of life measurements. Stated preference methods have a structured and quantitative approach consistent with theory and can consider attributes related to both processes and outcomes. One study examining patient preferences in Crohn’s disease treatments found patients were more willing to accept higher side effects for improved outcomes [4]. This kind of information can have huge impacts on utilization, especially when trying to encourage preventative care. Dr. Deborah Marshall cited her colorectal screening study from Canada, which showed that screening uptake would increase if all testing modalities were offered to patients instead of one alternative [5]. In a separate study, she also illustrated how providers’ and patients’ preferences differ for colorectal screening tests. Patients were willing to pay somewhere between 1.5 and 4 times more for colorectal cancer screenings and were more concerned with test accuracy than physicians expected [6]. Dr. Marshall argued that if we want to maximize up-take of appropriate health care interventions, we need to better understand the choices patients make with stated-preference studies like these. These methods can also allow researchers to measure the rates of substitution between various attributes of a treatment process and calculate patients’ maximum acceptable risk or willingness-to-pay.

In negotiating this space between allocating scarce resources and allowing access to needed treatments, Drs. Bridges and Marshall think a possible third way forward would be to allow for increased use of co-pays. While centralized agencies may decide not to fund a new technology or only fund it to a point, co-pays would allow patients access to the technologies if they can afford it. Dr. Sculpher cautioned this can become an equity issue which could obstruct the goals of population health if used incorrectly. Despite the differences, all three researchers agree that patients should play a bigger role in evaluating technologies in the future.

REFERENCES

Thoughts on the Validity of a New QALY Estimator
Jason N. Doctor PhD, Clinical Pharmacy & Pharmaceutical Economics & Policy, University of Southern California, Los Angeles, CA, USA
ISPOR Connections January/February 2010
(A response by Benjamin Craig follows this article)

INTRODUCTION
Quality-adjusted life years (QALYs) are the most widely used measure of health in the economic evaluation of health care. They are intended to reflect preference toward health. However, problems with aggregating better-than-death (BTD) and worse-than-death (WTD) health states in general population studies have led to unorthodox approaches to QALY estimation, where between 26% and 34% of responses are adjusted by the researchers [1, 2]. These approaches may contribute to differences between QALY utilities and conjoint analysis estimates for health. The ISPOR workshop, titled “Revisiting the estimation of QALYs” led by Benjamin Craig and co-authored by M. Oppe and J.J. Busschbach at the ISPOR 14th Annual International Meeting, Orlando, FL, USA, May 2009 addressed the problem of aggregating better- and worse-than-death TTO responses to estimate health state utility. The workshop was notable for its effort to make QALY estimation consistent with QALY theory and to unify preference scales (QALY utilities and conjoint analysis). With respect to these concerns, the workshop was constructive in highlighting current deficiencies in QALY estimation. Further, the workshop was a refreshing departure from work that seeks to discredit rather than unify measurements obtained by different approaches. There are, however, two concerns with the new estimator. First, one of the recently developed estimators presented by the authors, the episodic random utility model or ERUM, while in some ways is more consistent with QALY theory than previous efforts, applies a stochastic choice model to judgment data. Second, the ERUM estimator does not distinguish utility conditioned on better-than-death and worse-than-death responses. In this essay, we reflect on the problem of aggregating better- and worse-than-death responses under the ERUM estimator.

QUALITY-ADJUSTED LIFE YEAR ESTIMATION
Notation & Assumptions
We first give some preliminary notation that facilitates discussion of QALY estimation in the presentation. Health profiles (q1,t1; … ;qntn) reflect a sequence of health state,q, and duration, t, pairs over some time horizon, t1 + … + tn = T*. For example, if we set the time horizon, T*, equal to 10 years, a typical health profile might be (q1,11; q212; q313) (“full health, 3 years”; “back pain, 6 years”; “wheelchair, 1 year”). Persons have preferences over health profiles. Quality-adjusted life years use mathematical formulae to represent the empirically observed preferences. The most basic QALY model (and the one relevant for the current workshop) assumes that for any health profile (q1,11; … ;qntn) there is a utility function U, that represents preference over health profiles:

$$U(q_1,t_1; \ldots ;q_{ntn}) = H(q_1)t_1 + \ldots + H(q_{ntn})t_{ntn}.$$  (1)

