to a threshold to be of concern when interpreting the results of a traditional PSTP SA. We contend that DMs are intrinsically better suited to assess the significance of the results of an EVPI analysis because they are in essence evaluating the worth of marginal improvements measured in the same units as the problem's payoffs.

### EVPI Sensitivity Analysis: An Example

As an example for comparing SA with EVPI, we present a decision analysis of treatment options for deep venous thrombosis in pregnancy (DVT) performed by Klein and Pauker (1981). In addition to the probability parameters, the authors employed an efficacy parameter to modify the base rate of pulmonary embolism for patients receiving anticoagulants, and a parameter to quantify the DM's utility for an adverse fetal outcome consequent to anticoagulation. Their baseline analysis showed the administration of anticoagulants to be the BOA, which yielded an expected utility of 96.5155 on a scale in which survival of both mother and infant had utility 100, and death of both mother and infant had utility 0.

To explore the effects of parametric variation, Klein and Pauker performed a series of PSTP SAs and concluded that the BOA was sensitive to the probability of a pulmonary embolism and the utility of an adverse fetal outcome. We compared these sensitivity conclusions with an EVPI analysis. For consistency in this problem (and others we examined), we assigned parameter distributions based on the authors' formal statements of plausible parameter ranges, the range of values used in SAs, informal commentary about parameter values, and our own inferences about the likelihood of threshold crossings based on authors' sensitivity conclusions (see a subsequent section for details). The distributions we used in our EVPI analysis were piecewise linear over a parameter's feasible range  $[A,B]$  (provided by the authors) with a mode equal to its base value and 95% of the probability mass within the plausible parameter range  $[L,U]$  (provided by the authors). The other 5% of probability density was equally divided between  $[A,L]$  and  $[U,B]$ . The exception was the efficacy of Warfarin, which was stated to be at least 75% with a base value equal to 75%. In this case we assigned a uniform  $(0.75, 1.0)$  distribution to the parameter. The single- and all-parameter results of our EVPI analysis are provided in table 1. We assumed that the parameters were independent random variables for these calculations. The proportion  $\pi$  of the parameter's plausible range lying over the threshold is provided for completeness and will be discussed later.

Parameter	Calculated EVPI	Proportion $\pi$ over threshold	Authors' Declaration
Utility of adverse fetal outcome	1.11154	0.628325	Sensitive
P[Pulmonary Embolism]	0.06846	0.16175	Sensitive
P[Fetopathy]	0.03285	0	Insensitive
P[Death   Pulmonary Embolism]	0.00259	0	Insensitive
Efficacy of Treatment	0	0	Insensitive
P[Fatal Maternal Bleeding]	0	0	Insensitive
All Parameters Jointly	2.9881	*	*

**Table 1.** EVPI results for the DVT problem (Klein and Pauker 1981).

To what parameters does EVPI indicate problem sensitivity? Here is a rough-cut answer: the total-parameter EVPI of 2.9881 constitutes 3.13% of the base-optimal utility of 95.5155, an amount that seems to us significant, so the problem appears sensitive to its parameter set as a whole. The EVPI of the utility of adverse fetal outcome represents a 1.16% improvement over the base optimal 95.5155. It is not clear to us whether an improvement on this order is significant, so it is questionable whether or not the problem is sensitive to this parameter. All remaining parameters have EVPI values less than 0.072%. Sensitivity to these parameters seems implausible. Granted, these sensitivity judgments are simply reasonable conjectures on our part – the final word would rest with the analysts and DMs. It is nonetheless interesting that while the parameters the authors judged sensitive by PSTP standards correspond to those with the highest EVPI values, these values themselves do not seem sufficiently compelling to support the authors' sensitivity declarations. Klein and Pauker appear to have *overestimated* problem sensitivity. We will have more to say about this later.

We performed this type of retrospective analysis on 25 problems from the literature, always selecting parameter distributions based on information authors provided with the goal being to compare authors' sensitivity labels with those based on EVPI values. In the following sections we discuss issues that arose in problem selection and re-analysis before presenting our results.

## EMPIRICAL ISSUES

### Selection of problems for re-analysis

We employed both computer and text-based resources to search for published decision analyses to which we might apply our technique. Our search resulted in 80 papers that appeared promising. Of these 80 papers, we selected 25 problems from 24 papers for re-analysis based upon four criteria: the problems were single attribute, non-Markov in nature (we opted to avoid the extra calculation burden associated with their solution); the authors provided sufficient information for us to reconstruct their model; the authors reported some type of SA on the problem; the authors' included their interpretation of their SAs results (This was required to enable us to compare our EVPI-based sensitivity conclusions with their SA results and useful in determining of parameter distributions and minimum significant improvements.). It was difficult to find published articles satisfying all of these criteria. Fortunately, there exists a large body of medical decision analysis literature in which the convention is to present and explicitly discuss sensitivity analyses. Most of our problems are taken from that literature.

The 25 problems we selected contained 277 parameters, of which 214 were non-remote and 63 were remote (see below). Problem size ranged from 5 (4 non-remote) to 24 parameters (all non-remote).

