Qualification of Discordant Responses in Utility Assessment

Duane Steward, DVM, MSIE[*]
Mark Davis, MD, MS[+]
[*]Clinical Decision Making Group, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, [+]Beth Israel Deaconess Medical Center, Division of Emergency Medicine, Boston, Massachusetts

In many studies of utility assessment, the discordant response rate is significantly high. Discordant responses suggest inconsistency and, in turn, suggest inaccurate measurement of personal values that can lead to erroneous medical recommendations. The most common method of dealing with these responses is to exclude them from the sample statistics as incoherent or confused respondents. This paper proposes another perspective on discordant responses. In a recent study eliciting utility values for states of health that follow stroke, we observed a high rate of discordant responses. Closer examination of these discordant responses reveals that discordant responses are not all alike. Simple qualitative and quantitative views of these differences suggest that there may be information outside the concordant population of responses, which is lost by their exclusion. In an effort to understand the elevated discordant response rate, the effect of relaxing the defining boundaries of a discordant response was explored.

INTRODUCTION

Decision analysis is a rigorous technology that provides a model of a decision to help identify the critical issues that face the decision-maker in choosing treatment strategies. The values of outcomes represented in the model are critical to the validity of the model's implications. Recognized methods used to determine the value of an outcome from a patient's perspective exist and are known as "utility assessments". However, different methods often result in conflicting responses for the same individual, which may be referred to as "discordant" responses. Such inconsistency either questions the validity of one or more methods employed or indicates a poor understanding by the patient and consequently suggests an invalid result. Erroneous measurement of values can lead to inappropriate medical recommendations when those value measurements are used to support decisions involving tradeoffs. Inconsistent responses are therefore an important issue in the use of decision analytic approaches to medical decision support, health care protocol development and resource allocation where utility assessment is employed.

The gold standard for utility measurement eludes us, but desperate for an understanding of the patient's preference, analysts, clinicians and policy makers wish to make the most of what technology has been developed for assessing utilities. Utility assessment is an expensive process. It requires trained analysts to interview the individual or patient population to be represented. If taken seriously, these interviews usually involve some consideration of grave outcomes, which are at least sobering if not traumatic for the person interviewed. Questions of reliability and stability of responses plague the widespread acceptance of the results as useful information. In an effort to standardize administration of the utility assessment, computer programs have been developed. These programs ease the demand for trained experts to perform utility assessment. A natural progression to Internet administration with web pages[1] has relaxed the constraints on when and where the assessments can be performed.

Data from utility studies recent and past (Torrance[2]; Bass, et al[3]) suggest that discordant responses are rarely absent and occur with a frequency 17-30% or more. Reporting on the use of web page administration of utility assessment using standard gamble and visual rating scale methodologies, Lenert, et al[1], observe a 20% and 36% rate, respectively, of discordant responses between these assessments and the ordering produced by pair-wise comparison. Furthermore, they found that neither self-assessment of understanding nor of confusion predicted the consistent or inconsistent responses. Little has been done to investigate this significant portion of the population of responses and the meaning or significance of the discord (see Dolan[4] for a notable exception investigating factors affecting the rate of discord within individuals). If the methods of utility assessment are only considered valid in some portion of the population, a rigorous method must be employed to qualify which responses are valid for use in decision support and which are not.

Lenert, et al[1], propose to validate responses by evaluating adherence to the axioms of utility theory[5,6] with what they refer to as "consistency across methods of preference assessment" (CAMPA). Based on our observations, we suggest that a flexible

definition of discord may be useful. We advocate a systematic approach that enables analysts to include some responses that contribute information that would otherwise be forfeited. This argument applies to the treatment of discordant responses in modeling the preferences of an individual patient, as well as in forming population models. This paper explores the impact of variable definitions of discord on utility assessment.

METHODS

To address the question of whether patients prefer to avoid death or disability as a consequence of stroke, rank order and standard gamble utility assessment were used to interview visitors to an emergency department. This pursuit was based on a simplified decision model in which the motivating scenario involves a tradeoff of decreased disability in patients treated with thrombolytic therapy (e.g., tissue plasminogen activator - tPA) for a slight increase in the risk of death within a few days[7]. In a model which compares treatment with tPA to treatment without tPA the outcomes to both would be the possibility of varying degrees of disability, death within days and, distinctly, death at a more distant interval.

