Judgment and Decision Making, Vol. 2, No. 1, February 2007, pp. 70–78.

On the appropriateness of appropriateness judgments: The case of interferon treatment for melanoma

Yoav Ganzach*
Faculty of Management
Tel Aviv University

Moshe Leshno
Sackler Faculty of Medicine and
Faculty of Management
Tel Aviv University

Abstract

We compare experts' judgments of the appropriateness of a treatment (interferon treatment for melanoma) on the basis of important attributes of this disease (thickness, ulceration, lymph node involvement and type of metastases) to a decision analytic model in which the probabilities of deterioration are derived from the medical literature and from epidemiological studies. The comparison is based on what we call *the linearity test*, which examines whether appropriateness judgments are a linear function of the epidemiological value of p_2, the probability of deterioration of the patient condition if he would have received the treatment. This comparison allows for the assessment of the validity of the experts' judgments under the assumption that the decision analytic model is valid, or alternatively, the assessment of the validity of the decision analytic model under the assumption that the experts' judgments are valid. Under the former assumption the results indicate that appropriateness judgments are by and large accurate. Under the latter assumption the results support the idea of a *constant treatment effect*, the idea that efficacy of a treatment is constant over various levels of severity of the disease. Our results also support the idea that experts' aggregate judgments far exceed individuals' judgments.

Keywords: aggregating judgment, medical treatment decisions, decision analylitic models of judgment, melanoma treatment, ecological validity, Brunswickian models.

1 Introduction

Appropriateness judgments such as "How appropriate it is to perform procedure X on a patient with symptoms Y and Z," which communicate information of how worthwhile it is to perform a medical procedure, play a major role in clinical guidelines systems (Audet, Greenfield, & Field, 1990; Brook, 1994). In producing such systems, expert clinicians are given scenarios of a disease (e.g., melanoma) that vary along a number of dimensions (e.g., size of tumor and number of nodes affected) and are asked to judge on the appropriateness of using a certain procedure (e.g., interferon treatment) for each of the cases. These judgments can later be used by practitioners in deciding whether or not the treatment should be administered to their patients.

In view of the growing importance of such methods for communicating expertise in general and medical expertise in particular (e.g., Field, & Lohr,1990; Shapiro, Lasker, Bindman, & Lee, 1993) this paper examines expert appropriateness judgments within the framework of a normative decision analytic model, evaluates the validity of these judgments, and assesses their usefulness in understanding clinical models of treatment. Our empirical work is based on reanalysis of expert panel judgment that had been used in creating an authoritative guideline on whether to use interferon as an adjunct treatment for melanoma.

There are three perspectives from which the relationship between a decision model and judgments of appropriateness could be understood. First, if the model is assumed to correctly describe the judgments, it could be used to uncover the implicit rules, or policies, underlying these judgments. This is a "policy capturing" view of judgment modeling (Sheldon, & Kafry, 1997; Sorum et al., 2002), primarily used to assess attribute weights in expert judgment, but also to determine the presence of configural (i.e., interactive) or other nonlinear rules underlying judgment. Second, if our decision analytic model is viewed as a prescriptive model of the appropriateness of a medical treatment, consistency between the model and actual appropriateness judgments could be viewed as supporting the validity of those judgments. Third, if a set of appropriateness judgments are viewed as prescriptively accurate, agreement between the model and the judgments could be viewed as supporting the normative stand of the model and the basic tenets on which it

*Mailing Address: Yoav Ganzach, Faculty of Management, Tel Aviv University, Tel Aviv 69978 Israel, Tel: +972-3-6406467, Fax: +972-3-6046982, E-mail:yoavgn@post.tau.ac.il.

https://doi.org/10.1017/S1930297500000292 Published online by Cambridge University Press

is based. Thus, whereas the second and third perspectives lend prescriptive status either to the model or to the judgment, the first perspective is merely descriptive, lending prescriptive status to neither.

1.1 A decision analytic model for appropriateness judgments

The term "appropriateness" is the common language analogue of the difference between the expected utility of taking an action and the expected utility of not taking that action. Thus, when rating the appropriateness of a treatment as 6 on a 1 (not appropriate at all) to 9 (very appropriate) scale, the clinician implies that the expected utility of administering the treatment is slightly higher than the expected utility of not administering it, whereas when rating this appropriateness as 9, the clinician implies that the expected utility of administering this treatment is much higher than the expected utility of not administering it. It is important to note that appropriate judgments are intended as a support tool for evaluating the utility of a treatment. As such, they should serve as a direct (i.e., linear) indicator of utility, and deviations from linearity should be viewed as inappropriate. To use an example, consider a panel of experts who are asked to judge water temperature by sensing the water. Appropriate temperature judgment in this case should be linearly related to temperature, and the a linearity test could be viewed as a test of their validity.