Note that in Equation 1, utility is linear in time and different time periods contribute additively to overall utility. Usually (but not always) expected utility is used to establish a utility function for QALYs. Expected utility theory hypothesizes that U represents preference in the sense that if we empirically observe that one health profile is “at least as preferable” as another, then utility for that profile is at least as great as it is for the other. Likewise, if we claim the utility of a profile is at least as great as another, then we expect to empirically observe that the health profile is at least as preferable” as the other profile. We call the aforementioned relationship between preference and utility the expected utility hypothesis. Under the episodic random utility model, the decision maker is assumed to select (with error) a choice so as to maximize utility. Thus, for individual utility functions over health profiles, like those in Equation 1, a choice probability (the probability of selecting one profile over another) is a function of the differences in utility between the choices and error. Standard random utility models treat the value of attributes as linear in their parameters. Thus, a linear regression in these models is often taken to represent the value of prospects.

When a respondent chooses between two options, we call this “choice data”. When a respondent judges how much of an attribute in one option needs to be changed to make two options equal (or matched) we call this “judgment data”. Matching involves making judgments as opposed to choices. Time tradeoff data are judgments because, after bracketing through a series of choices, >
respondents judge how many years in full health are equal to a longer period in poor health. One concern in applying a random utility framework to the time tradeoff exercise then is that it is an application of a choice model to judgment data. Considerable empirical research has shown that preferences between two options collected by observing choice behavior are often reversed when collected by having subjects form matches [3-7]. In fact, a recent monograph review of decision research cautions, to mix choices and matches “is a recipe for trouble.” [8, p. 61] That said, the choice paradigm offers greater flexibility with the error term allowing ERUM to combine utility across BTD and WTD responses as a single value.

**THE ERUM ESTIMATOR**

In this section I derive the ERUM estimator from the perspective of fitting linear regression to judgments of stimuli (i.e. a perspective that treats judgments (dependent variable), as a linear function of stimuli to be judged (independent variable) and error). This deviates from how error is treated in the choice model, but is more consistent with how responses are obtained in the TTO exercise. The original article notes that there are two different elicitation procedures for the time tradeoff (TTO): One for health states better-than-death (BTD) and one for health states worse-than-death (WTD). The BTD-TTO asks a subject to form a match between say 10 years in poor health, q, and t (<= 10) years in full health, q*. Setting the utility of “full health” equal to 1, the estimator correctly reflects (via the expected utility hypothesis) that the BTD-TTO assigns a utility H_BTD(q) such that:

$$t = H_{BTD}(q) 10 + \text{error}$$ (2)

We have included an error term, “error”, to denote response error may be present and note that Equation 2 naturally reflects a regression equation with coefficient $H_{BTD}(q)$.

Sometimes there is no equivalence match reached in comparing (q*,t) with (q, 10 years). In this case, we find that death (q*, t= 0) is strictly preferred to poor health, q, for 10 – t’ years. Again working with Equation 1 and setting the utility of “full health” equal to 1, the estimator correctly reflects (via the expected utility hypothesis) that the WTD-TTO assigns a utility $H_{WTD}(q)$ such that:

$$t' = H_{WTD}(q)(10 - t')$$ (3)

Here the “WTD” subscript on the function H indicates that the worse-than-death exercise was completed. Notice that in Equation 3 (unlike Equation 2) I have excluded the error term. This is because Equation 3 is not properly described by regression theory. Linear regression requires that the regressor, (10 – t’), be error free. In other words, the regressor must not be contaminated by measurement error. Yet, clearly (10 – t’) is a response given by the subject so (10 – t’) contains measurement error at least to the extent that t’ is unreliable. Measurement error in the TTO is well documented [9]. This concern does not arise when the variables in a regression equation constitute separate responses from the subject, or when the regressor is a stimulus (e.g., health state) fixed by the experimenter (as in Equation 2). Though to derive the ERUM estimator, Equation 3 is needed. We follow published approach here only to show how the ERUM estimator is derived.