This study employed the standard gamble method of utility assessment along with rank order on one mild, one moderate and three grave states of health, based on disjunctive combinations of the Rankin Scale of Disabilities[8]:

1. Either no symptoms at all or only mild symptoms, for example: slurred speech, numbness in face, or reduced strength in an arm or leg, but... Still able to carry out all your usual duties and activities.
2. You are unable to carry out activities you could participate in prior to the stroke;… You require some help looking after you own affairs, but... You are able to walk without assistance
3. You are unable to walk without assistance and unable to attend to own bodily needs, or bedridden, incontinent, requiring constant nursing care
4. Sudden painless death within two days
5. Death within six months following the stroke

To be included in our study, the volunteer had to be English-speaking, 18 years or older, coherent, not experiencing signs of stroke, a non-psychiatric admission to the emergency department with a triage rating of 3 or 4 out of four and present long enough to conduct the interview. Two MIT undergraduates were specifically trained to conduct these interviews. Patients were assured that this study was confidential and in no way related to their case or care.

Patients were asked to rank the five health states which were presented to them on separate strips of paper. The order of presentation of the states was systemically randomized. The state ranked worst was then used by the research assistant as the negative side of the lottery in all standard gamble assessments. The remaining four states were then evaluated in randomized order. Standard gamble assessment was performed posing a choice between two treatments: one who's outcome was the health state being assessed and another who's outcome was full recovery but at a specified risk of their worst ranked health state. The risk was simply expressed in some number of people out of 100.

To minimize confusion, three visual aids were employed: (1) a plainly visible printed description of the health state being assessed, (2) a pie graph which could be adjusted to display any degree of risk and (3) a scoreboard displaying the five health states as ordered by the patient with utility values assessed to that point in the interview written beside the their corresponding label. With this aid, it was presumed that the patient would be able to readily notice and remedy any discordant response as it occurred. If the patient did not remedy a utility value that was out of order with their rank order of health states, it was pointed out to them. The interviewers were trained to briefly do their best to make sure the patient understood why others would see that something was out of order and ask them if they wished to change their response. However, they were to permit the patient to leave it unchanged if that is what they desired.

The impact of variable definitions for concordance was explored by retrospectively examining patient's utilities. Patient's responses were labeled as "**severity-concordant**" if none of the following occurred: 1) One of three grave states (Immediate Death, Death after 6 months, or Severe Disability) were scored as equally preferable or more preferable to those with mild or moderate disability; or 2) The moderate disability state was scored as equally preferred or more preferred to the mild disability state. A "**rank-concordant**" response pattern was defined as one where there was consistency between initial rank order and elicited utility values and the highest ranked state was preferred over at least two other states (the number of states involving death). Patients who gave responses that were both severity- and rank-concordant were labeled "**strictly-concordant**." Those responses that were not severity-concordant because patients assigned to different health states equal - but not disordered - utility values were classified as "**weakly-**

severity-discordant." Similarly, those responses that were not rank concordant because patients scored the best outcome and two or more other outcomes as equally preferable were classified as "**weakly-rank-discordant**." The union of rank-concordant, severity-concordant, weakly-severity-discordant and weakly-rank-discordant responses were classified as "**at-least-weakly-concordant**" responses. Those responses that were not concordant by severity or rank but whose "erroneous" margin was of 10 or less percentile points were classified as "**nearly-concordant**."

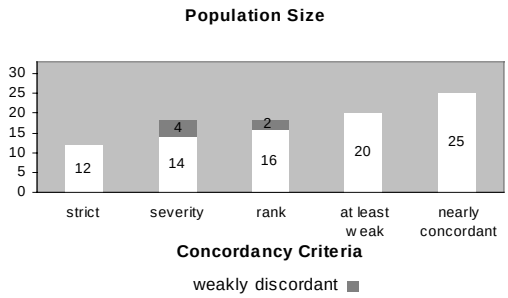

Figure 1. Population size for different concordancy criteria. Each bar represents the number of responses from among 33 completed interviews that satisfy the criteria for different definitions of concordancy. "At least weak" refers to responses at least weakly concordant by either severity or rank.

RESULTS

Of the 63 patients approached, there were 36 consenting participants (52.4%) ranging in age from 19-79 (average 42.5; 20 female; 16 male). One patient did not complete the interview because he could not understand the questions. Two patients were unable to complete the interview because the hospital staff became available for services interrupting the interview.

As seen in Figure 1, 12/33 (36%) patients were classified "strictly-concordant." Using the proposed definitions described: 16/33 (48%) were "rank-concordant;" 14/33 (42%) were "severity-concordant;" 20/33 (61%) were "at-least-weakly-concordant" by either rank or severity and 25/33 (76%) were "nearly-concordant."

Further characterizing the phenomenon of discord, 9/33 (27%) responses did not meet the criteria for strict-concordance where the disorder involved an erroneous margin of 10 percentile points or less.

The impact of changing the definition of discord on the range and mean utility values is shown in Figures 2. In this study there was overlap of the 95% confidence intervals for the different discord definition within each outcome state.

DISCUSSION

These results show that the definition of concordance impacts those eligible for inclusion in utility calculations and therefore in mean utility value calculations. The "correct" definition is of course unknown.