### Remote versus Local DMs and Parameters

We draw an important distinction between *remote* and *non-remote* parameters. In prescriptive decision analyses designed to address common problems, it is not uncommon for the analyst to provide a sort of "template analysis", where the analysis is complete for the most part, except that the values of certain parameters are left open to be supplied by a remote DM. Data or subjective estimates pertaining to the likely range of such remote parameters are not gathered. For our retrospective analyses, it was impossible to assign probability distributions to such parameters, or to compute EVPI. We therefore omit all remote parameters from the results we present here.

### Matching Sensitivity Labels with EVPI Results

Although increasing EVPI indicates increasing sensitivity, it is not clear where the line between “insensitive” and “sensitive” should be drawn. There are several obvious approaches to this question which we feel are *not* appropriate. One might normalize all EVPI values to [0%,100%] by dividing by the EVPI of the entire parameter set  $\Pi$ , and fix a percentage cutoff for sensitivity. However, this would result in every problem being declared sensitive to  $\Pi$ , regardless of the magnitude of  $EVPI(\Pi)$ . Normalization by the EVPI of all problem uncertainties  $\Omega$  (e.g.,  $\Omega$  would contain events as well as the parameters in  $\Pi$ ) is possible; however,  $EVPI(\Omega)$  is typically much larger than the EVPI of any parameter set, and no obvious percentage demarcation between sensitivity and insensitivity suggests itself for such a ratio. One might normalize by the optimal payoff value for the problem – or rather, since the latter can be zero, by the difference between the best and worse payoff values – but this scheme results in a sensitivity measure which is not invariant under the addition of dominated alternatives. Perhaps normalize by the payoff difference between the best and second-best alternatives? But adding randomized alternatives makes this difference zero.

Because  $EVPI(\xi)$  represents the *additional* payoff obtainable were  $\xi$  known, it seems clear that sensitivity to  $\xi$  should be recognized when  $EVPI(\xi)$  represents a significant additional payoff. To this end, we sought to quantify what the DM considered a *minimum significant improvement* (MSI), a practice that has been championed by advocates of Bayesian statistical approaches to clinical trials (Spiegelhalter and Freedman 1988, Spiegelhalter *et al.* 1994), and likely to become more frequent as Bayesian approaches gain acceptance. In some cases, authors explicitly specified such a quantity. For example, in the choice of therapy for patients suffering from idiopathic membranous nephropathy (Piccoli *et al.* 1994), the authors identified 0.5 QALY (quality adjusted life years remaining) as the minimum improvement they considered significant. Unfortunately, in 17 of our sample of 25 cases, we could not determine the authors’ MSI from their commentary. In order to compare SA and EVPI results in these cases, we adopted the following

hypothesis: a payoff increase constitutes a significant improvement if it exceeds a canonical percentage  $\delta$  (e.g., 1%, 5%, 10%) of the difference between the optimal problem payoff and the status quo suggested by problem context. For each problem in which the authors did not explicitly provide a MSI, we selected what we felt was a reasonable status quo and conducted sensitivity comparisons using  $\delta=1\%$ , 5% and 10%. These values for  $\delta$  were selected based upon our interpretation of what improvement the authors might consider significant based upon their analysis and commentary and the magnitudes of the  $\delta$  values we calculated *post-hoc* for the 8 problems in which the authors provided sufficient information for us to determine their MSI. We will use the term “relevant MSI” to designate the relevant value for the problem, either the authors’ stated MSI or the appropriate  $\delta$  hurdle.

### **Selection of Parameter Distributions**

EVPI calculation requires explicit distributions for parameter values. As none of our authors provided such distributions, we were forced to estimate them based upon information available from each article. This included: a parameter’s base value; a parameter’s feasible range (the set of logically possible values for the parameter); a parameter’s plausible range (the set of values the authors believed the parameter most likely to obtain); the range of values the authors used in their SAs; and the authors’ commentary where applicable. Our general approach was to select distributions for non-remote parameters so that (1) the distribution mode conformed with the authors’ base parameter value, (2) 95% of the probability mass fell within the plausible range, and (3) probability mass not contained in the plausible range spilled over equally either side. If there was no room beyond one endpoint of the plausible range (because it coincided with an endpoint of the feasible range), then we assigned 97.5% of the probability mass within the plausible range and 2.5% beyond the opposite plausible endpoint. If the feasible and plausible range coincided, then the plausible range received 100% of the probability mass.

We limited our distribution selection to piecewise linear and uniform distributions for bounded parameters (e.g., probabilities), and to generalized gamma distributions for unbounded parameters (e.g.,

days hospitalized). For remote parameters we assigned a degenerate distribution at the base value. Our selection of distributions for non-remote parameters was subject to the following considerations (for reference, we include the number of parameters falling into each category). If the authors explicitly specified a plausible range (94 parameters), or we could infer one from their SA (68 parameters) or commentary (22 parameters), then we assigned a distribution with 95% of its mass within the plausible range and equal amounts of mass beyond that range's upper and lower limits whenever possible. The exception to this general procedure was when the authors specified a base value at an extreme limit of the plausible range (11 parameters), where we used a uniform distribution over the plausible range.

We used a different approach if we could not infer a plausible range in these ways but the authors based their sensitivity judgments on PSTP. If the feasible range was a bounded interval (13 parameters), we set the initial plausible range to be the interval half-way between the base parameter value and the feasible range endpoints; if the feasible range was  $[0, \infty)$  (6 parameters), we used the interval [50% base value, 150% base value]. Where appropriate, the initial plausible range was expanded to accommodate sensitive thresholds and contracted to accommodate insensitive thresholds. The expansion was to halfway between the nearest threshold and the limit of the plausible range; the contraction was to the threshold point. We encountered no case in which expansion was required toward  $\infty$ .