In regression theory, a regression equation (y = Bx + error), under the usual assumptions has a least squares coefficient estimate

$$\hat{B} = \sum xy / \sum x^2$$

([10], Eq. 4.7, p. 163). Therefore, treating Equations 2 and 3 as regression equations, for any fixed health state q, we may estimate health state utility for better-than-death and worse-than-death responses as follows:

$$H_{BTD}(q) = \Sigma BTDq / \Sigma BTDq^2$$ (4)

and,

$$H_{WTD}(q) = \Sigma WTDq' / \Sigma WTDq'^2$$ (5)

Moving the denominators on the left of Equations 4 and 5 to the right we have:

$$H_{BTD}(q) \Sigma BTDq = \Sigma BTDq \Sigma BTDq'^2$$ (4')

and,

$$H_{WTD}(q) \Sigma WTDq' = \Sigma WTDq' \Sigma WTDq'^2$$ (5')

ERUM does not distinguish utility under the BTD and WTD exercise. For example, if a person’s true utility for a health state is greater than zero (i.e., better-than-death), but they completed the worse-than-death exercise then the random utility model accounts for this in the error term. Thus the random utility model assumes that for any health state q, across individuals (for which there is at least one BTD and WTD response) $H(q) = H_{BTD}(q) = H_{WTD}(q)$. Substituting $H(q)$ for $H_{BTD}(q)$ and $H_{WTD}(q)$ in Equations 4’ and 5’ respectively, adding the resultant equations together, and solving for $H(q)$ gives the ERUM estimator (also found in their published paper [11] as Equation 4):

$$H(q) = (\Sigma BTDq / \Sigma BTDq'^2 + \Sigma WTDq' / \Sigma WTDq'^2)$$ (6)

To explain the ERUM estimator in the most simple terms it is a linear regression over all responses. For any health state, q, each respondent completes either a BTD or WTD exercise giving a y value (either t (Eq. 2) or -t’ (Eq. 3)) and an x value (either 10 (Eq. 2) or 10 – t’ (Eq. 3)) and for a large sample of such responses these values are regressed to obtain a regression coefficient $\hat{B}$ that purports to be the utility for health state q.

A key advantage of the estimator is that small numbers of severe WTD responses do not drastically affect the utility H(q) and thus researchers are not compelled to “adjust” these utilities manually as has occurred elsewhere. A concern though is whether or not it is valid to assume that information as to whether the BTD or WTD exercise was completed does not affect one’s estimate of utility. Consider that linear regression is often understood in terms of conditional expectations, the expected value of y for the ith subject given x is E[y|xi] = Bxi. ERUM assumes knowledge that the subject completed either the better-than-death or worse-than-death exercise does not influence the expected y, E[y|xi, BTD] = E[y|xi, WTD] = Bxi. Yet it is conceivable that information about the completion of the BTD or WTD exercise does affect the conditional expectation of y. For starters, if you completed the worse-than-death exercise then y is negative or zero and if you completed the better-than-death exercise your response is positive or zero. It is easy to show that if one does not begin with the assumption that E[y|xi, BTD] and E[y|xi, WTD] are equal then the ERUM estimator is only a valid estimator for health states that are equivalent in preference to death. Further research will have to determine if distinguishing BTD and WTD responses leads to different conditional expectations.

**THE PROBLEM OF WORSE-THAN-DEATH UTILITY ESTIMATION**

When better-than-death and worse-than-death responses are combined via calculation of an arithmetic mean, the severity of even a small number of worse-than-death responses can bring down the mean estimate substantially, leading to severe quality-of-life reductions for moderate health states. The ERUM paper argues that the reason worse-than-death responses are a problem is because they are an estimate of a ratio: -t’/t’ - t’. And, as with cost-effectiveness analysis, an estimate of ratios is unstable. We note though that unlike in cost-effectiveness analysis, the error in the numerator and denominator in a worse-than-death time tradeoff is from the same source and, for each response, is the same magnitude and direction. For extreme errors, these facts would act to stabilize the ratio near -1 rather than force it to an extreme negative number.
Here I offer an alternative explanation to the problem of worse-than-death responses. One difference between better-than-death and worse-than-death elicitation is that the latter requires explicitly that utility be computed assuming that value of health is additive over disjoint time periods. Let “~” denote equivalence in preference between two health profiles. Recall that the utility of a worse-than-death state is computed from the indifference:

\[(\text{death, 10 years}) \sim (\text{full health, } t\text{'years; } q, 10-t\text{'years}, \text{for } t' < 10\text{ years}),\]

this by the expected utility hypothesis and Equation 1 implies:

\[H(\text{death})10\text{ years} = 0 = H(\text{full health})t'\text{ years} + H(q)(10 - t')\text{ years} (7)\]