Consider now a clinician's judgment of the appropriateness of a treatment of a condition that has a probability of p_1 of deteriorating (e.g., death) and $1-p_1$ of remitting. Assume that the treatment is associated with probability $p_2$of deteriorating $p_2 < p_1$, and a probability $1-p_2$ of remitting. Figure 1 depicts the decision tree facing the clinician. In our model we assume that the probability of adverse events under treatment equals one. We denote by u_r the utility for remission and u_d for deterioration (death). We also assume that the utility for remission under treatment is equal to $u_r - u_a$ where u_a is the disutility of the adverse event associated with the treatment[1]. The expected utility of administering the treatment (EU_T) and the expected utility of not administering it (EU_{NT}) is given by:

$$EU_T = p_2(u_d - u_a) + (1-p_2)(u_r - u_a) \qquad (1)$$

and

$$EU_{NT} = p_1 u_d + (1-p_1)u_r, \qquad (2)$$

respectively.

[1]This assumption suggests a fixed treatment protocol, e.g., independence of the treatment on p_2. This, indeed is the case in interferon treatment (where the dose of the treatment depends only on the surface area of the patients and not on the severity of the disease).

Thus, the difference between the expected utility of administering the treatment (EU_T) and the expected utility of not administering it (EU_{NT}) is given by:

$$(p_2 - p_1)(u_d - u_r) - u_a \qquad (3)$$

If appropriateness judgment is a linear representation of $\Delta U = EU_T - EU_{NT}$ (this assumption is further discussed below), then it could be expressed as:

$$APP = \alpha\left[(p_2 - p_1)(u_d - u_r) - u_a\right] \qquad (4)$$

where APP represents the level of appropriateness and α is a positive constant. Denoting $p_2/p_1 = K$ we obtain

$$APP = \alpha\left[p_1(K-1)(u_d - u_r) - u_a\right] \qquad (5)$$

The efficacy of a treatment is defined by $\frac{p_1 - p_2}{p_1} = 1 - \frac{p_2}{p_1} = 1 - \frac{1}{K}$. The assumption that $p_2/p_1 = K$ is constant is equivalent to asserting that the efficacy of a treatment is constant over various levels of severity of the disease or that the effect of the treatment in reducing mortality is constant over various levels of severity of the disease. For example, if treatment reduces the probability of mortality of patient A, whose initial probability of mortality is 0.2, by 10% (to 0.18) it will also reduce the probability of mortality of patient B, with an initial probability of 0.8, by 10% (to 0.72). The constant treatment effect, although not necessarily universally true, may reasonably describe the effect of treatment in many situations. This assumption is made in many epidemiological studies. Moreover, it is mandatory in epidemiological studies where the relative risk reduction is estimated by regression.

Whereas our decision analytic model represents appropriateness judgments as a function of p_1, they are usually obtained in response to clinical scenarios (indications) that include information about the severity, or levels, of various symptoms. Therefore, policy capturing studies usually model appropriateness judgments as a function of the level of symptoms rather than p_1 (or any other relevant probabilities) (Kee et al., 2002). This approach has two disadvantages. First, it does not allow for relating the descriptive policy capturing model, based on symptoms, to a prescriptive decision analytic model, based on probabilities and utilities. Second, the scales of the symptom levels may not be linear, thus introducing distortion into the interpretation of the results. In particular, it is not clear whether nonlinear relationships between the symptom and the judgment represent a nonlinear clinical rule or nonlinearity in the scale of the symptoms. To overcome these difficulties, our study models the judgment in terms of both the "raw" symptom scale and in terms of a transformed symptom scale in which the levels of the symptom are expressed using an epidemiological p_2 yardstick. For example, if the severity of the symptom is

https://doi.org/10.1017/S1930297500000292 Published online by Cambridge University Press