Here are some examples of our method of distribution assignment. Plante *et al.* (1987) used 5% as their base value for the probability of death resultant from primary surgery and cited previous research locating this probability somewhere between 1% and 10%. We assigned  $d_s$  a piecewise linear density possessing a mode of 0.05, 95% of its probability mass within  $[0.01, 0.10]$ , and 5% of its probability mass split equally between  $[0, 0.01]$  and  $[0.10, 1]$ . Velanovich *et al.* used a base value of 0.02 for the probability of rebleed after splenorrhaphy. They provided no plausible range, but performed a SA over the range 0 to 0.22. We assigned the parameter a piecewise linear density having mode 0.02 and 97.5% of its probability mass within  $[0, 0.22]$  and 2½% within  $[0.22, 1]$ . Pitcher and Galligan (1990) used a base value of 0.85 for

RMPA test sensitivity, and noted that all commercially available RMPA tests possessed sensitivities above 0.41. We assigned a piecewise linear density having mode 0.85 and 97.5% of its probability mass within  $[0.41, 1.0]$  and 2.5% within  $[0, 0.41]$ .

## METHODOLOGY FOR PROBLEM RE-ANALYSIS

### Platform and Procedure

EVPI calculations can be expressed in closed form only for simple models and low-order parameter sets. Higher dimensional problems require techniques such as numerical integration or Monte Carlo simulation. Fortunately, multiparametric EVPI calculations can be performed today on desktop computers using commercially available software. To stress this capability and encourage DMs to apply EVPI-based sensitivity measures to their own problems, we performed our analysis using common hardware and off-the-shelf software. We used a 60-MHz Pentium desktop computer running Microsoft Excel and @Risk to obtain Monte Carlo estimates of  $EVPI(\xi)$  for the parameter sets  $\xi$  in which we were interested. In our retrospective analyses, we imposed independent parameter distributions *post hoc* and measured expected foregone payoff as improvement over the *base* optimal rather than *average* optimal alternative (see Appendix).

### True and False Sensitivity

To succinctly compare the results of the authors' SAs and our EVPI analyses, we characterized an author's sensitivity label for a parameter (or parameter set) to be false when it conflicted with the results of our comparison of the parameter's EVPI with the problem's MSI. For example, a false sensitive occurred when authors labeled a parameter sensitive but its EVPI did not exceed the relevant MSI.

By examining the proportion of false sensitive and false insensitive labels, we could directly compare the results of conventional SAs and EVPI analyses. A high false sensitive rate would indicate that the SA

performed overestimated the sensitivity of the problem to its parameters; a high false insensitive rate would indicate underestimation. When comparing the results of the authors' SAs and our corresponding EVPI analyses, we often refer to parameters as being "sensitive" or "insensitive" for ease of communication. This should be understood to mean that the underlying decision problem is "sensitive" or "insensitive" to the parameter in question.

## **RESULTS OF PROBLEM RE-ANALYSES**

### **Sensitivity Overestimation**

The overwhelming conclusion in the 25 problems we analyzed is that conventional SA procedures *dramatically overestimate* problem sensitivity as measured by EVPI. Consider, for instance, the comparison between EVPI analysis for single parameters and the corresponding SA declarations summarized in table 3A. Even with the MSI hurdle set at only 1% of the difference between optimal value and the status quo value, 39 (76%) of the 51 parameters declared sensitive in the 25 analyses were falsely declared, that is, they had information value below the relevant MSI. Of these 51 parameters, 47 (92%) had information value below the relevant MSI at the  $\delta=5\%$  level, and 49 (96%) at the  $\delta=10\%$  level. In contrast, the corresponding false insensitivity rates were very low (8.6%, 1.8%, and 1.2% for relevant MSIs at the  $\delta=1\%$ , 5%, and 10% levels).

---

Insert Table 3 about here

---

The same high false sensitivity rate shows up in for the two-way SAs in our sample. Table 3B summarizes the comparisons we made. There were 20 instances in our sample in which a two-way SA was conducted on two non-remote parameters. In 15 cases, at least one of those parameters had been identified as sensitive by the authors' one-way SA. We labeled these 15 2-way SAs as sensitive from the authors'

perspective and then calculated the joint EVPI of the two parameters. Of these, 9 (60%) had information values which did not exceed the relevant MSI at the  $\delta=1\%$  level. At the  $\delta=5\%$  and 10% level, 13 (87%) failed to exceed the relevant MSI.

The observed high false sensitivity rate is not due to the manner in which we chose the relevant MSI. As can be seen in table 3A, there were 16 (6+10) parameters declared sensitive in problems in which the authors specified a MSI hurdle: *all 16* were false declarations. In Table 3B, there are 4 (4+0) parameter pairs declared sensitive for such problems: *all 4* are false.