In order to separate \(H(q)\) via an addition operation from the health profile on the right requires a strong preference condition known as “generalized marginality”–a condition that requires preference to depend on marginal probabilities and not on the joint probability distribution of risky choices [12]. In a recent study, only about 4 out of 57 subjects satisfied this condition [13]. Thus, worse-than-death utility elicitation may produce biased utilities because their computation depends on a preference condition that is not often satisfied. The ERUM estimator places less value on worse-than-death outliers which may result in more stable estimates.

What are other alternative for dealing with worse-than-death outlier responses? Because classical statistical methods are highly influenced by outliers, worse-than-death utilities (which may be large negative numbers), may lead to very biased health state estimates. One simple approach to handle this problem is to use a trimmed mean instead of a standard arithmetic mean estimate. Statistical theory indicates that for large samples an interquartile mean (a mean with the lowest 25% and highest 25% of scores discarded) will approximate an arithmetic mean. Most importantly, an interquartile mean is insensitive to outliers. Since most health states are better-than-death a trimmed mean should greatly reduce bias for these states. A limitation of this approach, however, is that states for which the trimmed mean is still worse-than-death remain problematic because of their bias.

**QALYS AND CONJOINT ANALYSIS**

One interesting aspect of Craig, Oppe and Busschbach’s workshop was the comparison of QALYs to conjoint analysis health value estimates. Conjoint analysis is a flexible probabilistic approach to modeling choice that allows many different attributes of the choice problem to be incorporated into a stated preference experiment. Although QALY utility estimation and conjoint analysis methods are distinct approaches, the similarities are often overlooked. Conjoint analysis and QALYs each investigate the possibility of using a numerical function to represent preference. Like QALYs, conjoint analysis uses preference conditions to achieve this goal (e.g., [14], Axiom 1, p. 5; [15], Section 2, p. 80). And, these preference conditions have been scrutinized, debated and criticized [16, 17]. Mathematically, the methods are very similar, because each assumes that utility is additively decomposable over attributes (“additively decomposable” means that there are numeric scales on components (where health status is a component) and a rule for combining them such that the resultant measure preserves the preference ordering and this decomposition is additive over factors).

Craig, Oppe and Busschbach found that the ERUM estimator improved agreement between QALY and conjoint analysis estimates. However, this comparison was made by examining correlations between value estimates. Unfortunately, using correlation analysis for establishing models of measurement can be as treacherous as using correlation as a means of establishing causation. For example, Birnbaum [18] generated data by one model and compared correlations under correct and incorrect models fit to the data. He found the incorrect model had a higher correlation with the generated data than did the correct one. Other evaluation approaches, such as factorial plots of data (e.g., plotting utility as a function of EQ-5D domain and severity level under two scaling procedures), may better elucidate when different measurement procedures agree and disagree. Further, some problems remain for scaling of quality of life values using conjoint analysis, at least for ranked data [19]. Of particular concern is that conditional (multinomial) regression on ranked data is strongly affected by the number of persons who conform to the theory [19]. Further, as pointed out by Flynn [19], the estimation of the death state in ranked data, represents the mean distance between the death state and the worst possible living state (confounded with the variance scale factor) and does not estimate the tradeoff between living and death states respondents are willing to make. A correlation analysis on ranked data does not reveal this problem. A true choice experiment (not based on ranked data) would surmount some of these problems.

**SUMMARY**

In summary, the workshop by Craig, Oppe and Busschbach brought to the light legitimate concerns with current QALY estimation procedures, showed how current methods deviate from QALY theory, and sought to improve agreement across different health valuation approaches. A limitation of the new ERUM estimator is that it was derived under a theory of discrete choice, but was applied to judgment data. In addition, there are concerns about applying regression theory to worse-than-death health states, as well as across both better-than-death and worse-than-death responses. An alternative explanation for the problem is that preferences in the worse-than-death exercise may not satisfy a necessary QALY preference condition and thus give biased results. Using methods that minimize the effect of outliers may go a long way in improving QALY estimation.