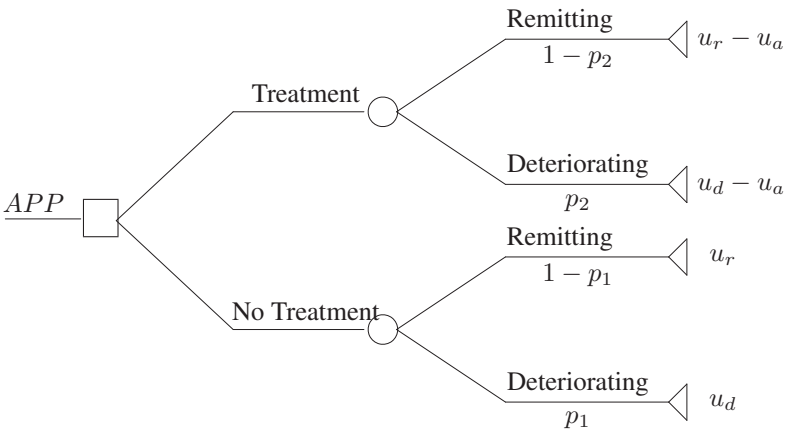

Figure 1: Decision Tree

measured on a 1 (low severity) to 4 (high severity) scale and the probability of mortality within five years is, respectively, q_1 to q_4, then the levels of the symptoms could be expressed in terms of the probability of mortality associated with each level, rather than the raw scale values. This process could be viewed as an intervalization of the symptom scale. Whereas the raw 1 to 4 scale is not necessarily an interval scale (equal changes on the scale are not necessarily equivalent with respect to their impact, e.g., a change from 1 to 2 may differ from a change from 2 to 3), the transformed scale is interval (equal changes on the scale could be viewed as equivalent in terms of their impact).

1.2 On the validity of appropriateness judgments

Assessment of validity in medical judgments has taken primarily either the approach of comparing methods (Bosch, Halpern, & Gazelle, 2002; Shackman et al., 2002), or examining whether the decision process suffers from biases (Chapman, & Sonnenberg, 2000; Stalmeier, 2002). A few studies have also examined the validity of appropriateness judgments by comparing them to normative models (Kuntz, Tsevat, Weinstein, & Goldman, 1999; Bernstein, Hofer, Meijler, & Rigter, 1997). In contrast to these approaches, our basic test for the validity of appropriateness judgments is based on a Brunswickian approach of comparing the function form in the environment model — the model that predicts the criterion from the cues — to the function form in the judgment model — the model that predicts the judgments from the cues (e.g., Stewart & Joyce, 1988, Wigton, 1996). In particular, our test, labeled *the linearity test*, involves an examination whether, in agreement with the model, appropriateness judgments are a linear function of the epidemiological value of p_1 (the probability derived from epidemiological studies). The linearity test is a test of the validity of appropriateness judgments, since to the extent that our decision analytic model is a correct model of appropriateness, valid judgments should satisfy this test. Thus, a linear relation supports (though it does not prove) the validity of appropriateness judgments, whereas a nonlinear relation provides some evidence against their validity. Note however that a nonlinear relationship does not necessarily suggest that appropriateness judgments are not valid. In particular, nonlinearity may be the result of our model being normatively incorrect (e.g., the assumption of a constant treatment effect is incorrect) rather than the appropriateness judgments being incorrect (e.g., judgments that rely on erroneous assessment of probability or utility, or on a correct integration of the two). Thus, our linearity test could be viewed as a joint test of the validity of our model for appropriateness judgment and the validity of the judgments themselves. Both need to be valid for linearity to occur.

1.3 The validity of individual judgments vs. the validity of the aggregated judgments

A basic question in medical decision making is whether aggregating the judgments of clinicians result in more valid clinical judgments. Despite the fundamental importance of this question, not much relevant empirical evidence is available, primarily because of problems associated with the establishment of criteria that will allow the evaluation of the utility of the aggregation.

https://doi.org/10.1017/S1930297500000292 Published online by Cambridge University Press

In the context of the current study, a criterion for the evaluation of the utility of the aggregation is available — whether or not judgments are linear. Thus our empirical test for the utility of aggregation of clinical judgments is whether or not the aggregated judgments conform with the linearity test better than the individual judgments.

1.4 The validity of expected utility models of treatment and the constant treatment effect hypothesis

Our discussion so far has focused on the validation of appropriateness judgments under the assumption that our decision analytic model is a valid model of the appropriateness of a medical treatment. However, as mentioned earlier, a complementary perspective emphasizes the validation of the model under the assumption that the appropriateness judgments are valid. In particular, if appropriateness judgments are assumed to be normatively valid and linearity is satisfied, the assumption of a constant treatment effect is supported.

1.5 Interferon treatment for malignant melanoma

In this study we examine the validity of appropriateness judgments in a specific clinical setting, adjuvant high-dose interferon alfa-2b in treating melanoma. Malignant melanoma is a common cancer in the western world. During the last 20 years, numerous agents have been evaluated in a series of both nonrandomized and randomized adjuvant therapy trials in melanoma. For patients who are in advanced stages of malignant melanoma, controversy abounds regarding high-dose adjuvant interferon alfa-2b therapy. Based on randomized clinical trials, it is currently agreed that high-dose interferon therapy is associated with approximately 10% improvement in relapse-free survival but also with high incidence of serious toxicity (Schuchter, 2004). In other words, relapse-free survival is "bought" at the price of increased frequency of serious toxicity. So the appropriateness judgments must revolve around the perceived tradeoff between harms and benefits.