It is possible that the high incidence of false sensitivity arises from the type of SA methodology employed. In addition to PSTP analyses, we also observed 4 problems (42 parameters) which used tornado diagrams as their SA tool and 5 (47 parameters) which employed scenario analyses. Because the SA methodology is consistent within a given problem, its effect might be revealed by examining sensitivity declarations for *problems* rather than parameters. We say a problem has been declared sensitive by conventional SA if the authors have declared sensitive to at least one parameter, otherwise we say the problem has been implicitly declared insensitive. We compared these declarations of problem sensitivity with the all-parameter EVPI for each problem and labeled the declaration of problem sensitivity (insensitivity) false if the corresponding all-parameter EVPI value did not exceed (did exceed) the canonical relevant MSI. The results are shown in Table 3D. The overall false sensitivity rate for problems (3 out of 15 at  $\delta=1\%$ ) is lower than for parameters, as might be expected. Nevertheless, it is still disturbingly high for larger values of  $\delta$ . The false sensitivity rate may be higher for VR analyses (2 false sensitivity declarations out of 5) than for PSTP analyses (1 false sensitivity declarations out of 10), although the number of problems considered is too small to come to a firm conclusion.

As an alternate comparison at the problem level, we examined the information value of the set of all parameters declared sensitive by authors. We called this set the *declared sensitivity set* (DSS). Its complement we called the *declared insensitivity set* (DIS). Ideally, nonempty DSSs should have high

information values and DISs should have low information values. Table 3C, however, shows that false declarations of sensitivity are disturbingly common (4 of 15 DSSs at  $\delta=1\%$  level, 12 of 15 DSS's at  $\delta=5\%$  and 10% level). False declarations of sensitivity again seem higher among the VR analyses, although the number of DSSs considered is small.

### **Threshold Proximity Effects in Sensitivity Overestimation**

We believe that the high rate of false sensitivity declarations we discovered is substantially due to the fact that conventional SA methods tend to ignore value issues and focus attention on a subjective likelihood that a parameter will cross a threshold. For instance, of the 22 problems we examined in which one-way SA was performed, 16 were PSTP analyses. In 12 of these 16 problems the authors explicitly used the term “threshold” to refer either to the point of decision change or the analysis itself. In 5 of these problems, the authors explicitly state that their criterion for sensitivity was threshold proximity or likelihood of threshold crossing, and in another 6 problems, circumstantial evidence strongly suggests their criterion was of this type. Among the 125 parameters included in these 16 problems, 45 were involved in an explicitly presented SA. Of these, 36 were included in two- or three-way graphical SAs, for which it would seem the only possible criterion for sensitivity is likelihood of threshold crossing.

For 20 parameters, one-way PSTP SA graphs were constructed similar to that presented in figure 1A. In such graphs, an analyst can pay attention to a variety of different elements. Examining the area  $A(\xi)$  above the BOA curve and below the payoff curve for some other alternative is equivalent to looking at  $EVPI(\xi)$  under an implicit uniform distribution for  $\xi$ . Under a uniform distribution, an analyst drawing sensitivity conclusions based on the proportion  $\pi(\xi)$  of the plausible range which lies to the other side of a threshold is implicitly using probability of threshold crossing as a sensitivity criterion. Correlation statistics for these 20 one-way PSTP analyses (Table 2A) strongly suggest that analysts are *not* examining  $A(\xi)$  to determine sensitivity. Because  $\pi$  has the higher unadjusted correlation and  $A$  is negatively correlated with declared sensitivity once its association with  $\pi$  is accounted for, we conclude that declared

sensitivity is explained by  $\pi$  rather than A.

(A)	PSTP correlation with declared sensitivity			(B)	Tornado correlation with declared sensitivity		
	Unadjusted	Adjusted for $\pi$	Adjusted for A		Unadjusted	Adjusted for $\pi$	Adjusted for L
$\pi$	0.462	-	0.456	$\pi$	0.800	-	0.739
A	0.234	-0.220	-	L	0.473	0.134	-

**Table 2.** (A) Correlation of  $\pi$  and A with declared sensitivity among 20 parameters for which analysts performed graphical one-way SA. (B) Correlation of  $\pi$  and L with declared sensitivity among 42 parameters for which analysts used tornado diagrams.

The DVT example in table 1 provides a good illustration of the sensitivity overestimation which can occur when sensitivity labels are based on the likelihood of threshold crossing without consideration of accompanying payoff changes. The authors did not reveal their specific criterion for declaring sensitivity, but did express that it was based on PSTP. It is evident that their declared sensitivity is strongly correlated with  $\pi$ . The corresponding EVPI values, however, are quite small compared to the base-optimal utility 96.5155, due to the small marginal benefits associated with deviation from the BOA. This accounts for the sensitivity overestimation we observe: the sensitive declaration for P[Pulmonary Embolism] cannot be justified in EVPI terms, and the sensitive declaration for Utility of Adverse Fetal Outcome is questionable.

Tornado diagrams were the primary tools for sensitivity analysis in 4 of the 25 problems we re-analyzed. Because a tornado diagram captures the effect of parameter variability by displaying a corresponding range of payoffs, one might expect that there would be less susceptibility for sensitivity overestimation compared to PSTP methods. However, we found sensitivity overestimation to be at least as prevalent among tornado diagram analyses. We believe the reason is the same: despite the display of value range, users in our sample tended to read the probability of a threshold crossing from tornado diagrams.