**ACKNOWLEDGEMENTS**

Thanks to Janet Shin for review of a previous version of this article.

**REFERENCES**

RESPONSE
Benjamin M. Craig PhD, Assistant Faculty Member, Moffitt Cancer Center, Tampa, FL, USA

In response, I hope to briefly clarify my take on QALY estimation using trade-off responses and to provide a few thoughts on Doctor’s insights concerning our ISPOR workshop.

In health valuation, there are three types of trade-off responses: time trade-off (TTO), standard gamble (SG), and person trade-off (PTO), each of which expresses the value of health by varying quantities of life (i.e., time, risk, or persons). Currently, there are only three QALY estimators using trade-off responses (x, y):

In TTO, x is 10 or 10^-t’ years and y is t or -t’ years (see Doctor’s equations 2 and 3). As with incremental cost-effectiveness ratios (ICER), the mean ratio (e.g., mean (cost/effectiveness)) is not robust, particularly when denominator (x) may be small. Although this limitation is well recognized in economic evaluations (Drummond et al. 2005; page 258), little recognition had yet to make its way into QALY estimation, which motivated our workshop [1].

After delineating this limitation, the workshop highlighted the fact that most prominent QALY studies changed a large proportion of their data to make their mean ratios look right. In the original UK QALY estimation, the largest possible ratio (y/x) is -39 (or -9.75 years/0.25 years). Because this extreme worse-than-dead (WTD) ratio swamped the mean ratio, all WTD responses were changed (e.g., -39 became -0.975), arbitrarily replacing 34% of the sample. Likewise, 26% of the US TTO responses were changed by dividing all WTD ratios by 39 (e.g., -39 became -1). Instead of arbitrarily changing (or trimming) ratios, I believe that ratio statistics (mean or medians) should be abandoned, like in economic evaluations.

In his interpretation, Doctor correctly asserts that Busschbach and I recommend the use of choice models in the analysis of TTO responses. When modeling a choice response (A>B), the inclusion of an additive error term is commonplace: U(A)-U(B)= error (e.g., probit). In TTO, A and B represent two alternative episodes of health, and instead of expressing preference (>), a respondent makes an equivalence, or match (~). The proposed episodic random utility model (eRUM) is: U(A)-U(B)=error. Its parameters may be estimated by minimizing the sum of squared error (i.e., Gauss-Markov Theory).

Doctor also asserts that the ERUM estimator (or coefficient) does not distinguish between BTD and WTD responses. This is also a criticism of the mean ratio, mean angle and most trade-off protocols. By separating BTD and WTD responses, Doctor also asserts that the ERUM estimator (or coefficient) does not distinguish between BTD and WTD responses. This is also a criticism of the mean ratio, mean angle and most trade-off protocols. By separating BTD and WTD responses, trade-off protocols (e.g., TTO, SG or PTO) may confound judgment in three ways. First, if a respondent errors in the initial question (BTD or WTD?), this error affects the down-stream equivalence statement or match (~). Compounding errors is a potential limitation in any adaptive survey design. Secondly, different trade-off statements are used for WTD and BTD responses, which may impose differential cognitive challenges (e.g., respondents simply do not understand WTD questions.). Third, judgments, unlike choices, induce ceiling effects by imposing a bounded scale (i.e., non-optimal gap). The mean angle, which was introduced after the workshop by Oppe and Craig, was not available to Doctor, but it does addresses some of his concerns regarding measurement error in x [2-3].

Overall, Doctor and I seem to agree on two primary points of the workshop. We each recognize that arbitrarily changing WTD responses is a question of scientific integrity, not science; and that the proposed alternatives to the mean ratio are simple and consistent with choice theory (for better or worse). The future in the QALY estimation may lie in discrete-choice responses (e.g., voting); however, the analysis of ordinal responses introduces further issues, such as anchoring on a QALY scale and distributional assumptions (e.g., EV-1 for Logit or Gaussian for Probit?). Busschbach, Oppe and I continue to explore innovative approaches in the analysis of trade-off responses. I look forward to these new challenges, now that the matters of arbitrary transformations are behind us.

REFERENCES
3. Wald A. The fitting of straight lines if both variables are subject to error. Ann Math Statist 1940;11:284-300.