2 Method

The judgments analyzed in this study were appropriateness judgments of high-dose interferon treatment of melanoma collected by Dubois et al. (2001) elicited from a panel of 13 experts (four dermatologists, four oncologist and five surgeons) using the RAND Delphi method (Park et al., 1986; Landrum & Normand, 1999). The judgments were given in response to 56 clinical scenarios based on permutations of four factors: thickness of the tumor, classified into four levels, level 1 ($\leq$1.00 mm), level 2 (1.01-2.00 mm), level 3 (2.01-4.00 mm) and level 4 (>4.0 mm); ulceration (present or absent); LNI, or lymph node involvement — the number of lymph nodes to which the tumor had spread (none, 1, 2, 3, or $\geq$4); and presence of micro metastases vs. macro metastases (for patients with LNI>0).

The judgments were given on a 9-point scale where 9 indicated extremely appropriate, 5 uncertain and 1 extremely inappropriate. Appropriateness was defined as "the expected health benefits of the therapy exceeding its expected negative health consequences by a sufficiently wide margin to justify giving the therapy" (Averbook et al., p. 1218), suggesting a difference model (e.g., Anderson, 1990; Rule, Curtis & Mullin, 1981).

Our analysis will focus on the effect of tumor thickness on judgment because of its central role in estimating the prognosis of primary melanoma in the clinical literature.[2] (Balch et al., 2000), and because good epidemiological data regarding this effect are available, in contrast to the lack of such data regarding the effect of LNI and ulceration. Our epidemiological source supplies a univariate probability of mortality for each level of thickness, but provides the probability of mortality only for a present/absent dichotomy with regard to LNI and ulceration and no data regarding presence of metastasis.

Our estimate of p_1 (the probability of deterioration given treatment) was based on the literature. In a recent epidemiological study (Averbook et al., 2002) p_1 of melanoma patients was reported as a function of thickness, (p_1 approximately equal to 0.1, 0.2, 0.4 and 0.65, respectively, for levels 1 through 4), LNI (p_1 is 0.447 when there is node involvement and 0.117 when there is no node involvement) and ulceration (p_1 is 0.443 when there is ulceration and 0.129 when there is no ulceration).

3 Results

We first analyzed the mean appropriateness judgments of the 13 panelists. We began by examining the correlations between the average appropriateness judgment and the severity of each symptom (aggregated over the levels of the symptoms and averaged over judges). Since by design the association between the symptoms was negligible, these correlations reflect the weight each symptom has in the judgment. (For linear relationships between the judgments and symptom, these correlations are a precise representation of the weight. For nonlinear but

[2]For example (Balch et al., 2001), "...it is well established that tumor thickness is the single most important prognostic feature of primary melanoma."

https://doi.org/10.1017/S1930297500000292 Published online by Cambridge University Press

monotonic relations they are an approximate, yet good, representation of these weights.) The values of the correlations are 0.68, 0.23 0.07 and 0.48 for LNI, thickness, ulceration, and presence of metastasis, respectively.[3]

Figure 2 presents the *average* aggregated appropriateness judgments (aggregated over the various levels of LNI and ulceration and averaged over judges) as a function of thickness, using the original "raw" 1-4 thickness scale. This figure suggests that the relationship between raw thickness and appropriateness judgment is not linear. Indeed, the functional relationship between the level of thickness and the average ratings of the 13 judges differed significantly from a linear function ($p < 0.005$).

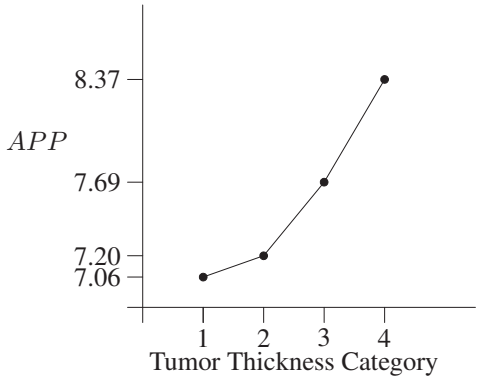

Figure 2: APP as a function of tumor thickness, averaging over all other variables

However, an appropriate test of linearity requires transforming raw thickness into an interval scale by positioning each level of thickness on a scale of probability of mortality, as estimated from epidemiological data (see above). This is presented in Figure 3. It is clear from this figure that on an interval thickness scale the relationship between thickness and the average aggregated appropriateness judgment is linear. Indeed, the functional relationship between the level of thickness after transformation and the average ratings of the 13 judges does not differ significantly from a linear function ($p > 0.2$). Thus, the linearity test is satisfied in our data.

Figures 4 and 5 present the individual aggregated judgments of the 13 judges (aggregated over the various levels of LNI and ulceration). By comparing Figures 4 and 5 to Figure 3, it is clear that the average appropriateness judgments are more linear than the individual appropriateness judgments. Out of the 13 judges, only four exhibit a linear

[3]For presence for metastasis the correlation was performed only within the scenarios for which there was lymph node involvement.