When a parameter is uniformly distributed over its plausible range, the proportion of the tornado

diagram bar for  $\xi$  which is shaded is the probability  $\pi(\xi)$  of a threshold crossing, and the length  $L(\xi)$  of the shaded portion of the bar for  $\xi$  is proportional to the expected foregone payoff given a threshold crossing. Again, we analyzed whether these quantities were good predictors of declared sensitivity. For the 42 parameters involved in the four problems, table 2B shows the correlations and partial correlations with declared sensitivity. Notice that  $\pi$  has a much higher correlation with declared sensitivity, and most of the correlation of  $L$  is explained by its correlation with  $\pi$ . We conclude that, as with the PSTP approach, the primary driver of sensitivity declaration for tornado diagram users is probability of threshold crossing.

Another possible explanation for the apparent sensitivity overestimation we observed is that the parameter distributions we assessed were too narrow compared to the implicit beliefs authors used in their SAs. Overly narrow parameter distributions would produce unjustifiably small EVPI values, compared to which authors would appear to be overestimating sensitivity. To explore this possibility, we recomputed EVPI values using *uniform* distributions for all parameters across their plausible ranges. The results were virtually identical to our original analysis, still indicating a significant overestimation of sensitivity on the part of authors (see parenthetical entries on table 3).

### Parameter Interaction

Granting that conventional SA significantly overestimates problem sensitivity, we can still ask about the importance of parameter interaction. For example, how misleading can a series of one-way SAs be compared to the corresponding ideal multi-way SA?

In our sample, the probability of decision change due to individual parametric variation was subadditive (i.e., the overall probability of decision change was greater than the sum of the one-way probabilities of decision change) in 7 of the 16 problems in which the authors employed PSTP SA, and in 3 of the 9 problems using VR SA. That is, in 40% of problems overall sensitivity to decision change was substantially more than would be indicated by simple one-way SA. Should one be interested only in decision sensitivity, it would appear from this small sample that one-way SA is not a reliable approach.

What about EVPI? Might it suffice to consider only one-way EVPI analysis, or should an all-parameter EVPI be routinely calculated? In our sample, EVPI was subadditive (i.e.,  $EVPI(\Pi)$  was greater than the sum of the EVPIs of individual problem parameters) in 17 of the 25 (68%) problems we considered. Once again, it would appear from this small sample that one may not receive an adequate picture of problem sensitivity unless all-parameter EVPI is calculated.

## DISCUSSION

The conclusion that conventional SA overestimates problem sensitivity is, on balance, reassuring – it means that analyses which identify “clear winners” most likely do so correctly. Moreover, high estimates of problem sensitivity may compensate for other sources of problem sensitivity (e.g., parameter interaction, problem structure) which are difficult to detect.

Von Winterfeldt and Edwards (1973) suggested the *flat maxima principle* (FMP) to characterize their practical experience with decision model robustness to small changes in input parameters. They argued that aggregate values arising from decision models (e.g., expected utilities and expected values) “are seldom much affected by the precision of the model form or of the input numbers” (von Winterfeldt and Edwards 1986). In essence, the FMP states that, upon the elimination of ordinally and cardinally dominated alternatives, moderately sized errors in probability assessments do not produce large changes in expected payoffs. The basic reason for this is threefold. First, eliminating dominated options reduces the risk of choosing a poor alternative by leaving available only relatively attractive options. Second, it generally makes little difference which option a DM selects when options are close call. Third, averaging is a flattening process that combines quantities into values lying between extremes. Supporting recognition that decision-theoretic maxima are flat may be found in Murphy and Winkler (1970).

There has been some contention that two of the assumptions upon which the FMP is founded, the elimination of dominated alternatives and convexity of the maximal expected utility function, may not hold

(Rios Insua 1990; Rios Insua and French 1991). Restricting the problem to include only non-dominated options, for example, can be prohibitively time-consuming, and consideration of problem sensitivity with respect to probabilities and utilities can result in a maximal expected utility function which is not convex. Nonetheless, our empirical results support and are explained by the FMP. While the probability of a threshold crossing may be high, our results indicate that the resulting average forgone payoff is often trivial relative to the value offered by the BOA

As we have noted, one additional requirement of an EVPI-based sensitivity analysis is the specification of probabilistic dependencies among parameters, where appropriate. Correlations between parameters are rarely accounted for in conventional sensitivity analyses. For example, as Littenberg and Moses (1993) point out, sensitivity and specificity of diagnostic tests are correlated through the threshold used to declare a positive test outcome.<sup>1</sup> However, of the ten articles we reanalyzed containing sensitivity and specificity parameters, only one performed a two-way analysis on these parameters. Some authors do build correlation directly into the model structure by specifying functional relationships between parameters. One method common in the medical literature involves the use of an efficacy parameter in problems where drug and/or treatment choice can affect the likelihood of specific events (e.g., Klein and Pauker (1981)).

For the purposes of our retrospective EVPI analyses, we assumed all parameters were probabilistically independent. Might this have biased our comparisons between EVPI-based and conventional SAs? Certainly not for our comparisons in the single parameter case – here parameter independence is irrelevant and the sensitivity overestimation effect is large. More generally, although the assumption of independence may be inaccurate for some parameter collections, the assumption was present *both* in our retrospective EVPI analyses *and* in the authors' original SAs. The worst that can be said is that our analysis may fail to account for yet another bias in conventional SA. Even so, extent of any undiscovered bias is not likely to

---

<sup>1</sup> Thanks to Dennis Fryback for drawing this issue to our attention.

be important, in our opinion, for several reasons. First, correct accounting for parameter correlations may increase or decrease the likelihood of threshold crossings, so the direction of potential bias is unclear. Second, correct parameter dependencies are often built into the parameter structure, as mentioned above, and these dependencies are accounted for by *both* EVPI-based and conventional SA. Third, some parameter independence's are clearly valid, such as independence between probability parameters and value parameters.