A *weakly-rank-concordant* response could be considered rational if one is willing to accept that the patient actually considers neither state to be better than the other. Alternatively stated, the states may not differ in a significant dimension in the patient's perspective. For example, if a patient with AIDS holds a compensatory high regard for mental health over physical incapacity, they might score any health state that did not affect mental health with relatively equal value. This is simply using an adaptive multidimensional model for health that heavily weighs the mental health dimension.

If disparate states can be given equal utility values, then it might be reasonable to consider disordered values with small margins as similarly explained. Dimensions of the patient's multivariate point of view that are only poorly captured in univariate utility assessment perspectives might offer an explanation. Or, perhaps the inversion of values is a matter of "noise" in the data, elicitation technique, or value system of the patient. It is compelling to view responses that are nearly concordant with more regard than those with vast margins of discord. The notion of a tolerance threshold for the margin of discord provides a parameter for gauging the degree of discord. The method for determining this threshold is not proven.

Relaxing the criteria for concordance allows a broader representation of the sample population that includes more of the unusual perspectives. It might more accurately represent the diversity of the population. In application of utility assessment to individual decision making, systematically relaxed concordance criteria enables more persons to benefit from the expression of their decision in rigorous terms. Health care providers can be informed by further discussing with the patient the meaning of weak or nearly discordant results. As with the traditional method for handling discord, decision support should not be based on this form of utility assessment in the case of responses that remain outside the nearly concordant population.

Patrick, et al[9,10], found in their work, assessing states worse than death, that there only 18% agreement between rank orders of four methods tested. The difficulty appears most severe for states

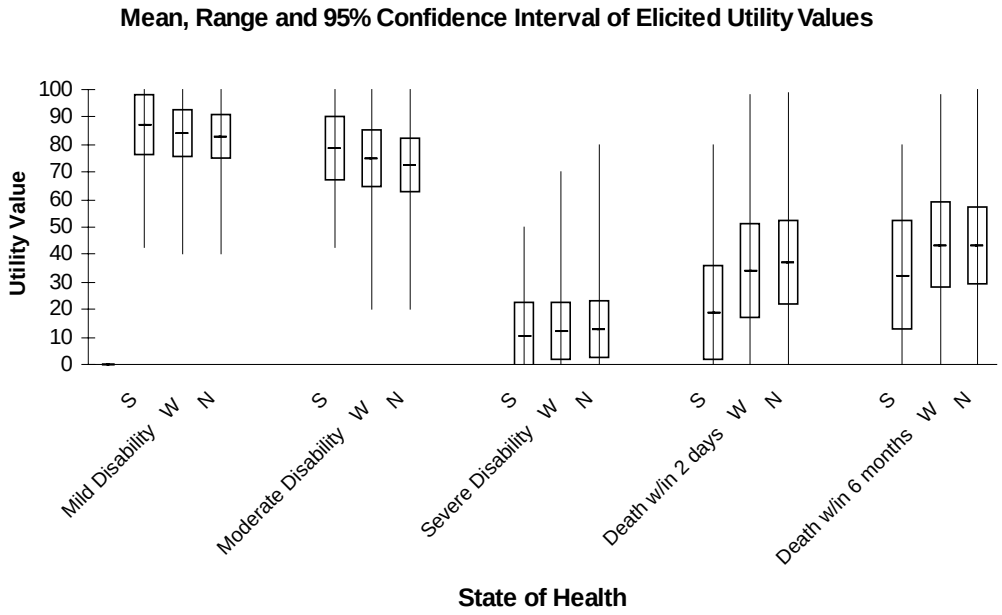

Figure 2. Mean (cross-hair), range (whisker), and 95% confidence interval (box) of the elicited utility values for mild disability, moderate disability, severe disability, death within two days and death within six months following a stroke. Triplets represent (from left to right) the "strictly-concordant" responses (S), "at-least-weakly-concordant" responses (W) and "nearly-concordant" responses (N).

near or worse than death. They conclude that the cognitive burden of these methods must be reduced if these assessments are to be used in frail older institutionalized adults. No mention is made of the degree of discord or any patient explanations offered for such responses.

The mere presence of results which can be explained yet lie outside the boundaries of strict concordance warn us to be careful in the interpretation of study results where they are discarded. What is being discarded is not those assessments which would make the descriptive statistics more normal. In our attempt to "normalize" the data, we may exclude patients who have legitimate reasons for differing from our expectations.

Two experiences in our study illustrate the potential for rational perspectives that explain responses that appear discordant. One patient explained her ranking of death within 2 days as the best possible outcome. She pointed out her devotion to her work in a lab and identified all the outcomes, even mild disability, as prohibiting her return to work in that lab. She would simply rather die than not be able to go to her lab. A second elderly lady, having her discordant response on the scoreboard pointed out, responded, "When you get as old as I am, you will understand these things." These patients clearly are convinced of their own rationality and will be asked to employ the same in their consent for any decision made.