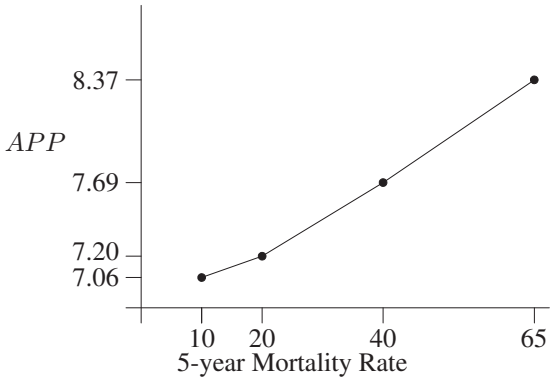

Figure 3: APP as a function of 5-year Mortality Rate, averaging over all other variables

relationship between judged appropriateness and the 5-year mortality rate as proposed by the analytical decision model — the other exhibit either marginally decreasing or marginally increasing functions. However, the average appropriateness rating of all the judges revealed a linear relationship between the appropriateness and the 5-year mortality rate.

Figure 6 presents the *average* appropriateness judgment as a function of thickness *separately for each level of LNI* (in this figure the judgments are aggregated only over the two levels of ulceration and averaged over the 13 judges). In Figure 6 the thickness scale is the original raw scale whereas in Figure 7 it is the transformed interval scale. One thing that is apparent from this figure is that, whereas raw thickness is not linearly related to appropriateness judgment within each level of LNI, after transformation, thickness is linearly related to these judgments within each of these levels. This finding is consistent with the idea that our transformation of raw thickness results in an interval thickness scale.[4]

Finally, both Figures 6 and 7 suggest that the relationship between thickness, LNI and appropriateness judgment is disjunctive: thickness has a larger impact on appropriateness when the LNI level is low than when it is high (repeated measures ANOVA with thickness and LNI as repeated measures revealed a significant interaction between the two, $F(1,108) = 108.9$, $p<0.0001$). This

[4]Note that in Figure 6 — unlike Figure 2 — the abscissa cannot be interpreted as a probability scale, and therefore linearity in each of the node levels is not directly associated with our decision analytic model. Note also that linearity between appropriateness judgments and symptom A (i.e., thickness) for each level of symptom B (i.e., number of nodes) is a sufficient condition for linearity between appropriateness judgments and symptom A on the aggregate level.

https://doi.org/10.1017/S1930297500000292 Published online by Cambridge University Press

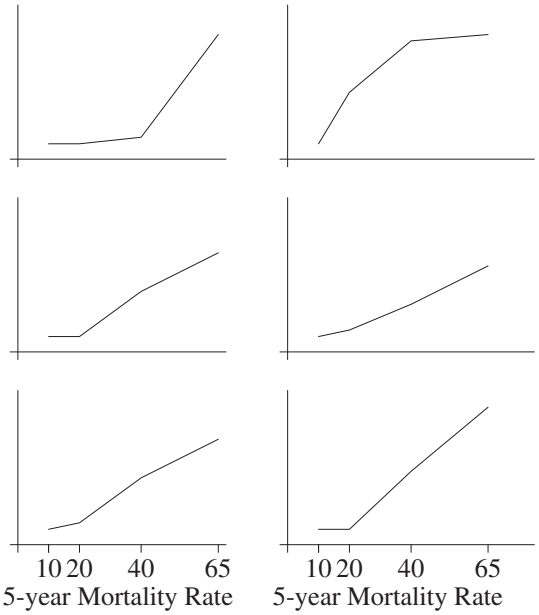

Figure 4: APP as a function of 5-year Mortality Rate, Judges 1-6

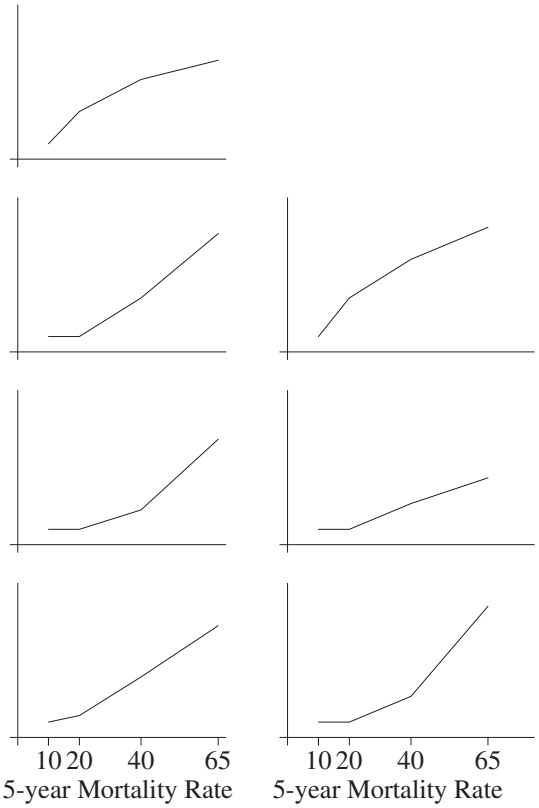

Figure 5: APP as a function of 5-year Mortality Rate, Judges 7-13

pattern is consistent with a policy in which once the evidence for a severe malignancy pass a certain threshold, treatment is universally recommended.[5]