In this paper, we have limited our attention to problems in which the authors' sensitivity conclusions were drawn from measures based on payoff spreads and threshold proximity in a parameter space. For a treatment of the comparative advantages of EVPI relative to probabilistic and entropy-based SA measures, see Felli and Hazen (1998). Although we believe EVPI to be superior to conventional SA methods, we have also established that one-way EVPI has limited value due to its tendency toward subadditivity relative to EVPI( $\Pi$ ). Hence, to more scrupulously explore the sensitivity of a decision problem to its parameters, we not only advocate the use of EVPI, but of all-parameter EVPI as well. Only in this way can the DM be assured that her assessments of sensitivity have been made in accordance with her beliefs regarding *both* the distributions of problem parameters *and* the value she places on marginal increases in payoffs.

## REFERENCES

- Abrams HB, Detsky AS, Roos LL, Wajda A. Is there a role for surgery in the acute management of infective endocarditis? *Medical Decision Making*. 1988; 8 (3): 165-74.
- Barrett BJ, Parfrey PS, Foley RN, Detsky, AS. An economic analysis of strategies for the use of contrast media for diagnostic cardiac catheterization. *Medical Decision Making*. 1994; 14: 325-35.
- Buchbinder R. and Detsky AS. Management of suspected giant cell arteritis: a decision analysis. *Journal of Rheumatology*. 1992; 19 (8):1220-8.
- Buckley JJ. A mathematical programming approach to sensitivity analysis in single-stage decision making. *Decision Sciences*. 1988; 19: 211-232.
- Clemen RT. Making hard decisions: an introduction to decision analysis. PWS-Kent Publishing Company, Boston, 1996.
- Dippel DWJ, Touw-Otten F, Habbema JDF. Management of children with acute pharyngitis: a decision analysis. *Journal of Family Practice*. 1992; 34 (2): 149-59.
- Djulbegovic B, Hendler FJ, Hamm J, Hadley T, Woodcock TM. Residual mediastinal mass after treatment of Hodgkin's disease: a decision analysis. *Medical Hypotheses*. 1992; 38: 166-75.
- Eschenbach, TG and Gimpel, RJ. Stochastic sensitivity analysis. *Engineering Economist*. 1990; 35 (4): 305-321.
- Evans, JR. 1984. Sensitivity Analysis in Decision Theory, *Decision Sciences* **15**: 239-247.
- Feliciano PD, Mullins RJ, Trunkey DD, Crass RA, Beck JR, Helfand M. A decision analysis of traumatic

splenic injuries. *Journal of Trauma*. 1992; 33 (3): 340-8.

Felli JC and Hazen GB. Information value and decision sensitivity. *International Engineering Consortium Annual Review of Communications*. 1996; 49: 243-53.

Felli JC and Hazen GB. Sensitivity analysis and the expected value of perfect information. *Medical Decision Making*. 1998; 18 (1): 95-109.

Fishburn, PC, Murphy, AH and Isaacs, HH. Sensitivity of decisions to probability estimation errors: a reexamination. *Operations Research*. 1968; 13: 217-237.

Howard, RA. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*. 1966; SCC-2 (1): 22-26.

Howard, RA. Value of information lotteries. *IEEE Transactions on Systems Science and Cybernetics*. 1967; SCC-3 (1): 24-60.

Isaacs, HH. Sensitivity of decisions to probability estimation errors. *Operations Research*. 1963; 11: 536-552.

Keefer DL and Bodily SE. Three-point approximations for continuous random variables. *Management Science*. 1983; 29 (5): 595-609.

Kiemeney LALM, Witjes JA, Heijbroek RP, Koper NP, Verbeek ALM, Debruyne FMJ, and the members of the Dutch South-East Co-Operative Urological Group. Should random urothelial biopsies be taken from patients with primary superficial bladder cancer? A decision analysis. *British Journal of Urology*. 1994; 73 (2): 164-71.

Klein K and Pauker SG. Recurrent deep venous thrombosis in pregnancy: analysis of the risks and benefits

of anticoagulation. *Medical Decision Making*. 1981; 1(2): 181-202.

Leblanc R, Worsley KJ, Melanson D, Tampieri D. Angiographic screening and elective surgery of familial cerebral aneurysms: a decision analysis. *Neurosurgery*. 1994; 35 (1): 9-19.

Levey AS, Pauker SG, Kassirer JP. Occult intracranial aneurysms in polycystic kidney disease. *New England Journal of Medicine*. 1983; 308 (17): 986-94.

Littenberg B and Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Medical Decision Making*. 1993; 13 (4): 313-321

Lyman GH, Lyman CG, Sanderson RA, Balducci L. Decision analysis of hematopoietic growth factor use in patients receiving cancer chemotherapy. *Journal of the National Cancer Institute*. 1993; 85 (6): 488-93.

O'Brien, BJ, Drummond, MF, Labelle, RJ, Willan, A. In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care*. 1994; 32 (2): 150-163.

Olak J and Detsky A. Surgical decision analysis: esophagectomy/esophagogastrectomy with or without drainage? *Annals of Thoracic Surgery*. 1992; 53: 493-7.