One important limitation to proving a lack of difference between different discord definitions is the power of a study to detect such differences if they exist. It is also unclear what magnitude of difference would be clinically significant.

It is clear from our study that all discordant responses are not created equal. By refining the definition of discord in light of the interplay between rank order and other utility assessment results, a taxonomy of discordant responses emerges. That taxonomy is extended by the notion of tolerance levels for the margin of discord. Strong discord should be distinguished from what we call "weak" discord. Strong discord can be further characterized by the magnitude of the margin of discord, identifying a population we call "nearly" concordant in their responses. With care, results assessed from these patients may be fruitfully used in representing the utility models of the individual and of the population.

CONCLUSION

The implications of this study are relevant to patient decision making in two ways: as affecting descriptive statistics used to design health care protocol or resource management and as affecting

decisions of individual patients. From the social perspective, decisions based upon the descriptive statistics of strict definitions for discord will not reflect the entire human population. Which sub-population is represented will vary with the method of utility assessment employed and it is not clear which protocol is most rational. Grading the degree of discord allows the population represented to be more inclusive with qualifications about the consistency of some marginally consistent segment of the represented population. Conclusions drawn concerning these fuller representations should be carefully stated to preserve the qualifications. As an impact on the support of individual decisions, the presented taxonomy of discord suggests a means of qualifying the utility assessment responses of an individual. If a patient is consistent without qualification, decision analytic technology may be used to support decisions with sound values for outcomes. Individuals with marginally discordant responses may be able to explain their reasoning if asked about the discord. This could yield better understanding or discovery of deficiencies in understanding on the part of the patient or the provider. Results of utility assessment with severe discord should be regarded invalid for use in decision analytic models.

This study reaffirms that we can use standard gamble assessments to elicit utility values for grave states of health that are considered by some to be near or worse than death. Addressing the clinical motivation for this study, this limited sample indicates that most, but not all, people interviewed consider a severe degree of disability as worse than sudden or delayed death. Of more general interest is the observation that responses judged as discordant solely on the basis of relative utility values may result in the exclusion of individuals with rational preference structures. We propose that rank order may be used not only to compare consistency across methods of preference assessment but also to substantiate responses as potentially rational that would otherwise be undistinguished from more severely discordant responses. We advise caution in the dismissal of results for which standard gamble results are in discord with health state severity without rank order information. We have shown how loosening the definition and interpretation of discordant responses can impact conclusions drawn from utility assessment interviews. Doing so results in a more accurate representation of atypical but rational individuals and the populations they are members of. Systematic methods for understanding marginally concordant utility values and preferences as rational are needed.

Acknowledgments

The authors gratefully acknowledge the support of Isaac Kohane, M.D., Ph.D., Professor Peter Szolovits at M.I.T. and Geoffrey Rutledge, M.D., Ph.D. in the form of helpful comments in the design of the underlying study and the specific findings reported here. The authors also acknowledge the hard work of Annie L. Thompson and Richard Chen conducting the interviews. This work has been supported by a Medical Informatics Research Training Grant (2 T15 LM07092) from the National Library of Medicine.

References

1. Lenert LA, Morss S, Goldstein MK, Bergen MR, Faustman WO, Garber AM. Measurement of the validity of utility elicitations performed by computerized interview. Med Care 1997;35(9):915-20.
2. Torrance GW, Boyle MH, Horwood SP. Application of Multi-Attribute Utility Theory to Measure Social Preferences for Health States. Operations Research 1982;30(6):1043-1069.
3. Bass E, Fink N, Wills S, Levev A, Sadler J, Powe N. Do Preference Values For End Stage Renal Disease Differ By Dialysis Type? Annual Meeting of the Society for Medical Decision Making. Houston, Texas, 1997.
4. Dolan P, Kind P. Inconsistency and health state valuations. Soc Sci Med 1996;42(4):609-15.
5. Luce RD, Raiffa H. Games and Decisions. New York: Wiley, 1957.
6. Torrance G. Toward a utility theory foundation for health status index models. Health Serv Res 1976;11(4):349-69.
7. Tissue plasminogen activator for acute ischemic stroke. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group [see comments]. N Engl J Med 1995;333(24):1581-7.
8. Rankin J. Cerebral Vascular Accidents In Patients Over The Age Of 60. Scottish Medical Journal 1957;2:200-15.
9. Patrick DL, Pearlman RA, Starks HE, Cain KC, Cole WG, Uhlmann RF. Validation of preferences for life-sustaining treatment: implications for advance care planning. Ann Intern Med 1997;127(7):509-17.
10. Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. Med Decis Making 1994;14(1):9-18.