4 Discussion

In this section we first discuss the results from the point of view of the three perspectives by which the relationship between our decision analytic model and the judgments of appropriateness could be understood. The first perspective, the policy capturing perspective, suggests that, if the model is assumed to correctly describe the judgment, insight regarding the implicit rules underlying the judgment could be revealed. Indeed two such insights are revealed by our analysis. First, the analysis reveals a discrepancy between the epidemiological data regarding the importance of LNI and thickness reported by Averbook et al. (2002) and the subjective weights assigned to these factors by the judges. According to the Averbook et al. (2002) data, thickness is the most important determinant of p1, whereas according to the judgments, LNI is the most important characteristic of p1. Second, the analysis reveals a configural (disjunctive) rule with respect to the integration of the severity of thickness and severity of LNI in the determination of appropriateness, in that thickness has a larger impact on appropriateness when the level of LNI is low than when it is high.

The second perspective suggests that, if our decision analytic model is a prescriptive model of the appropriateness of a medical treatment, consistency between the model and appropriateness judgment could be viewed as supporting the validity of the judgment. Within this context it is worthwhile to distinguish between three types of validity of appropriateness judgment. Ecological validity refers to valid perception of the probabilities (and utilities[6]) associated with the judgment. Normative validity refers to reliance on normative rules (e.g., rules for integrating probabilities and utilities) in arriving at a judgment. Scale validity refers to accurate use of judgment scales, in our case valid use of the appropriateness scale, and in particular to the notion that appropriateness judgments are a *linear* representation of the difference between subjective expected utility of treatment versus no-

[5]Note that because of the crossing of the simuli in the design, the averages over other values of the cues and over LNI in particular, lead to a linear function in thickness, even though the multi-cue pattern is disjunctive

[6]Note, however, that our data are only relevant to testing the valid perception of probabilities. Furthermore, the test of valid perception of utilities is much more ambiguous since utilities vary across people.

https://doi.org/10.1017/S1930297500000292 Published online by Cambridge University Press

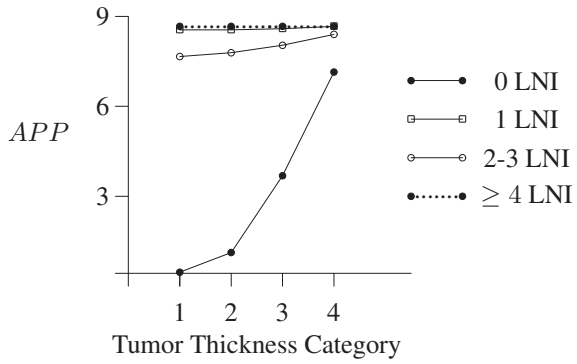

Figure 6: APP as a function of tumor thickness, various LNI levels

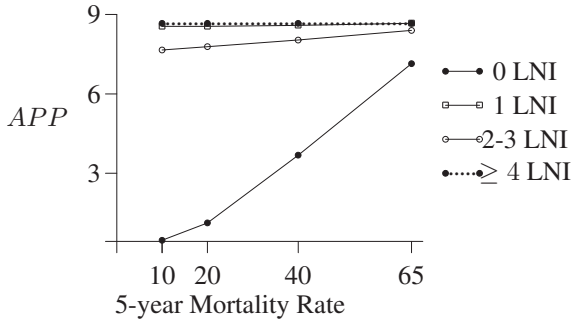

Figure 7: APP as a function of 5-year Mortality Rate, various LNI levels

treatment. Although none of these three validities is directly demonstrated by the results, they are all supported by the data, since all are necessary for appropriateness judgment to be a linear function of thickness under the assumption that the decision analytic model is valid.

The third perspective suggests that, if appropriateness judgments are viewed as prescriptively accurate, a consistency between the model and the judgments could be viewed as supporting the prescriptive stand of the model and the basic tenets on which it is based, in particular the assumption of constant treatment effect.