Piccoli A, Pillon L, Passerini P, Ponticelli C. Therapy for idiopathic membranous nephropathy: tailoring the choice by decision analysis. *Kidney International*. 1994; 45: 1193-1202.

Pitcher PM and Galligan DT. Decision analysis and economic evaluation of the use of the rapid milk progesterone assay for early detection of pregnancy status of cows. *Journal of the American Veterinary Medical Association*. 1990; 197 (12): 1586-90.

- Plante DA, Piccirillo JF, Sofferman RA. Decision analysis of treatment options in pyriform sinus carcinoma. *Medical Decision Making*. 1987; 7 (2): 74-83.
- Raiffa H and Schlaifer R. Applied Statistical Decision Theory. Harvard University. 1961.
- Rios Insua D and French S. A framework for sensitivity analysis in discrete multi-objective decision-making. *European Journal of Operational Research*. 1991; 54: 176-190.
- Ruegg PL and Carpenter TE. Decision-tree analysis of treatment alternatives for left displaced abomasum. *Journal of the American Veterinary Medical Association*. 1989; 195 (4): 464-7.
- Schapira MM, Matchar DB, Young MJ. The effectiveness of ovarian cancer screening. *Annals of Internal Medicine*. 1993; 118 (11): 838-43.
- Schneller GO and Sphicas GP. On sensitivity analysis in decision theory. *Decision Sciences*. 1985; 16: 399-409.
- Smith RD. Decision analysis in the evaluation of diagnostic tests. *Journal of the American Veterinary Medical Association*. 1993; 203 (8): 1184-92.
- Sonnenberg FA, Roberts MS, Tsevat J, Wong JB, Barry M., Kent DL. Toward a peer review process for medical decision analysis models. *Medical Care*. 1994; 32 (7): JS52-JS64.
- Spiegelhalter DJ and Freedman LS. Bayesian approaches to clinical trials. Bayesian Statistics 3. Bernardo JM, Degroot MH, Lindley DV, and Smith AFM (Eds.). New York: Oxford University Press. 1988: 453-477.
- Spiegelhalter DJ, Freedman LS, and Parmer MKB. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Association (A)*. 1994; 157: 357-416.

- Starr MK. A discussion of some normative criteria for decision-making under uncertainty. *Industrial Management Review*. 1966; 8: 71-78.
- van Crevel H, Habbema JDF, Braakman R. Decision analysis of the management of incidental intracranial saccular aneurysms. *Neurology*. 1986; 36: 1335-9.
- Velanovich V and Tapper D. Decision analysis in children with blunt splenic trauma: the effects of observation, splenorrhaphy, or splenectomy on quality-adjusted life expectancy. *Journal of Pediatric Surgery*. 1993; 28 (2): 179-85.
- Venta LA, Venta ER, Mumford LM. Value of diagnostic tests for deep venous thrombosis: a decision analysis model. *Radiology*. 1990; 174 (2): 433-9.
- von Winterfeldt D and Edwards W. Flat maxima in linear optimization models. Technical report Number 011-313-4-T. University of Michigan, Engineering Psychology Laboratory, Ann Arbor, MI. 1973.
- von Winterfeldt D and Edwards W. Decision analysis and behavioral research. Cambridge University Press. 1986.
- Wadland WC and Plante DA. Screening for asymptomatic bacteriuria in pregnancy. *Journal of Family Practice*. 1989; 29 (4): 372-376.
- Zwetsloot-Schonk JHM and Leer JWH. Decision analysis -- a helpful tool for clinicians to establish diagnostic-therapeutic guidelines. *Acta Oncologica*. 1993; 32 (4): 379-91.

## APPENDIX

The central justification of EVPI as a sensitivity measure for a parameter set is its interpretation as an expected foregone gain induced by parametric uncertainty. However, the notion of foregone gain takes on a different meaning in the *post hoc* setting in which our EVPI-SA comparisons were performed. To maintain the legitimacy of these comparisons, we felt that the form of EVPI we used had to be modified. Here we describe the specifics of those modifications.

Given our distribution choices and assumption of probabilistic independence, we found in five of our 25 sample problems that the alternative  $a = a^*$  maximizing  $E[V|a]$  was different from the BOA  $a_0$  specified by the authors. To explain how this can occur, we suppose for simplicity that the payoff function  $E[V|a, \Pi]$  is *multilinear* in the vector  $\Pi$  of problem parameters. (This was the case in all 25 of our sample problems.) From this and parameter independence it follows that  $E[V|a] = E_{\Pi}[E[V|a, \Pi]] = E[V|a, \bar{\Pi}]$  where  $\bar{\Pi} = E[\Pi]$  is the mean of  $\Pi$ . The alternative  $a = a^*$  maximizes  $E[V|a, \bar{\Pi}]$ , and for this reason we call  $a^*$  the *average-optimal alternative* (AOA). However, if  $\Pi_0$  is the base value of  $\Pi$  chosen by the authors, then the BOA maximizes  $E[V|a, \Pi_0]$ . So if  $\bar{\Pi} \neq \Pi_0$ , then it is possible that  $a^* \neq a_0$ .