It is important to note that the constant treatment effect stipulates linearity between appropriateness judgment and thickness only with regard to the average (over the various levels of the other symptoms) appropriateness judgment. It does not necessarily stipulate linearity between thickness and appropriate judgments at each level of LNI. The latter requirement, labeled multivariate linearity, is weaker than the former, labeled univariate linearity. In fact, multivariate linearity is a sufficient but not necessary condition for univariate linearity. Conceptually, the difference between the two types of linearity is that whereas univariate linearity suggests that the effect of interferon treatment does not depend on the severity of the melanoma, multivariate linearity suggest that, *other things being equal*, the effect of interferon treatment is constant for various levels of thickness.[7]

Within this context, note that the difference in slopes of the effect of thickness on the appropriateness of interferon treatment is explained in terms of different values of K for various levels of node involvement. A larger slope for high node involvement than for low node involvement will occur if $K = p_2/p_1$ is larger for high node involvement than for low node involvement; that is, if the effect of treatment is generally (e.g., across levels of thickness) higher for high node involvement than for low node involvement.

One particular interesting aspect of our analysis is the comparison between the individual judgments and the average judgments. Assuming that the average judgments represent a "true" model of the medical community's view of the relationship between the symptom (thickness) and the appropriateness of interferon treatment, deviations from these judgments — or for that matter deviation from linearity — could be viewed as an error. There are two plausible sources for this error. First, it could stem from an individual judge's idiosyncratic models, dissimilar to the clinical community's mode. And second, it could stem from a random noise. The systematic nature of the individual judges' deviations from linearity (i.e., the deviations are either marginally decreasing or marginally increasing) is consistent with a systematic, but not with a random, deviation from the true model. In particular, our analysis of the individual judgments (Figures 4 and 5) suggest two types of idiosyncratic models underlying judges systematic errors, a marginally increasing model associated with a threshold above which increase in probability does not lead to much change in appropriateness, and a marginally decreasing model associated with a threshold below which increase in probability does not lead to much change in appropriateness.

Another explanation for the individual judges' deviations from linearity is nonlinearity in the appropriateness scale. Within this context it is important to note that appropriate judgments are aimed as a support tool to help rank and file physicians evaluate the utility of a treatment. As such they should serve as a direct (i.e. linear) indicator of utility, and deviations from linearity should be viewed as inappropriate. To use an example, consider a panel of experts who are asked to judge water temperature by sensing the water. Appropriate temperature judgment in this case should be linearly related to temperature, and the

[7] To see why multivariate linearity is not necessary for univariate linearity consider a case in which the relationship between the level of symptom A and appropriateness is marginally increasing at one level of symptom B and marginally decreasing at another level of this symptom. This could lead to univariate linearity in regard to A, associated with multivariate *non*linearity.

https://doi.org/10.1017/S1930297500000292 Published online by Cambridge University Press

a linearity test could be viewed as a test for their validity.

What are the implications of this study for the status of appropriateness judgments in medical decision making? By and large, the results of the study highlight the importance of combining the clinical judgments of individual experts, and strengthen our confidence in the appropriateness of averaged appropriateness judgments. The appropriateness judgments examined in this study, which is based on averaged or consensus judgments, appear to be valid in that they reflect accurate perception of probabilities, reliance on normative strategies in incorporating these probabilities into clinical evaluation, and adequate expression of this evaluation in manifested judgment (i.e., in an appropriateness scale). Furthermore, our results also highlight the utility of aggregation over judges, since the average judgments are more linear than the individual judgments, which, in terms of our model, implies better judgment. (This finding is consistent with Goldberg, 1970. See also Hammond, Hamm, & Grassia, 1986).

Second, the study provides an example of how the analysis of appropriateness judgments can be used to capture clinical intuition, by revealing the implicit rules underlying clinical judgment about treatment effects. In our case, the analysis of these judgments suggests rules such as the constant treatment effect in its univariate and multivariate forms, and shifts in the judged effectiveness of treatment at various levels of LNI.

Finally, the lack of linearity in the thickness scale raises a question regarding the appropriateness of the scales by which medical information is communicated to clinicians in general and to experts making appropriateness judgments in particular. Nonlinearity of a symptom scale is an undesirable feature, since, in comparison to a linear scale, it does not permit the natural assessment of the implications of clinical information for treatment. Thus, even though moving from simple, non-epidemiological, scales (such as length in millimeters for thickness) is cumbersome, a stronger emphasis on the construction of linear, or interval, scales based on epidemiological information seems a desirable direction for improved communication of clinical information.

References

Anderson, N. H. (1990). *Contributions to information integration theory, Vol. 1: Cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Audet, A. M., Greenfield S. and Field M. (1990). Medical practice guidelines:current activities and future directions. *Annals of Internal Medicine, 113*, 709–714.