If a distribution for  $\Pi$  has been estimated or assessed, then on normative grounds the AOA  $a^*$  should be preferred to the BOA  $a_0$  when they are different. Moreover, the foregone payoff due to uncertainty in  $\Pi$  should be calculated using  $a^*$  as well. That is, average foregone payoff should be  $EVPI(\Pi) = E[\max_a E[V|a, \Pi] - E[V|a^*, \Pi]]$ . However, in the problems we considered, authors neither assessed nor estimated distributions for  $\Pi$ , and no AOA  $a^*$  was calculated. Instead, authors estimated only base values  $\Pi_0$  and computed the corresponding BOA  $a_0$ . In this *post-hoc* setting, we did not feel it is reasonable to measure foregone payoff by comparison to the AOA. Instead, we felt that foregone payoff should be measured in comparison to the alternative actually chosen by the DM, namely

the BOA. Thus, average foregone payoff should be calculated as  $EVPI_0(\Pi) = E[\max_a E[V|a, \Pi] - E[V|a_0, \Pi]]$ . We therefore chose to use  $EVPI_0$  as the measure of information value in our retrospective analyses.

Notice that  $EVPI_0$  can be written as a sum of conventional  $EVPI(\Pi)$  and a nonnegative term  $E[V|a^*] - E[V|a_0]$ , consisting of the average improvement in moving from the BOA to the AOA. Therefore  $EVPI_0(\Pi) \geq EVPI(\Pi)$ . It follows that our fundamental conclusion that conventional SA overestimates sensitivity would have been even more strongly supported for the entire parameter set  $\Pi$  had we used  $EVPI(\Pi)$  rather than  $EVPI_0(\Pi)$ . It should be mentioned that although we have assumed multilinearity and independence in this discussion, none of these conclusions depend on these assumptions.

For parameter sets  $\xi$  which are proper subsets of the entire parameter set  $\Pi$ , the situation is somewhat more delicate. Suppose  $\eta$  is the set of parameters complementary to  $\xi$ , so  $\Pi = (\xi, \eta)$ . Let  $a^*(\xi, \eta)$  be the optimal alternative when both  $\xi$  and  $\eta$  are revealed prior to selecting an alternative. Then assuming independence and multilinearity,  $a_0 = a^*(\xi_0, \eta_0)$  and  $a^* = a^*(\bar{\xi}, \bar{\eta})$ . Again, in regard to the information value of  $\xi$  *post-hoc*, we argue that the foregone payoff should be the difference between the payoffs under alternatives  $a^*(\xi, \eta_0)$  and  $a_0$ , rather than under alternatives  $a^*(\xi, \bar{\eta})$  and  $a^*$ . Therefore, in our retrospective analyses, we measured the information value of  $\xi$  using  $EVPI_0(\xi) = E_\xi[\max_a E[V|a, \xi, \eta_0] - E[V|a_0, \xi, \eta_0]]$  rather than the canonical  $EVPI(\xi) = E_\xi[\max_a E[V|a, \xi, \bar{\eta}] - E[V|a^*, \xi, \bar{\eta}]]$ .

Unlike the all-parameter case,  $EVPI_0(\xi)$  may or may not exceed  $EVPI(\xi)$ , so the effect of using  $EVPI_0$  instead of  $EVPI$  in our retrospective analyses is unclear. However, the degree of overestimation of problem sensitivity by conventional SA was so high and so clearly explainable in terms of the flat-maximum principle, that we doubt whether the choice between  $EVPI_0$  and  $EVPI$  would have a major effect.

	A				B		C				D			
	No. of single parameters which authors declared				No. of 2-parameter sets authors declared		Number of				Number of problems which authors declared			
	S	False S	I	False I	S	False S	DSS $\neq\emptyset$	False DSS	DIS	False DIS	S	False S	I	False I
<b>Threshold analyses</b>														
in which authors specified MSI	6	6	14	1	4	4	2	1 (+1)	3	0 (+1)	2	1	1	0
for which we set MSI equal to $\delta=1\%$ of optimal improvement	26	19	79	6 (-1)	7	3	8	1	13	5 (-2)	8	0	5	2
<b>Non-threshold analyses</b>														
in which authors specified MSI	10	10	27	0	0	0	2	2	4	0	2	2	2	0
for which we set MSI equal to $\delta=1\%$ of optimal improvement	9	4 (-1)	43	7 (-5)	4	2	3	0 (+1)	5	3 (-1)	3	0 (+1)	2	1
<b>Totals for all analyses</b>														
MSI by author or $\delta=1\%$	51	39 (-1)	163	14 (-6)	15	9	15	4 (+2)	25	8 (-2)	15	3 (+1)	10	3
MSI by author or $\delta=5\%$	51	47 (-1)	163	3 (-1)	15	13 (-1)	15	12	25	2	15	10 (-3)	10	0 (+1)
MSI by author or $\delta=10\%$	51	49 (-2)	163	2 (-1)	15	13 (-1)	15	12	25	1	15	12 (-1)	10	0

**Table 3.** Results of our EVPI re-analysis of 214 parameters from 25 problems from the literature. The columns contain the number of parameters/sets/problems the authors designated sensitive (S) and insensitive (I). Columns labels prefixed “False” designate the number of those declarations that were not substantiated by our EVPI analysis. DSS sets are presumed to be labeled “S” by the authors; DIS sets are presumed to be labeled “I”. Numbers in parentheses designate the change in the number when the parameter distributions used were uniform over their plausible range (see text).