Averbook, B. J. Fu P., Rao, J. S. and Mansour, E. G. (2002). A long term analysis of 1018 patients with melanoma by classic cox regression and tree-structured survival analysis at a major referral center: Implications for the future of cancer staging. *Surgery, 132*, 589–604.

Balch, C. M. Buzaid, A. C. Atkins, M. B. et al. (2000). A new ajcc staging system for cutaneous melanoma. *Cancer, 88*, 1484–1491.

Balch, C. M. Soong, S. J. Gershenwald, J. E. Thompson, JF. et al. (2001). Prognostic factors analysis of 17,600 melanoma patients: validation of the american joint committee on cancer melanoma staging system. *Journal of Clinical Oncology, 19*, 3622–3634.

Bernstein, S. J. Hofer, T. P. Meijler, A. P. and Rigter H. (1997). Setting standards for effectiveness: a comparison of expert panels and decision analysis. *International Journal for Quality in Health Care, 9*, 255–263.

Bosch, J. L., Halpern, E. F. and Gazelle, G. S. (2002). Comparison of preference-based utilities of the short-form 36 health survey and health utilities index before and after treatment of patients with intermittent caludication. *Medical Decision Making, 22*, 403–409.

Brook, R. H. (1994). Appropriateness: The next frontier [editorial]. *BMJ, 308*, 218–219.

Chapman, G. B. and Sonnenberg FA. (Eds). (2000). *Decision making in health care: Theory, psychology, and applications*, volume 22. Cambridge University Press.

Dubois, R. W. Swetter, S. M. Atkins M. McMasters, K. et al. (2001). Developing indications for the use of sentinel lymph node biopsy and adjuvant high-dose interferon alfa-2b in melanoma. *Archives of Dermatology, 137*, 1217–1224.

Field, M. J. and Lohr, KN. eds. (1990). *Clinical Practice Guidlines: Directions for a New Agency: Institute of Medicienc*. Washington, DC: National Acacemy Press.

Goldberg, L. W. (1970). Man versus model of man: A rationale plus some evidence for a method of improving on clinical inference. *Psychological Bulletin*, 73, 422–432.

Hammond, K. R., Hamm, R. M., & Grassia, J. (1986). Generalizing over conditions by combining the multitrait-multimethod matrix and the representative design of experiments. *Psychological Bulletin*, 100, 257–269.

Kee F. Patterson, C. C. Wilson, A. E. McConnell J. M. et al. (2002). Judgment analysis of prioritization decision within a dialysis program in one united kngdom region. *Medical Decision Making, 22*, 140–151.

Kuntz, K. M. Tsevat J. Weinstein, M. C. and Goldman L. (1999). Expert panet vs decision –analysis recommendations for postdischarge coronary angiography after myocardial infarction. *JAMA, 282*, 2246–2251.

Landrum, M. B. Normand, S. L. (1999). Applying bayesian ideas to the development of medical guidelines. *Statistics in Medicine, 18*, 117–137.

Park, R. E. Fink A. Brook, R. H. Chassin MR. et al.

https://doi.org/10.1017/S1930297500000292 Published online by Cambridge University Press

(1986). Physician ratings of appropriate indications for sixe mecical and surgical procedures. *American Journal of Public Health, 76*, 766–772.

Rule, S. J. Curtis, D. W and Mullin, L. C. (1981). Subjective ratios and differences in perceived heaviness. *Journal of Experimental Psychology: General, 7*, 459–466.

Schuchter, L. M. (2004). Adjuvant interferon therapy for melanoma: High-dose, low-dose, no-dose, which dose? *Journal of Clinical Oncology, 22*, 7–10.

Shackman, B. R. Goldie, S. J. Freedberg, K. A. Losina E. et al. (2002). Comparison of health state utilities using community and patients preference weights derived from survey of patients with hiv/aids. *Medical Decision Making, 22*, 27–38.

Shapiro, D. W. Lasker, R. D. Bindman, A. B and Lee, P. R. (1993). Containing costs while improving quality of care: the role of profiling and practice guidelines. *Annual Review of Public Health, 14*, 219–241.

Sheldon Z. and Kafry D. (1997). Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance, 18*, 269–294.

Sorum, P. C., Stewart, T. R., Mullet E., González-Vallejo, C., et al. (2002). Does choosing a treatment depend on making a diagnosis? US and french physicians' decision making about acute otitis media. *Medical Decision Making, 22*, 394–402.

Stalmeier, P. F. (2002). Discrepancies between chained and classic utilities induced by anchoring and occasional adjustment. *Medical Decision Making, 22*, 53–64.

Stewart, T. R., & Joyce, C. R. B. (1998). Increasing the power of clinical trials through judgment analysis. *Medical Decision Making, 8*, 33–38.

Wigton, R. S. (1996). Social judgment theory and medical judgment. *Thinking and Reasoning, 2*, 175–190.